# COMPUTATIONAL APPROACHES FOR BIOMARKER DETECTION AND PRECISION THERAPEUTICS IN CANCERS

EDITED BY: Suman Ghosal, Shaoli Das and Rosalba Giugno
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-861-7
DOI 10.3389/978-2-88974-861-7

## 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

# COMPUTATIONAL APPROACHES FOR BIOMARKER DETECTION AND PRECISION THERAPEUTICS IN CANCERS

Topic Editors:
**Suman Ghosal,** National Institutes of Health (NIH), United States
**Shaoli Das,** National Institutes of Health (NIH), United States
**Rosalba Giugno,** University of Verona, Italy

**Citation:** Ghosal, S., Das, S., Giugno, R., eds. (2022). Computational Approaches for Biomarker Detection and Precision Therapeutics in Cancers. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-861-7

# Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 November 2020
doi: 10.3389/fgene.2020.595733

# RNF183 Is a Prognostic Biomarker and Correlates With Tumor Purity, Immune Infiltrates in Uterine Corpus Endometrial Carcinoma

*Rong Geng[1,2,3†], Yuhua Zheng[2†], Lijie Zhao[2], Xiaobin Huang[2], Rong Qiang[2], Rujian Zhang[2], Xiaoling Guo[2*] and Ruiman Li[1*]*

*[1] Department of Gynecology and Obstetrics, The First Affiliated Hospital, Jinan University, Guangzhou, China, [2] Department of Gynecology, Affiliated Foshan Maternity & Child Healthcare Hospital, Southern Medical University, Foshan, China, [3] Foshan Maternal and Children Healthy Research Institute, Affiliated Foshan Maternity & Child Healthcare Hospital, Southern Medical University, Foshan, China*

OPEN ACCESS

***Edited by:***
*Rosalba Giugno, University of Verona, Italy*

***Reviewed by:***
*Milind B. Ratnaparkhe, ICAR Indian Institute of Soybean Research, India*
*Tanima Bose, Ludwig Maximilian University of Munich, Germany*

****Correspondence:***
*Xiaoling Guo*
*gxl_gr@163.com*
*Ruiman Li*
*hqyyck@126.com*

*† These authors have contributed equally to this work*

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

***Received:*** *17 August 2020*
***Accepted:*** *02 November 2020*
***Published:*** *26 November 2020*

***Citation:***
*Geng R, Zheng YH, Zhao LJ, Huang XB, Qiang R, Zhang RJ, Guo XL and Li RM (2020) RNF183 Is a Prognostic Biomarker and Correlates With Tumor Purity, Immune Infiltrates in Uterine Corpus Endometrial Carcinoma. Front. Genet. 11:595733. doi: 10.3389/fgene.2020.595733*

RNF183, a member of the E3 ubiquitin ligase, has been shown to involve in carcinogenesis and proposed as one of the biomarkers in Uterine Corpus Endometrial Carcinoma (UCEC). However, no research focused on the role of RNF183 in UCEC. We analyzed the expression and immune infiltration of RNF183 in UCEC. TIMER, UALCAN, and GEPIA were used to analyze the gene expression of RNF183. We emplored Kaplan-Meier Plotter to examine the overall survival and progression-free survival of RNF183, and applied GeneMANIA to identify RNF183-related functional networks. LinkedOmics was helpful to identify the differential gene expression of RNF183, and to further analyze gene ontology and the genome pathways in the Kyoto Protocol. Finally, we used TIMER to investigate the immune infiltration of RNF183 in UCEC. Otherwise, we partly verified the results of bioinformatics analysis that RNF183 controlled ERα expression in ERα-positive Ishikawa cells dependent on its RING finger domain. We also found that ERα increased the stability of RNF183 through the post-translational mechanism. Together, patients with a high level of RNF183 harbor favorable overall and progression-free survival. High expression of RNF183 was associated with a low stage, endometrioid, and TP53 Non-Mutant status in endometrial cancer. The RNF183 expression was greater at higher expression and the tumor stage was greater at the lower level. On the side of immunization, high level of RNF183 in UCEC is negatively related to tumor purity, infiltrating levels of CD4 + T cells, neutrophils, and dendritic cells. Besides, the expression of RNF183 in UCEC is significantly correlated with the expression of several immune cell markers, including B cell, M1 macrophage marker, M2 Macrophage, Dendritic cell, Th1 markers, Th2 markers, Treg markers, and T cell exhaustion markers, indicating its role in regulating tumor immunity. These results suggested that RNF183 may be considered as a novel prognostic factor in endometrial cancer and an early diagnostic indicator for patients with UCEC.

**Keywords: estrogen receptor alpha, immune infiltration, prognosis, uterine corpus endometrial carcinoma, RNF183**

## INTRODUCTION

Uterine Corpus Endometrial Carcinoma (UCEC) is the fourth most common gynecological malignancy in developed countries (Kandoth et al., 2013), and the incidence has been raised rapidly in China, increasing 63,400 new cases a year (Chen et al., 2016). According to biological and histopathological variables, endometrial cancer is classified into two types. Type II tumors are usually poorly differentiated, non-endometrioid, and more likely to metastasis, relapse even after aggressive clinical intervention. By contrast, type I endometrial cancer is often endometrioid and well-differentiated, presumably owing to greater exposure to a long history of unopposed estrogen or other risk factors inducing hyperestrogenism such as obesity. Endometrial cancer is one of the few human malignant tumors for which mortality is increasing (Berg et al., 2017), which underlines the urgency to develop more effective methods for the early diagnosis and treatment of this disease.

The RNF183 (RING finger 183) is served as an E3 ubiquitin ligase ($E3_s$) belonged to the RING finger protein family. RING finger domain has been characterized by the sequence of CX2CX(9–39)CX(1–3)HX(2–3)C/HX2CX(4–48) primarily responsible for substrate specific identification in ubiquitylation (Joazeiro and Weissman, 2000; Lipkowitz and Weissman, 2011). RING finger ubiquitin ligases are involved in the process of essential cellular functions, such as maintaining the integrity of genomic, cell cycle, cell signal, and DNA repair. For example, the FANC core complex containing RING finger-like PHD domain. Its mutation induces Fanconi anemia which increases the risk of cancer (Moldovan and D'Andrea, 2009). Besides, MDM2 targets tumor suppressor p53 for degradation (Oliner et al., 1992; Wade et al., 2010). Inactivated RING finger E3s BRCA1 destroys the DNA repair pathway in breast and ovarian cancer (Hashizume et al., 2001; Ruffner et al., 2001). Properly, RING finger E3s are involved both in the promotion and the suppression of cancers. Distinct RING finger E3s are particular therapeutic targets. Small molecular inhibitors suppress the MDM2–p53 interaction in preclinical studies. Accumulation functional and controlled pathway data from RING finger E3s are helpful for developing new targeted therapy.

RNF183, RNF182, RNF186, and RNF152, are further identified as the RNF183 family, which share the similar structure RING finger domain (C3HC4) at their N-terminus and transmembrane domains at their C-terminus with high homology (Kaneko et al., 2016; Okamoto et al., 2020a). As common features, members of RNF183 family have exhibited a broad range of functions in diverse biological and pathological processes such as prolonged endoplasmic reticulum stress, apoptosis, ischemia-reperfusion injury, oxygen, and glucose metabolism, immune and inflammatory response (Liu et al., 2008; Nectoux et al., 2010; Wang et al., 2018; Wu et al., 2018; Cao et al., 2019; Maeoka et al., 2019a). It was proposed that RNF183 could be as one of the potential biomarkers for endometrial cancer through gene expression screening (Colas et al., 2011). However, the RNF183 involvement of molecular mechanisms underlining the disease remains unclear.

Here we find that RNF183 is upregulated in endometrial cancer and mostly higher in endometrioid, low-grade, TP53-Non-Mutant samples. It is also negatively related to tumor purity, infiltrating levels of $CD4^+$ T cells, neutrophils, and dendritic cells. Besides, the RNF183 in UCEC is significantly correlated with the expression of several immune cell markers, including B cell, M1 macrophage marker, M2 Macrophage, Dendritic cell, Th1 markers, Th2 markers, Treg markers, and T cell exhaustion markers. For mechanism, RNF183 shows a significant correlation with ERα. We prove that RNF183 regulates ERα and ERα target genes under the existence of the RING finger domain. Furthermore, ERα promotes the stability of RNF183.

## MATERIALS AND METHODS

### UALCAN Database

UALCAN[1] (Chandrashekar et al., 2017) is a cancer data online analysis, mainly based on the TCGA level 3 RNA—seq and clinical data of 31 types of cancer in 74 samples of normal and tumor by the relative expression of genes. The database can be spectrum identification of target gene expression, DNA promoter region methylation analysis, survival analysis and correlation analysis. It also can check other related information in the database through the link. For example, gene modification and miRNA prediction were examined.

### GEPIA Database

Gene Expression Profiling Interactive Analysis (GEPIA) database[2] is used to analyze the RNA sequencing expression data of 8,587 healthy and 9,736 tumor tissue samples from TCGA and GTEx projects (Tang et al., 2017) including single-gene analysis, cancer type analysis, and polygene analysis. By inputting the target gene on this website, the differential expression, survival analysis, correlation analysis and PCA of the target gene can be obtained. We generated the expression of the RNF183 gene through GEPIA.

### Kaplan-Meier Plotter Database

Kaplan-Meier survival curve analysis is used to evaluate the correlation between the expression of 54,000 genes in 10,000 cancer samples and the survival rates of 21 different cancers. The samples include 371 livers, 1,440 stomachs, 2,190 ovaries, 3,452 lungs, 6,234 breast cancer and 543 UCEC samples. Use the Kaplan-Meier diagram[3] to analyze the relationship between gene expression and survival rates of endometrial cancer through hazard ratio (HR), and Logarithmically sort the P value (Lánczky et al., 2016).

### GeneMANIA Database

GeneMANIA[4] is mainly used to construct a protein-protein interaction (PPI) network, generate hypotheses about gene

[1]http://ualcan.path.uab.edu

[2]http://gepia.cancer-pku.cn/index.html

[3]http://kmplot.com/analysis/

[4]http://www.genemania.org

function, and determine the priority of genes by analyzing the gene list (Warde-Farley et al., 2010). Entering the target gene in the site to generate protein—protein interaction network, each small circles represent different proteins in the network, the size of the circle represents the strength of the interaction, different colors of the attachment has the validation of different means of interaction, the validation includes a variety of bioinformatics methods: physical interaction, gene co-expression, gene co-localization, gene enrichment analysis, and website prediction. Besides, the annotation information of the protein can also be queried in the target network.

## LinkedOmics Database

The LinkedOmics database[5] is mainly used for comprehensive data analysis related to TCGA cancer 32 sets (Vasaikar et al., 2018). It also includes mass spectro-based proteomic data generated by the Clinical Proteomics Oncology Analysis Association (CPTAC) for TCGA breast, colorectal, and ovarian tumors. The LinkFinder module of LinkedOmics was used to study the differentially expressed genes related to RNF183 in the TCGA UCEC cohort ($n = 176$). The results provided by the database are shown in the form of a volcano map, heat map or scatter plot by Pearson correlation coefficient analysis. Besides, biological processes, cellular components, molecular functions, and enrichment, and analysis of KEGG pathways were performed through genomic enrichment analysis (GSEA). The grade standard is FDR $< 0.05$, and 500 simulations have been performed.

## TIMER Database

The TIMER database runs more than 10,000 samples from the Cancer Genome Atlas (TCGA) to systematically analyze the tumor infiltrating immune cells (TIIC) of 32 kinds of cancers[6] (Li et al., 2017). TIMER determines the abundance of tumors by statistical analysis of gene expression profile, 106 infiltrated immune cells (TIIC) were analyzed (Li et al., 2016). The gene module is mainly used to explore the correlation between gene expression and immunoglobulin content. The survival module is applied to seek the relationship between clinical outcomes and immune infiltration or gene expression richness. Correlation between the mutated gene and the content of immune infiltration fluid from the mutation module. SCNA model is adopted to explore the correlation between somatic CNA and immune infiltration richness. The Diff Exp module is selected to examine the differential gene expression between tumor and normal tissues. The correlation module is used to research the correlation between genes. The Go Estimatio module can run private samples of users with the TIMER algorithm. We analyzed the relationship between RNF183 gene expression level and infiltrating immune cells by Spearman analysis (including B cells, CD4 + T cells, CD8 + T cells, neutrophils, dendritic cells, and macrophages).

[5]http://www.linkedomics.org/login.php

[6]https://cistrome.shinyapps.io/timer/

## Plasmids and Antibodies

Anti-RNF183 antibody (1:1,000, NBP1-74192, Novus Biologicals, Colorado, United States), anti-ERα antibody (1:1,000, ab267512, Abcam, Cambridge, United Kingdom), anti-GAPDH antibody (1:3,000, 10494-1-AP, Proteintech Group, Chicago, United States), HRP-conjugated Affinipure Goat Anti-Mouse IgG (H + L) (1:10,000, SA00001-1 Proteintech Group, Chicago), HRP-conjugated Affinipure Goat Anti-Rabbit IgG (H + L) (1:10,000, SA00001-2 Proteintech Group) were used for western blot. RNF183 (pcDNA4-myc/his-RNF183) and RNF183 without amino acids 1–60 were illustrated previously (Geng et al., 2017). The ERE-TK-Luc and the pRL-TK plasmids were constructed by the Genewiz Company (Suzhou, China).

## Cell Culture

Ishikawa cells were cultured in RPMI-1640 (Gibco, Carlsbad, CA, United States) with 10% fetal bovine serum (FBS) (Gibco) plus 100 U/ml penicillin G, and 100μg/ml streptomycin (Gibco) in a humidified atmosphere of 5% CO2 at 37°C. Ishikawa cells were treated with 100μg/mL cycloheximide after transfected with siNC or siERα for 48 h. Ishikawa cells were transfected with siNC or siRNF183 followed by administrating 100 nM MG132 6 h.

## siRNA Transfection

The package of si-h-RNF183 and si-h-ESR1 were designed by RIBOBIO company (siRNA for RNF183 ID: SIGS0015614-1, siRNA for ESR1 ID: SIGS0005356-1, Beijing, China). 50% fusion Cells were transfected with 75 nM siRNAs using 5 μL Lipofectamine 2000 (Invitrogen, Grand Island, NY) in per six well-cell plates. The samples were collected after transfected 48 h.

## Quantitative PCR

RNAs were extracted using Trizol (Invitrogen). The cDNA was reversed from 1 μg RNA using M-MLV reverse transcriptase (Promega, Madison, WI, United States). qPCR was examined using SYBR Green (BIO-RAD, Hercules, CA, United States) for 40 cycles (95°C for 15 s, 60°C for 30 s). The primer sequence of mRNA for qPCR are available in **Supplementary Table S1** and synthesized by Sangon Biotech (Shanghai, China).

## Luciferase Reporter Assay

Luciferase activity was assessed by the Dual-Luciferase Reporter Assay (Promega). Briefly, Ishikawa cells were transfected with siRNF183 or siNC or with pcDNA4-myc/his-RNF183 or RNF183△t or pcDNA4-myc/his vector along with ERE reporter plasmids. Cells were treated with E2 (10 nM) after 24 h post-transfection. After another 24 h, samples were collected for Luciferase activity measure.

## Western Blotting

Cells were lysed with RIPA lysis buffer (G-Clone) containing protease inhibitor (G-Clone). The concentration of protein was examined by BCA Protein Assay Kit (KeyGen BioTECH, Jiangsu, China). Collected lysates were resolved in 12% SDS-polyacrylamide gel and the protein was detected with the indicated antibodies.

## Statistical Analysis

Data were revealed as mean ± standard deviation (SD). The survival curve was generated by Kaplan-Meier plots relation analysis. The expression of related genes were evaluated using Pearson correlations. Other data were assessed using Student's *t*-test. $P < 0.05$ was considered statistically significant.

# RESULTS

## Clinical Relevance of RNF183 Expression in Endometrial Cancer

From the TIMER database of the Diff Exp module across all the cancer genome atlas (TCGA) tumors, our studying showed that a high proportion of RNF183 exists in the majority of human cancer tissues (**Figure 1A**). Among all the cancer types, RNF183 is remarkably upregulated in endometrial cancer compared with normal endometrium. To investigate the role of RNF183 in endometrial cancer, we utilized UALCAN website to assess RNA-seq in 546 primary endometrial tumors and 35 normal endometrial tissues. RNF183 was shown to be elevated in cancerous tissues compared to normal endometrium (**Figure 1B**), which was accordant with statistics documented in GEPIA (**Figure 1C**).

Next, we were encouraged to apply the Kaplan-Meier Plotter online tool to explore the clinical importance of RNF183 in extensive RNA-seq data classifying patients based on the "best cut-off" value. RNF183 high expression was associated with favorable overall survival (OS, **Figure 1D**) and progression-free survival (PFS, **Figure 1E**) in 542 patients.

The conclusion above made us search RNF183 expression in different subtypes and tumor grades of endometrial cancer, which results in diversification of the disease and specific clinical outcomes. The results from UALCAN showed that RNF183 expression was significantly increased at stage 1 in comparison with other high-grade stages (**Figure 2A**). Additionally, we found that RNF183 level was considerably higher in endometrioid adenocarcinomas compared to non-endometrioid adenocarcinomas (**Figure 2B**). Meanwhile, TP53-Non-Mutant patients harbored high RNF183 expression compared to TP53-mutated patients (**Figure 2C**).

## RNF183 Co-expression Networks in UCEC

For gaining insight into RNF183 biological meaning in UCEC, We used the function module of LinkedOmics to examine RNF183 co-expression mode in the UCEC cohort. As shown in **Figure 3C**, 8,777 genes (dark red dots) were demonstrated significant positive correlations with RNF183, whereas 11,121 genes (dark green dots) were shown significant negative associations (false discovery rate, FDR < 0.01). The top 50 significant genes positively and negatively correlated with RNF183 were shown in the heat map (**Figures 3A,B**). The statistical scatter plots for individual genes are shown in **Figure 3D**. Besides, we discussed the protein-protein interaction (PPI) network and the function of RNF183 through GeneMANIA (**Figure 4**).

## Enrichment Analysis of RNF183 Functional Networks in UCEC

GO term analysis by gene set enrichment analysis (GSEA) showed that genes differentially expressed in correlation with RNF183 were located mainly in the membrane and nucleus, where they participate in biological regulation, metabolic process, and response to the stimulus. They act as protein binding, ion binding, and nucleic acid binding (**Figure 5A**). KEGG pathway analysis showed enrichment in the drug metabolism, Huntington disease, fatty acid degradation, peroxisome, IL-17 signaling pathway, and PPAR signaling pathway (**Figure 5B**).

## RNF183 Correlates With Tumor Purity and Immune Infiltration Level in UCEC

We investigated whether RNF183 expression was correlated with immune infiltration levels in UCEC from TIMER database. The results show that RNF183 expression has negatively correlations with tumor purity ($r = -0.063$, $p = 2.85E{-}01$), infiltrating levels of CD4 + T cells ($r = -0.064$, $p = 2.74E{-}01$), neutrophils ($r = -0.126$, $p = 3.17E{-}02$), and dendritic cells ($r = -0.042$, $p = 4.78E{-}01$) (**Figure 6A**). In addition, RNF183 CNV has significant correlations with infiltrating levels of CD8 + T cells, macrophages, and dendritic cells (**Figure 6B**).

## Correlation Analysis Between mRNA Levels of RNF183 and Markers of Different Subsets of Immune Cells

We further evaluated the relationship between the RNF183 level and immune infiltrating cells through the TIMER database based on the expression level of immune marker genes in UCEC tissues. The immune cells analyzed include $CD8^+$ T cells, $CD4^+$ T cells, B cells, monocytes, tumor-associated macrophages (TAM), M1 and M2 macrophages, neutrophils, and natural killer (NK) cells, dendritic cells, and besides, different subgroups of T cells, namely T helper 1 (Th1), Th2, Th17, regulatory T (Tregs), and T cell exhaustion. Because tumor purity will affect the level of immune infiltration of clinical samples, the purity of the relevant analysis was adjusted (**Table 1**).

Specifically, RNF183 expression showed significant correlation with the expression of markers of specific immune cells such as B cell, CD79A ($r = -0.121$; $P = 3.84e{-}02$), M1 macrophage marker, iNOS ($r = -0.256$; $P = 9.01e{-}06$), M2 Macrophage, CD163 ($r = -0.182$; $P = 1.77e{-}03$), VSIG4 ($r = -0.122$; $P = 3.71e{-}02$), MS4A4A ($r = -0.146$; $P = 1.21e{-}02$), Dendritic cell, HLA-DRA ($r = -0.182$; $P = 3.30e{-}02$), BDCA-1 ($r = -0.249$; $P = 1.60e{-}05$), BDCA-4 ($r = -0.172$; $P = 3.15e{-}03$). The expression of RNF183 correlated significantly with the expression of the marker genes of different subsets of T cells in UCEC, namely, Th1 markers, STAT1 ($r = -0.252$; $P = 1.25e{-}05$), IFN-γ ($r = -0.138$; $P = 1.80e{-}02$), Th2 markers, GATA3 ($r = -0.217$; $P = 1.77e{-}04$), STAT6 ($r = -0.197$; $P = 6.96e{-}04$), Treg markers, TGFβ ($r = -0.251$; $P = 1.37e{-}05$),

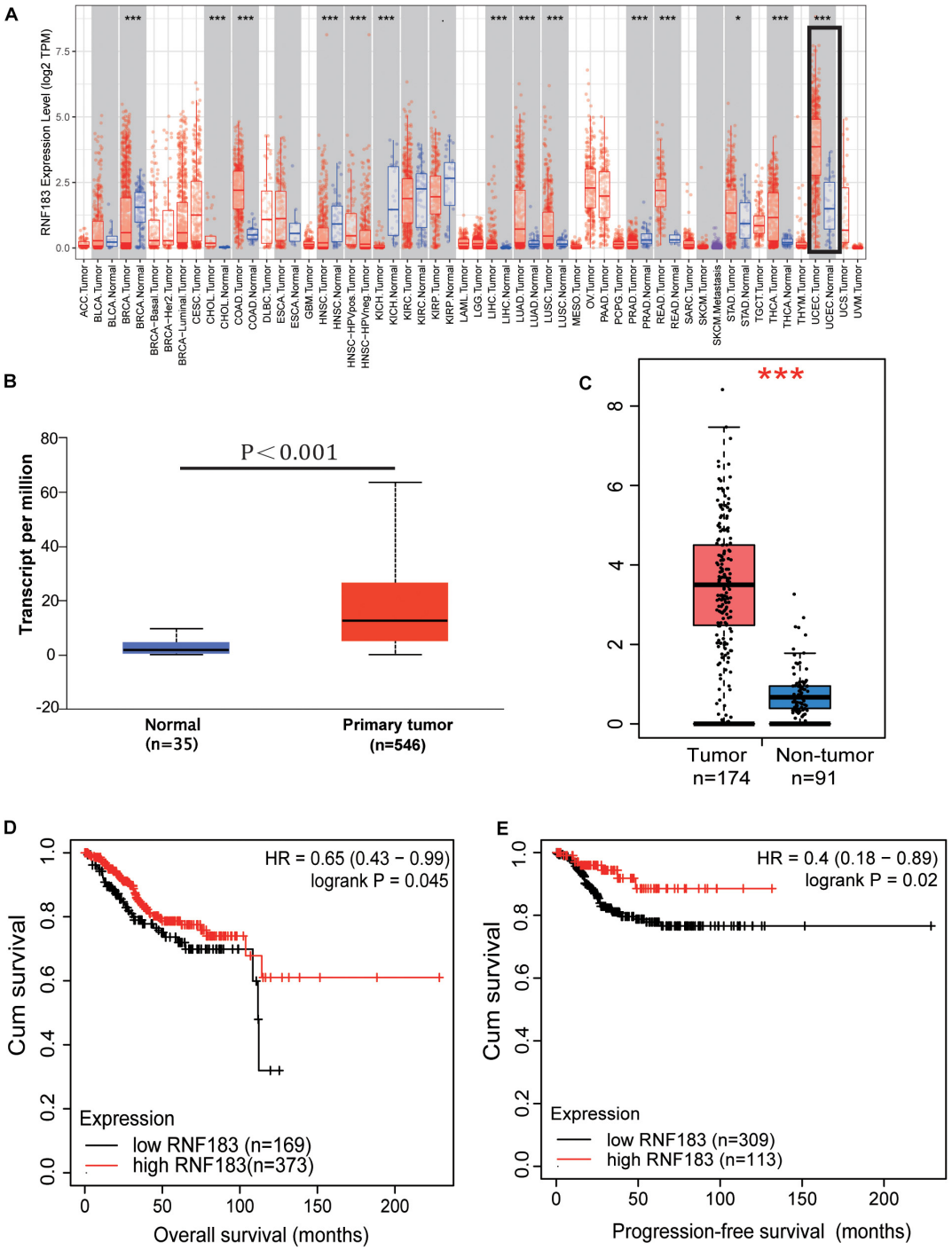



**FIGURE 1 |** Expression of RNF183 in human endometrial cancer. **(A)** The differential expressions of RNF183 between normal and tumor tissues exist in the majority of human cancers (TIMER). **(B,C)** RNF183 was elevated in endometrial cancer tissues compared to normal endometrium (UALCAN and GEPIA). Increased expression of RNF183 is associated with favorable prognosis of overall survival **(D)** and progression-free survival **(E)** in TCGA patients stratified at "best cut-off" (Kaplan-Meier Plotter Database). *$P$<0.05, ***$P$<0.001.

T cell exhaustion markers, PD-1 ($r = -0.217$; $P = 1.34e{-}02$), LAG3 ($r = -0.223$; $P = 5.83e{-}02$), GZMB ($r = -0.172$; $P = 3.31e{-}03$). RNF183 expression did not show any significant correlation with the expression of marker genes for $CD8^+$ T cells, T cell (general), Monocyte, TAM, Neutrophils, Natural killer cell and Th17 cells. These results demonstrated RNF183 expression were associated with infiltration of immune cells in UCEC (**Table 1**).

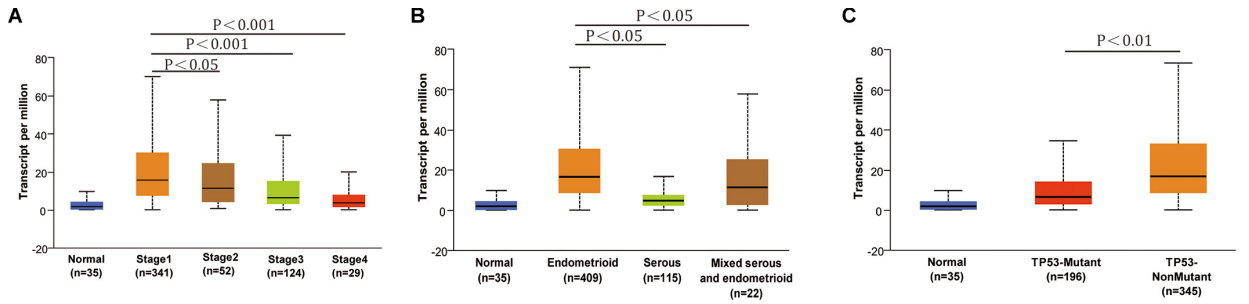



**FIGURE 2 |** The expression of RNF183 in particular subtypes of endometrial cancer (UALCAN). **(A)** RNF183 mRNA level in normal individuals or endometrial tumor stage 1,2,3, or 4. **(B)** RNF183 mRNA expression in normal individuals or histological subtypes (Endometrioid, Serous, or Mixed serous and endometrioid). **(C)** RNF183 mRNA level in normal individuals, in TP53-Mutant or TP53-Non-Mutant tumors.

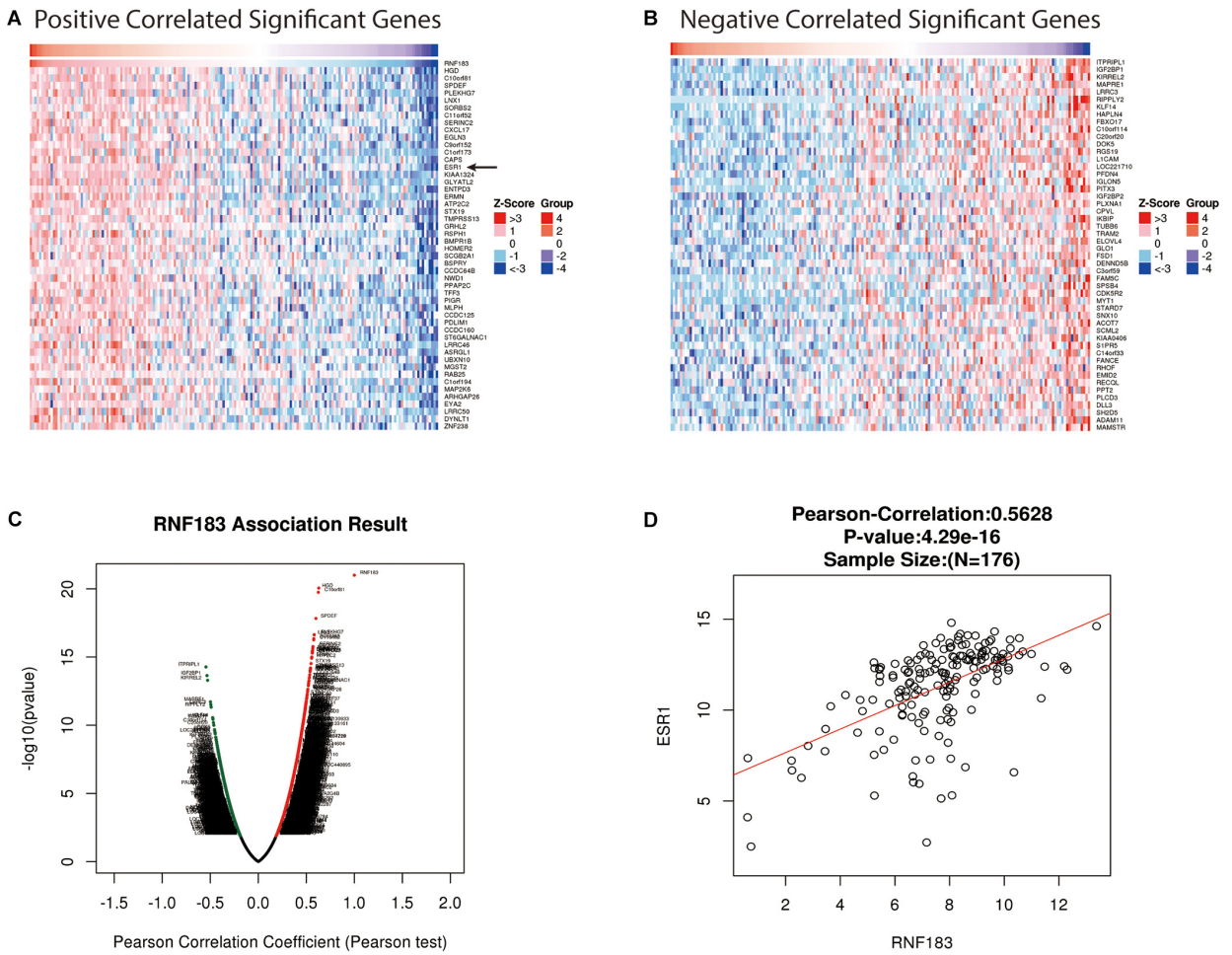



**FIGURE 3 |** Genes differentially expressed in correlation with RNF183 in UCEC (LinkedOmics). **(A–B)** Heat maps showing genes positively and negatively correlated with RNF183 in UCEC (TOP 50). **(C)** A Pearson test was used to assess correlations between RNF183 and genes differentially expressed in UCEC. Red indicates positively correlated genes and green indicates negatively correlated genes. **(D)** The scatter plot shows the Pearson correlation of RNF183 expression with ESR1.

## RNF183 Modulates ERα Expression of ERα Positive Endometrial Cancer Cell

Bioinformatics analysis via LinkedOmics, We found RNF183 was markedly positively correlated with ESR1 (**Figure 3D**). To verify this finding, we used the ERα-positive Ishikawa cell line as a model. Upon silencing of RNF183 using two different individual small interfering RNA, we detected a noticeable reduction in ESR1 mRNA levels (**Figure 7B**), and the knockdown

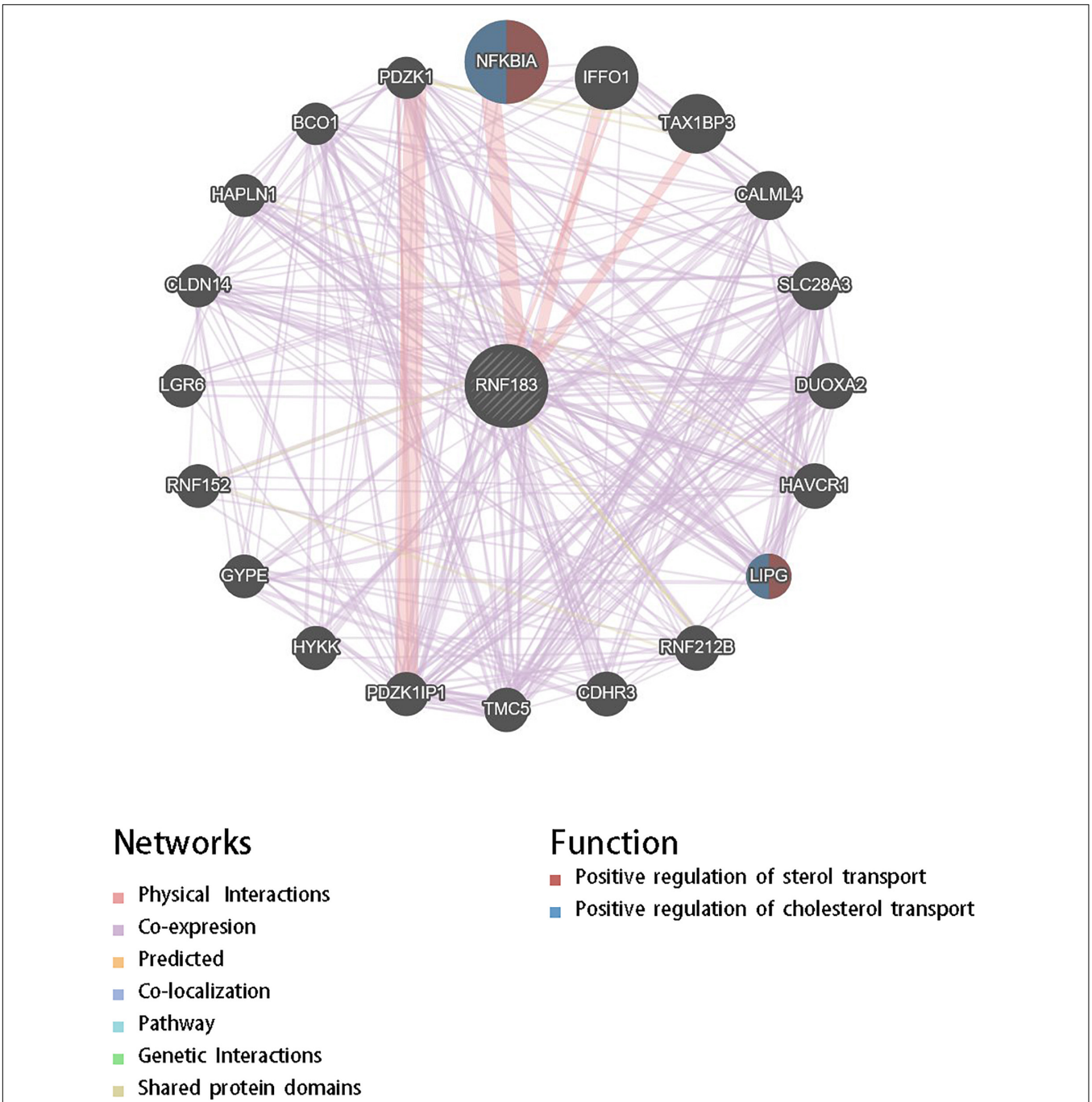



**FIGURE 4 |** Protein-protein interaction network of RNF183 networks (GeneMANIA). Protein-protein interaction (PPI) network and functional analysis revealed the enrichment of the target gene set of RNF183. The different colors at the edges of the network represent the applied enrichment methods: physical interactions, co-expression, predicted, co-localization, pathways, genetic interactions, and shared protein domains. The different colors of the network nodes represent the biological functions of the enriched gene set.

efficiency was shown in **Figure 7A**. Under stimulating E2 (17β-estradiol) or vehicle (absolute ethanol) conditions, ERα protein levels were also diminished following RNF183 silencing (**Figure 7C**). To determine the mechanism through which RNF183 regulates ERα, we assayed ERα luciferase reporter activity following RNF183 depletion or RNF183 overexpression. **Figure 7D** shows that the RNF183 knockdown suppressed the activity of the ERα reporter gene. While overexpression of RNF183 resulted in the raised activity of the ERα reporter gene no matter existence or absence of E2 stimulation (**Figure 7E**). As ubiquitin ligase has been reported to engage in the process of transcription upon the structure of the zinc finger domain

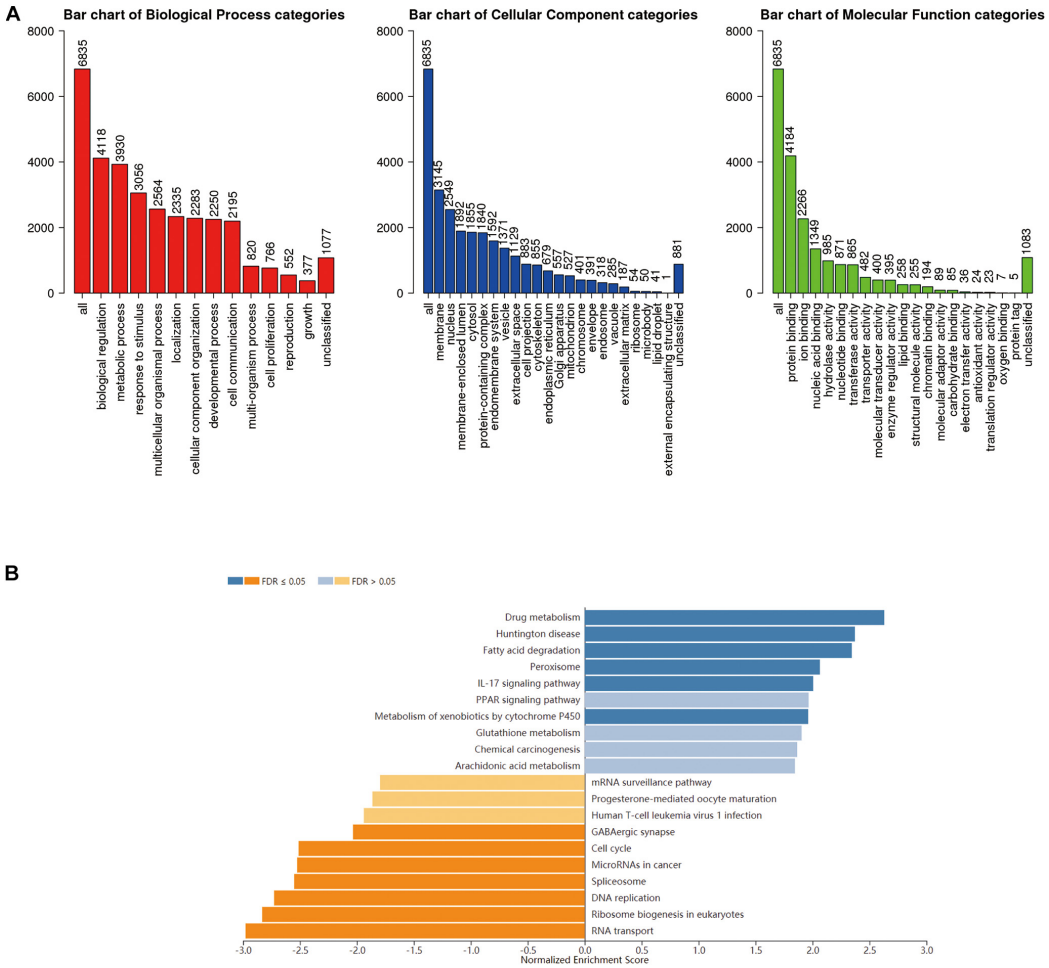



**FIGURE 5 |** Enriched GO annotations and KEGG pathways of RNF183 correlated genes in UCEC (LinkedOmics). **(A)** Biological process, Cellular Component and Molecular function analysis. **(B)** KEGG pathway analysis. Dark blue and orange indicate FDR ≤ 0.05, light blue and orange indicate FDR > 0.05 in **(B)**. FDR, false discovery rate.

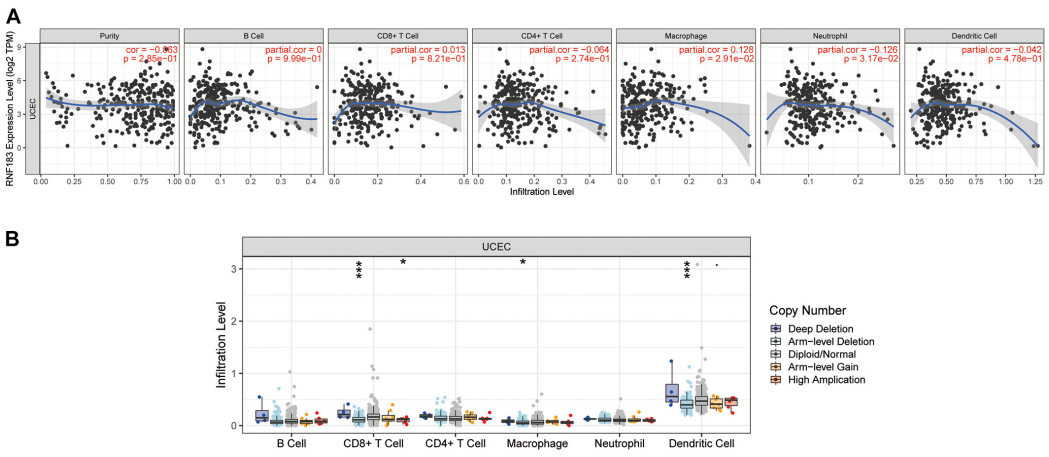



**FIGURE 6 |** Correlations of RNF183 expression with immune infiltration level in UCEC (TIMER). **(A)** RNF183 expression is negatively related to tumor purity, infiltrating levels of CD4 + T cells, neutrophils, and dendritic cells and has positively correlations with infiltrating levels of macrophages in UCEC. **(B)** RNF183 CNV affects the infiltrating levels of CD8 + T cells, macrophages, and dendritic cells in UCEC. $^{*}P < 0.05$, $^{***}P < 0.001$.

**TABLE 1 |** Correlation analysis between RNF183 and relate genes and markers of immune cells in TIMER.

| Description | Gene markers | UCEC | | | |
|---|---|---|---|---|---|
| | | None | | Purity | |
| | | Cor | P | Cor | P |
| CD8 + T cell | CD8A | −0.052 | 2.22e−01 | −0.119 | 4.16e−02 |
| | CD8B | −0.027 | 5.22e−01 | −0.031 | 6.02e−01 |
| T cell (general) | CD3D | −0.046 | 2.82e−01 | 0 | 9.98e−01 |
| | CD3E | −0.037 | 3.86e−01 | −0.027 | 6.49e−01 |
| | CD2 | −0.015 | 7.22e−01 | −0.018 | 7.64e−01 |
| B cell | CD19 | −0.026 | 5.52 e−01 | −0.051 | 3.81e−01 |
| | CD79A | −0.023 | 5.88e−01 | −0.121 | * |
| Monocyte | CD86 | −0.094 | * | −0.087 | 1.36e−01 |
| | CD115 | −0.025 | 5.57e−01 | −0.011 | 8.54e−01 |
| TAM | CCL2 | −0.013 | 7.85e−01 | 0.048 | 4.09e−01 |
| | CD68 | −0.07 | 1.01e−01 | −0.066 | 2.60e−01 |
| M1 Macrophage | iNOS | −0.26 | 7.61e−01 | −0.256 | *** |
| | IRF5 | −0.074 | 8.38e−02 | −0.107 | 6.65e−02 |
| | COX2 | −0.12 | ** | 0.085 | 1.46e−01 |
| M2 Macrophage | CD163 | −0.183 | *** | −0.182 | ** |
| | VSIG4 | −0.114 | ** | −0.122 | * |
| | MS4A4A | −0.131 | ** | −0.146 | * |
| Neutrophils | CD66b | 0.07 | 1e−01 | 0.023 | 6.92e−01 |
| | CD11b | 0.089 | * | 0.092 | 1.16e−01 |
| | CCR7 | 0.014 | 7.44e−01 | −0.047 | 4.18e−01 |
| Natural killer cell | KIR2DL1 | 0.028 | 5.16e−01 | 0.005 | 9.34e−01 |
| | KIR2DL3 | 0.028 | 5.17e−01 | −0.029 | 6.27e−01 |
| | KIR2DL4 | 0.039 | 3.67e−01 | −0.032 | 5.82e−01 |
| | KIR3DL1 | 0.045 | 2.94e−01 | 0.016 | 7.81e−01 |
| | KIR3DL2 | −0.74 | 8.43e−02 | −0.027 | 6.41e−01 |
| | KIR3DL3 | −0.004 | 9.23e−01 | −0.034 | 5.62e−01 |
| | KIR2DL4 | 0.039 | 3.67e−01 | −0.032 | 5.82e−01 |
| Dendritic cell | HLA-DPB1 | 0.083 | 5.27e−02 | 0.055 | 3.44e−01 |
| | HLA-DQB1 | 0.134 | ** | 0.099 | 9.00e−02 |
| | HLA-DRA | 0.144 | *** | 0.125 | * |
| | HLA-DPA1 | 0.051 | 2.33e−01 | 0.018 | 7.60e−01 |
| | BDCA-1 | 0.275 | *** | 0.249 | *** |
| | BDCA-4 | 0.219 | *** | 0.172 | ** |
| | CD11c | 0.094 | * | 0.098 | 9.52e−02 |
| Th1 | T-bet | 0.025 | 5.53e−01 | −0.048 | 4.16e−01 |
| | STAT4 | 0.067 | 1.18e−01 | −0.037 | 5.29e−01 |
| | STAT1 | −0.256 | *** | −0.252 | *** |
| | IFN-γ | −0.104 | * | −0.138 | ** |
| | TNF-α | 0.044 | 3.06e−01 | 0.004 | 9.50e−01 |
| Th2 | GATA3 | −0.113 | ** | −0.217 | *** |
| | STAT6 | 0.25 | *** | 0.197 | *** |
| | STAT5A | 0.014 | 7.47e−01 | −0.009 | 8.73e−01 |
| | IL13 | −0.021 | 6.18e−01 | 0.007 | 9.90e−01 |
| | IL21 | −0.007 | 8.69e−01 | −0.058 | 3.23e−01 |
| Th17 | STAT3 | 0.202 | *** | −0.156 | 7.43e−03 |
| | IL17A | 0.005 | 9.15e−01 | 0 | 9.96e−01 |
| Treg | FOXP3 | 0.008 | 8.6e−01 | −0.083 | 1.57e−01 |
| | CCR8 | 0.067 | 1.18e−01 | −0.001 | 9.93e−01 |
| | STAT5B | 0.035 | 4.17e−01 | −0.076 | 1.97e−01 |
| | TGFβ | −0.172 | *** | −0.251 | *** |
| T cell exhaustion | PD-1 | −0.12 | ** | −0.144 | ** |
| | CTLA4 | 0.038 | 3.72e−01 | −0.002 | 9.78e−01 |
| | LAG3 | −0.222 | *** | −0.296 | *** |
| | TIM-3 | −0.078 | 6.81e−02 | −0.097 | 9.62e−02 |
| | GZMB | −0.151 | *** | −0.172 | ** |

*TAM, tumor-associated macrophage; Th, T helper cell; Treg, regulatory T cell; Cor, P value of Spearman's correlation; None, correlation without adjustment. Purity, correlation adjusted by purity. *P < 0.05, **P < 0.01, ***P < 0.001.*

(Molloy et al., 2018), we generated a curtailed form of RNF183 (RNF183△t) without E3 ubiquitin ligase activity by deleting zinc finger domain (amino acids 1–60). This truncation mostly canceled the function of RNF183 in stimulating the activity of the ERα luciferase report gene (**Figure 7E**). RNF183 depletion also reduced the expression of endogenous ERα target genes dependent on E2 stimulation such as TFF1, PGR, FOXA1, and XBP1 (**Figure 7F**). Furthermore, TFF1, PGR, FOXA1, and XBP1 showed markedly positive correlation with RNF183 from TIMER database (**Figure 7G**).

## ERα Mediates RNF183 Stability in ERα Positive Endometrial Cancer Cell

Given that ERα has been shown to participate in the feedback loop with some enzymes and transcription factors (Eeckhoute et al., 2007; Molloy et al., 2018), the impact of ERα on the expression of RNF183 was analyzed in the Ishikawa cell line. We noticed that ERα knockdown had little effect on the mRNA level of RNF183 (**Figure 8A**). However, there was a marked decline in the RNF183 protein level (**Figure 8B**). Next, in the presence of proteasome inhibitor MG132, RNF183 was in a stable state, even being with siERα (**Figure 8C**). Furthermore, ERα inhibition expedited the reduction of RNF183 protein expression in the presence of protein synthesis inhibitor cycloheximide (**Figure 8D**). In sum, these data indicate that ERα raises RNF183 protein stability in the ERα-positive endometrial cancer cells.

# DISCUSSION

RNF183 has been reported to occur in diverse diseases such as colorectal cancer (CRC), kidney disease, inflammatory bowel disease and various biological processes. RNF183 stimulated inflammatory bowel disease progression (Yu et al., 2016). RNF183 was also identified as an oncogene promoting proliferation, metastasis, and a resistance gene for trametinib in CRC cells via activating the NF-κB signal (Geng et al., 2017). In the renal medullary collecting duct, specific RNF183 controlled

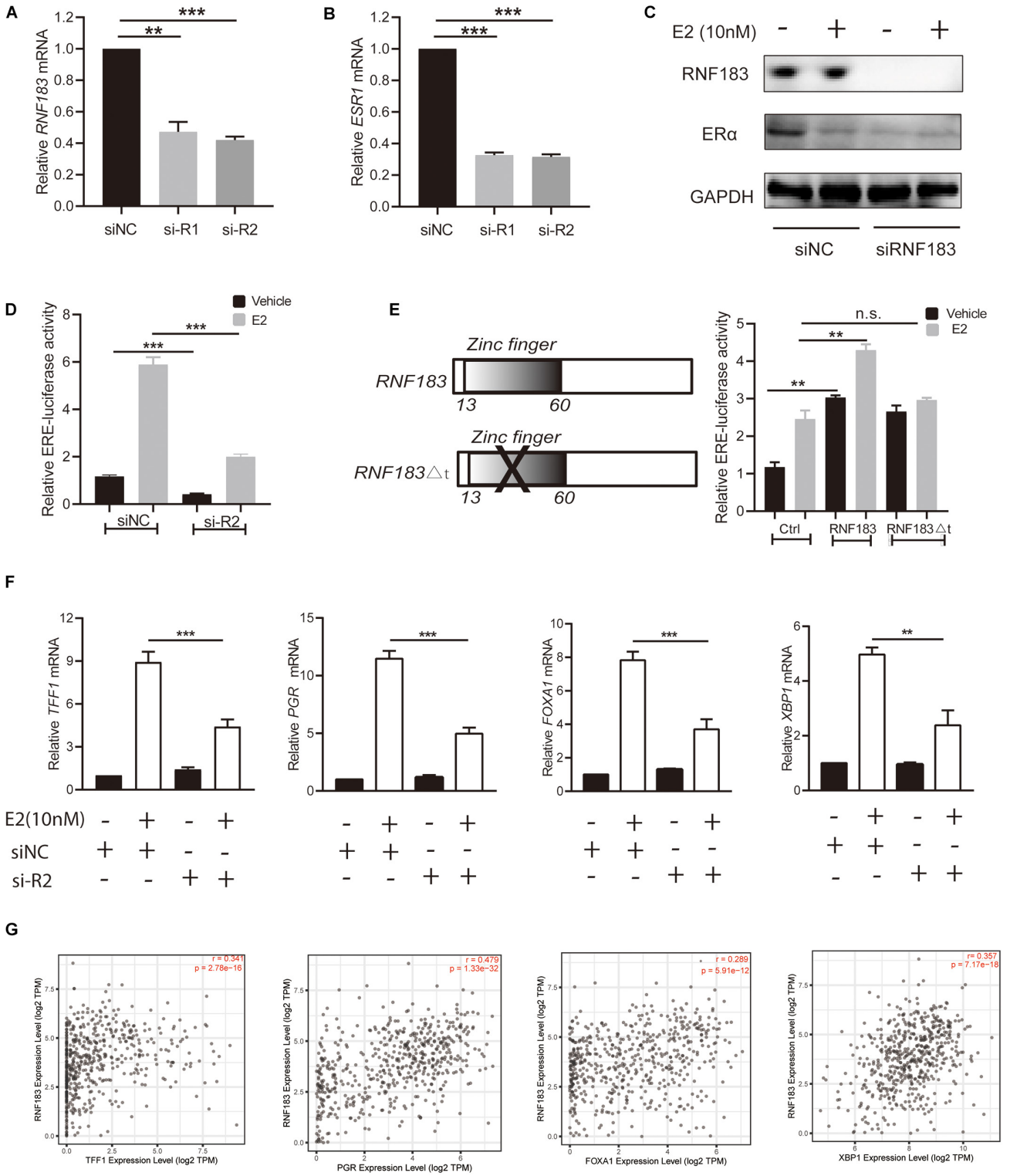



**FIGURE 7 |** RNF183 controls ERα expression in endometrial cancer. **(A)** Ishikawa cells transfected with siRNF183 or siNC and RNF183 knockdown efficiency were examined by RT-PCR. **(B)** ESR1 mRNA level decreased in the Ishikawa cell line after transfection with siRNF183. **(C)** The protein level of ERα was downregulated based on RNF183 deletion. **(D)** RNF183 deletion decreased ERα-dependent expression of the ERE-luciferase activity. **(E)** The ERE-Luciferase activity was evaluated in Ishikawa cells with overexpression of pcDNA4-myc/his-RNF183 or pcDNA4-myc/his vector or truncated RNF183 without E3 ubiquitin ligase activity (13–60 amino acids). **(F)** Diminished E2 induced reduction of ERα target genes following inhibition of RNF183 with siRNA. **(G)** RNF183 positively associated with TFF1, PGR, FOXA1 and XBP1 from the TIMER database. Experiments were repeated in triplicates. Mean ± S.D. ($n = 3$). $^{**}P<0.01$, $^{***}P<0.001$.

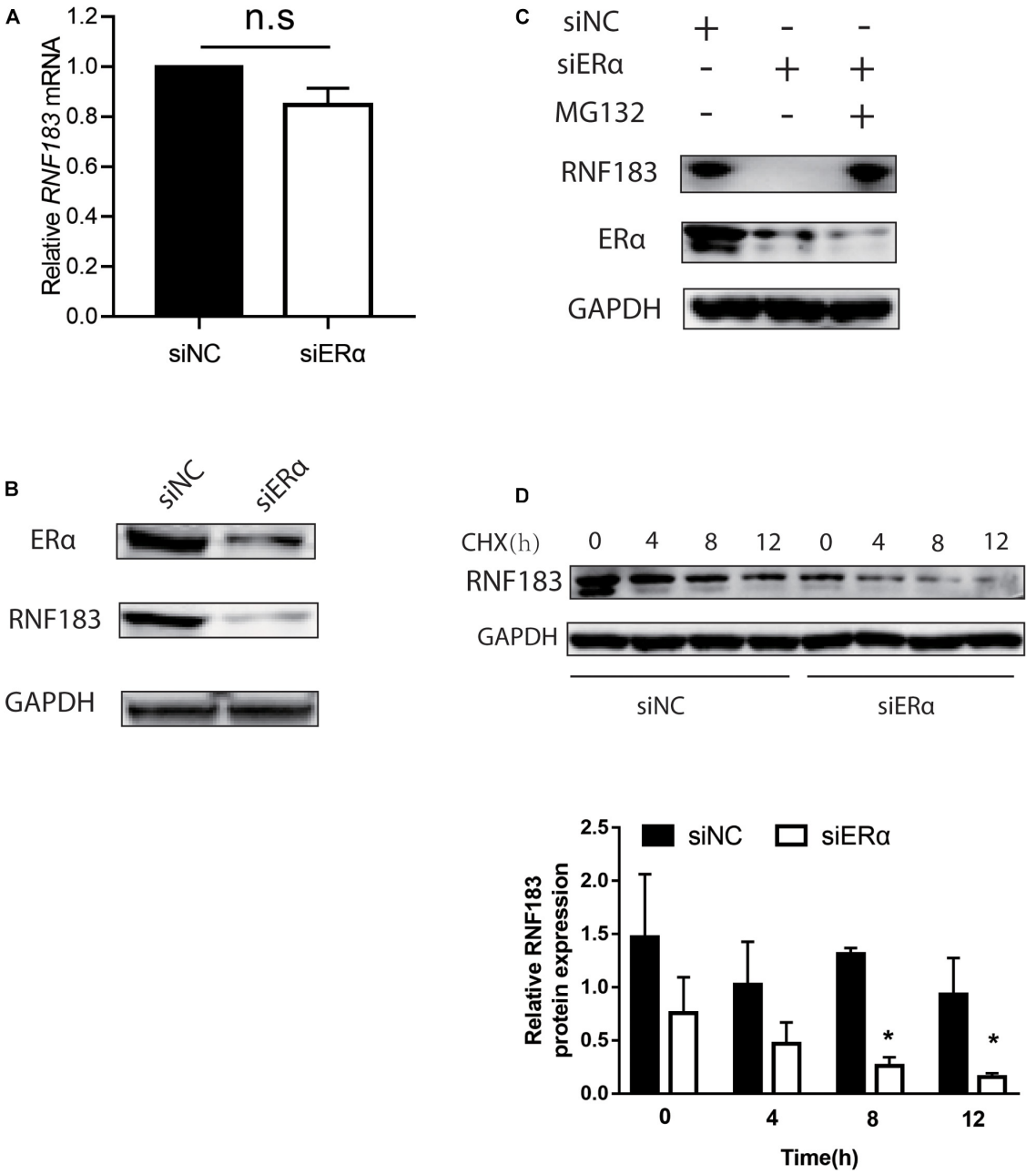



**FIGURE 8 |** ERα regulates RNF183 expression and increases its stability. The expression of endogenous RNF183 mRNA **(A)** and protein levels **(B)** in Ishikawa cells after transfecting with siERα or siNC. **(C)** Downregulated RNF183 protein level reduced by siERα was recovered based on MG132 treatment. **(D)** Depletion ERα weakens RNF183 stability. RNF183 protein level examined at indicated time after transfected with siERα 48h followed 100μg/mL cycloheximide treatment. Experiments were repeated in triplicates. Mean ± S.D. ($n = 3$). *$P$<0.05.

cell adaption to hypertonic stress by regulating Na, K-ATPase level (Maeoka et al., 2019b; Okamoto et al., 2020b). Under physiological condition, RNF183 localizing on the endoplasmic reticulum, interacted and ubiquitin-mediated degradation of Bcl-xL, suggesting a crucial role of RNF183 in executing programmed cell death (Wu et al., 2018). The results of our study showed that significantly amplification of RNF183 was considered as a prognostic marker in endometrial cancer. Analysis implied that among endometrial cancer, High RNF183 expression seems to associate with low stage, endometrioid and TP53-Non-Mutant status, which are usually with a good prognosis. Also, the RNF183 expression was greater at higher expression and the tumor stage was greater at the lower level, implying the early role of RNF183 in the development of endometrial cancer.

Based on the marker levels of different immune cell types in UCEC, RNF183 mRNA level is correlated with the number of

tumor infiltrating immune cells, which indicates that RNF183 plays a vital role in regulating tumor immunity. We observed that the expression level of RNF183 mRNA was negatively correlated with CD4 + T cells, neutrophils, and dendritic cells. We also observed the correlation between the levels of RNF183 mRNA and the expression of the B cell (CD79A), M1 macrophage marker (iNOS), M2 Macrophage (CD163, VSIG4, and MS4A4A), Dendritic cell (HLA-DRA, BDCA- 1, and BDCA-4). The expression of RNF183 is also related to the markers in different subgroups of T helper (Th) cells, including Th1 (STAT-1, and IFN-γ), Th2 (GATA3 and STAT6), Treg (TGF-β), T cell exhaustion markers (PD-1, LAG3, and GZMB). Above indicate the role of RNF183 in regulating tumor invasion of T helper cells.

Moreover, the depleted T cell markers PD-1, LAG3 and GZMB, which are critical inhibitory immune checkpoint proteins, are negatively correlated with the expression of RNF183. The expression of PD-1 (Kucukgoz Gulec et al., 2019) is considered a sign of poor prognosis of endometrial cancer and it has been widely studied as a target of immunotherapy. LAG3 (Friedman et al., 2020) can be used as a target for immunotherapy in endometrial cancer and in conjunction with other immune checkpoints, such as PD-1. Besides, Granzyme B + cells (Pakish et al., 2017) have increased expression in high microsatellite instability (MSI-H) endometrial cancer, providing a therapeutic target for immunotherapy. We speculate that RNF183, which is highly expressed in the tumor microenvironment, leads to a better prognosis of UCEC by regulating the expression of inhibitory immune checkpoint proteins PD-1, LAG3 and GZMB on exhausted T cells. However, this assumption needs further verification.

Through heatmap about the top 50 genes positively correlated with RNF183, we found ESR1 was one of the most notable positive genes with RNF183. Most endometrial cancers are estrogen-related endometrioid adenocarcinomas. Beyond 90% endometrioid carcinoma express moderate to high levels of the ERα (gene symbol ESR1) (Lebeau et al., 2008; Smith et al., 2018). A consensus that patients with tumor positive ERα expression have a favorable prognosis of endometrial cancer (Creasman et al., 1980; Iversen et al., 1988; Jongen et al., 2009; **Supplementary Figure S1**). To confirm the regulatory relationships between RNF183 and ERα, We used ERα-positive cell line Ishikawa as a model to examine. We clarify a character for RNF183 in promoting ERα expression at the transcript and protein level in endometrial cancer. ERα is a substrate for E3 ubiquitin ligase (Byun and Jung, 2008; Zhang et al., 2015). We proved that RNF183 controls ERα activity determined by the RING finger domain. The transcriptional activation of estrogen bound ERα is tissue-specific (Kushner et al., 2000; Castro-Rivera and Safe, 2003). PGR (Progesterone receptor), FOXA1 (Forkhead box protein A1), XBP1 (X-box binding protein 1), and TFF1 (Trefoil factor 1) were reported to involve in the estrogen signal in endometrial cancer (Baxter et al., 2019). Clinical samples favor RNF183 positively correlated with TFF1, FOXA1, XBP1, and PGR.

Noticeable studies showed that ERα often participates in the positive regulation with the related gene, such as autocrine loop of the CXCR4/SDF-1 and ERα/ERβ signaling pathways (Sauvé et al., 2009), S6K1-ERα and ER-α36/EGFR positive feed-forward loop (Zhang et al., 2011; Maruani et al., 2012). We found that although the essential role of ERα is a transcription factor, knockdown ERα does not affect the mRNA level of RNF183, but a decreased in protein level. ERα could mediate ubiquitination and protein degradation had been reported (Lai et al., 2019). Our further results demonstrated that RNF183 protein level recovery followed proteasome inhibitor MG132 treated signifying ERα impeded RNF183 reduction by the inhibition of the proteasome, and ERα depletion stimulated RNF183 degradation.

## CONCLUSION

Our results indicate that RNF183 is a potential independent prognostic biomarker of UCEC, which can also be used to assess the level of immune cell infiltration in tumor tissues. Furthermore, ERα plays a vital role in the histology and progression of endometrial cancer. We found that RNF183 seems to be a new marker associated with ERα in ERα-positive endometrial cancer. Furthermore, the crosstalk between RNF183 and ERα may be the reason for the abnormally high expression of RNF183 in endometrial cancer.

## 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

LRM, GXL, and GR: conceptualization. LRM and GXL: resources. GR, ZLJ, HXB, QR, and ZRJ: formal analysis and investigation. GR and ZYH: data curation. GR: writing—original draft preparation, funding acquisition. ZYH: writing—review and editing, supervision. All authors have read and agreed to the published version of the manuscript.

## FUNDING

This study was supported by the Science and Technology Planning Project of Foshan City of China (2018AB000281) and the National Natural Science Foundation of China (81901453).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.595733/full#supplementary-material

**Supplementary Figure 1 |** ERα is a strong predictor for prognosis in endometrial cancer. Patients with high expression of ERα harbor good OS **(A)** and PFS **(B)**.

**Supplementary Table 1 |** The primer sequence of mRNA for qPCR in article.

## REFERENCES


Baxter, E., Windloch, K., Kelly, G., Lee, J. S., Gannon, F., and Brennan, D. J. (2019). Molecular basis of distinct oestrogen responses in endometrial and breast cancer. *Endocr. Relat. Cancer* 26, 31–46. doi: 10.1530/erc-17-0563

Berg, A., Gulati, A., Ytre-Hauge, S., Fasmer, K. E., Mauland, K. K., Hoivik, E. A., et al. (2017). Preoperative imaging markers and PDZ-binding kinase tissue expression predict low-risk disease in endometrial hyperplasias and low grade cancers. *Oncotarget* 8, 68530–68541. doi: 10.18632/oncotarget.19708

Byun, B., and Jung, Y. (2008). Repression of transcriptional activity of estrogen receptor alpha by a Cullin3/SPOP ubiquitin E3 ligase complex. *Mol. Cell.* 25, 289–293.

Cao, Y., Sun, Y., Chang, H., Sun, X., and Yang, S. (2019). The E3 ubiquitin ligase RNF182 inhibits TLR-triggered cytokine production through promoting p65 ubiquitination and degradation. *FEBS Lett.* 593, 3210–3219. doi: 10.1002/1873-3468.13583

Castro-Rivera, E., and Safe, S. (2003). 17 beta-estradiol- and 4-hydroxytamoxifen-induced transactivation in breast, endometrial and liver cancer cells is dependent on ER-subtype, cell and promoter context. *J. Steroid Biochem. Mol. Biol.* 84, 23–31. doi: 10.1016/s0960-0760(03)00010-4

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia (New York, NY)* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Colas, E., Perez, C., Cabrera, S., Pedrola, N., Monge, M., Castellvi, J., et al. (2011). Molecular markers of endometrial carcinoma detected in uterine aspirates. *Int. J. Cancer* 129, 2435–2444. doi: 10.1002/ijc.25901

Creasman, W. T., McCarty, K. S. Sr., Barton, T. K., and McCarty, K. S. Jr. (1980). Clinical correlates of estrogen- and progesterone-binding proteins in human endometrial adenocarcinoma. *Obstet. Gynecol.* 55, 363–370. doi: 10.1097/00006250-198003000-00019

Eeckhoute, J., Keeton, E. K., Lupien, M., Krum, S. A., Carroll, J. S., and Brown, M. (2007). Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res.* 67, 6477–6483. doi: 10.1158/0008-5472.can-07-0746

Friedman, L. A., Ring, K. L., and Mills, A. M. (2020). LAG-3 and GAL-3 in Endometrial Carcinoma: emerging candidates for immunotherapy. *Int. J. Gynecol. Pathol.* 39, 203–212. doi: 10.1097/pgp.0000000000000608

Geng, R., Tan, X., Wu, J., Pan, Z., Yi, M., Shi, W., et al. (2017). RNF183 promotes proliferation and metastasis of colorectal cancer cells via activation of NF-κB-IL-8 axis. *Cell Death Dis.* 8:e2994. doi: 10.1038/cddis.2017.400

Hashizume, R., Fukuda, M., Maeda, I., Nishikawa, H., Oyake, D., Yabuki, Y., et al. (2001). The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. *J. Biol. Chem.* 276, 14537–14540. doi: 10.1074/jbc.c000881200

Iversen, O. E., Utaaker, E., and Skaarland, E. (1988). DNA ploidy and steroid receptors as predictors of disease course in patients with endometrial carcinoma. *Acta Obstet. Gynecol. Scand.* 67, 531–537. doi: 10.3109/00016348809029865

Joazeiro, C. A., and Weissman, A. M. (2000). RING finger proteins: mediators of ubiquitin ligase activity. *Cell* 102, 549–552.

Jongen, V., Briët, J., de Jong, R., ten Hoor, K., Boezen, M., van der Zee, A., et al. (2009). Expression of estrogen receptor-alpha and -beta and progesterone receptor-A and -B in a large cohort of patients with endometrioid endometrial cancer. *Gynecol. Oncol.* 112, 537–542. doi: 10.1016/j.ygyno.2008.10.032

Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., et al. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature* 497, 67–73. doi: 10.1038/nature12113

Kaneko, M., Iwase, I., Yamasaki, Y., Takai, T., Wu, Y., Kanemoto, S., et al. (2016). Genome-wide identification and gene expression profiling of ubiquitin ligases for endoplasmic reticulum protein degradation. *Sci. Rep.* 6:30955.

Kucukgoz Gulec, U., Kilic Bagir, E., Paydas, S., Guzel, A. B., Gumurdulu, D., and Vardar, M. A. (2019). Programmed death-1 (PD-1) and programmed death-ligand 1 (PD-L1) expressions in type 2 endometrial cancer. *Archiv. Gynecol. Obstet.* 300, 377–382. doi: 10.1007/s00404-019-05180-2

Kushner, P. J., Agard, D., Feng, W. J., Lopez, G., Schiau, A., Uht, R., et al. (2000). Oestrogen receptor function at classical and alternative response elements. *Novartis Found. Symp.* 230, 20–26. discussion 27-40, doi: 10.1002/04708708 18.ch3

Lai, Y. J., Zhu, B. L., Sun, F., Luo, D., Ma, Y. L., Luo, B., et al. (2019). Estrogen receptor α promotes Cav1.2 ubiquitination and degradation in neuronal cells and in APP/PS1 mice. *Aging Cell* 18:e12961.

Lánczky, A., Nagy, Á, Bottai, G., Munkácsy, G., Szabó, A., Santarpia, L., et al. (2016). miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res. Treat.* 160, 439–446. doi: 10.1007/s10549-016-4013-7

Lebeau, A., Grob, T., Holst, F., Seyedi-Fazlollahi, N., Moch, H., Terracciano, L., et al. (2008). Oestrogen receptor gene (ESR1) amplification is frequent in endometrial carcinoma and its precursor lesions. *J. Pathol.* 216, 151–157. doi: 10.1002/path.2405

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17:174.

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.

Lipkowitz, S., and Weissman, A. M. (2011). RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis. *Nat. Rev. Cancer* 11, 629–643. doi: 10.1038/nrc3120

Liu, Q. Y., Lei, J. X., Sikorska, M., and Liu, R. (2008). A novel brain-enriched E3 ubiquitin ligase RNF182 is up regulated in the brains of Alzheimer's patients and targets ATP6V0C for degradation. *Mol. Neurodegener.* 3:4. doi: 10.1186/1750-1326-3-4

Maeoka, Y., Okamoto, T., Wu, Y., Saito, A., Asada, R., Matsuhisa, K., et al. (2019a). Renal medullary tonicity regulates RNF183 expression in the collecting ducts via NFAT5. *Biochem. Biophys. Res. Commun.* 514, 436–442. doi: 10.1016/j.bbrc.2019.04.168

Maeoka, Y., Wu, Y., Okamoto, T., Kanemoto, S., Guo, X. P., Saito, A., et al. (2019b). NFAT5 up-regulates expression of the kidney-specific ubiquitin ligase gene Rnf183 under hypertonic conditions in inner-medullary collecting duct cells. *J. Biol. Chem.* 294, 101–115. doi: 10.1074/jbc.ra118.002896

Maruani, D. M., Spiegel, T. N., Harris, E. N., Shachter, A. S., Unger, H. A., Herrero-González, S., et al. (2012). Estrogenic regulation of S6K1 expression creates a positive regulatory loop in control of breast cancer cell proliferation. *Oncogene* 31, 5073–5080. doi: 10.1038/onc.2011.657

Moldovan, G. L., and D'Andrea, A. D. (2009). How the fanconi anemia pathway guards the genome. *Ann. Rev. Genet.* 43, 223–249. doi: 10.1146/annurev-genet-102108-134222

Molloy, M. E., Lewinska, M., Williamson, A. K., Nguyen, T. T., Kuser-Abali, G., Gong, L., et al. (2018). ZBTB7A governs estrogen receptor alpha expression in breast cancer. *J. Mol. Cell Biol.* 10, 273–284. doi: 10.1093/jmcb/mjy020

Nectoux, J., Fichou, Y., Rosas-Vargas, H., Cagnard, N., Bahi-Buisson, N., Nusbaum, P., et al. (2010). Cell cloning-based transcriptome analysis in Rett patients: relevance to the pathogenesis of Rett syndrome of new human MeCP2 target genes. *J. Cell. Mol. Med.* 14, 1962–1974. doi: 10.1111/j.1582-4934.2010.01107.x

Okamoto, T., Imaizumi, K., and Kaneko, M. (2020a). The role of tissue-specific Ubiquitin Ligases, RNF183, RNF186, RNF182 and RNF152, in disease and biological function. *Int. J. Mol. Sci.* 21:3921. doi: 10.3390/ijms21113921

Okamoto, T., Wu, Y., Matsuhisa, K., Saito, A., Sakaue, F., Imaizumi, K., et al. (2020b). Hypertonicity-responsive ubiquitin ligase RNF183 promotes Na, K-ATPase lysosomal degradation through ubiquitination of its β1 subunit. *Biochem. Biophys. Res. Commun.* 521, 1030–1035. doi: 10.1016/j.bbrc.2019.11.001

Oliner, J. D., Kinzler, K. W., Meltzer, P. S., George, D. L., and Vogelstein, B. (1992). Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* 358, 80–83. doi: 10.1038/358080a0

Pakish, J. B., Zhang, Q., Chen, Z., Liang, H., Chisholm, G. B., Yuan, Y., et al. (2017). Immune microenvironment in microsatellite-instable endometrial cancers: hereditary or sporadic origin matters. *Clin. Cancer Res.* 23, 4473–4481. doi: 10.1158/1078-0432.ccr-16-2655

Ruffner, H., Joazeiro, C. A., Hemmati, D., Hunter, T., and Verma, I. M. (2001). Cancer-predisposing mutations within the RING domain of BRCA1: loss of

ubiquitin protein ligase activity and protection from radiation hypersensitivity. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5134–5139. doi: 10.1073/pnas.081068398

Sauvé, K., Lepage, J., Sanchez, M., Heveker, N., and Tremblay, A. (2009). Positive feedback activation of estrogen receptors by the CXCL12-CXCR4 pathway. *Cancer Res.* 69, 5793–5800. doi: 10.1158/0008-5472.can-08-4924

Smith, D., Stewart, C. J. R., Clarke, E. M., Lose, F., Davies, C., Armes, J., et al. (2018). ER and PR expression and survival after endometrial cancer. *Gynecol. Oncol.* 148, 258–266. doi: 10.1016/j.ygyno.2017.11.027

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.

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963.

Wade, M., Wang, Y. V., and Wahl, G. M. (2010). The p53 orchestra: Mdm2 and Mdmx set the tone. *Trends Cell Biol.* 20, 299–309. doi: 10.1016/j.tcb.2010.01.009

Wang, J. H., Wei, Z. F., Gao, Y. L., Liu, C. C., and Sun, J. H. (2018). Activation of the mammalian target of rapamycin signaling pathway underlies a novel inhibitory role of ring finger protein 182 in ventricular remodeling after myocardial ischemia-reperfusion injury. *J. Cell. Biochem* doi: 10.1002/jcb.28038 [Epub ahead of print].

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220.

Wu, Y., Li, X., Jia, J., Zhang, Y., Li, J., Zhu, Z., et al. (2018). Transmembrane E3 ligase RNF183 mediates ER stress-induced apoptosis by degrading Bcl-xL. *Proc. Natl. Acad. Sci. U. S. A.* 115, E2762–E2771.

Yu, Q., Zhang, S., Chao, K., Feng, R., Wang, H., Li, M., et al. (2016). E3 Ubiquitin ligase RNF183 Is a novel regulator in inflammatory bowel disease. *J. Crohns colitis* 10, 713–725. doi: 10.1093/ecco-jcc/jjw023

Zhang, P., Gao, K., Jin, X., Ma, J., Peng, J., Wumaier, R., et al. (2015). Endometrial cancer-associated mutants of SPOP are defective in regulating estrogen receptor-α protein turnover. *Cell Death Dis.* 6:e1687. doi: 10.1038/cddis.2015.47

Zhang, X. T., Kang, L. G., Ding, L., Vranic, S., Gatalica, Z., and Wang, Z. Y. (2011). A positive feedback loop of ER-α36/EGFR promotes malignant growth of ER-negative breast cancer cells. *Oncogene* 30, 770–780. doi: 10.1038/onc.2010.458

**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: 02 December 2020
doi: 10.3389/fgene.2020.596078

# Genome-Wide Analysis of Cell-Free DNA Methylation Profiling for the Early Diagnosis of Pancreatic Cancer

*Shengyue Li[1†], Lei Wang[2†], Qiang Zhao[3], Zhihao Wang[1], Shuxian Lu[1], Yani Kang[3], Gang Jin[4*] and Jing Tian[1*]*

[1] Key laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, School of Medicine, Northwest University, Xi'an, China, [2] Department of Gastroenterology, Changhai Hospital, Second Military Medical University, Shanghai, China, [3] School of Biomedical Engineering, Bio-ID Center, Shanghai Jiao Tong University, Shanghai, China, [4] Department of General Surgery, Changhai Hospital, Second Military Medical University, Shanghai, China

As one of the most malicious cancers, pancreatic cancer is difficult to treat due to the lack of effective early diagnosis. Therefore, it is urgent to find reliable diagnostic and predictive markers for the early detection of pancreatic cancer. In recent years, the detection of circulating cell-free DNA (cfDNA) methylation in plasma has attracted global attention for non-invasive and early cancer diagnosis. Here, we carried out a genome-wide cfDNA methylation profiling study of pancreatic ductal adenocarcinoma (PDAC) patients by methylated DNA immunoprecipitation coupled with high-throughput sequencing (MeDIP-seq). Compared with healthy individuals, 775 differentially methylated regions (DMRs) located in promoter regions were identified in PDAC patients with 761 hypermethylated and 14 hypomethylated regions; meanwhile, 761 DMRs in CpG islands (CGIs) were identified in PDAC patients with 734 hypermethylated and 27 hypomethylated regions ($p$-value $< 0.0001$). Then, 143 hypermethylated DMRs were further selected which were located in promoter regions and completely overlapped with CGIs. After performing the least absolute shrinkage and selection operator (LASSO) method, a total of eight markers were found to fairly distinguish PDAC patients from healthy individuals, including *TRIM73*, *FAM150A*, *EPB41L3*, *SIX3*, *MIR663*, *MAPT*, *LOC100128977,* and *LOC100130148*. In conclusion, this work identified a set of eight differentially methylated markers that may be potentially applied in non-invasive diagnosis of pancreatic cancer.

**Keywords: pancreatic ductal adenocarcinoma, cfDNA, MeDIP-seq, methylation, biomarkers**

OPEN ACCESS

***Edited by:***
*Suman Ghosal, National Institutes of Health (NIH), United States*

***Reviewed by:***
*Cuncong Zhong, University of Kansas, United States*
*Enrique Medina-Acosta, Darcy Ribeiro North Fluminense State University, Brazil*

***Correspondence:**
*Jing Tian*
*tianjing@nwu.edu.cn*
*Gang Jin*
*Jingang@smmu.edn.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:*** *18 August 2020*
***Accepted:*** *05 November 2020*
***Published:*** *02 December 2020*

***Citation:***
*Li S, Wang L, Zhao Q, Wang Z, Lu S, Kang Y, Jin G and Tian J (2020) Genome-Wide Analysis of Cell-Free DNA Methylation Profiling for the Early Diagnosis of Pancreatic Cancer. Front. Genet. 11:596078. doi: 10.3389/fgene.2020.596078*

## INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is one of the most highly aggressive diseases in the world. Due to the hard challenge of detecting the disease at an early stage, poor prognosis often occurs. The morbidity of PDAC is approximately close to that of mortality. Nearly 80% of PDAC patients have no early symptoms before the advanced stage (Kaur et al., 2012) with a 5-year survival rate as low as 9% (Siegel et al., 2019). Accordingly, PDAC is the fourth leading cause of cancer-related death worldwide and is predicted to rise to second place by 2030 (Rahib et al., 2014). Currently, ultrasonography, computed tomography, positron emission tomography, magnetic

resonance imaging, and endoscopic ultrasonography are the most commonly used diagnostic methods for PDAC (Kamisawa et al., 2016; Chu et al., 2017). However, operator experience, patient obesity and intestinal gas, and other factors affect the accuracy of diagnosis (Kamisawa et al., 2016). In addition, due to the location of the pancreas, it is not easy to make an early diagnosis compared to other digestive tract tumors (Lowenfels and Maisonneuve, 2004). Therefore, it would be very valuable to identify both sensitive and specific non-invasive biomarkers for the early diagnosis of PDAC.

Epigenetic regulation, especially DNA methylation, plays an important role in the regulation of gene expression and the development of cancers. Genome-wide hypomethylation is common in cancer cells, leading to genomic instability. Some tumor suppressor genes with promoter hypermethylation are observed to cause gene silencing (Hanahan and Weinberg, 2000; Esteller, 2007). Hypermethylation of CpG islands (CGIs) in the promoters of tumor suppressor genes is a major and early event during tumorigenesis (Hanahan and Weinberg, 2000; Park et al., 2011; Udensi and Tchounwou, 2016; Liu et al., 2019). Aberrant methylation of promoter CGI regions in some genes has been proven to be associated with tumorigenesis and tumor growth (Cai et al., 2011; Pistore et al., 2017). Therefore, it is vital to detect the hypermethylation of promoter CpG islands for early diagnosis. This may contribute to the early detection of cancer and improve the therapeutic effect.

In recent years, circulating cell-free DNA (cfDNA), known as liquid biopsy, has attracted much more attention from the medical community due to its clinical advantages. As small double-stranded DNA fragments, cfDNA is released by necrotic or apoptotic cells and is circulated in the peripheral blood (Jahr et al., 2001; Stroun et al., 2001). During tumorigenesis, the increase of cell necrosis and apoptosis leads to the accumulation of cfDNA, which can be detected at a relatively early stage. Furthermore, cfDNA not only contains the same mutations as tumor cells, but also has the same methylation pattern, making it possible and convenient for early cancer diagnosis, even for those hidden organs such as the pancreas and bile ducts (Schwarzenbach et al., 2011).

Methylated DNA immunoprecipitation coupled with high-throughput sequencing (MeDIP-seq) is a sensitive technology for the detection of DNA methylation, which can even detect an initial DNA amount as low as 1 ng (Taiwo et al., 2012; Zhao et al., 2014). Genome-wide detection of cfDNA methylation profiling using the MeDIP-seq method has been developed recently for screening potential biomarkers of cancers in early stages. Based on cfDNA methylation patterns by MeDIP-seq analysis, (Shen et al., 2018) identified different potential biomarkers in pancreatic ductal adenocarcinoma, colorectal cancer, breast cancer, lung cancer, renal cancer, bladder cancer, and acute myeloid leukemia for early-stage detection. Xu et al. (2019) also identified a set of potential biomarkers that could be served in lung cancer clinical diagnosis by screening cfDNA methylation profiling using MeDIP-seq.

Therefore, in this study, we aimed to investigate the potential cfDNA methylation biomarkers in the diagnosis of PDAC. By MeDIP-seq analysis, we compared the differentially methylated regions (DMRs) of PDAC cfDNA with that of normal control, and identified 143 hypermethylated DMRs which were located in promoter regions and completely overlapped with CGIs in PDAC patients. After cross-validation with publicly available DNA methylation data, including 339 pancreatic adenocarcinoma (PAAD) patients and 357 normal controls, we successfully identified eight probes from six differentially methylated genes, containing *TRIM73*, *FAM150A*, *EPB41L3*, *SIX3*, *MIR663*, *MAPT*, *LOC100128977,* and *LOC100130148*, which could be used as potential biomarkers for early detection for PDAC patients.

# MATERIALS AND METHODS

## Sample Collection

A total of six samples including four PDAC patients and two healthy controls were used for this study. Four serum samples from PDAC patients were supplied by ChangHai Hospital. All of them signed informed consent forms. Specimens were collected and analyzed with the approval of the ethics committees of ChangHai Hospital and School of Medicine, Northwest University, respectively.

## cfDNA Extraction

First, 5 ml peripheral blood was collected using EDTA anticoagulant tubes before surgery and drug treatment. The plasma was purified by centrifuge for 15 min at 1500 × *g* within 6 h of collection. cfDNA was extracted from 800 μl aliquots of plasma using a QIAamp Circulating Nucleic Acid Kit (Qiagen, 55114) according to manufacturer's protocol and quantified with Bioanalyzer 2100 (Agilent Technologies).

## MeDIP-seq Library Construction and Sequencing

The cfDNA MeDIP-seq library was prepared as we described previously (Xu et al., 2019). In short, approximately 20 ng cfDNA was ligated with Illumina barcode adapters using a KAPA Hyper Prep Kit (KAPA, KK8502). The constructed cfDNA libraries were denatured at 95°C for 10 min. The methylated cfDNA was separated from the cfDNA libraries by immunoprecipitation using the 5-Methylcytosine (5mC) Monoclonal Antibody (Epigentek, A-1014). MeDIP DNA was further amplified using a Q5 High-Fidelity DNA Polymerase (NEB, M0491). After quality assessment using Bioanalyzer 2100 (Agilent Technologies), amplified libraries were subjected to deep sequencing by the Illumina HiSeq 2000 platform.

## Data Processing and Analysis

MeDIP-seq raw data were processed using the Trimmomatic software (version 0.38) to filter out low-quality reads and Illumina adapters. The clean reads were mapped to the human reference genome GRCh37/hg19 (UCSC) using the Bowtie software (version 2.3.3.1) (Langmead et al., 2009). The differentially methylated regions (DMRs) between pancreatic cancer patients and healthy controls were calculated with the R package MEDIPS (version 1.36.0) (Lienhard et al., 2014), the

coupling factor for CpG density was generated based on the normalization of the patient MeDIP-seq data. The function of region of interest (ROI) analysis in the MEDIPS package was specifically used to investigate the DNA methylation levels in UCSC CpG islands, CpG shore (∼2 Kb from islands), and CpG shelf regions (∼4 Kb from islands)[1]. Mapping results were visualized using Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2013). Pathway analysis was carried with the Ingenuity Pathway Analysis (IPA) software (Qiagen).

Illumina Infinium HumanMethylation 450K BeadChip Array (HM450K) data from The Cancer Genome Atlas (TCGA) project and Gene Expression Omnibus (GEO) were used to validate our MeDIP-seq results. A total of 696 HM450K sample sets including 339 PAAD patients and 357 normal controls were assembled from the TCGA[2] and GEO (GSE49149 and GSE40279) databases. The information about the patient age and gender of 696 HM450K sample sets are supplied in **Supplementary Table 1**. The bioinformatics pipeline and R codes are available as supplementary code in zenodo[3]. The variable selection was performed using the LASSO method (Xu et al., 2017). We subsampled 75% of the dataset for model building. After 500 iterations, we selected the probes that appeared more than 450 times as covariates, and obtained a total of eight probes. We fitted a logistic regression model with these candidate markers and measured the classification performance of the binary classifier using an area under the ROC curve (AUC).

The Paired Student's *t*-test was performed using the processed beta (β) values (proportion of the methylated signal over the total signal) to compare the DNA methylation levels in the probe regions between 339 PAAD sample and 357 normal samples, the *p*-value for each maker was corrected by multiple testing with the Benjamini-Hochberg procedure (Benjamini and Yekutieli, 2001).

Multivariate Cox regression analysis was performed to construct the prognostic model based on the AIC value. Kaplan-Meier curves were generated and used to perform survival analysis using GEPIA[4].

# RESULTS

## Analysis of Global cfDNA Methylation Profiling in Pancreatic Cancer by MeDIP-seq

Four plasma samples of PDAC patients and two of healthy controls were collected, the clinical information of patients is shown in **Table 1**. The four PDAC samples were in the IB or IIB stage which had entered into the early or middle stage of pancreatic cancer (**Table 1**) (van Roessel et al., 2018). After being subjected to quality testing, the size of the cfDNA fragments was mainly distributed in the range of 150–200 bp with a main peak of 172 bp, which met the previous criteria where cfDNA showed a specific size of ∼167 bp (Lo et al., 2010; Thierry et al., 2010). After immunoprecipitation and amplification, the size distribution profiles of all cfDNA libraries showed a range from 172 to 292 bp with a main peak of ∼292 bp including ∼120 bp sequencing adapters (**Supplementary Figure 1**). The cfDNA MeDIP libraries were sequenced with Illumina HiSeq 2000 (a flow chart of the steps in the analysis is presented in **Figure 1**). A total of 41 million raw sequenced reads were obtained from PDAC patients, 72.7% of which was mapped to the reference genome (Human hg19), and 32 million reads from healthy controls of which 54.8% was mapped. After quality filtering, there were approximately 24 million unique reads of patients and 17 million unique reads of healthy controls (**Table 2**).

**TABLE 1 |** Clinical information of PDAC patients.

| Sample | Gender | Age | Stage | Histology |
|---|---|---|---|---|
| P1 | Male | 59 | pT3N1Mx | Ductal adenocarcinoma |
| P2 | Male | 79 | pT3N1Mx | Ductal adenocarcinoma |
| P3 | Female | 67 | pT2N0Mx | Ductal adenocarcinoma |
| P4 | Female | 56 | pT2N0Mx | Ductal adenocarcinoma |

In order to analyze the whole-genome methylation patterns between PDAC patients and healthy controls, we performed the principal component analysis (PCA) to investigate the genome-wide methylation profiles in the two groups. The methylation patterns in PDAC patients exhibited a significant difference from the healthy control groups (**Figure 2A**). The unsupervised clustering analysis result further showed that there was a dramatic change in methylation patterns between PDAC patients and healthy controls (**Figure 2B**). This indicates that there are epigenetic differences between PDAC patients and healthy people.

## Differentially Methylated Regions of Promoters in Pancreatic Cancer Patients

A total of 5,205 differentially methylated regions (DMRs) were identified through MeDIP-seq analysis in PDAC patients ($p < 0.05$), which included 5,117 hypermethylated regions (98.3%) and 88 hypomethylated regions (1.7%) as shown in **Supplementary Table 2**. The clustering analysis also exhibited a significant alteration between PDAC patients and controls (**Figure 3A**). Previous studies have revealed that aberrant methylation patterns in the promoter region of tumor suppressor genes may cause transcriptional silencing which could be a driving force for cancer development (Herman and Baylin, 2003). We focused on promoter regions and recognized 775 different DMRs ($p < 0.0001$), including 761 hypermethylated regions (98.2%) from 532 genes and 14 hypomethylated regions (1.8%) from 14 genes (**Figure 3B** and **Supplementary Table 3**). These data suggest that most of the promoter regions are hypermethylated in pancreatic cancer samples, which is consistent with previous findings that specific hypermethylation occurring at specific promoter sites likely leads to cancer (Park et al., 2011; Liu et al., 2019; Zhang et al., 2020).

[1]http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cpgIslandExt.txt.gz
[2]https://portal.gdc.cancer.gov/projects/TCGA-PAAD
[3]https://doi.org/10.5281/zenodo.4066412
[4]http://gepia.cancer-pku.cn/index.html

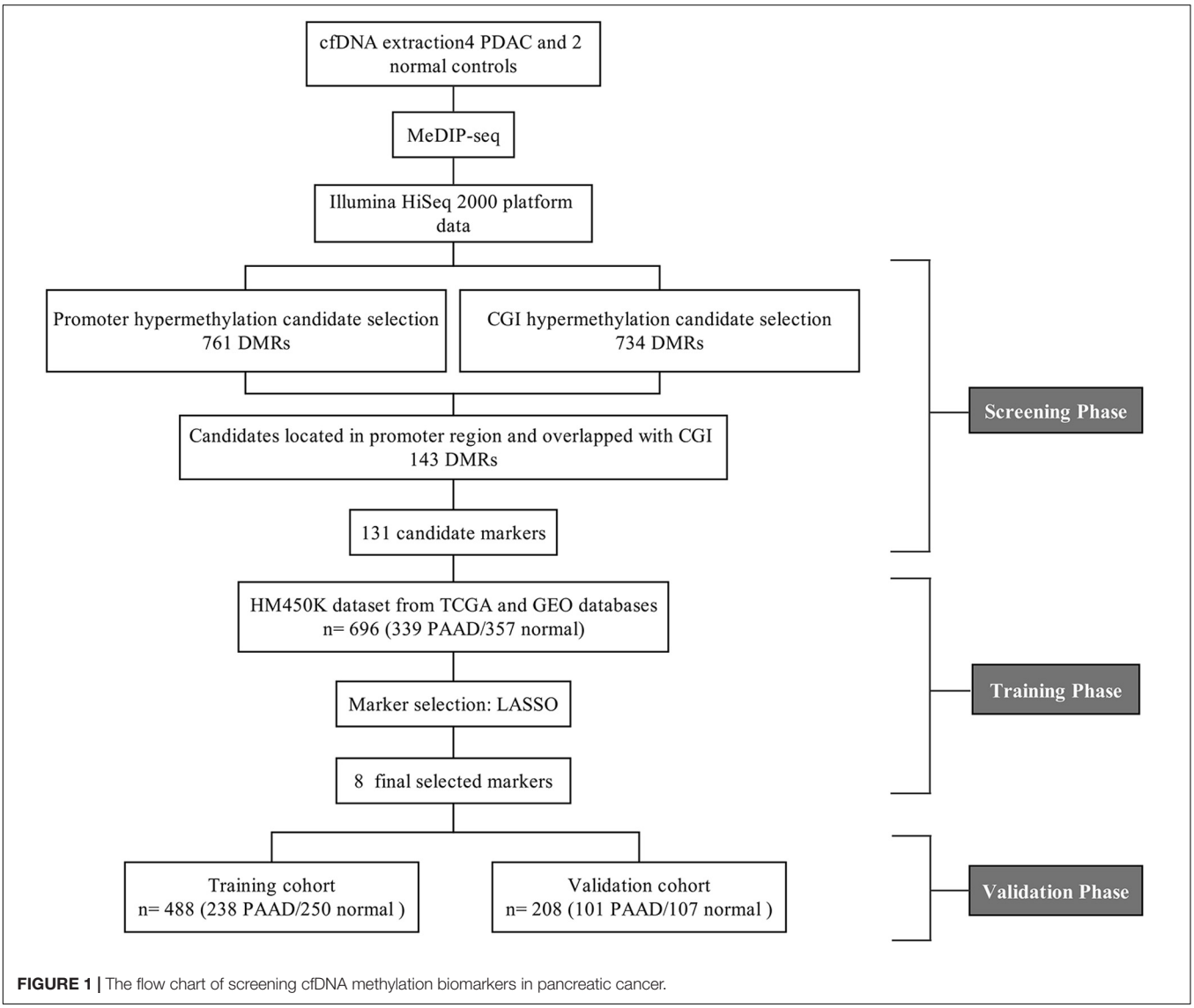



**FIGURE 1 |** The flow chart of screening cfDNA methylation biomarkers in pancreatic cancer.

## Differentially Methylated Regions (DMRs) of CpG Regions in Pancreatic Cancer Patients

According to the division of the CG content, some areas in the genome can be determined as CpG islands (CG content > 50%) (Gardiner-Garden and Frommer, 1987), CpG shores (up to 2 kb from CpG islands) (Irizarry et al., 2009), and CpG shelfs (≥2 kb from CpG islands) (Nones et al., 2014). It is reported that 72% of promoters are unmethylated GC-rich (Saxonov et al., 2006). Here we found that the general methylation levels of CpG regions in pancreatic cancer patients were higher than those in normal controls, which showed the median methylation levels in CGI, CpG shore, and CpG shelf to be 0.39, 0.57, and 0.5475, respectively, compared with 0.265, 0.45, and 0.41, respectively in controls (**Figure 4A**). Hypermethylation of CGI sites in promoter regions is considered as a risk marker for cancer development and progression (Costello et al., 2000; Esteller et al., 2001; Widschwendter and Jones, 2002), therefore, only DMR in CGIs were in focus and used for further analysis.

**TABLE 2 |** Statistics summary of MeDIP-seq data.

| Sample | Number of total reads | Number of mapped reads | Mapped read rate | Number of unique reads | Unique read rate |
|---|---|---|---|---|---|
| P1 | 41,251,616 | 30,099,978 | 73.0% | 25,187,519 | 83.7% |
| P2 | 37,618,679 | 27,811,589 | 73.9% | 22,962,727 | 82.6% |
| P3 | 54,836,822 | 40,245,896 | 73.4% | 33,248,824 | 82.6% |
| P4 | 31,699,187 | 22,318,731 | 70.41% | 17,448,398 | 78.18% |
| C1 | 12,247,801 | 6,219,267 | 50.78% | 5,547,597 | 89.20% |
| C2 | 53,490,488 | 31,510,659 | 58.91% | 29,241,359 | 92.80% |

*C, healthy control; P, PDAC patient.*

A total of 761 DMRs was identified in CGIs of the whole genome in PDAC patients ($p$ value < 0.0001).

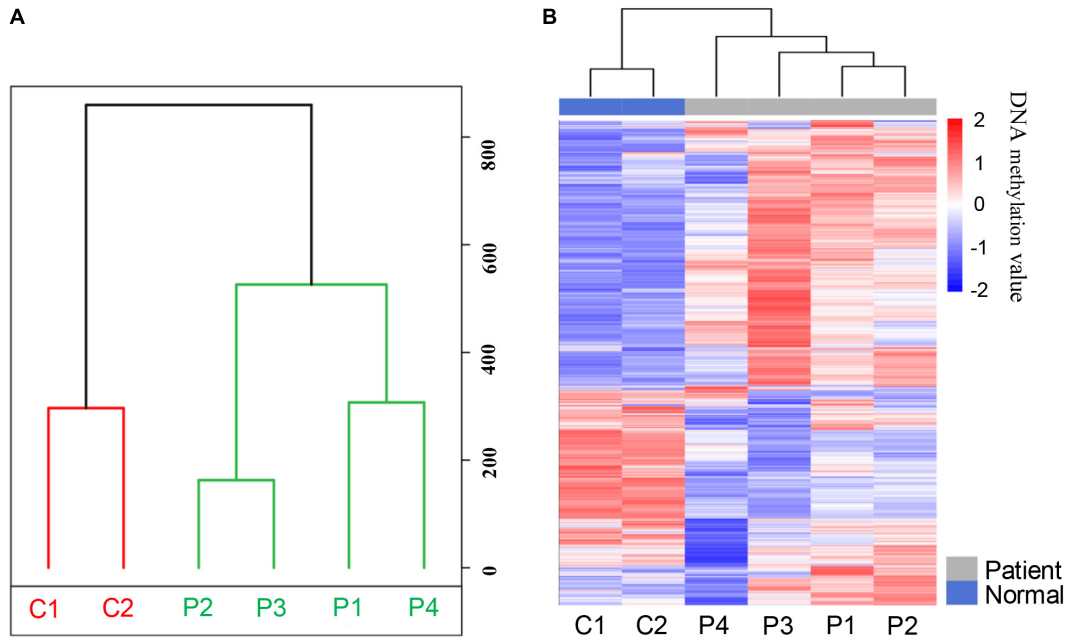



**FIGURE 2 |** The methylation patterns of pancreatic cancer patients and healthy controls after MeDIP-seq datasets analysis. **(A)** Principal component analysis (PCA) of the methylation profiles between patients and controls. **(B)** The unsupervised cluster analysis of the genome-wide methylation profiles in patients and controls.

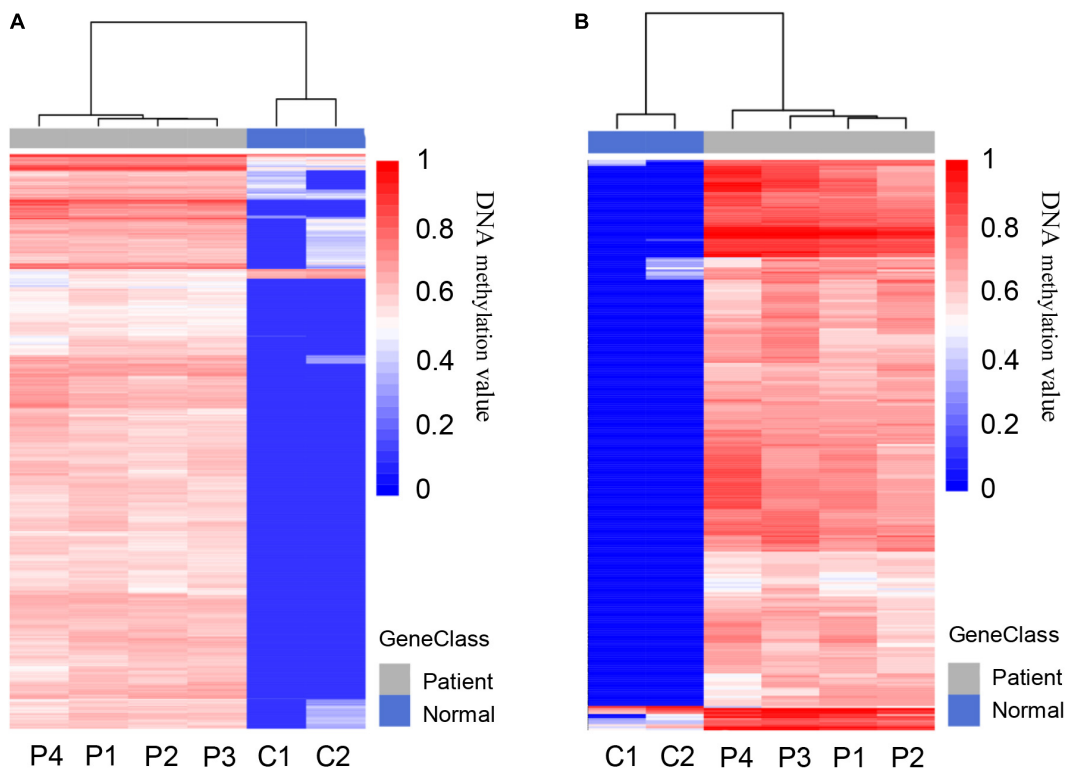



**FIGURE 3 |** Differentially methylated regions (DMRs) in pancreatic cancer patients and healthy controls. **(A)** Heat map of total 5,205 DMRs located in the whole genome of PDAC patients compared to healthy controls, including 5,117 hypermethylated and 88 hypomethylated regions. **(B)** Heat map of total 775 DMRs located in the promoter regions of patients compared to healthy controls, including 761 hypermethylated and 14 hypomethylated regions.

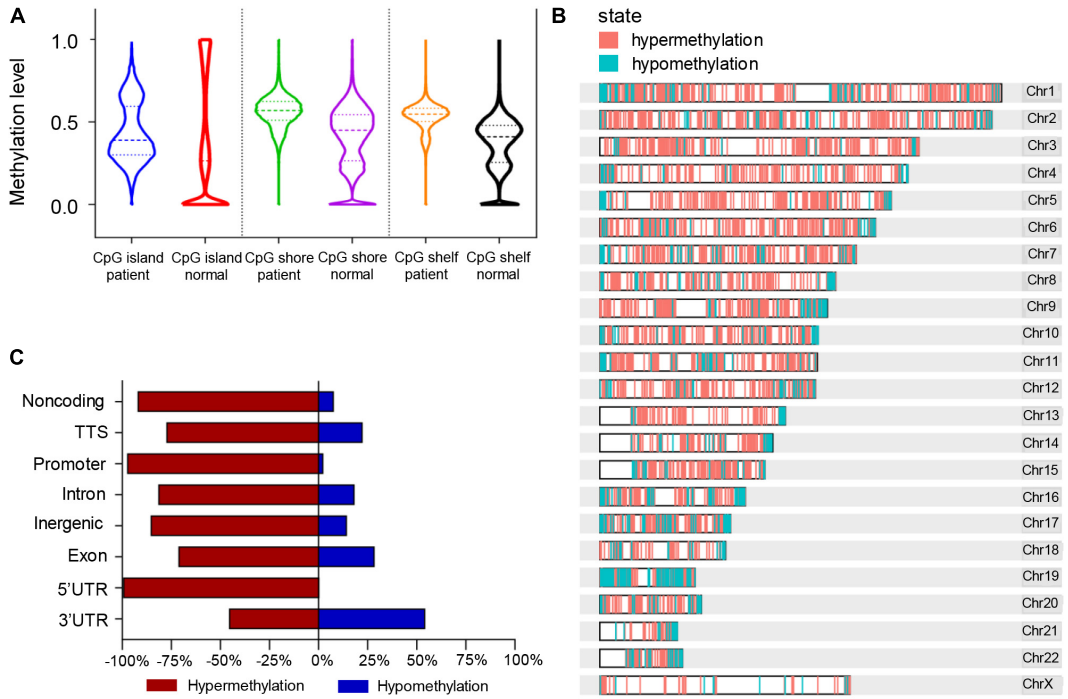



**FIGURE 4 |** Differentially methylated regions (DMRs) of the CpG regions in pancreatic cancer patients and healthy controls. **(A)** Violin plots of DMRs located in CpG islands, CpG shores, and CpG shelfs of PDAC patients compared to controls. **(B)** Whole genomic and chromosomal location of DMRs in CGIs. **(C)** The different features of CGI distribution according to hypermethylated and hypomethylated regions.

Among them, there were 734 (96.5%) hypermethylation regions from 507 genes and 27 (3.5%) hypomethylation regions from 26 genes (**Supplementary Table 4**). The visual DMR signals of hypermethylation and hypomethylation in CGIs mapped to the whole genome are shown in **Figure 4B**. The distribution features of hypermethylated and hypomethylated regions in CGIs were further classified as shown in **Figure 4C**. A predominant hypermethylation of DMRs in CGIs was observed, except in the 3′UTR region (**Figure 4C**).

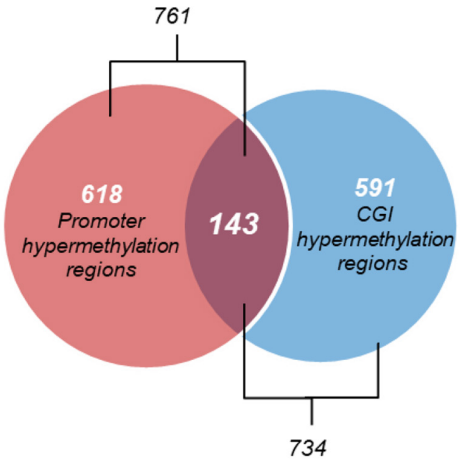
## Identification of Differentially Methylated Genes Located in Promoter CGIs in Pancreatic Cancer Patients

It is reported that the hypermethylation of promoter CGIs is supposed to be an indicator of the risk of progression or development of cancers which is associated with the silencing of tumor suppressor genes (Feinberg, 2005; Park et al., 2011). We further screened those DMRs which were located in CGIs promoters. A total of 143 hypermethylated DMRs located in promoter regions that completely overlapped with CGIs were identified as candidate DMRs (**Figure 5A**). The 143 candidate DMRs were derived from 70 genes. To further understand the biological associations of the 70 genes, ingenuity pathway analysis (IPA) was performed and showed that cancer was included in the top diseases (**Figure 5B**).

## Cross-Validation of Potential Candidate Genes With Publicly Available DNA Methylation Data

The 143 candidate DMRs were further annotated to 131 probes on an Illumina HM450K BeadChip Array (**Supplementary Table 5**) and were analyzed by the Least Absolute Shrinkage and Selection Operator (LASSO) method to select the most discriminating markers. The 75% HM450K datasets were randomly selected each time for loop modeling. Eventually, eight probes were identified as a final selection of markers which were required to appear over 450 times out of a total of 500 repetitions in the model (**Table 3**). To evaluate the diagnostic value of the eight markers, we built a risk prediction model in training and validation dataset using the logistic regression method. The HM450K datasets were then divided into a training cohort of 488 individuals (238 PAAD patients and 250 normal controls) and a validation cohort of 208 individuals (101 PAAD patients and 107 normal controls). The final prediction model achieved a sensitivity of 97.1% and a specificity of 98.0% on the training cohort, the sensitivity and specificity of the validating cohort was 93.2 and 95.2%, respectively (**Figure 6A**). This model could distinguish PAAD patients from the normal controls both in the training dataset (the area under the ROC curve, AUC = 0.975) and the validation dataset (AUC = 0.943). The prediction performance of the model in two datasets is shown in **Figure 6B**. To further characterize the methylation status of the eight markers in PAAD patients and normal controls,

A




B

**Top Diseases and Bio Functions**

**Diseases and Disorders**

| Name | p-value range | # Molecules |
|---|---|---|
| Cancer | 1.83E-02 - 1.23E-03 | 2 |
| Connective Tissue Disorders | 4.97E-02 - 1.23E-03 | 2 |
| Developmental Disorder | 4.97E-02 - 1.23E-03 | 3 |
| Gastrointestinal Disease | 4.97E-02 - 1.23E-03 | 2 |
| Hereditary Disorder | 4.92E-02 - 1.23E-03 | 2 |

**Molecular and Cellular Functions**

| Name | p-value range | # Molecules |
|---|---|---|
| Cell Morphology | 1.35E-02 - 1.23E-03 | 3 |
| Cell-To-Cell Signaling and Interaction | 4.82E-02 - 1.23E-03 | 2 |
| Cellular Growth and Proliferation | 2.44E-02 - 1.23E-03 | 4 |
| Cellular Assembly and Organization | 4.82E-02 - 2.46E-03 | 4 |
| Cellular Function and Maintenance | 4.82E-02 - 2.46E-03 | 4 |

**FIGURE 5 |** Selection and definition of differentially methylated genes in both the CGIs and promoter regions. **(A)** Hypermethylated DMRs in the overlap of promoter regions and CGIs. **(B)** Top disease and bio functions by IPA analysis for genes derived from hypermethylated DMRs located in both the promoter regions and CGIs.

unsupervised hierarchical clustering was performed in 696 cases of the HM450K datasets (**Figure 6C**). The result demonstrated that these eight markers were able to distinguish PAAD patients from normal controls with high sensitivity and specificity.

## Analysis of Relative Methylation Levels of the Eight Markers Between PAAD Patients and Normal Controls

To further address whether the eight markers we identified can distinguish pancreatic cancer patients from the healthy individuals, we next assessed the methylation levels of the eight markers in 696 cases including 339 PAAD patients and 357 normal controls. For all eight markers, there was a significantly difference in the overall methylation levels between the PAAD patients and normal controls (BH-adjusted $p < 0.0001$) (**Figure 7**). It suggested that the eight markers: *MAPT*, *SIX3*, *MIR663*, *EPB41L3*, *FAM150A*, *TRIM73*, *LOC100128977*, and *LOC100130148* may serve as potential biomarkers for the early diagnosis of pancreatic cancer.

# DISCUSSION

Here, we performed a genome-wide epigenetic profiling assessment in pancreatic cancer patients for screening potential biomarkers using MeDIP-seq technology in cfDNA. Our analysis exhibited global changes in cfDNA methylation

patterns in pancreatic cancer patients. In our study, we found 761 hypermethylated DMRs in promoter regions and 734 hypermethylated DMRs in CGIs derived from pancreatic cancer patients, furthermore, a total of 143 candidate DMRs were identified, located in both the promoter regions and CGIs. For subsequent analysis, tissue-derived data from TCGA and GEO was used due to the lack of cfDNA metalation data in public datasets. Finally, the diagnostic prediction model of the eight probes was established, including *MAPT*, *SIX3*, *MIR663*, *EPB41L3*, *FAM150A*, *TRIM73*, *LOC100128977*, and *LOC100130148*. Among these, *MAPT*, *LOC100128977*, and *LOC100130148* are the three differentially methylated CpG sites that hit only one gene locus. The diagnostic prediction model could effectively distinguish between PAAD patients and normal controls according to both the training cohort (AUC = 0.975) and validation cohort (AUC = 0.943). These results represented promising novel methylation markers for the early diagnosis of pancreatic cancer.

To determine the prognostic value of the eight markers in pancreatic cancer patients, Kaplan–Meier survival analysis was performed (**Supplementary Figure 2**). Pancreatic cancer patients

**TABLE 3 |** Characteristics of the eight methylation markers and their coefficients in PAAD diagnosis prediction.

| Markers | Ref Gene | Coefficients | SE | *z* value | *p*-value |
|---|---|---|---|---|---|
| cg00394725 | TRIM73 | −3.1937 | 0.6835 | −4.673 | <0.05 |
| cg09442654 | FAM150A | 0.3357 | 0.4777 | 0.703 | <0.05 |
| cg26170805 | EPB41L3 | 1.8672 | 0.781 | 2.391 | <0.05 |
| cg19186145 | SIX3 | 1.877 | 0.871 | 2.155 | <0.05 |
| cg11220245 | MIR663 | 0.4137 | 0.4914 | 0.842 | <0.05 |
| cg11909912 | MAPT | 0.9288 | 0.5346 | 1.737 | <0.05 |
| cg10780632 | LOC100128977 | 8.7616 | 3.4814 | 2.517 | <0.05 |
| cg19670923 | LOC100130148 | 0.7289 | 0.8566 | 0.851 | <0.05 |

*SE: standard errors of coefficients; z value: Wald z-statistic value.*

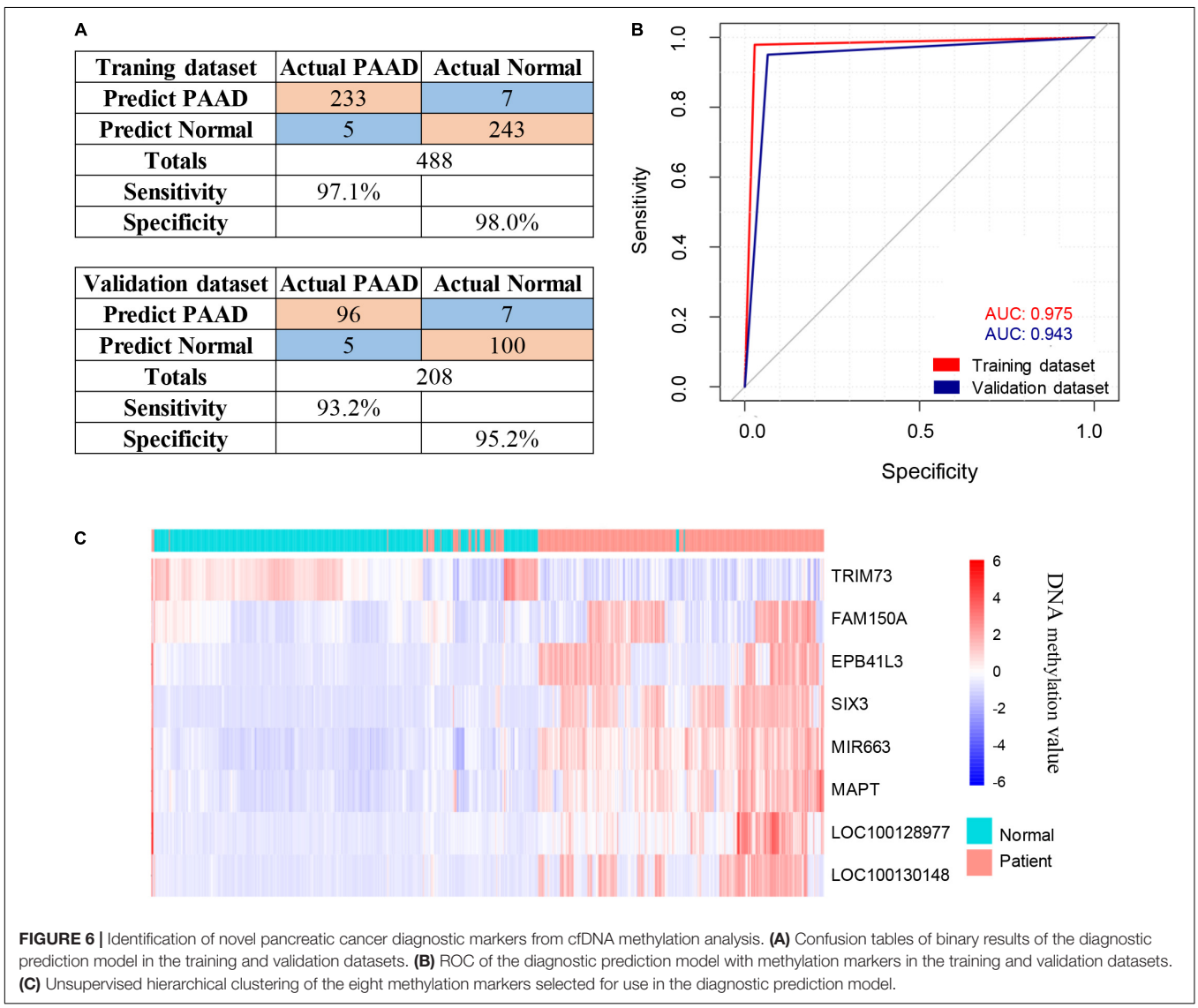



**FIGURE 6 |** Identification of novel pancreatic cancer diagnostic markers from cfDNA methylation analysis. **(A)** Confusion tables of binary results of the diagnostic prediction model in the training and validation datasets. **(B)** ROC of the diagnostic prediction model with methylation markers in the training and validation datasets. **(C)** Unsupervised hierarchical clustering of the eight methylation markers selected for use in the diagnostic prediction model.

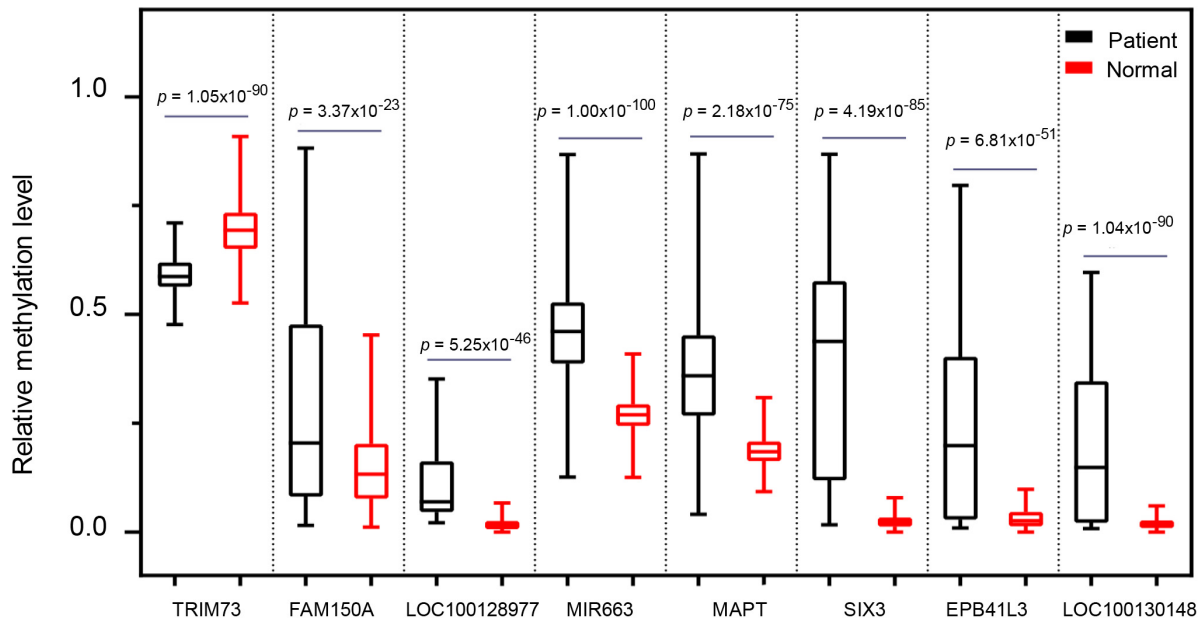



**FIGURE 7 |** The comparison of the methylation level of the eight selected markers between pancreatic cancer patients and healthy controls.

with a high expression of *MAPT*, *EPB41L3*, *LOC100128977,* and *LOC100130148* had an evidently higher overall survival as compared with those with a low expression of *MAPT* ($p = 0.0034$), *EPB41L3* ($p = 0.0088$), *LOC100128977* ($p = 0.0077$), and *LOC100130148* ($p = 0.0017$). However, the multivariate Cox regression analysis indicated that *TRIM73*, *FAM150A*, *EPB41L3*, *SIX3*, *MAPT*, *LOC100128977,* and *LOC100130148* might not be independent factors for the prognosis of pancreatic cancer patients (**Supplementary Table 6**). This may indicate that gene expression is not only regulated by methylation, but also under a complex regulatory system. Therefore, these eight markers may be effective biomarkers for the diagnosis of pancreatic cancer, but they can not be used as prognostic indicators.

In recent years, there have been a few studies into the genome-wide detection of cfDNA methylation profiling using the MeDIP-seq method to screen potential tumor biomarkers. Shen et al. (2018) collected seven kinds of cancer samples for MeDIP-seq data analysis and took transcription factors into consideration while processing the biomarker analysis. Xu et al. (2019) identified hypermethylated DMRs in the promoter region for finding early diagnosis markers of lung cancer. In this study, we aimed to identify biomarkers in cfDNA which were located both in promoter regions and CGIs. CGIs are closely related to tumor epigenome, especially in promoter regions. Lay et al. (2015) demonstrated that compared to non-CGI promoters, methylation in CGI promoters had a greater impact on nucleosome phasing and histone modifications which have an influence on directing the functional organization of cancer epigenome. Tumorigenesis often coincides with CGI hypermethylation, leading to the inactivation of tumor suppressor genes (Namba et al., 2019). In a study of the genome-wide search for identifying potentially methylated changes during the progression of colorectal neoplasia, (Gu et al., 2019) found that hypermethylation occurred mainly in the overlap regions of CGIs and promoters, while hypomethylation tended to be far away from functional regions. Studies in hepatocellular carcinoma and ovarian cancer also revealed that the methylation status of some genes in the promoter and CGI regions can be used as prognosis markers for cancer patients (Dai et al., 2013; Lee et al., 2016).

Allele-specific methylation (ASM) has been well documented in imprinted loci. The parental allele $5^mC$ asymmetry would create allele-specific imprinted differentially methylated regions (iDMRs). Moreover, it has been recently reported that some ASM loci undergo cancer-associated epigenetic changes in hematopoietic cancer. de Sa Machado Araujo et al. (2018) reported that the maternally inherited $5^mCpG$ imprints for one gametic (*PARD6GAS1*) and one somatic (*GCSAML*) iDMRs are dysregulated in hematopoietic cancers. Among the eight methylated probes that could potentially serve as diagnosis markers in this study, we found four markers that were allele-specific methylated, including *EPB41L3, SIX3, MIR663, and MAPT*, suggesting that ASM also occurs in solid malignancies. Unlike whole-genome bisulfite sequencing (WGBS), which could detect the methylation state of nearly each CpG site, MeDIP technology uses an anti-methylcytosine antibody at a resolution of 100–300 bp. Therefore, MeDIP could not distinguish DNA methylation at a single base resolution (Yong et al., 2016). So ASM could not be included in the current study. Pancreatic cancer is a highly lethal disease, the lack of early detection and optional treatment is the main reason. Therefore, as a non-invasive micro diagnostic technology, cfDNA combined with MeDIP-seq is expected to be an effective

method for early clinical diagnosis. In our analysis, *MAPT*, *SIX3*, *MIR663*, *EPB41L3*, *FAM150A*, *TRIM73*, *LOC100128977,* and *LOC100130148* exhibited statistically significant differences between pancreatic cancer patients and the healthy controls (**Figure 7**). *MAPT* is a potential predictive biomarker of the efficacy of SG410, a benzoylphenylurea sulfur analog for pancreatic cancer treatment (Jimeno et al., 2007). Tumor suppressor *SIX3* is reported to inhibit cell proliferation, migration, and invasion in glioblastoma and breast cancer (Zhang et al., 2017; Zheng et al., 2018; Yu et al., 2020). *MIR663* could act as a tumor suppressor in gastric cancer (Pan et al., 2010) and glioblastoma (Shi et al., 2014). *FAM150A* is a potential prognostic marker of clear cell renal cell carcinomas (Tian et al., 2014). Taken together, these markers, which we identified in the plasma of pancreatic cancer, may have potential clinical values.

## CONCLUSION

In summary, by analyzing genome-wide cfDNA methylation profiling using the MeDIP-seq method, we established a set of eight potential biomarkers which might be applied in non-invasive diagnosis of early-stage pancreatic cancer.

## DATA AVAILABILITY STATEMENT

Publicly available datasets in the TCGA (https://portal.gdc.cancer.gov/projects/TCGA-PAAD) and GEO (GSE49149 and GSE40279) databases were used in this study. The pancreatic cancer patients' raw data of MeDIP-seq in this study are available in the EMBL database (http://www.ebi.ac.uk/arrayexpress/) under accession number E-MTAB-9678, and the healthy controls C1 (ERS2672506, ERS2672505) and C2 (ERS2672508, ERS2672507) are available in the EMBL database under accession number E-MTAB-7163

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committees of ChangHai Hospital and School of Medicine, Northwest University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

JT and GJ conceived the study and were in charge of the overall direction and planning. JT and SyL wrote the manuscript with input from all authors. SyL and LW collected the samples. QZ and YK performed the computational framework. SyL, LW, and QZ analyzed the data. ZW and SxL provided the technical support. JT and GJ provided the funding support. All authors reviewed the manuscript and approved the final version for publication.

## FUNDING

This work was supported by the Shaanxi Key Industry Innovation Chain (Group) Foundation in the Social Development Field, China (2019ZDLSF02-05) and the National Key Research and Development Project (2019YFC1315904).

## ACKNOWLEDGMENTS

We sincerely thank all participants in the study.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.596078/full#supplementary-material

**Supplementary Figure 1 |** Representative bioanalyzer profiles of cfDNA and MeDIP-seq libraries.

**Supplementary Figure 2 |** Kaplan–Meier analysis of PAAD patients revealed the prognosis ability of seven markers.

**Supplementary Table 1 |** Clinical information of 696 samples from Illumina HM450K.

**Supplementary Table 2 |** Genome DMRs identified in cfDNA of pancreatic cancer patient plasma.

**Supplementary Table 3 |** Promoter DMRs identified in cfDNA of pancreatic cancer patient plasma.

**Supplementary Table 4 |** CGI DMRs identified in cfDNA of pancreatic cancer patient plasma.

**Supplementary Table 5 |** 131 Illumina HM450K BeadChip Array probes corresponding to 143 candidate DMRs.

**Supplementary Table 6 |** Multivariate Cox regression analysis shows the prognosis value of the seven markers.

## REFERENCES

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.

Cai, H. H., Sun, Y. M., Miao, Y., Gao, W. T., Peng, Q., Yao, J., et al. (2011). Aberrant methylation frequency of TNFRSF10C promoter in pancreatic cancer cell lines. *Hepatobiliary Pancreat. Dis. Int.* 10, 95–100. doi: 10.1016/s1499-3872(11)60014-3

Chu, L. C., Goggins, M. G., and Fishman, E. K. (2017). Diagnosis and detection of pancreatic cancer. *Cancer J.* 23, 333–342.

Costello, J. F., Fruhwald, M. C., Smiraglia, D. J., Rush, L. J., Robertson, G. P., Gao, X., et al. (2000). Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* 24, 132–138. doi: 10.1038/72785

Dai, W., Zeller, C., Masrour, N., Siddiqui, N., Paul, J., and Brown, R. (2013). Promoter CpG island methylation of genes in key cancer pathways associates with clinical outcome in high-grade serous ovarian cancer. *Clin. Cancer Res.* 19, 5788–5797. doi: 10.1158/1078-0432.CCR-13-1217

de Sa Machado Araujo, G., Da Silva Francisco Junior, R., Dos Santos Ferreira, C., Mozer Rodrigues, P. T., Terra Machado, D., Louvain De Souza, T., et al. (2018). Maternal 5(m)CpG imprints at the PARD6G-AS1 and GCSAML differentially

methylated regions are decoupled from parent-of-origin expression effects in multiple human tissues. *Front. Genet.* 9:36. doi: 10.3389/fgene.2018.00036

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* 8, 286–298. doi: 10.1038/nrg2005

Esteller, M., Corn, P. G., Baylin, S. B., and Herman, J. G. (2001). A gene hypermethylation profile of human cancer. *Cancer Res.* 61, 3225–3229.

Feinberg, A. P. (2005). Cancer epigenetics is no mickey mouse. *Cancer Cell* 8, 267–268. doi: 10.1016/j.ccr.2005.09.014

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282. doi: 10.1016/0022-2836(87)90689-9

Gu, S., Lin, S., Ye, D., Qian, S., Jiang, D., Zhang, X., et al. (2019). Genome-wide methylation profiling identified novel differentially hypermethylated biomarker MPPED2 in colorectal cancer. *Clin. Epigenet.* 11:41. doi: 10.1186/s13148-019-0628-y

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/s0092-8674(00)81683-9

Herman, J. G., and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med.* 349, 2042–2054. doi: 10.1056/NEJMra023075

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z. J., Montano, C., Onyango, P., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* 41, 178–186. doi: 10.1038/ng.298

Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F. O., Hesch, R. D., et al. (2001). DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* 61, 1659–1665.

Jimeno, A., Hallur, G., Chan, A., Zhang, X., Cusatis, G., Chan, F., et al. (2007). Development of two novel benzoylphenylurea sulfur analogues and evidence that the microtubule-associated protein tau is predictive of their activity in pancreatic cancer. *Mol. Cancer Ther.* 6, 1509–1516. doi: 10.1158/1535-7163.MCT-06-0592

Kamisawa, T., Wood, L. D., Itoi, T., and Takaori, K. (2016). Pancreatic cancer. *Lancet* 388, 73–85. doi: 10.1016/s0140-6736(16)00141-0

Kaur, S., Baine, M. J., Jain, M., Sasson, A. R., and Batra, S. K. (2012). Early diagnosis of pancreatic cancer: challenges and new developments. *Biomark. Med.* 6, 597–612. doi: 10.2217/bmm.12.69

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:R25. doi: 10.1186/gb-2009-10-3-r25

Lay, F. D., Liu, Y., Kelly, T. K., Witt, H., Farnham, P. J., Jones, P. A., et al. (2015). The role of DNA methylation in directing the functional organization of the cancer epigenome. *Genome Res.* 25, 467–477. doi: 10.1101/gr.183368.114

Lee, H. G., Kim, H., Son, T., Jeong, Y., Kim, S. U., Dong, S. M., et al. (2016). Regulation of HK2 expression through alterations in CpG methylation of the HK2 promoter during progression of hepatocellular carcinoma. *Oncotarget* 7, 41798–41810. doi: 10.18632/oncotarget.9723

Lienhard, M., Grimm, C., Morkel, M., Herwig, R., and Chavez, L. (2014). MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* 30, 284–286. doi: 10.1093/bioinformatics/btt650

Liu, Y. Z., Qiao, Y. C., Zhang, H. Y., Li, W. T., and Zheng, J. (2019). Wnt7a, frequently silenced by CpG methylation, inhibits tumor growth and metastasis via suppressing epithelial-mesenchymal transition in gastric cancer. *J. Cell. Biochem.* 120, 18142–18151. doi: 10.1002/jcb.29118

Lo, Y. M., Chan, K. C., Sun, H., Chen, E. Z., Jiang, P., Lun, F. M., et al. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* 2:61ra91. doi: 10.1126/scitranslmed.3001720

Lowenfels, A. B., and Maisonneuve, P. (2004). Epidemiology and prevention of pancreatic cancer. *Jpn. J. Clin. Oncol.* 34, 238–244. doi: 10.1093/jjco/hyh045

Namba, S., Sato, K., Kojima, S., Ueno, T., Yamamoto, Y., Tanaka, Y., et al. (2019). Differential regulation of CpG island methylation within divergent and unidirectional promoters in colorectal cancer. *Cancer Sci.* 110, 1096–1104. doi: 10.1111/cas.13937

Nones, K., Waddell, N., Song, S., Patch, A. M., Miller, D., Johns, A., et al. (2014). Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling. *Int. J. Cancer* 135, 1110–1118. doi: 10.1002/ijc.28765

Pan, J., Hu, H., Zhou, Z., Sun, L., Peng, L., Yu, L., et al. (2010). Tumor-suppressive mir-663 gene induces mitotic catastrophe growth arrest in human gastric cancer cells. *Oncol. Rep.* 24, 105–112. doi: 10.3892/or_00000834

Park, S. Y., Kwon, H. J., Lee, H. E., Ryu, H. S., Kim, S. W., Kim, J. H., et al. (2011). Promoter CpG island hypermethylation during breast cancer progression. *Virchows Archiv.* 458, 73–84. doi: 10.1007/s00428-010-1013-6

Pistore, C., Giannoni, E., Colangelo, T., Rizzo, F., Magnani, E., Muccillo, L., et al. (2017). DNA methylation variations are required for epithelial-to-mesenchymal transition induced by cancer-associated fibroblasts in prostate cancer cells. *Oncogene* 36, 5551–5566. doi: 10.1038/onc.2017.159

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74, 2913–2921. doi: 10.1158/0008-5472.CAN-14-0155

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1412–1417. doi: 10.1073/pnas.0510310103

Schwarzenbach, H., Hoon, D. S., and Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* 11, 426–437. doi: 10.1038/nrc3066

Shen, S. Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M. H. A., Chadwick, D., et al. (2018). Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563, 579–583. doi: 10.1038/s41586-018-0703-0

Shi, Y., Chen, C., Zhang, X., Liu, Q., Xu, J. L., Zhang, H. R., et al. (2014). Primate-specific miR-663 functions as a tumor suppressor by targeting PIK3CD and predicts the prognosis of human glioblastoma. *Clin. Cancer Res.* 20, 1803–1813. doi: 10.1158/1078-0432.CCR-13-2284

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

Stroun, M., Lyautey, J., Lederrey, C., Olson-Sand, A., and Anker, P. (2001). About the possible origin and mechanism of circulating DNA apoptosis and active DNA release. *Clin. Chim. Acta* 313, 139–142. doi: 10.1016/s0009-8981(01)00665-9

Taiwo, O., Wilson, G. A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., et al. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protoc.* 7, 617–636. doi: 10.1038/nprot.2012.012

Thierry, A. R., Mouliere, F., Gongora, C., Ollier, J., Robert, B., Ychou, M., et al. (2010). Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nucleic Acids Res.* 38, 6159–6175. doi: 10.1093/nar/gkq421

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Tian, Y., Arai, E., Gotoh, M., Komiyama, M., Fujimoto, H., and Kanai, Y. (2014). Prognostication of patients with clear cell renal cell carcinomas based on quantification of DNA methylation levels of CpG island methylator phenotype marker genes. *BMC Cancer* 14:772. doi: 10.1186/1471-2407-14-772

Udensi, U. K., and Tchounwou, P. B. (2016). Oxidative stress in prostate hyperplasia and carcinogenesis. *J. Exp. Clin. Cancer Res.* 35:139. doi: 10.1186/s13046-016-0418-8

van Roessel, S., Kasumova, G. G., Verheij, J., Najarian, R. M., Maggino, L., de Pastena, M., et al. (2018). International validation of the eighth edition of the American Joint Committee on Cancer (AJCC) TNM staging system in patients with resected pancreatic cancer. *JAMA Surg.* 153:e183617. doi: 10.1001/jamasurg.2018.3617

Widschwendter, M., and Jones, P. A. (2002). DNA methylation and breast carcinogenesis. *Oncogene* 21, 5462–5482. doi: 10.1038/sj.onc.1205606

Xu, R. H., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., et al. (2017). Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat. Mater.* 16, 1155–1161. doi: 10.1038/nmat4997

Xu, W., Lu, J., Zhao, Q., Wu, J., Sun, J., Han, B., et al. (2019). Genome-wide plasma cell-free DNA methylation profiling identifies potential biomarkers for lung cancer. *Dis. Markers* 2019:4108474. doi: 10.1155/2019/4108474

Yong, W. S., Hsu, F. M., and Chen, P. Y. (2016). Profiling genome-wide DNA methylation. *Epigenet. Chromatin* 9:26. doi: 10.1186/s13072-016-0075-3

Yu, Z., Feng, J., Wang, W., Deng, Z., Zhang, Y., Xiao, L., et al. (2020). The EGFR-ZNF263 signaling axis silences SIX3 in glioblastoma epigenetically. *Oncogene* 39, 3163–3178. doi: 10.1038/s41388-020-1206-7

Zhang, B., Shen, C., Ge, F., Ma, T., and Zhang, Z. (2017). Epigenetically controlled Six3 expression regulates glioblastoma cell proliferation and invasion alongside modulating the activation levels of WNT pathway members. *J. Neurooncol.* 133, 509–518. doi: 10.1007/s11060-017-2476-y

Zhang, L., Meng, X., Pan, C., Qu, F., Gan, W., Xiang, Z., et al. (2020). piR-31470 epigenetically suppresses the expression of glutathione S-transferase pi 1 in prostate cancer via DNA methylation. *Cell. Signal.* 67:109501. doi: 10.1016/j.cellsig.2019.109501

Zhao, M. T., Whyte, J. J., Hopkins, G. M., Kirk, M. D., and Prather, R. S. (2014). Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA. *Cell Reprogram* 16, 175–184. doi: 10.1089/cell.2014.0002

Zheng, Y., Zeng, Y., Qiu, R., Liu, R., Huang, W., Hou, Y., et al. (2018). The homeotic protein SIX3 suppresses carcinogenesis and metastasis through recruiting the LSD1/NuRD(MTA3) complex. *Theranostics* 8, 972–989. doi: 10.7150/thno.22328

**Conflict of Interest:** The authors declare that there is no conflict of interest regarding the publication of this paper.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2020
doi: 10.3389/fgene.2020.595361

# Identification of Hub Prognosis-Associated Oxidative Stress Genes in Pancreatic Cancer Using Integrated Bioinformatics Analysis

*Xin Qiu[1], Qin-Han Hou[2], Qiu-Yue Shi[1], Hai-Xing Jiang[1] and Shan-Yu Qin[1]**

*[1] Department of Gastroenterology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [2] Department of Neurosurgery, Affiliated Tumor Hospital of Guangxi Medical University, Nanning, China*

**Background:** Intratumoral oxidative stress (OS) has been associated with the progression of various tumors. However, OS has not been considered a candidate therapeutic target for pancreatic cancer (PC) owing to the lack of validated biomarkers.

**Methods:** We compared gene expression profiles of PC samples and the transcriptome data of normal pancreas tissues from The Cancer Genome Atlas (TCGA) and Genome Tissue Expression (GTEx) databases to identify differentially expressed OS genes in PC. PC patients' gene profile from the Gene Expression Omnibus (GEO) database was used as a validation cohort.

**Results:** A total of 148 differentially expressed OS-related genes in PC were used to construct a protein-protein interaction network. Univariate Cox regression analysis, least absolute shrinkage, selection operator analysis revealed seven hub prognosis-associated OS genes that served to construct a prognostic risk model. Based on integrated bioinformatics analyses, our prognostic model, whose diagnostic accuracy was validated in both cohorts, reliably predicted the overall survival of patients with PC and cancer progression. Further analysis revealed significant associations between seven hub gene expression levels and patient outcomes, which were validated at the protein level using the Human Protein Atlas database. A nomogram based on the expression of these seven hub genes exhibited prognostic value in PC.

**Conclusion:** Our study provides novel insights into PC pathogenesis and provides new genetic markers for prognosis prediction and clinical treatment personalization for PC patients.

**Keywords: pancreatic cancer, oxidative stress, prognosis, integrated bioinformatics analysis, risk model**

OPEN ACCESS

***Edited by:***
*Suman Ghosal,*
*National Institutes of Health (NIH),*
*United States*

***Reviewed by:***
*Hamed Bostan,*
*North Carolina State University,*
*United States*
*Vishal Midya,*
*Icahn School of Medicine at Mount Sinai, United States*

***Correspondence:**
*Shan-Yu Qin*
*qsy0511@163.com*

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

***Received:*** *16 August 2020*
***Accepted:*** *17 November 2020*
***Published:*** *08 December 2020*

***Citation:***
*Qiu X, Hou Q-H, Shi Q-Y, Jiang H-X and Qin S-Y (2020) Identification of Hub Prognosis-Associated Oxidative Stress Genes in Pancreatic Cancer Using Integrated Bioinformatics Analysis. Front. Genet. 11:595361. doi: 10.3389/fgene.2020.595361*

## INTRODUCTION

Pancreatic cancer (PC) is one of the most common tumors worldwide and is a severe threat to human health (Kamisawa et al., 2016). The 5-year overall survival rate of patients with PC is estimated at only 2–9% (Ilic and Ilic, 2016), and by 2030, PC is expected to become the second leading cause of cancer-associated death after lung cancer, ranking above breast and colorectal

cancers (Rahib et al., 2014). The poor outcomes of patients with PC are mainly associated with early metastasis, rapid progression, and a lack of sensitive screening tools for early diagnosis (Singhi et al., 2019). To date, surgical resection of cancer tissues remains the most common choice for PC treatment, which effectively increases patients' 5-year overall survival rate to 20–30%; however, less than 20% of PC patients are eligible for surgical treatment because of advanced-stage diagnoses, at which point cancer has already metastasized (Kamisawa et al., 2016).

In recent years, new developments in targeted molecular therapy, immunotherapy, and neoadjuvant therapy have demonstrated certain beneficial effects for PC; however, several side effects and questionable curative benefits for individual treatment must be addressed (Wu et al., 2019). Therefore, many studies have focused on constructing more effective prediction models that could better clarify the factors contributing to the prognosis and progression of PC, aiming to provide more evidence for individual treatment strategies. Despite these efforts, few screening biomarkers and tools have shown sufficient significance for widespread clinical application in PC. Thus, it is necessary to uncover additional biomarkers and construct novel tools with validated diagnostic value predicting individual diagnosis and prognosis in PC cases.

Oxidative stress (OS) is a pathological phenomenon in which an imbalances between oxidants and antioxidants production that results in the production of high levels of reactive oxygen species (ROS), which represent a potentially critical factor driving tumorigenesis and cancer progression (Brown and Wilson, 2004; Zhou et al., 2017; Kangari et al., 2018). ROS include several reactive non-radical and free radical species, such as singlet oxygen, hydrogen peroxide, and superoxide anion (Lü et al., 2010), which are dramatically elevated in patients with PC (Martinez-Useros et al., 2017). Previous studies have shown that as the scavenging potential is reduced, excessive ROS could damage the DNA causing genotoxicity (Zhou et al., 2013; Wang et al., 2017), eventually inducing genomic mutations that may initiate tumorigenesis (Oates and Gilkeson, 2006; Smith et al., 2010). In PC, ROS are linked to different factors, such as high alcohol intake, cigarette smoking, obesity, and inflammatory conditions (Nöthlings et al., 2005). ROS accumulation can significantly suppress apoptosis in PC cells and contributes to PC tumorigenesis and progression (Vaquero et al., 2004; Yu and Kim, 2014; Martinez-Useros et al., 2017). Accordingly, some compounds targeting OS, such as vitamins (Monti et al., 2012; Patacsil et al., 2012), curcumin (Dhillon et al., 2008; Bimonte et al., 2016), and coenzyme Q10 (Hertz and Lister, 2009) have been proposed as novel chemotherapeutic treatments for PC. Together, the studies discussed above indicate that OS is closely associated with PC progression. Nevertheless, the value of OS-related genes in PC prognosis prediction remains largely unclear, and the underlying mechanisms require further validation.

With the recent development of genomic technologies, bioinformatics analysis has been widely employed for identifying the interaction between gene signatures and tumors (Haqq et al., 2005; Qiu et al., 2015); however, a few studies have focused on identifying gene expression signatures to construct predictive models for patients with PC. Moreover, no systematic study has aimed to discover specific OS-related hub genes that correlate with cancer prognosis or progression. In the present study, we aimed to identify candidate OS genes that are significantly differentially expressed between PC and normal pancreatic tissues based on publicly available data obtained from The Cancer Genome Atlas (TCGA) and Genome Tissue Expression (GTEx) databases. Subsequently, protein-protein interaction (PPI) network construction, univariate Cox regression analysis, least absolute shrinkage and selection operator (LASSO) analyses were performed to identify hub genes among differentially expressed OS-related genes (DEOGs) that were significantly related to PC prognosis. Furthermore, we constructed a prognostic risk model based on hub gene expression and systematically explored each gene function and clinical significance in patients with PC. To the best of our knowledge, this is the first OS-associated risk model for prognostic prediction, which might provide novel insight into PC pathogenesis to tailor personalized treatment and improve the outcome for PC patients.

# MATERIALS AND METHODS

## Raw Data Acquisition

RNA-sequencing data of 178 PC samples and four normal tissues with corresponding clinical information were acquired from TCGA[1] (Liu et al., 2019). In addition, the transcriptome data of 167 whole normal pancreatic tissue samples were retrieved from the Genome Tissue Expression (GTEx) database[2] (Gentles et al., 2015; The GTEx Consortium, 2015). Gene expression profiles and clinical information of patients with PC from the Gene Expression Omnibus (GEO) GSE28735 (including 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues) and GSE62452 (including 69 pancreatic tumor and 61 adjacent non-tumor tissues) cohorts[3] were downloaded and merged as validation group (Huang et al., 2020). Detailed characteristics of the datasets are listed in **Supplementary Table 1**. The averages expression values of the probe sets were calculated for the same gene with multiple probe sets (Li et al., 2014). OS genes detected in over 80% of samples were identified, and the minimum non-zero value replaced zero-values in the corresponding gene in the expression matrix (Yan et al., 2019).

To screen out OS-associated genes, 1399 protein domains of OS, with a relevance score $\geq 7$ (approximately top 10% OS-related genes), were acquired from the GeneCards database[4] and subsequently applied for further exploration.

## Differential Gene Expression Analysis

To avoid inaccurate differential expression analysis caused by the small sample size of normal tissues, DEOGs between PC and normal pancreas tissues were identified from the TCGA and GTEx database. Original gene expression data were measured

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

[2]https://gtexportal.org/home/datasets

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

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

as fragments per kilobase of transcript per million mapped reads (FPKM) and log2-transformed. Furthermore, the RNA expression profiles were normalized with the R package "sva" to remove batch effects, as previously reported (Xiao et al., 2020; Zhang et al., 2020b). Then, the "limma" package in R was applied, and genes with an average count value lower than 1 were all excluded from further analyses. OS-related genes with a false discovery rate (FDR) $< 0.05$ and $|\log2$ fold change (FC)$| \geq 2$, which was calculated utilizing gene expression levels, were regarded as DEOGs in accordance with previously reported methods (Li et al., 2020) and visualized as a volcano plot and heatmap using the "ggplot2" and "pheatmap" packages in R (Wickham, 2009).

## Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Enrichment Analyses

Gene ontology and KEGG enrichment analyses of the identified DEOGs were performed to systematically understand the biological functions of the selected OS genes (Pathan et al., 2017). All analyses were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) 6.8 tool (Huang et al., 2009). Genes associated with GO terms and KEGG pathways with P and FDR values $< 0.05$ were considered to indicate significant enrichment.

## Construction of the PPI Network and Screening of Key Modules

The STRING platform[5] (Szklarczyk et al., 2019) was used to obtain PPI information for the DEOGs, and then explore the functional interactions between proteins (Szklarczyk et al., 2015). Subsequently, the interaction data were submitted to the Cytoscape 3.7.0 software to construct a PPI network. The Molecular Complex Detection (MCODE) plug-in was used to select the virtual modules and hub genes in the PPI network, with an MCODE score and node count $> 5$ and $P < 0.05$ (Bader and Hogue, 2003).

## Prognostic Model Construction and Efficacy Evaluation

To identify the prognosis-associated OS genes, hub genes identified in the PPI network were subjected to univariate Cox regression analysis using the "survival" package in R to identify genes that are highly crucial for patients' survival (Zhang et al., 2020a), with a cut-off criterion of $P < 0.05$. After that, genes identified to be significantly associated with the overall survival of PC patients through the univariate Cox regression analysis were integrated for analysis using LASSO, a widely used machine-learning algorithm, which can preserve valuable variables and avoid overfitting (Jiang et al., 2018), to complete the shrinkage of prognostic OS genes and categorizes patients into high- or low-risk subgroups. In the regression analysis, the normalized gene expression profile of candidate prognosis-associated DEOGs was set as the independent variable, whereas the response variables were the status and overall survival of PC patients. The optimal penalty parameter ($\lambda$) was identified via the minimum criteria (i.e., the value of $\lambda$ was accompanied with the lowest partial likelihood deviance), and 1000 iterations and ten-fold cross-validation was also applied to reduce the coefficient instability. The risk score for each sample was calculated using the following formula:

$$riskscore = \sum_{i=1}^{n} (Exp_i^* \beta_i)$$

where $Exp_i$ represents the relative expression value of the $i^{th}$ OS gene, and $\beta$ represents the regression coefficient. Genes screened through the LASSO analysis were selected as hub OS genes.

Based on the median risk score, PC patients in the TCGA cohort were stratified into low- and high-risk subgroups. The Kaplan-Meier method and log-rank test using the Kaplan-Meier "survival" package in R were further used to compare survival between two risk subgroups in PC samples (Klein and Moeschberger, 1997). The R packages "survivalROC" and "timeROC" were also applied to validate the predictive accuracy of the gene signature (Heagerty and Zheng, 2005). Univariate and multivariate Cox regression analyses were conducted to evaluate the relationship between clinical characteristics and risk scores. Besides, the same formula and regression coefficients described above were applied to the GSE28735 and GSE62452 validation cohorts to confirm the predictive applicability of our OS-related hub gene prognostic PC signature. Patients in the validation set were also stratified into low- and high-risk groups by the same median risk score calculated from the TCGA database.

## Hub Gene Evaluation

To validate the differential expression of the hub OS genes at the protein level, data from the Human Protein Atlas (HPA) online database[6] were used to compare the protein levels between normal pancreas and PC tumor tissues (Thul et al., 2017). The expression profile of these OS genes in PC was also verified in TCGA and validation cohorts. Furthermore, the Kaplan-Meier method was applied to estimate each gene's prognostic value in the TCGA-PC cohort. Finally, a nomogram incorporated with calibration plots was constructed based on the expression of hub prognosis-associated OS genes to be used as a predictive tool for the clinical outcome of patients with PC using the "rms" package in R (Gu et al., 2020).

# RESULTS

## Identification of DEOGs

Bioinformatics analysis of publicly available datasets was performed according to the workflow shown in **Figure 1**. A total of 1399 OS genes were obtained from the GeneCards database, and their differential expression between PC samples and normal tissues was explored. Of these, 148 genes were screened out as DEOGs in PC (FDR $< 0.05$ and $|\log_2 FC| \geq 2$), including

[5]http://www.string-db.org/

[6]http://www.proteinatlas.org/

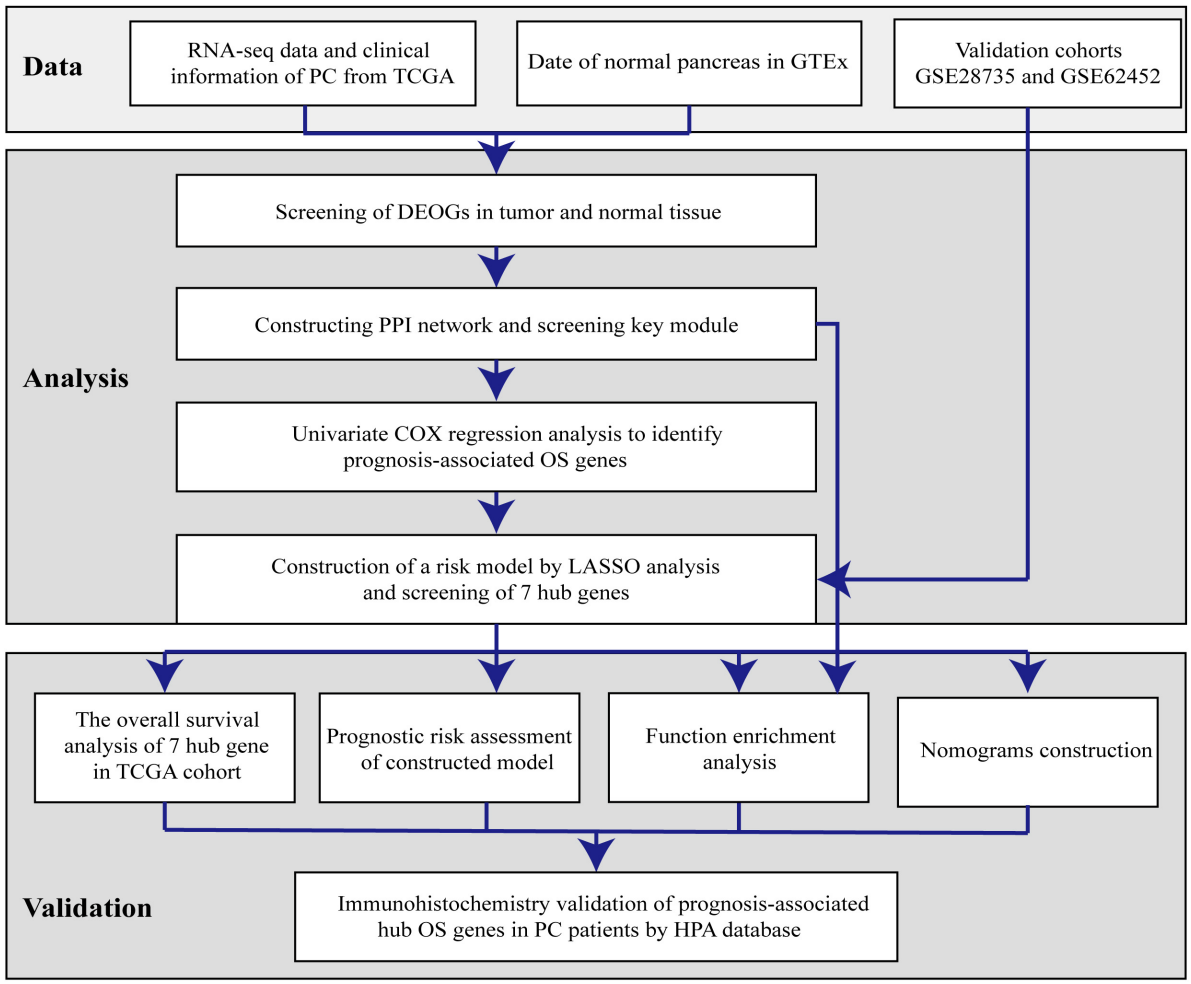



**FIGURE 1 |** Flowchart describing the schematic overview of the study design.

66 upregulated and 82 downregulated genes. The distribution of these genes is shown in **Figures 2A–C**.

## Functional Enrichment Analysis of DEOGs

Gene ontology analysis showed that, with respect to the upregulated DEOGs, the most enriched biological processes included the response to lipopolysaccharide, leukocyte migration, and extracellular structure organization (**Figure 3A**), whereas relative to the downregulated DEGs, intrinsic apoptotic signaling pathway, cellular oxidant detoxification, and cellular detoxification were most enriched terms (**Figure 3B**). In terms of cellular components, the upregulated genes were linked to enriched terms such as collagen-containing extracellular matrix, COPII-coated endoplasmic reticulum to Golgi transport vesicle, and focal adhesion (**Figure 3A**), whereas downregulated genes were associated with cytoplasmic vesicle lumen, vesicle lumen, and secretory granule lumen (**Figure 3B**). With regard to the molecular function GO terms, upregulated OS genes were linked to enriched terms including cytokine activity, receptor-ligand activity, and chemokine activity (**Figure 3A**), whereas the downregulated OS genes were associated with glutathione transferase activity, antioxidant activity, and transferase activity (**Figure 3B**). KEGG pathway analysis showed that the upregulated genes were enriched in viral myocarditis, proteoglycans in cancer, and fluid shear stress and atherosclerosis (**Figure 4A**), whereas the downregulated genes were mainly enriched in non-alcoholic fatty liver disease, platinum drug resistance, and drug metabolism-cytochrome P450 pathways (**Figure 4B**).

## Construction of the PPI Network for DEOGs and Screening of Key Modules

To further understand the inter-relationship among the DEOGs, we constructed a PPI network with 131 nodes and 934 edges (**Figure 5A**); in this network, the most significant module was identified to have 25 nodes and 235 edges (**Figure 5B**). Functional enrichment analysis indicated that the genes in the key module were mainly enriched in leukocyte migration, positive chemotaxis, and cell chemotaxis, whereas KEGG

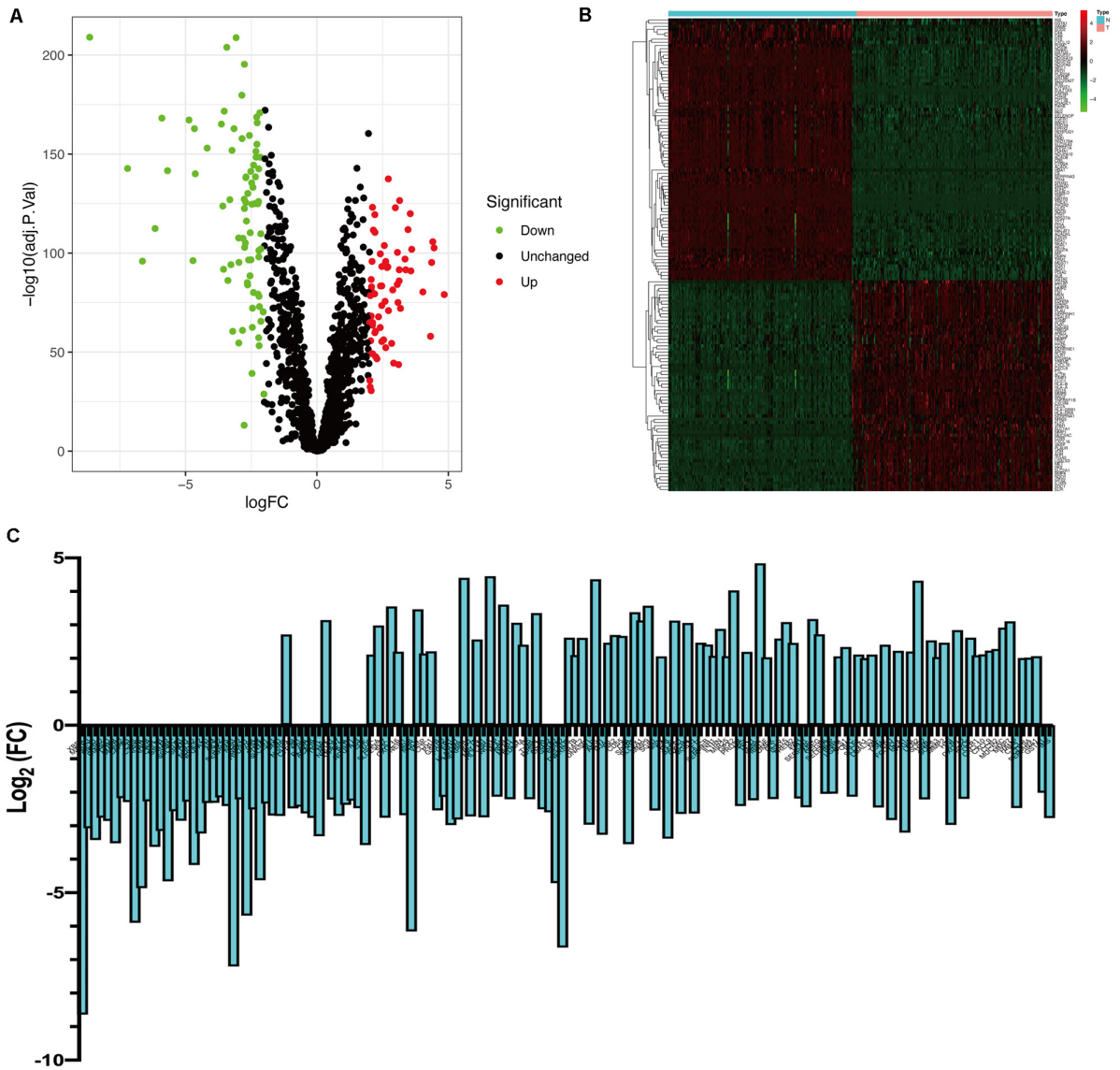



**FIGURE 2 |** Identification of differently expressed OS genes. **(A)** Volcano plot of DEOGs between TCGA-PC and GTEx-pancreas cohorts. **(B)** Heatmap of DEOGs. **(C)** Histogram of DEOGs.

analysis indicated that these genes were significantly enriched in pathways associated with bladder cancer, proteoglycans in cancer, and AGE-RAGE signaling pathway in diabetic complications (**Table 1**).

## Screening of Prognosis-Related OS Genes and Construction of a Genetic Risk Score Model for Patients With PC

To further identify prognosis-associated OS genes, the 131 DEOGs identified from the PPI network were further analyzed using univariate Cox regression analysis, revealing 25 OS genes demonstrating significant ($P < 0.05$) associations with patient overall survival (**Figure 6A**). Thereafter, a LASSO algorithm was employed for specific OS gene range shrinkage (**Figures 6B,C**), and seven hub OS genes (PLAU, CXCL10, CXCL9, MET, IL1RN, PAH, and PKD1) were ultimately selected to compute the risk score. All PC patients in the TCGA (**Figure 6D**) or validation (**Figure 6E**) cohorts were separated into low- and high-risk subgroups according to the median risk score. The coefficients of the seven hub genes are shown in **Table 2**.

## Associations Between Prognostic Risk Score and Clinical Characteristics of PC Patients

Univariate and multivariate Cox regression analyses (**Figures 7A,B**) showed that our identified risk score was

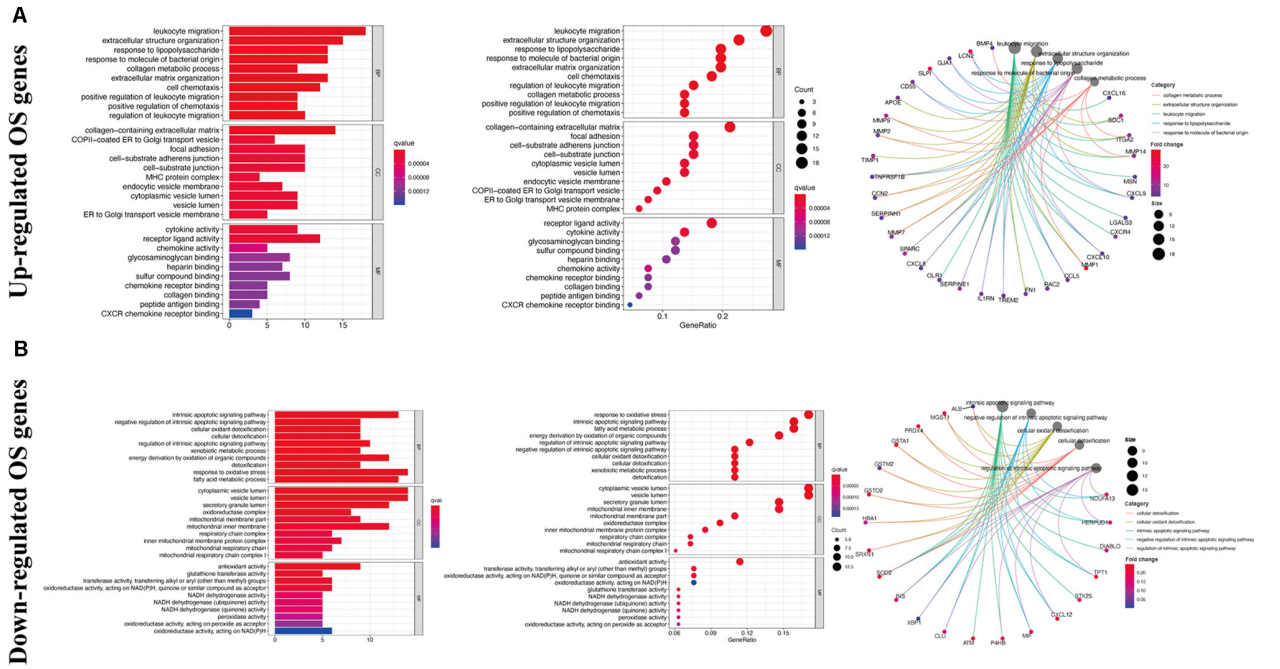



**FIGURE 3 |** Gene ontology enrichment analysis of DEOGs. **(A)** Top 10 classes of GO enrichment terms about up-regulated DEOGs in biological process (BP), cellular component (CC), and molecular function (MF). **(B)** Top 10 classes of GO enrichment terms about down-regulated DEOGs in BP, CC, and MF.

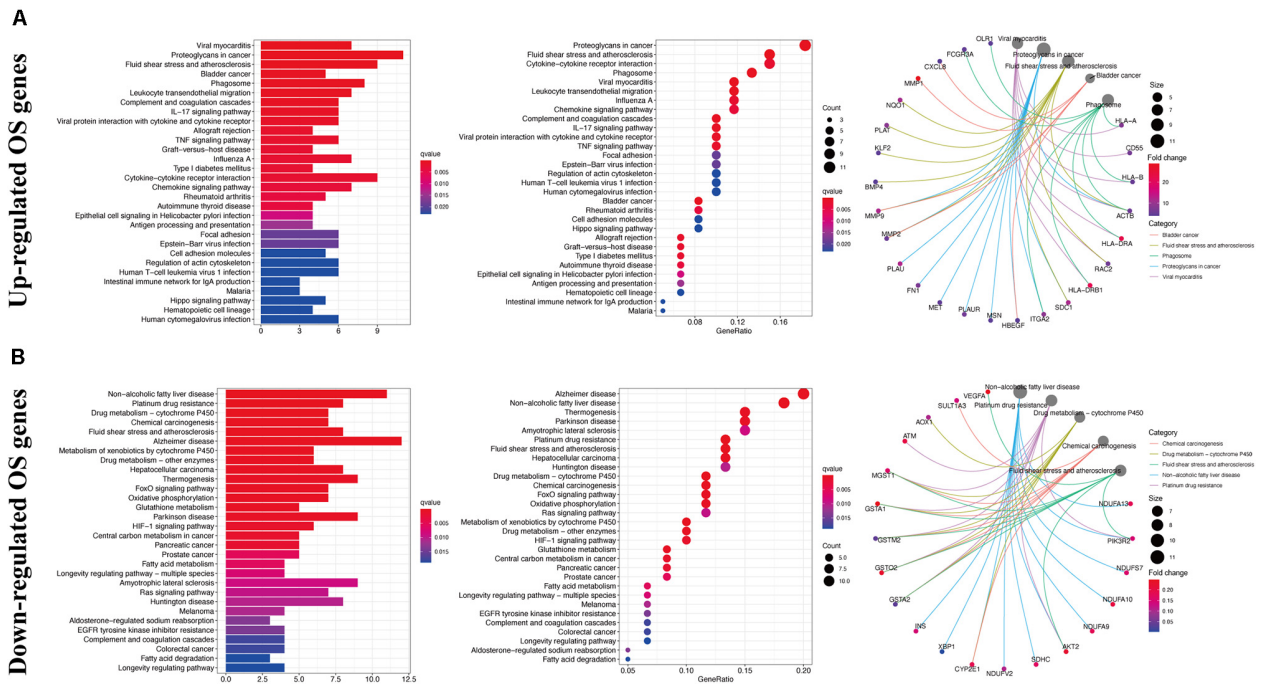



**FIGURE 4 |** Kyoto encyclopedia of genes and genomes enrichment analysis of DEOGs. **(A)** Top 30 classes of KEGG enrichment terms about up-regulated DEOGs. **(B)** Top 30 classes of KEGG enrichment terms about down-regulated DEOGs.

significantly connected with PC patient prognosis and emerged as an independent prognostic feature. Expectedly, the predictive value analysis of our risk score model in the TCGA cohort showed that it was significantly associated with the overall survival of patients with PC ($P < 0.05$), and the AUC (area under the receiver operating characteristic curve) reached 0.798 and

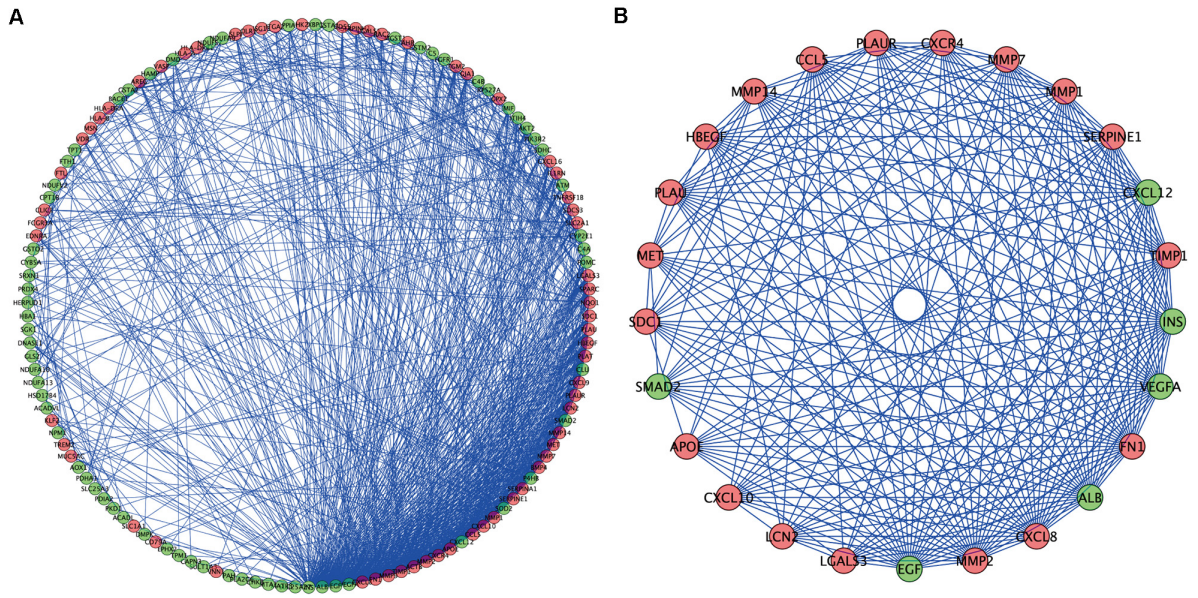



**FIGURE 5 |** Protein–protein interaction network and modules screening. **(A)** PPI network of DEOGs. **(B)** Critical module from PPI network. Green circles represent down-regulated genes, and red circles represent up-regulated genes.

**TABLE 1 |** Kyoto encyclopedia of genes and genomes pathway and GO enrichment analysis of OS genes in key module.

| | Terms | *P*-value | FDR |
|---|---|---|---|
| **GO enrichment** | | | |
| **Biological processes** | Leukocyte migration | 7.68E-10 | 3.68E-10 |
| | Positive chemotaxis | 1.67E-09 | 7.98E-10 |
| | Cell chemotaxis | 1.67E-09 | 7.98E-10 |
| | Positive regulation of leukocyte migration | 1.67E-09 | 7.98E-10 |
| | Positive regulation of chemotaxis | 2.05E-09 | 9.83E-10 |
| **Cellular component** | Cytoplasmic vesicle lumen | 5.43E-09 | 2.82E-09 |
| | Vesicle lumen | 5.43E-09 | 2.82E-09 |
| | Platelet alpha granule lumen | 5.43E-09 | 2.82E-09 |
| | Platelet alpha granule | 2.67E-08 | 1.38E-08 |
| | Secretory granule lumen | 6.08E-08 | 3.16E-08 |
| **Molecular function** | Receptor ligand activity | 6.77E-08 | 3.08E-08 |
| | Heparin binding | 7.72E-06 | 3.51E-06 |
| | Chemoattractant activity | 1.08E-05 | 4.89E-06 |
| | CXCR chemokine receptor binding | 1.32E-05 | 6.01E-06 |
| | Cytokine activity | 1.32E-05 | 6.01E-06 |
| **KEGG pathway** | Bladder cancer | 1.73E-07 | 1.04E-07 |
| | Proteoglycans in cancer | 2.05E-07 | 1.24E-07 |
| | AGE-RAGE signaling pathway in diabetic complications | 1.36E-05 | 8.24E-06 |
| | Rheumatoid arthritis | 1.90E-04 | 1.15E-04 |
| | Viral protein interaction with cytokine and cytokine receptor | 2.17E-04 | 1.31E-04 |

0.898 for 3- and 5-year survival, respectively (**Figures 7C,D**). Of note, the same prognostic capacity of seven genes' prognostic signature was also validated in the GEO validation cohort. The survival analysis results also indicated that the overall survival of patients with PC was significantly decreased, as evidenced by an increased risk score in the validation cohort ($P$ = 0.029; **Figure 7E**). In addition, time-dependent receiver operating characteristic (ROC) curve analysis of overall survival in patients with PC indicated that our prediction model had moderate predictive accuracy with an AUC value of 0.819 and 0.872 for 3- and 5-year survival, respectively, in the GEO cohorts (**Figure 7F**), which demonstrated that our prognostic model had reliable specificity and sensitivity for patients with PC. Moreover, while compared with other clinicopathological

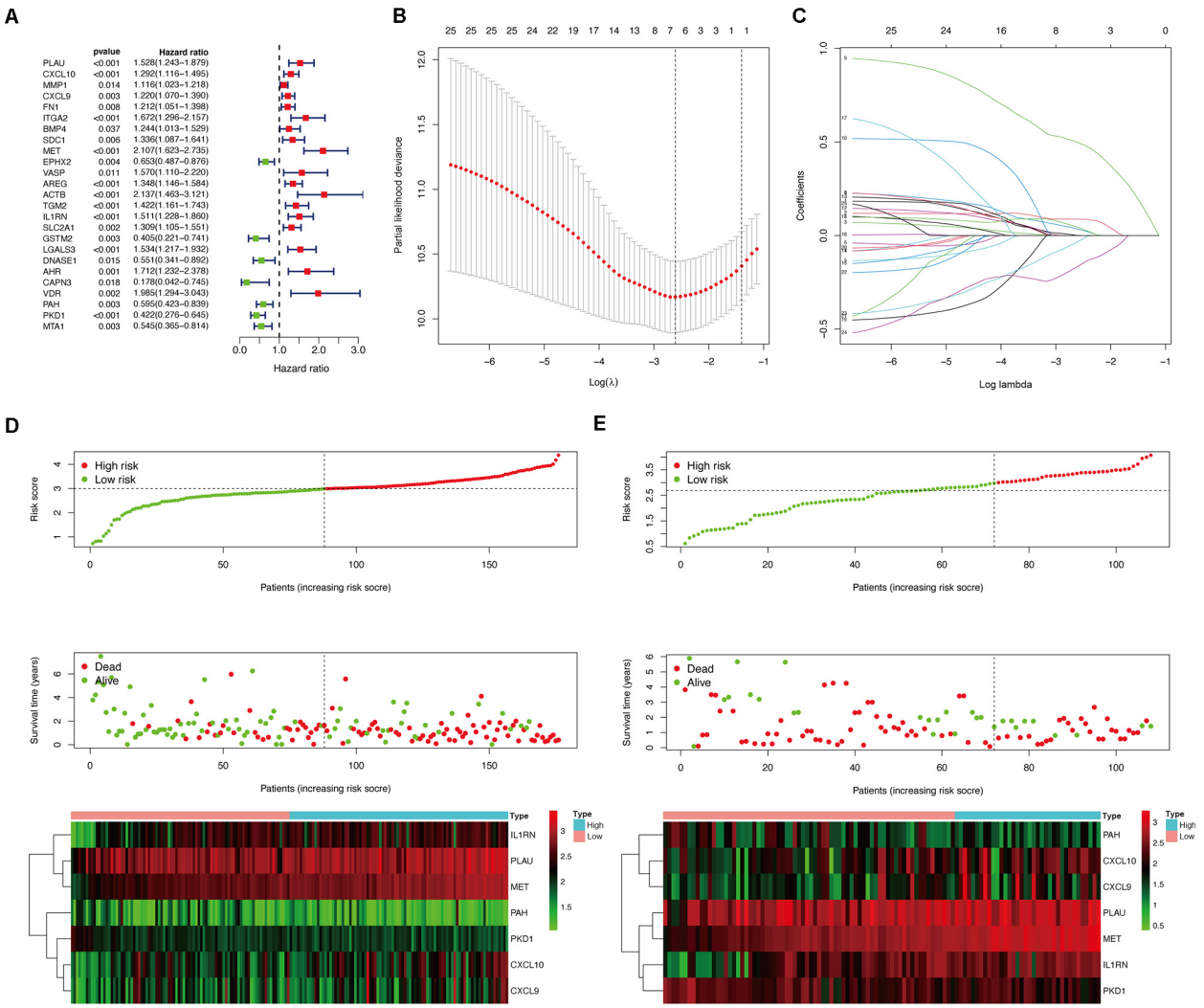



**FIGURE 6 |** Construction of prognostic model in the TCGA and validation cohorts. **(A)** Univariate Cox regression analysis for identifying prognosis-related OS genes in TCGA cohort. **(B,C)** LASSO analysis for determining the number of factors and constructing the prognosis prediction model. **(D)** Risk score distribution, survival status, and expression heat map of TCGA cohort. **(E)** Risk score distribution, survival status, and expression heat map of validation cohort.

**TABLE 2 |** Seven prognosis-associated OS genes with PC in the TCGA dataset were identified by LASSO analysis.

| OS name | Univariate Cox regression analysis | | | | LASSO coefficient | Value of log lambda |
|---|---|---|---|---|---|---|
| | HR | Lower 95% CI | Upper 95% CI | *P*-value | | |
| PLAU | 1.5283 | 1.2431 | 1.8790 | 0.0001 | 0.0077 | −2.6226 |
| CXCL10 | 1.2917 | 1.1164 | 1.4947 | 0.0006 | 0.0879 | −1.7799 |
| CXCL9 | 1.2196 | 1.0701 | 1.3900 | 0.0029 | 0.0477 | −2.1978 |
| MET | 2.1067 | 1.6230 | 2.7345 | 2.1552 | 0.5167 | −1.1219 |
| IL1RN | 1.5113 | 1.2278 | 1.8602 | 0.0001 | 0.0620 | −2.1525 |
| PAH | 0.5952 | 0.4225 | 0.8386 | 0.0030 | −0.0620 | −2.3475 |
| PKD1 | 0.4222 | 0.2762 | 0.6454 | 0.0001 | −0.1978 | −1.6825 |

characteristics in the validation cohort, our ROC curve analysis indicated that our risk model outcompeted other diagnostic features in terms of reliably and accurately predicting 3- and 5-year survival (**Figure 7H**). Of course, this improved predictive value was also calculated in the TCGA cohort at 3 and 5 years (**Figure 7G**).

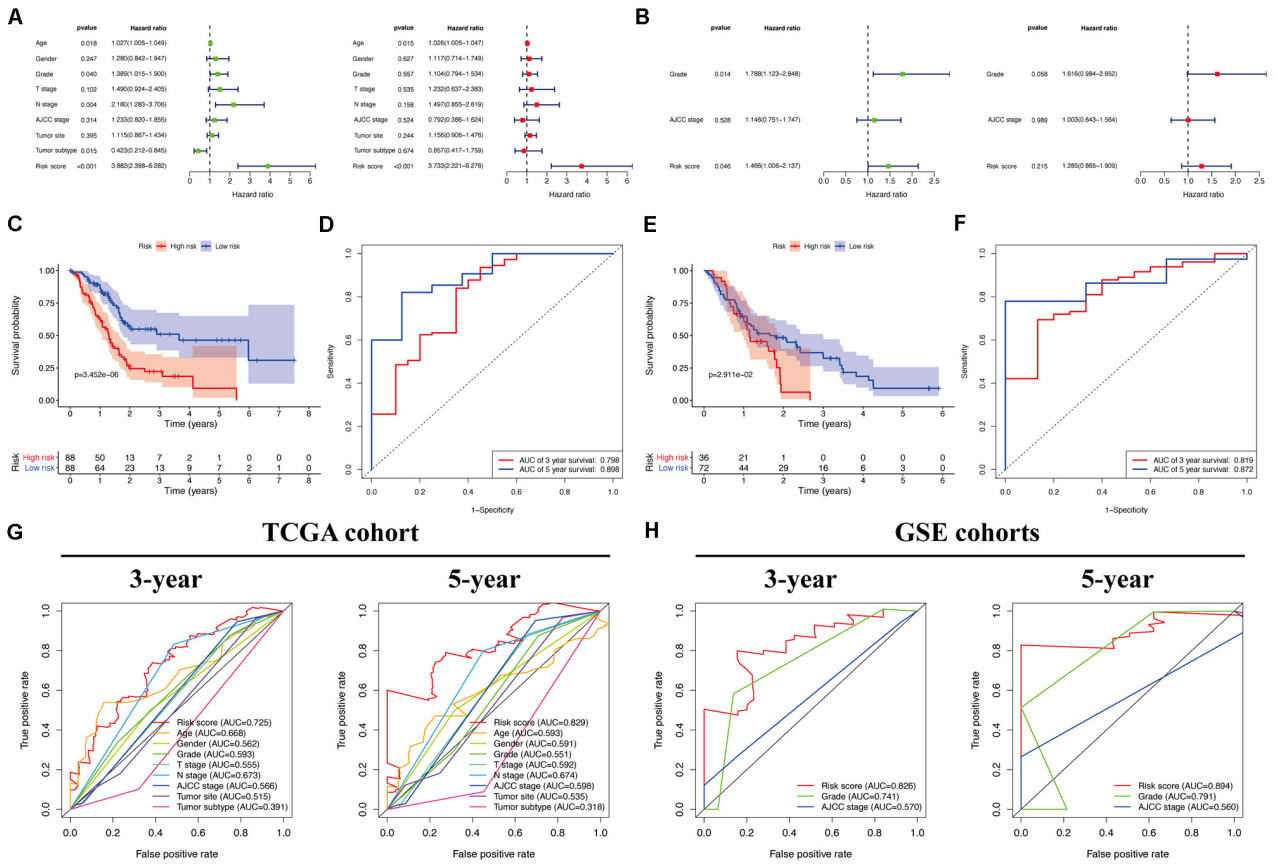



**FIGURE 7 |** Efficacy evaluation of constructed prognostic model. Univariate and multivariate Cox regression analysis of the clinicopathological features in TCGA **(A)** and validation **(B)** cohorts. **(C)** Survival curve of TCGA cohort. **(D)** TimeROC curves for forecasting overall survival in TCGA cohort. **(E)** Survival curve of validation cohort. **(F)** TimeROC curves for forecasting overall survival in validation cohort. ClinicalROC curves for forecasting overall survival in TCGA **(G)** and validation **(H)** cohort.

Pancreatic cancer at a higher T stage (**Figure 8G**), American Joint Committee on Cancer (AJCC) stage (**Figure 8H**), or tumor grade (**Figure 8I**) had a significantly increased risk score ($P < 0.05$), indicating that our risk model reliably predicted PC progression. Interestingly, tumors of grade 4 had the lowest risk score, which may be due to the minimal number of grade 4 PC tissues in the analyzed sample. Cancers histologically diagnosed as pancreatic ductal adenocarcinoma (PDAC) were significantly associated with higher risk scores than other PC subtypes in the TCGA cohort ($P < 0.05$; **Figure 8J**). In the validation cohort, patients with a higher tumor grade also had a higher risk score ($P < 0.05$; **Figure 8K**).

To further clarify the modulation mechanism of the risk score in predicting the overall survival of patients with PC, we also determined the relationship between the clinicopathological features and overall survival. The results indicated significantly poorer outcomes for PC samples of the PDAC subtype or PC samples with higher T stage, N stage, AJCC stage, and tumor grade (**Figures 8A–F**), suggesting that our risk model is strongly associated with the overall survival of PC patients by accurately predicting cancer progression and subtypes. Heatmaps constructed using the TCGA and validation cohorts for the expression levels of the seven hub OS-related genes in the two risk subgroups (**Figures 8L,M**) and showed significant differences in tumor grade between groups, in both cohorts ($P < 0.05$). These results indicated that our prognostic model has remarkable potential for predicting PC outcomes and clinical features.

## Prognostic Value of Hub OS-Related Genes

As shown in **Figures 9A,B**, among the seven hub genes, the expression levels of PLAU, CXCL10, CXCL9, MET, and IL1RN were significantly elevated, whereas the expression levels of PAH and PKD1 were significantly decreased in PC samples compared with those in the normal pancreas samples. Similar results were obtained by analyzing these hub OS-related genes' protein expression levels using the immunohistochemistry results from the HPA database (**Figures 9C–G**). Kaplan-Meier analysis further showed that the overall survival of patients with PC was inversely associated with the gene expression levels of PLAU, CXCL10, CXCL9, MET, and IL1RN ($P < 0.05$; **Figures 10A–E**); however, the expression levels of PAH and PKD1 had positive associations with the prognosis of patients with PC ($P < 0.05$; **Figures 10F,G**).

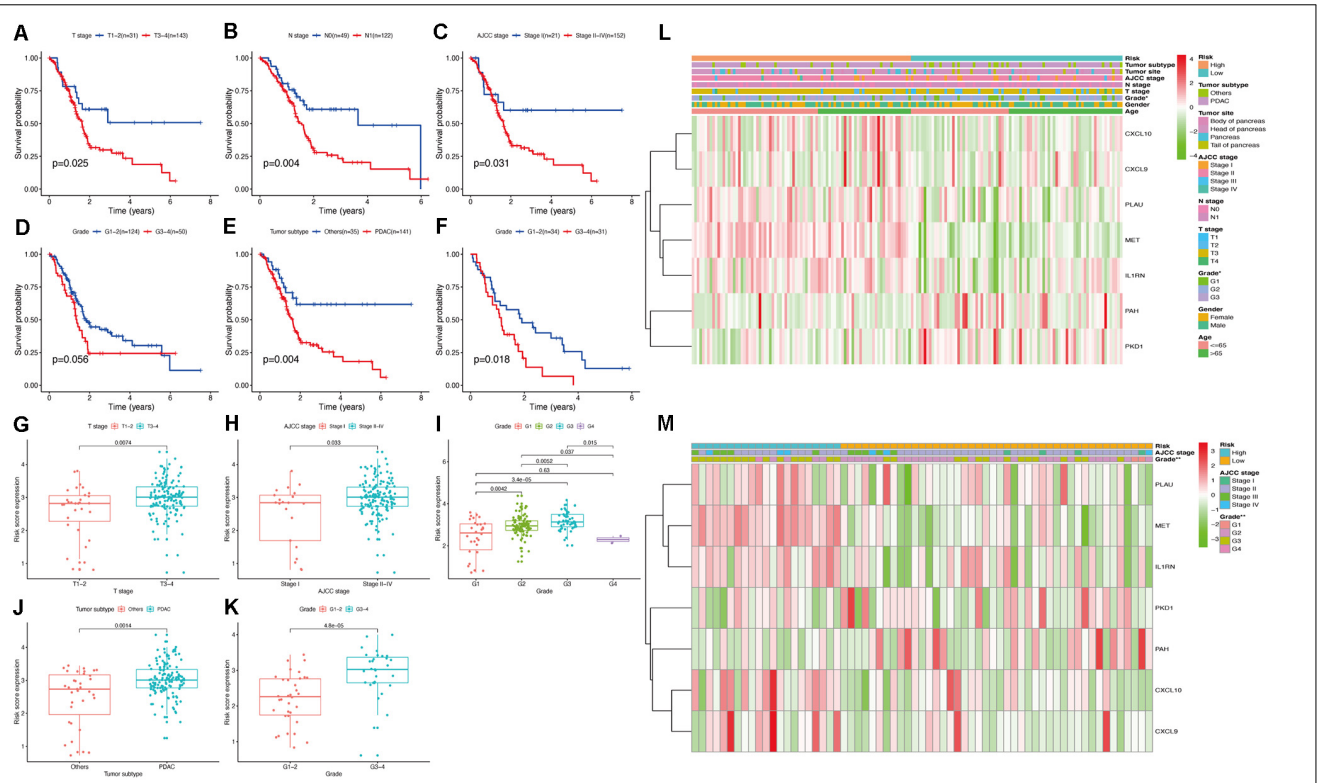



**FIGURE 8 |** Evaluation the relationship between the risk score and clinicopathological parameters in patients with PC. **(A–E)** Survival curve of clinicopathological characters in TCGA cohort. **(F)** Survival curve of clinicopathological characters in validation cohort. **(G–J)** Correlation analysis between the risk score and clinicopathological characters in TCGA cohort. **(K)** Correlation analysis between the risk score and clinicopathological characters in validation cohort. The heatmap shows the distribution of clinicopathological features and OS genes expression in two risk subgroups from the TCGA **(L)** and validation **(M)** cohorts.

A similar prognostic trend was also discovered in the validation cohort (**Supplementary Figure 1**), whereas only genes PLAU and MET genes were significantly associated with the prognosis of patients with PC ($P < 0.05$), which might be due to the small number of PC samples and the unequal composition of patients with PC (in the validation cohort, no PC patient had an overall survival of more than 3 years). Therefore, further experiments are warranted to validate the specific role of these seven hub OS-related genes in the prognosis of patients with PC.

## Nomogram Construction

Finally, to enable the identified hub genes to be applied for predicting the overall survival of patients with PC in a practical setting, the nomogram plots based on the expression levels of the seven hub genes were constructed to predict the clinical outcome of patients with PC in the TCGA-PC (**Figure 11A**) and validation cohorts (**Figure 11C**). The calibration plots demonstrated that our nomograms showed good agreement between the predicted and observed outcomes (**Figures 11B,D**).

# DISCUSSION

Pancreatic cancer is one of the most common malignancies and a major cause of cancer-related deaths worldwide (Kamisawa et al., 2016). Although many novel diagnostic techniques and molecular biomarkers have been recently discovered, they have not sufficiently improved the early diagnosis and prognosis of patients with PC (Yan et al., 2019). Therefore, it is imperative to identify more PC prognosis-associated biomarkers and elucidate the precise mechanism underlying cancer progression. In the present study, we aimed to identify reliable molecular biomarkers for the prognostic assessment of PC and provide a basis for treatment decisions. To this end, we focused on OS as a confirmed mechanism of cancer progression and applied differential expression analysis to identify candidate DEOGs between PC and healthy pancreatic samples. A total of 148 DEOGs were selected for further exploration. In addition, the KEGG pathway enrichment analysis indicated that our identified DEOGs were not only significantly associated with the prognosis of pancreatic cancer, but also played a critical role in numerous other tumors, including bladder cancer, hepatocellular carcinoma, prostate cancer, melanoma, and colorectal cancer, prompting us to further explore the potential role of OS genes in other tumors.

The PPI network, univariate Cox regression, and LASSO analysis of the DEOGs identified a total of seven genes (PLAU, CXCL10, CXCL9, MET, IL1RN, PAH, and PKD1) as hub prognosis-associated genes for further exploration. The mRNA and protein expression profiles of these seven genes using

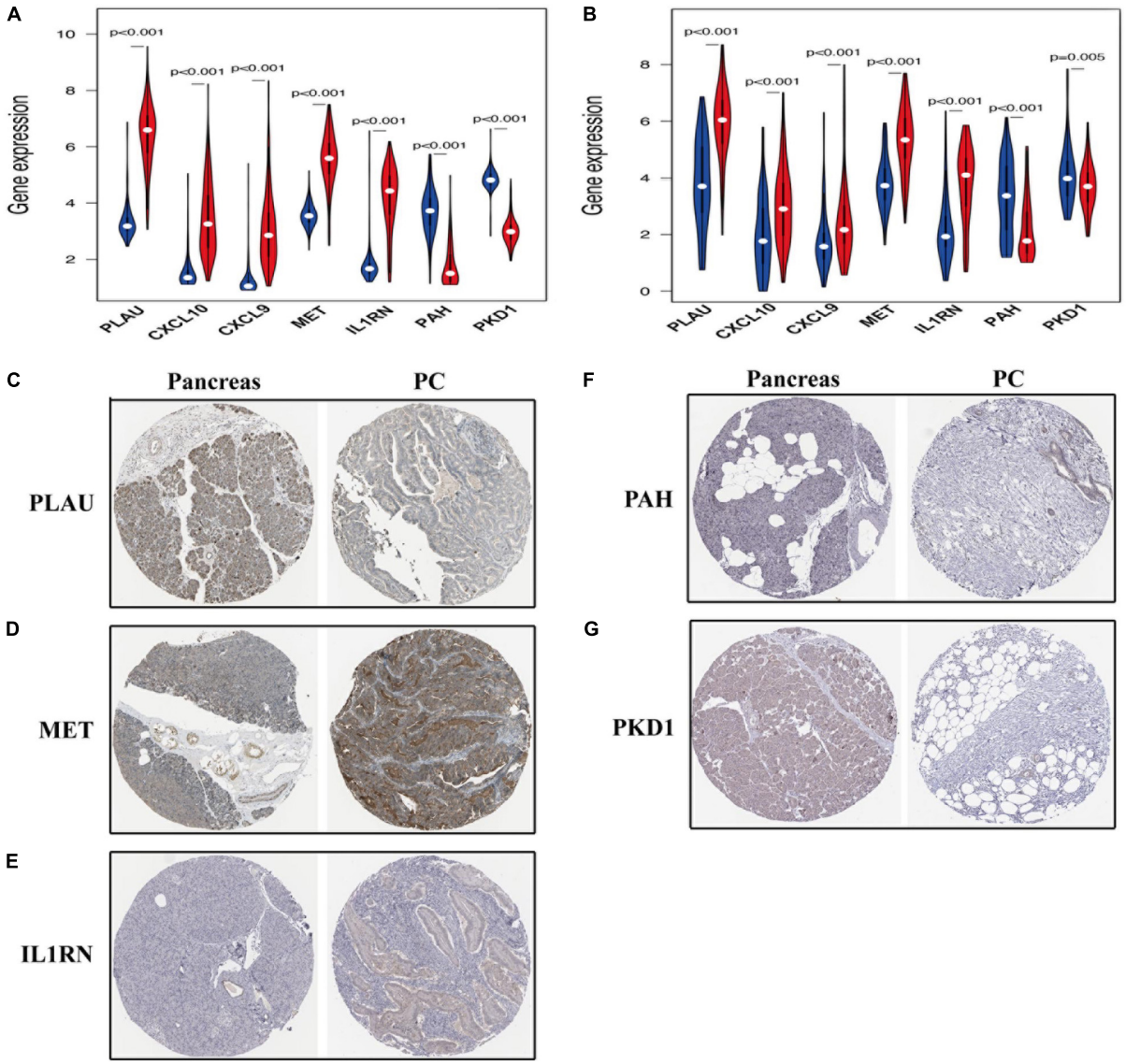



**FIGURE 9 |** The expression of prognosis-related OS genes in patients with PC. The violin plot reveals the transcription expression of OS genes in TCGA **(A)** and validation **(B)** cohorts. HPA database verifies the protein expression of PLAU **(C)**, MET **(D)**, IL1RN **(E)**, PAH **(F)**, and PKD1 **(G)** in PC.

the expression data from TCGA-PC and GEO (GSE28735 and GSE62452) cohorts and the HPA database revealed that PLAU, CXCL10, CXCL9, MET, and IL1RN were overexpressed, whereas PAH and PKD1 were downregulated in PC tissues. Kaplan-Meier analysis further revealed that these overexpressed hub genes were negatively associated with the overall survival of patients with PC, whereas PAH and PKD1 expression levels positively correlated with patient outcomes. These results might correspond with the modulation effects of these genes in PC metastasis and growth, as previously reported.

PLAU is reportedly significantly overexpressed in PC samples and associated with pancreatic cell invasive ability (Bournet et al., 2012; Liu et al., 2016). Several bioinformatics analyses also indicated that PLAU has prognostic value in PC (Lu et al., 2018; Chen et al., 2019). ELR-negative CXC chemokines, CXCL9 and CXCL10 were shown to induce lymphocytic migration and attenuate angiogenesis, leading to longer overall survival in patients with advanced PDAC (Qian et al., 2019). However, some studies also indicated that these chemokines might play tumorigenic roles by promoting tumor metastasis and proliferation (Mir et al., 2015; Wightman et al., 2015); thus, the specific roles of CXCL9 and CXCL10 in PC remain unclear. MET was originally identified as an oncogene that displayed 7-fold increased expression levels in PC samples, and its overexpression directly correlated with tumor grade and an aggressive PC phenotype (Modica et al., 2018). Protein kinase D1 (PKD1) is one of three members of the PKD family of serine/threonine kinases, which can be activated by intracellular

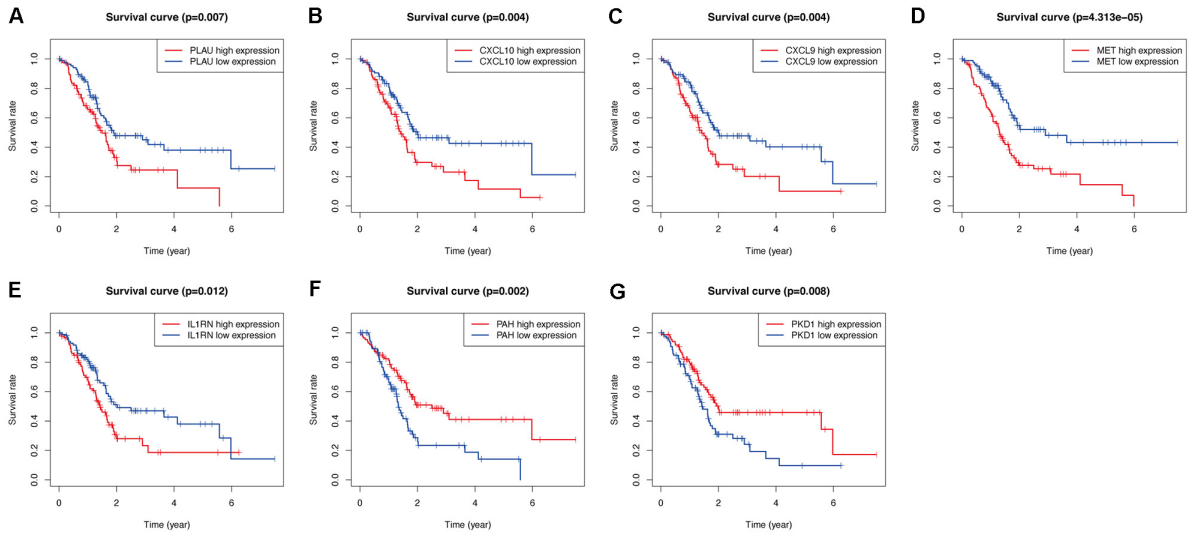



**FIGURE 10 |** Validation the prognostic value of the prognosis-related OS genes of PLAU **(A)**, CXCL10 **(B)**, CXCL9 **(C)**, MET **(D)**, IL1RN **(E)**, PAH **(F)**, and PKD1 **(G)** in TCGA-PC cohort by Kaplan-Meier analysis.

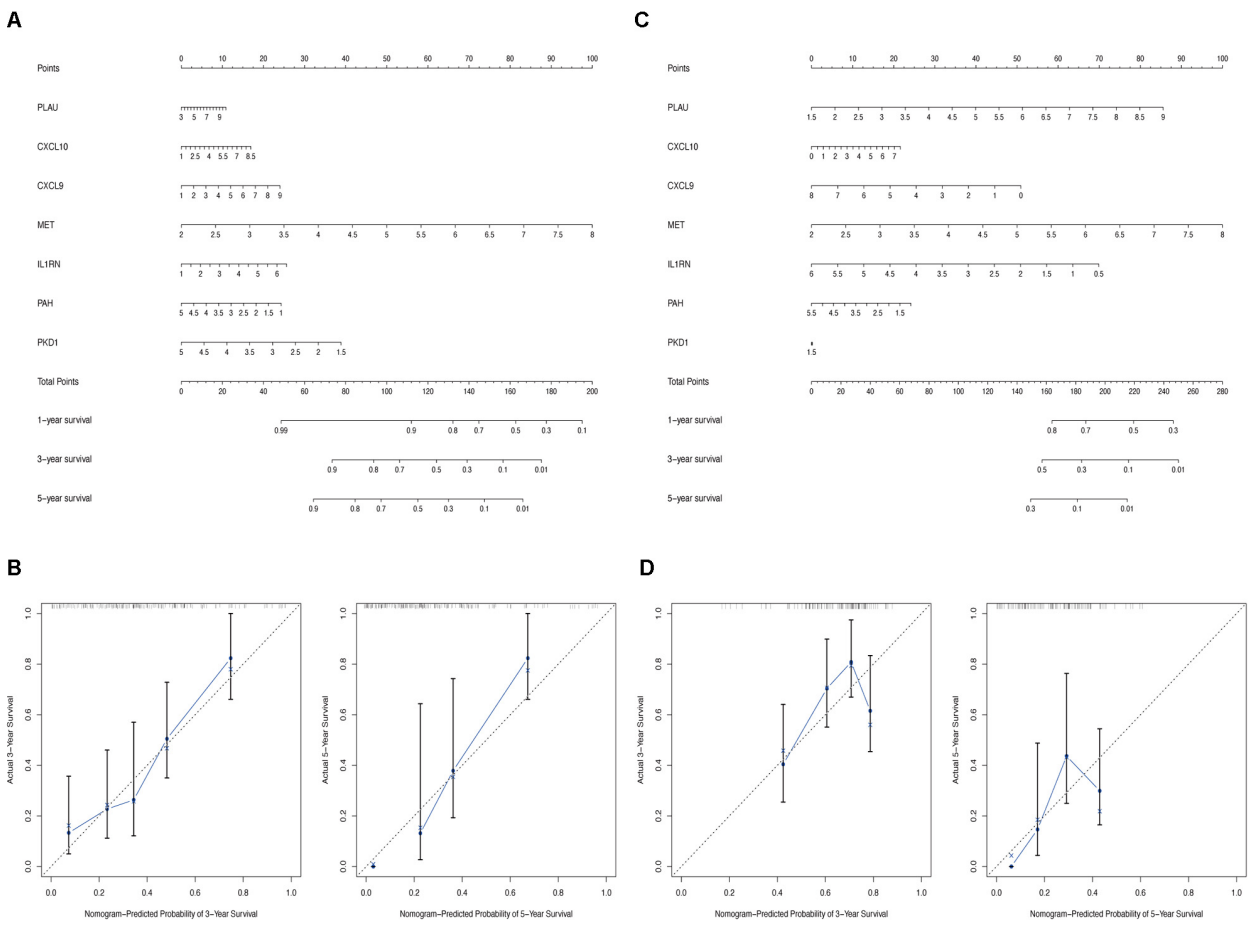



**FIGURE 11 |** Construction of nomogram based on the expression of 7 hub OS genes. The nomogram **(A)** and calibration plots **(B)** of hub OS genes in TCGA cohort. The nomogram **(C)** and calibration plots **(D)** of hub OS genes in validation cohort.

OS (Waldron and Rozengurt, 2000), and its activation has proven to contribute to the initiation of PC (Döppler and Storz, 2017). Although some of our identified hub genes were previously reported to be significantly associated with PC progression, no study has systematically analyzed the specific prognostic role of OS genes in PC. In the current study, we demonstrated that the differential expression of seven hub OS-related genes is significantly associated with patients' overall survival and involved in PC development. Nevertheless, to validate these OS-related genes as potential prognostic biomarkers for PC, more experimental evidences from prospective clinical and pre-clinical studies are needed. Future studies must verify whether PC patients could benefit from the modulation of these genes and the exact relationship between these genes and PC cells.

Moreover, to identify whether these specific OS genes could be used as prognostic factors, we constructed a novel prognostic prediction model based on the expression of the seven hub genes. To our knowledge, this is the first OS-associated risk model for prognostic prediction. Univariate and multivariate Cox regression analyses revealed that our risk model had reliable prognostic value for PC and could be used as an independent prognostic factor in PC. Survival and ROC analyses confirmed the advantage of the biological implications of our OS hub genes-related risk model for predicting PC prognosis. They showed improved predictive accuracy compared with conventional clinicopathological features, such as age, sex, AJCC stage, tumor grade, tumor site and tumor subtype. In addition, considering the critical role of OS in various stages of cancer progression and carcinogenesis (Reuter et al., 2010; Hecht et al., 2016), we further assessed the connections between risk score and PC clinical factors and discovered that the constructed risk model was significantly associated with T stage, AJCC stage, grade, and subtype of cancer samples, which was consistent with the prognostic effects of clinical features in overall survival. The AJCC staging system is one of the most widely used clinicopathological parameters for PC prognosis prediction (Kamarajah et al., 2017). However, the AJCC staging model is still not suitable for elucidating comprehensive PC behaviors and does not have sufficient diagnostic accuracy for PC (Yan et al., 2019). A similar conclusion was made in this study. Compared with the AJCC stage, our risk model not only showed a stronger relationship with PC prognosis but also could effectively predicted other PC features, including tumor grade and subtypes. These results indicate that our risk model has great advantages in the prognosis prediction of patients with PC. Our nomogram analysis confirmed the credibility of the identified OS genes in predicting the overall survival of patients with PC. Taken together, our results demonstrate the prognostic value of an OS-related gene-based risk model for patients with PC and suggest a novel method for evaluation the survival rate of PC patients.

Nonetheless, this study has some limitations. First, this study was designed as a retrospective analysis; thus, more prospective research should be performed to verify our results. Second, our results lack *in vitro* or *in vivo* exploration to confirm the reliability of the mechanism analysis. Therefore, several experiments are needed to prove the mechanistic connections between the identified hub genes and PC progression.

## CONCLUSION

In conclusion, through a series of bioinformatics analyses, we identified seven hub OS-related genes that are significantly associated with the overall survival of patients with PC. We also successfully established a prognostic model with powerful predictive effects and developed an effective nomogram composed of the gene signature in PC patients. Thus, our study foretells that these OS genes will greatly contribute to explain the pathogenesis and progression mechanism of PC and may serve as potential therapeutic targets to treat PC patients.

## 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/s.

## AUTHOR CONTRIBUTIONS

H-XJ and S-YQ conceived and designed the research, conducted the experiments, analyzed the data, and wrote the manuscript. XQ and Q-HH participated in the collection of clinical samples. Q-YS participated in the experimental design and provided financial and instrumental support. All authors read and approved the final manuscript.

## FUNDING

This work has been financially supported by the National Natural Science Foundation of China (Grant Nos. 31560257 and 81960439) and the "139" Plan for Training High Level Cadre Talents in Guangxi Medicine (G201903004).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.595361/full#supplementary-material

**Supplementary Figure 1 |** Validation the prognostic value of 7 prognosis-related OS genes in validation cohort by Kaplan-Meier analysis.

**Supplementary Table 1 |** Details of datasets in this study.

## REFERENCES


Bader, G. D., and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *Bmc Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Bimonte, S., Barbieri, A., Leongito, M., Piccirillo, M., Giudice, A., Pivonello, C., et al. (2016). Curcumin AntiCancer studies in pancreatic cancer. *Nutrients* 8:433. doi: 10.3390/nu8070433

Bournet, B., Pointreau, A., Souque, A., Oumouhou, N., Muscari, F., Lepage, B., et al. (2012). Gene expression signature of advanced pancreatic ductal adenocarcinoma using low density array on endoscopic ultrasound-guided fine needle aspiration samples. *Pancreatology* 12, 27–34. doi: 10.1016/j.pan.2011.12.003

Brown, J. M., and Wilson, W. R. (2004). Exploiting tumour hypoxia in cancer treatment. *Nat. Rev. Cancer* 4, 437–447. doi: 10.1038/nrc1367

Chen, Q., Yu, D., Zhao, Y., Qiu, J., Xie, Y., and Tao, M. (2019). Screening and identification of hub genes in pancreatic cancer by integrated bioinformatics analysis. *J. Cell. Biochem.* 120, 19496–19508. doi: 10.1002/jcb.29253

Dhillon, N., Aggarwal, B. B., Newman, R. A., Wolff, R. A., Kunnumakkara, A. B., Abbruzzese, J. L., et al. (2008). Phase II trial of curcumin in patients with advanced pancreatic cancer. *Clin. Cancer Res.* 14, 4491–4499. doi: 10.1158/1078-0432.CCR-08-0024

Döppler, H., and Storz, P. (2017). Mitochondrial and oxidative stress-mediated activation of protein kinase D1 and its importance in Pancreatic Cancer. *Front. Oncol.* 7:41. doi: 10.3389/fonc.2017.00041

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* 21, 938–945. doi: 10.1038/nm.3909

Gu, H. Y., Zhang, C., Guo, J., Yang, M., Zhong, H. C., Jin, W., et al. (2020). Risk score based on expression of five novel genes predicts survival in soft tissue sarcoma. *Aging* 12, 3807–3827. doi: 10.18632/aging.102847

Haqq, C., Nosrati, M., Sudilovsky, D., Crothers, J., Khodabakhsh, D., Pulliam, B. L., et al. (2005). The gene expression signatures of melanoma progression. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6092–6097. doi: 10.1073/pnas.0501564102

Heagerty, P. J., and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* 61, 92–105. doi: 10.1111/j.0006-341X.2005.030814.x

Hecht, F., Pessoa, C. F., Gentile, L. B., Rosenthal, D., Carvalho, D. P., and Fortunato, R. S. (2016). The role of oxidative stress on breast cancer development and therapy. *Tumour. Biol.* 37, 4281–4291. doi: 10.1007/s13277-016-4873-9

Hertz, N., and Lister, R. E. (2009). Improved survival in patients with end-stage cancer treated with coenzyme Q(10) and other antioxidants: a pilot study. *J. Int. Med. Res.* 37, 1961–1971.

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, 44–57. doi: 10.1038/nprot.2008.211

Huang, R., Mao, M., Lu, Y., Yu, Q., and Liao, L. (2020). A novel immune-related genes prognosis biomarker for melanoma: associated with tumor microenvironment. *Aging* 12, 6966–6980. doi: 10.18632/aging.103054

Ilic, M., and Ilic, I. (2016). Epidemiology of pancreatic cancer. *World J. Gastroenterol.* 22, 9694–9705.

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

Kamarajah, S. K., Burns, W. R., Frankel, T. L., Cho, C. S., and Nathan, H. (2017). Validation of the American Joint Commission on Cancer (AJCC) 8th edition staging system for patients with pancreatic adenocarcinoma: a surveillance, epidemiology and end results (SEER) Analysis. *Ann. Surg. Oncol.* 24, 2023–2030. doi: 10.1245/s10434-017-5810-x

Kamisawa, T., Wood, L. D., Itoi, T., and Takaori, K. (2016). Pancreatic cancer. *Lancet* 388, 73–85. doi: 10.1016/S0140-6736(16)00141-0

Kangari, P., Zarnoosheh Farahany, T., Golchin, A., Ebadollahzadeh, S., Salmaninejad, A., Mahboob, S. A., et al. (2018). Enzymatic antioxidant and lipid peroxidation evaluation in the newly diagnosed Breast Cancer Patients in Iran. *Asian Pac. J. Cancer Prev.* 19, 3511–3515. doi: 10.31557/apjcp.2018.19.12.3511

Klein, J. P., Moeschberger, M. L. (1997). "Refinements of the semiparametric proportional hazards model," in *Survival Analysis*. (New York, NY: Springer). doi: 10.1007/978-1-4757-2728-9_9

Li, W., Gao, L. N., Song, P. P., and You, C. G. (2020). Development and validation of a RNA binding protein-associated prognostic model for lung adenocarcinoma. *Aging* 12, 3558–3573. doi: 10.18632/aging.102828

Li, W., Li, K., Zhao, L., and Zou, H. (2014). Bioinformatics analysis reveals disturbance mechanism of MAPK signaling pathway and cell cycle in Glioblastoma multiforme. *Gene* 547, 346–350. doi: 10.1016/j.gene.2014.06.042

Liu, J., Li, R., Liao, X., Hu, B., and Yu, J. (2019). Comprehensive investigation of the clinical significance and molecular mechanisms of plasmacytoma variant translocation 1 in sarcoma using genome-wide RNA sequencing data. *J. Cancer* 10, 4961–4977. doi: 10.7150/jca.31675

Liu, P., Weng, Y., Sui, Z., Wu, Y., Meng, X., Wu, M., et al. (2016). Quantitative secretomic analysis of pancreatic cancer cells in serum-containing conditioned medium. *Sci. Rep.* 6:37606. doi: 10.1038/srep37606

Lü, J. M., Lin, P. H., Yao, Q., and Chen, C. (2010). Chemical and molecular mechanisms of antioxidants: experimental approaches and model systems. *J. Cell Mol. Med.* 14, 840–860. doi: 10.1111/j.1582-4934.2009.00897.x

Lu, Y., Li, C., Chen, H., and Zhong, W. (2018). Identification of hub genes and analysis of prognostic values in pancreatic ductal adenocarcinoma by integrated bioinformatics methods. *Mol. Biol. Rep.* 45, 1799–1807. doi: 10.1007/s11033-018-4325-2

Martinez-Useros, J., Li, W., Cabeza-Morales, M., and Garcia-Foncillas, J. (2017). Oxidative stress: a new target for pancreatic cancer prognosis and treatment. *J. Clin. Med.* 6:29. doi: 10.3390/jcm6030029

Mir, M. A., Maurer, M. J., Ziesmer, S. C., Slager, S. L., Habermann, T., Macon, W. R., et al. (2015). Elevated serum levels of IL-2R, IL-1RA, and CXCL9 are associated with a poor prognosis in follicular lymphoma. *Blood* 125, 992–998. doi: 10.1182/blood-2014-06-583369

Modica, C., Tortarolo, D., Comoglio, P. M., Basilico, C., and Vigna, E. (2018). MET/HGF co-targeting in pancreatic cancer: a tool to provide insight into the tumor/stroma crosstalk. *Int. J. Mol. Sci.* 19:3920. doi: 10.3390/ijms19123920

Monti, D. A., Mitchell, E., Bazzan, A. J., Littman, S., Zabrecky, G., Yeo, C. J., et al. (2012). Phase I evaluation of intravenous ascorbic acid in combination with gemcitabine and erlotinib in patients with metastatic pancreatic cancer. *PLoS One* 7:e29794. doi: 10.1371/journal.pone.0029794

Nöthlings, U., Wilkens, L. R., Murphy, S. P., Hankin, J. H., Henderson, B. E., and Kolonel, L. N. (2005). Meat and fat intake as risk factors for pancreatic cancer: the multiethnic cohort study. *J. Natl. Cancer Inst.* 97, 1458–1465.

Oates, J. C., and Gilkeson, G. S. (2006). The biology of nitric oxide and other reactive intermediates in systemic lupus erythematosus. *Clin. Immunol.* 121, 243–250. doi: 10.1016/j.clim.2006.06.001

Patacsil, D., Osayi, S., Tran, A. T., Saenz, F., Yimer, L., Shajahan, A. N., et al. (2012). Vitamin E succinate inhibits survivin and induces apoptosis in pancreatic cancer cells. *Genes Nutr.* 7, 83–89. doi: 10.1007/s12263-011-0242-x

Pathan, M., Keerthikumar, S., Chisanga, D., Alessandro, R., Ang, C. S., Askenase, P., et al. (2017). A novel community driven software for functional enrichment analysis of extracellular vesicles data. *J. Extracell Vesicles* 6:1321455. doi: 10.1080/20013078.2017.1321455

Qian, L., Yu, S., Yin, C., Zhu, B., Chen, Z., Meng, Z., et al. (2019). Plasma IFN-γ-inducible chemokines CXCL9 and CXCL10 correlate with survival and chemotherapeutic efficacy in advanced pancreatic ductal adenocarcinoma. *Pancreatology* 19, 340–345. doi: 10.1016/j.pan.2019.01.015

Qiu, T., Wang, H., Wang, Y., Zhang, Y., Hui, Q., and Tao, K. (2015). Identification of genes associated with melanoma metastasis. *Kaohsiung J. Med. Sci.* 31, 553–561. doi: 10.1016/j.kjms.2015.10.002

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74, 2913–2921. doi: 10.1158/0008-5472.CAN-14-0155

Reuter, S., Gupta, S. C., Chaturvedi, M. M., and Aggarwal, B. B. (2010). Oxidative stress, inflammation, and cancer: how are they linked? *Free Radic. Biol. Med.* 49, 1603–1616. doi: 10.1016/j.freeradbiomed.2010.09.006

Singhi, A. D., Koay, E. J., Chari, S. T., and Maitra, A. (2019). Early detection of pancreatic cancer: opportunities and challenges. *Gastroenterology* 156, 2024–2040. doi: 10.1053/j.gastro.2019.01.259

Smith, J., Tho, L. M., Xu, N., and Gillespie, D. A. (2010). The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Adv. Cancer Res.* 108, 73–112. doi: 10.1016/b978-0-12-380888-2.00003-0

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

The GTEx Consortium. (2015). Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017). A subcellular map of the human proteome. *Science* 356:eaal3321. doi: 10.1126/science.aal3321

Vaquero, E. C., Edderkaoui, M., Pandol, S. J., Gukovsky, I., and Gukovskaya, A. S. (2004). Reactive oxygen species produced by NAD(P)H oxidase inhibit apoptosis in pancreatic cancer cells. *J. Biol. Chem.* 279, 34643–34654.

Waldron, R. T., and Rozengurt, E. (2000). Oxidative stress induces protein kinase D activation in intact cells. Involvement of Src and dependence on protein kinase C. *J. Biol. Chem.* 275, 17114–17121.

Wang, J. Y., Liu, G. Z., Wilmott, J. S., La, T., Feng, Y. C., Yari, H., et al. (2017). Skp2-mediated stabilization of MTH1 promotes survival of melanoma cells upon oxidative stress. *Cancer Res.* 77, 6226–6239. doi: 10.1158/0008-5472.Can-17-1965

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wightman, S. C., Uppal, A., Pitroda, S. P., Ganai, S., Burnette, B., Stack, M., et al. (2015). Oncogenic CXCL10 signalling drives metastasis development and poor clinical outcome. *Br. J. Cancer* 113, 327–335. doi: 10.1038/bjc.2015.193

Wu, M., Li, X., Zhang, T., Liu, Z., and Zhao, Y. (2019). Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of Pancreatic Cancer. *Front. Oncol.* 9:996. doi: 10.3389/fonc.2019.00996

Xiao, Y., Zhu, Z., Li, J., Yao, J., Jiang, H., Ran, R., et al. (2020). Expression and prognostic value of long non-coding RNA H19 in glioma via integrated bioinformatics analyses. *Aging* 12, 3407–3430. doi: 10.18632/aging.102819

Yan, X., Wan, H., Hao, X., Lan, T., Li, W., Xu, L., et al. (2019). Importance of gene expression signatures in pancreatic cancer prognosis and the establishment of a prediction model. *Cancer Manag. Res.* 11, 273–283. doi: 10.2147/cmar.S185205

Yu, J. H., and Kim, H. (2014). Oxidative stress and cytokines in the pathogenesis of pancreatic cancer. *J. Cancer Prevent.* 19, 97–102. doi: 10.15430/JCP.2014.19.2.97

Zhang, M., Wang, X., Chen, X., Guo, F., and Hong, J. (2020a). Prognostic value of a stemness index-associated signature in primary lower-grade glioma. *Front. Genet.* 11:441. doi: 10.3389/fgene.2020.00441

Zhang, M., Wang, X., Chen, X., Zhang, Q., and Hong, J. (2020b). Novel immune-related gene signature for risk stratification and prognosis of survival in lower-grade glioma. *Front. Genet.* 11:363. doi: 10.3389/fgene.2020.00363

Zhou, F., Pan, Y., Wei, Y., Zhang, R., Bai, G., Shen, Q., et al. (2017). Jab1/Csn5-thioredoxin signaling in relapsed acute monocytic leukemia under oxidative stress. *Clin. Cancer Res.* 23, 4450–4461. doi: 10.1158/1078-0432.Ccr-16-2426

Zhou, F., Shen, Q., and Claret, F. X. (2013). Novel roles of reactive oxygen species in the pathogenesis of acute myeloid leukemia. *J. Leukoc Biol.* 94, 423–429. doi: 10.1189/jlb.0113006

**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: 13 January 2021
doi: 10.3389/fgene.2020.604655

# A Six-lncRNA Signature for Immunophenotype Prediction of Glioblastoma Multiforme

*Ming Gao[1,2,3], Xinzhuang Wang[1,2,3], Dayong Han[1,2,3], Enzhou Lu[1,2,3], Jian Zhang[4], Cheng Zhang[5], Ligang Wang[1,2,3], Quan Yang[1,2,3], Qiuyi Jiang[1,2,3], Jianing Wu[1,2,3], Xin Chen[1,2,3]* and Shiguang Zhao[1,2,3]**

[1] Department of Neurosurgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Key Colleges and Universities Laboratory of Neurosurgery in Heilongjiang Province, Harbin, China, [3] Institute of Neuroscience, Sino-Russian Medical Research Center, Harbin Medical University, Harbin, China, [4] Department of General Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [5] North Broward Preparatory School, Coconut Creek, FL, United States

OPEN ACCESS

**Edited by:**
Suman Ghosal,
National Institutes of Health,
United States

**Reviewed by:**
Provas Das,
Baylor University, United States
Jayshree Advani,
National Institutes of Health,
United States

***Correspondence:**
Xin Chen
chenxin_tracy@yeah.net
Shiguang Zhao
guangsz@hotmail.com

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

**Received:** 10 September 2020
**Accepted:** 26 October 2020
**Published:** 13 January 2021

**Citation:**
Gao M, Wang X, Han D, Lu E, Zhang J, Zhang C, Wang L, Yang Q, Jiang Q, Wu J, Chen X and Zhao S (2021) A Six-lncRNA Signature for Immunophenotype Prediction of Glioblastoma Multiforme. Front. Genet. 11:604655. doi: 10.3389/fgene.2020.604655

Glioblastoma multiforme (GBM) is the most aggressive primary tumor of the central nervous system. As biomedicine advances, the researcher has found the development of GBM is closely related to immunity. In this study, we evaluated the GBM tumor immunoreactivity and defined the Immune-High (IH) and Immune-Low (IL) immunophenotypes using transcriptome data from 144 tumors profiled by The Cancer Genome Atlas (TCGA) project based on the single-sample gene set enrichment analysis (ssGSEA) of five immune expression signatures (IFN-γ response, macrophages, lymphocyte infiltration, TGF-β response, and wound healing). Next, we identified six immunophenotype-related long non-coding RNA biomarkers (im-lncRNAs, USP30-AS1, HCP5, PSMB8-AS1, AL133264.2, LINC01684, and LINC01506) by employing a machine learning computational framework combining minimum redundancy maximum relevance algorithm (mRMR) and random forest model. Moreover, the expression level of identified im-lncRNAs was converted into an im-lncScore using the normalized principal component analysis. The im-lncScore showed a promising performance for distinguishing the GBM immunophenotypes with an area under the curve (AUC) of 0.928. Furthermore, the im-lncRNAs were also closely associated with the levels of tumor immune cell infiltration in GBM. In summary, the im-lncRNA signature had important clinical implications for tumor immunophenotyping and guiding immunotherapy in glioblastoma patients in future.

**Keywords: long non-coding RNA, biomarker, immunophenotype, machine learning, glioblastoma multiforme**

## INTRODUCTION

Glioblastoma multiforme (GBM) is the most aggressive type of primary brain tumor in adults, with a median survival of 14.6 months (Kaffes et al., 2019). The emergence of tumor immunotherapy has revolutionized GBM treatment and its success is highly dependent on the development and activation of immune cells in the host microenvironment (Pardoll, 2012; Daud et al., 2016). In the GBM microenvironment, the non-neoplastic cells are mainly from the innate immune system,

which can interact with neoplastic tumor cells and play an important role in tumor growth and progression (Engler et al., 2012; Feng et al., 2015; Hu et al., 2015). Therefore, evaluation of GBM tumor immunoreactivity is critical in determining personalized treatment.

Long non-coding RNAs (lncRNAs) are defined as non-coding RNAs of more than 200 nt in length (Zhai et al., 2018). The discovery of lncRNAs has revealed a new dimension to the pathological processes of many diseases, including the occurrence and development of cancer (Martens-Uzunova et al., 2014; Zhou et al., 2014). Moreover, recent studies showed that lncRNAs play an important role in tumor immune escape (Pei et al., 2018; Wang et al., 2019; Jin et al., 2020). For example, UCA1 is able to promote proliferation, migration, immune escape, and inhibit apoptosis in gastric cancer (Wang et al., 2019); SNHG1 can regulate the differentiation of Treg cells and affect the immune escape of breast cancer (Pei et al., 2018). Besides, immune-associated lncRNAs can also serve as improving prognosis and immunotherapy response biomarkers (Zhou et al., 2018; Sun et al., 2020). Therefore, identification of lncRNA biomarkers for tumor immunoreactivity may provide new insights into the treatment of GBM patients.

In this study, we systemically characterized the GBM tumor immune microenvironment in the TCGA GBM cohort. Moreover, we defined the GBM Immune-High (IH) and Immune-Low (IL) subtype based on five immune expression signatures including macrophages, lymphocyte infiltration, TGF-β response, IFN-γ response, wound healing. Furthermore, we identified six immunophenotype-related lncRNA signatures (im-lncRNAs, including USP30-AS1, HCP5, PSMB8-AS1, AL133264.2, LINC01684, and LINC01506) using the minimum redundancy maximum relevance (mRMR) feature selection method and the random forest model. The im-lncRNAs showed good performance in distinguishing tumor immunophenotypes and were closely associated with the levels of tumor immune cell infiltration. These results suggested the im-lncRNAs had the promising potential for clinical diagnosis of GBM immunophenotypes.

# MATERIALS AND METHODS

## Data Acquisition and Pre-processing

All glioblastoma multiforme tissue samples were obtained from the surgical resection tissue of GBM patients ($n = 10$), non-tumor brain tissue was used as the negative control group ($n = 5$). The tissue samples were stored in liquid nitrogen. All patients have signed informed consent, and the study was supervised and approved by the Ethics Committee of The First Affiliated Hospital of Harbin Medical University.

The Cancer Genome Atlas (TCGA) level 3 gene/lncRNA expression data, and clinical data of GBM ($n = 149$, 144 cancer samples, 5 normal samples) were obtained from the Genomic Data Commons (GDC, available at https://www.cancer.gov/tcga). Two independent datasets GSE79671 (Urup et al., 2017) and GSE121810 (Cloughesy et al., 2019) were used for the validation of im-lncRNAs. For the gene/lncRNA expression data, we removed the genes/lncRNAs that were not expressed over 70% of the samples. The remaining 18,094 genes and 18,567 lncRNAs were used for subsequent analysis.

## Total RNA Extraction and Quantitative Real-Time PCR

According to the manufacturer's instructions, total RNA was extracted from the GBM tissues and non-tumor brain tissues using TRIzol Reagent (Invitrogen, Carlsbad, CA, United States). The concentration of the total RNA was detected by spectrophotometer (Thermo Scientific$^{TM}$ NanoDrop 2000c). Total RNA (1000 ng) was reverse transcribed into cDNA using qPCR RT Kit (TOYOBO, Japan). The relative level of lncRNAs to the housekeeping gene GAPDH was determined by qRT-PCR using FastStart Universal SYBR Green Master (ROX) (Roche, Germany). All primers used in this study is showed in **Supplementary Table 1**. Analysis between the two groups was performed by an unpaired $t$-test, $P < 0.05$ was considered statistically significant.

## Identification of Tumor Immune Subtypes of GBM

Based on five immune expression signatures reorganized by Vesteinn et al. (Lek et al., 2016) including IFN-γ response (Wolf et al., 2014), macrophages/monocytes (Beck et al., 2009), overall lymphocyte infiltration (dominated by T and B cells) (Calabro et al., 2009), TGF-β response (Teschendorff et al., 2010), wound healing (Chang et al., 2004), we evaluated the enrichment scores (ESs) of GBM samples using the single-sample gene set enrichment analysis (ssGSEA) (Barbie et al., 2009). The ssGSEA was based on the R package "GSVA." Furthermore, we used the ESs of immune expression signatures to perform a consensus clustering on 149 cancer samples using the R package "ConsensusClusterPlus" (Monti et al., 2003).

## Evaluation of Tumor Purity, Tumor-Infiltrating Immune Cells, and Cytolytic Activity

The tumor purity of corresponding TCGA samples was evaluated using the ESTIMATE score calculated by the R package "ESTIMATE" (Yoshihara et al., 2013). The higher ESTIMATE score, the lower tumor purity. The tumor immune cell infiltration levels were estimated based on the gene expression profile by Tumor Immune Estimation Resource (TIMER) (Li et al., 2017). Here, six tumor-infiltrating immune cells (B cells, CD4 T cells, CD8 T cells, macrophages, neutrophils, and myeloid dendritic cells) were considered. Cytolytic activity (CYT) was calculated as the geometric mean of the *GZMA* gene and *PRF1* gene (as expressed in FPKM) (Rooney et al., 2015).

## Differential Expression Analysis of lncRNAs

We first calculated the $\log_2$(fold change) ($\log_2$(FC)) of each lncRNA between the IL and normal samples, and between the IH and normal samples, respectively. Then we scaled the expression

level ($log_2$FPKM) of each lncRNA and into a Z-score. Next, we compared lncRNA expression differences between the IL and normal samples, and between the IH and normal samples, using the Student's *t*-test, respectively. The *P-values* were corrected using the Benjamini-Hochberg adjustment. The lncRNAs with $FDR < 0.01$ and $|\ log_2FC| > 2$ were considered as the differentially expressed lncRNAs.

## Identification of im-lncRNAs

We first divided the GBM cancer samples into three parts (two "training" sets and one "test" set) to apply three-fold cross-validation. Next, we screened the lncRNA features with minimal redundancy using the mRMR feature selection method in the training set (Hanchuan et al., 2005). Further, we trained a random forest model based on the top 5% mRMR lncRNA features. The performance of the random forest model was assessed through prediction making in the test set and the computation of the balanced error rate (BER). For a more robust estimation of the BER, three-fold cross-validation was applied 1,000 times and for each run, randomized data were used as the negative control. The signature size, for which no more improvement of the BER was observed (6 features signature size), was selected as the final size. This process generated 3 × 1000 output signatures. The distance (D) between these signatures was defined as (Jeschke et al., 2017):

$$D = 1 - \frac{\sum_{i=1}^{6} cor\,(F1_i, F2_i)}{6} \quad (1)$$

where cor represents the Spearman's correlation coefficient (Rho); $F1_i$ to the $i^{th}$ feature from signature 1 and $F2_i$ to the $i^{th}$ feature from signature 2 after sorting the features to maximize the sum of the Rho based on the changes in the Gini index. For each signature, the sum of its pairwise distance to all other output signatures was computed, and the signature with the smallest sum was assumed to be the most representative and chosen as the final lncRNA signature (im-lncRNA).

## Construction of im-lncScore

To conveniently evaluate the GBM tumor immunophenotypes, we constructed the im-lncScore. Firstly, we applied principal component analysis (PCA) on the Z-scores of im-lncRNAs. Then the first component was used as the final im-lncScore for the cancer samples.

## Analysis of Association Between im-lncRNAs and Tumor Immune Cell Infiltration

Firstly, we calculated the median infiltration levels of each immune cell; if the sample infiltration level was higher than the median level, the sample was defined as a high immune infiltration sample; if the sample infiltration level was lower than the median level, the sample was defined as a low immune infiltration sample. Then, the univariate logistic regression was performed to assess the association between each im-lncRNA and the infiltration levels of each immune cell. The OR, 95% confidence level of the OR, and *P-values* were calculated for each immune cell. The logistic regression was based on the R package "epiDisplay."

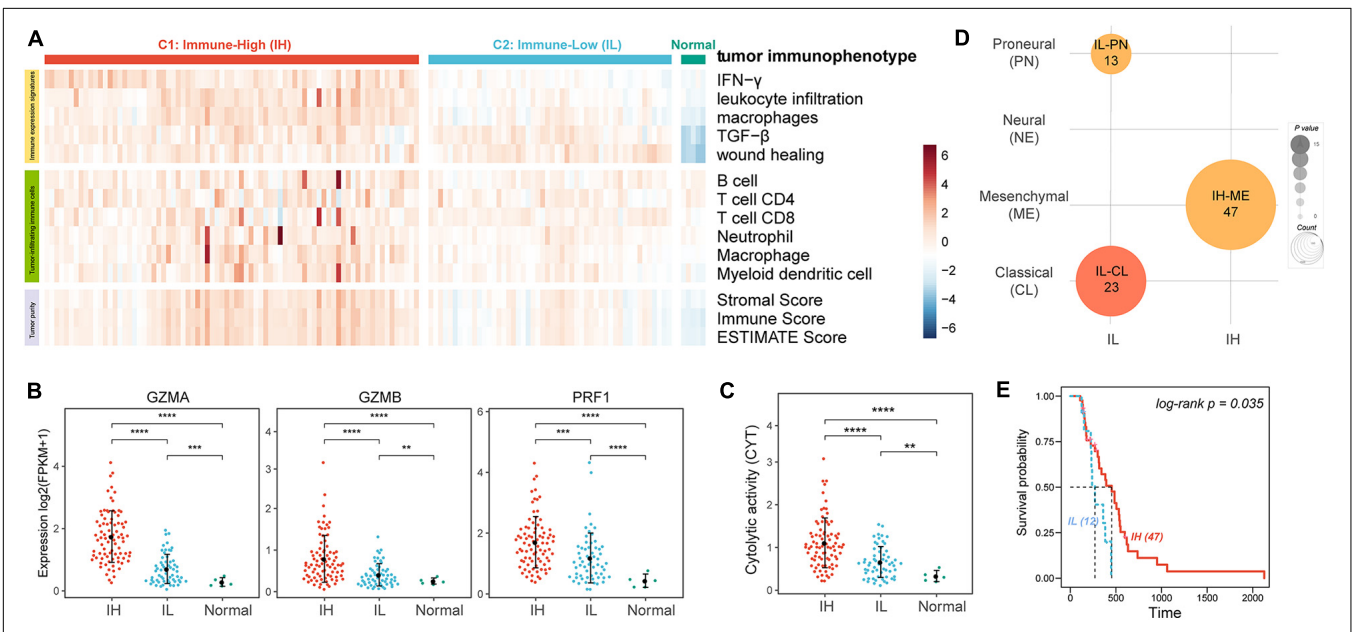



**FIGURE 1 |** The tumor immune microenvironment of GBM. **(A)** Heatmaps showing the ssGSEA of immune expression signatures (first five lines), the levels of tumor-infiltrating immune cells (sixth to twelfth lines), and tumor purity (last three lines) in GBM patients. **(B)** The expression levels of effectors of immunity log2(FPKM+1). **(C)** The CYT of GBM patients. **(D)** Enrichment analysis of GBM subtypes and GBM immunophenotypes. **(E)** The KM survival curve showing the impact of immune subtypes on the survival of ME GBM patients. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

## Identification of Co-expressed Genes With im-lncRNAs

Based on the expression profiles of im-lncRNAs and genes, we calculated the Spearman's correlation coefficient (Rho) between im-lncRNAs and genes. The raw *P-values* ($P_r$) were adjusted by multiple hypotheses using a permutation method. For each gene, the expression value (FPKM) was held consistent, and 1,000 random im-lncRNAs were used to perform the same Spearman's correlation test, generating a set of 1,000 permutation *P-values* ($P_p$). Finally, an empirical *P-value* ($P_e$) was corrected using the following formula (which introduces a pseudo-count of 1). The gene with the Rho>0.6 and $P_e$ <0.01 were treated as the co-expressed genes of im-lncRNAs.

$$P_e = \frac{num\left(P_p \leq P_r\right) + 1}{1001} \quad (2)$$

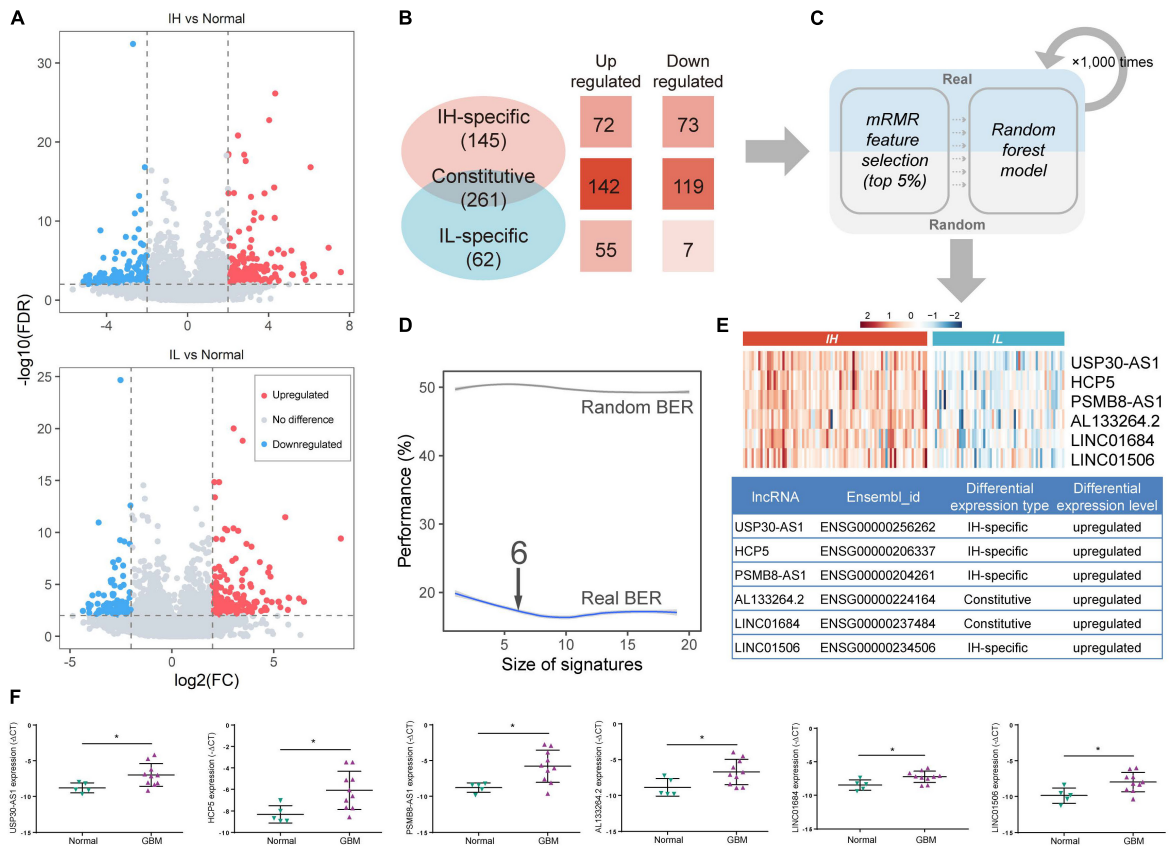
## Functional Enrichment Analysis

To annotated the biological functions of im-lncRNAs, we performed functional enrichment analysis on the co-expressed genes of im-lncRNAs using Metascape (Zhou et al., 2019). For each co-expressed gene list, pathway and process enrichment analysis have been carried out with the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, and Hallmark Gene Sets.

# RESULTS

## Characterizing the Immune Microenvironment of GBM

We analyzed 149 GBM RNA-seq expression profiles from TCGA. To evaluate the tumor immune activity, we used a previously described technique employing ssGSEA (Barbie et al., 2009) based on the five immune expression signatures reorganized by Vesteinn et al. (Lek et al., 2016) including IFN-γ response (Wolf et al., 2014), lymphocyte infiltration (Calabro et al., 2009), macrophages/monocytes (Beck et al., 2009), TGF-β response (Teschendorff et al., 2010), wound healing (Chang et al., 2004). The result showed that there were higher ESs of all immune expression signatures in cancer than in normal samples (**Figure 1A**, IFN-γ *P = 1.01e-03*, leukocyte infiltration *P = 1.83e-04*, macrophages *P = 5.80e-08*, TGF-β *P = 1.55e-05*, and wound healing *P = 3.33e-08*). Moreover, based on the ESs of immune




| lncRNA | Ensembl_id | Differential expression type | Differential expression level |
|---|---|---|---|
| USP30-AS1 | ENSG00000256262 | IH-specific | upregulated |
| HCP5 | ENSG00000206337 | IH-specific | upregulated |
| PSMB8-AS1 | ENSG00000204261 | IH-specific | upregulated |
| AL133264.2 | ENSG00000224164 | Constitutive | upregulated |
| LINC01684 | ENSG00000237484 | Constitutive | upregulated |
| LINC01506 | ENSG00000234506 | IH-specific | upregulated |

**FIGURE 2 |** Identification of im-lncRNAs in GBM. **(A)** Volcano plots showing the differentially expressed lncRNAs between IH/IL and normal samples, respectively. The *y*-axis shows the -$\log_{10}$(FDR). The *x*-axis shows the $\log_2$(FC) **(B)** The classification of differentially expressed lncRNAs. **(C)** The pipeline of identifying im-lncRNAs. **(D)** The evaluation of model BER performance. **(E)** Heatmap showing the im-lncRNAs expression in IH and IL samples, respectively. **(F)** The up-regulated six lncRNAs expression in glioblastoma multiforme tissues was confirmed by qRT-PCR (*$P < 0.05$).

expression signatures, we subclassified the cancer samples using the consensus clustering method. The analysis resulted in 2 robust clusters: C1 and C2. Notably, the ESs of IFN-γ, leukocyte infiltration, macrophages in C1 were significantly higher than in C2 (**Figure 1A**, IFN-γ *P = 1.46e-29*, leukocyte infiltration *P = 3.43e-18*, and macrophages *P = 2.60e-16*). And, there was no significant difference in the ESs of TGF-β (*P = 3.46e-01*) and wound healing (*P = 1.09e-01*) between the two clusters. Furthermore, we evaluated the levels of tumor purity and tumor-infiltrating immune cells between the two clusters. There were lower tumor purity (*P = 4.75e-16*) and higher percent of tumor-infiltrating immune cells (B cell *P = 1.48e-05*, T cell CD4 *P = 8.26e-4*, Neutrophil *P = 3.58e-06*, and macrophage *P = 7.14e-10*) in C1 than C2 (**Figure 1A**). Therefore, we annotated the C1 sample was the Immune-High (IH) subtype, and the C2 sample was the Immune-Low (IL) subtype.

To further verify the levels of immune activation in different immune subtypes, we examined the expression levels of common effectors of immunity, such as granzyme A (GZMA), granzyme B (GZMB), and perforin (PRF1) (**Figure 1B**; Mandal et al., 2016) and the immune cytolytic activity (CYT, an indicator of tumor local immunity, **Figure 1C**; Rooney et al., 2015). Remarkably, these effectors of immunity and CYT were much higher in the IH subtype compared with the IL subtype.

Glioblastoma multiforme can be subclassified into distinct molecular subtypes based on their expression profiles: classical (CL), mesenchymal (ME), neural (NE), and proneural (PN) (Verhaak et al., 2010; Ceccarelli et al., 2016). Here, we also enriched the tumor immune subtypes into the GBM molecular subtypes using Fisher's exact test. The previous study indicated ME GBM was the most immunogenic among the four subclasses while the PN subtype was the least immunogenic (Doucette et al., 2013). Our result also showed that ME GBM was significantly enriched in the IH subtype, while CL and PN GBM tumors were significantly enriched in the IL subtype (**Figure 1D**). Besides, by analyzing the survival of ME GBM patients between IL and

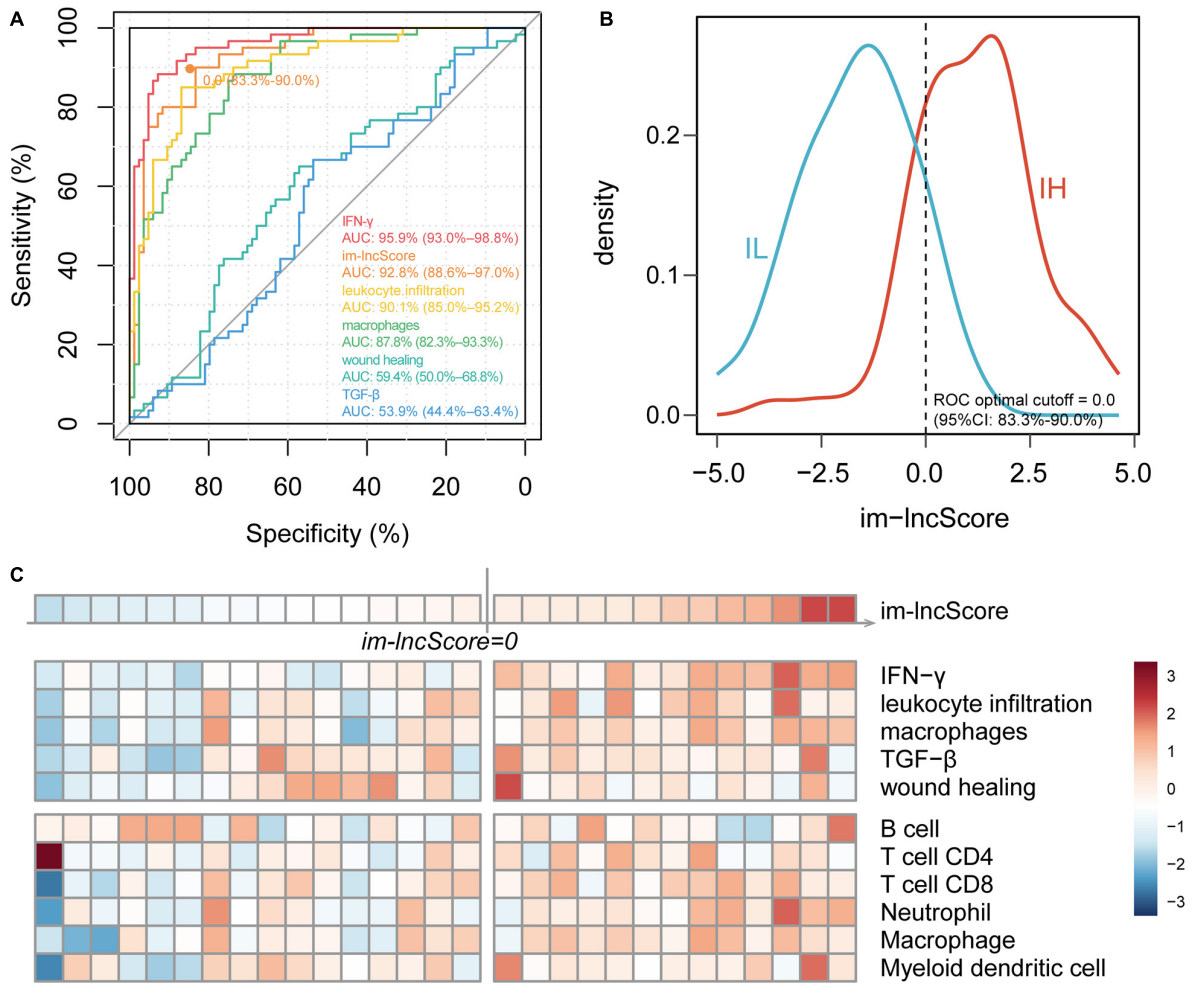



**FIGURE 3 |** im-lncRNAs enable evaluate GBM immunophenotypes. **(A)** ROC curves for prediction of GBM tumor immunophenotypes based on the im-lncScore (orange) or immune expression signatures (other colors) in the TCGA GBM cohort. **(B)** Density plot showing the distribution of im-lncScore in IH and IL samples. **(C)** Heatmap showing the im-lncScore (the first line), ESs of immune expression signatures (second to sixth lines), and infiltration levels of tumor immune cells (last six lines) in GSE121810 cohort.

IH subtypes, we found the survival of ME with IH patients was significantly better than ME with IL patients (**Figure 1E**).

## Identification of Immunophenotype-Related lncRNA Biomarkers in GBM

lncRNA, an emerging biomarker, plays an important role in tumor immune regulation (Li et al., 2020). However, few studies focus on the ability of lncRNA in tumor immunophenotyping. To identify the immunophenotype-related lncRNA biomarkers (im-lncRNAs), we first characterized the differentially expressed lncRNAs between the IH/IL and normal samples, respectively ($FDR < 0.01$ and $|log_2FC| > 2$, see section "Materials and Methods," **Figure 2A**). We identified 261 "Constitutive" lncRNAs differentially expressed in both immune subtypes (142 upregulated and 119 downregulated), 145 "IH-specific" lncRNAs only differentially expressed in IH subtype (72 upregulated and 73 downregulated), and 70 "IL-specific" lncRNAs only differentially expressed in IL subtype (55 upregulated and 7 downregulated, **Figure 2B**).

Next, we applied a machine learning method in differentially expressed lncRNAs to identify the im-lncRNAs (**Figure 2C**). Firstly, under three-fold cross-validation (dividing 144 cancer samples into three parts, two "training" sets [96 samples], and one "test" set [48 samples]), the mRMR feature selection method was used to establish a small signature with minimal redundancy and selected the top 5% lncRNA features to train the random forest models. Next, in the test set, the balanced-error rate (BER) was calculated to evaluate the model performance. For a more robust estimation of the BER, three-fold cross-validation was applied 1,000 times. In each run, randomized data were used as the negative control. The signature size, for which no more improvement of the BER was observed (6 features signature size), was selected as the final size (**Figure 2D**). This pipeline generated 3 × 1000 output signatures and the signature with the minimum distance summed was assumed to be the most representative (see

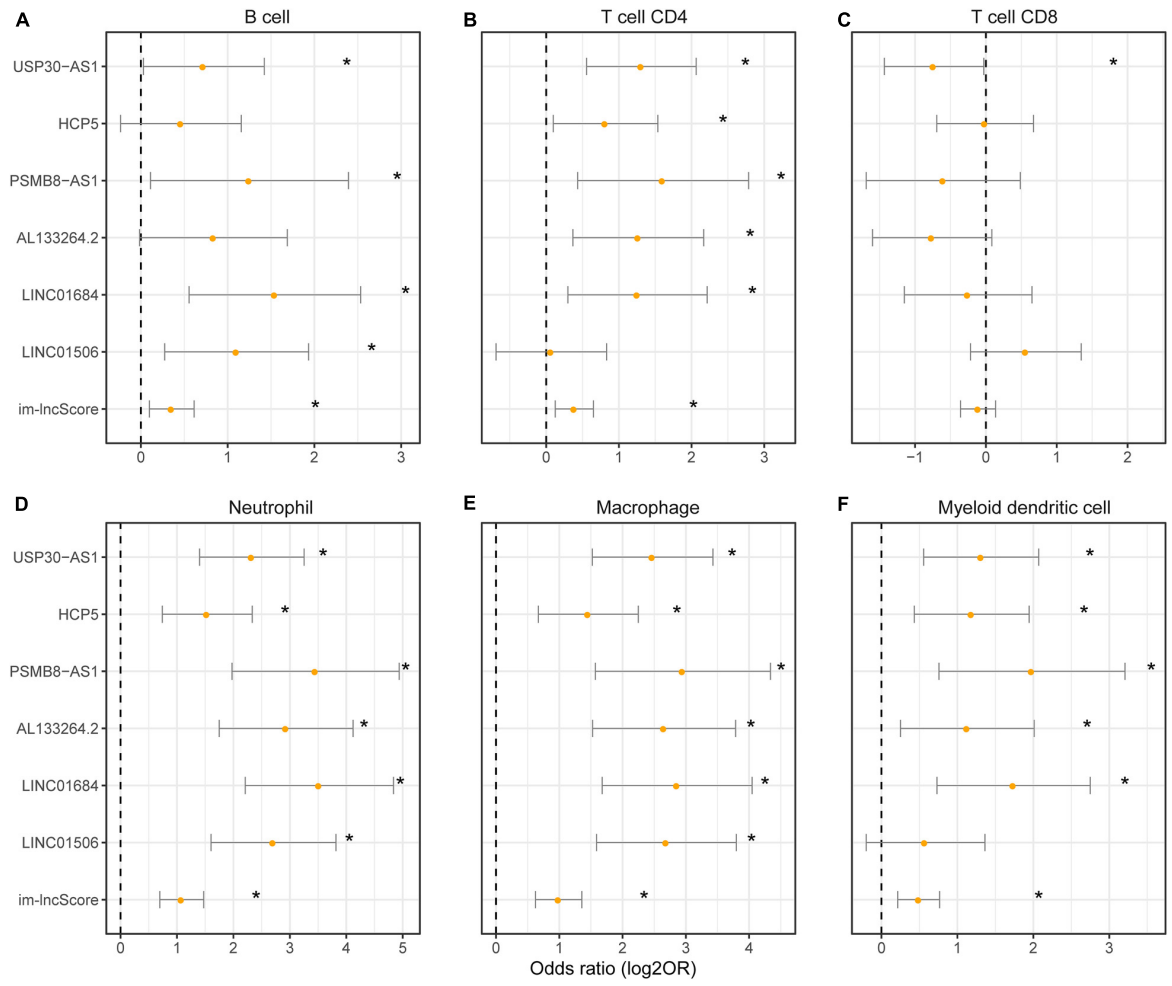



**FIGURE 4 |** The association between im-lncRNAs and tumor immune cell infiltration. **(A–F)** The im-lncRNAs were correlated with immune cell infiltration. The dots represent the odds ratio (OR) of the Wald test and the error bars show the 95% confidence intervals of the OR. **(A,B)** cells; **(B)** CD4 T cells; **(C)** CD8 T cells; **(D)** neutrophils; **(E)** macrophages; and **(F)** myeloid dendritic cells.

section "Materials and Methods"). Based on the approach, we identified 6 im-lncRNAs (USP30-AS1, HCP5, PSMB8-AS1, AL133264.2, LINC01684, LINC01506). Notably, USP30-AS1, HCP5, PSMB8-AS1, and LINC01506 were "IH-specific" lncRNAs, and AL133264.2, LINC01684 were "Constitutive" lncRNAs. The expression levels of all im-lncRNAs were significantly higher in IH than IL samples (USP30-AS1 *P = 1.12e-18*, HCP5 *P = 8.07e-15*, PSMB8-AS1 *P = 1.15e-15*, AL133264.2 *P = 4.07e-10*, LINC01684 *P = 1.50e-11*, and LINC01506 *P = 3.00e-10*, **Figure 2E**). Besides, all of 6 im-lncRNAs were also upregulated in GBM cancer than normal samples, which had been validated by RT-qPCR in five non-tumor brain tissues and ten GBM tissues (**Figure 2F**). To ensure that the recognized im-lncRNAs were robust, we also employed the same way on an independent dataset [GSE79671 (Urup et al., 2017)]. The results showed that six im-lncRNAs closely associated with the GBM immunophenotypes and four of the six (USP30-AS1, HCP5, AL133264.2, and LINC01506) were consistent with the im-lncRNAs identified in the TCGA GBM cohort (**Supplementary Figures 1A–C**).

## Evaluation of GBM Tumor Immunophenotyping Efficacy of im-lncRNAs

To further evaluate the relationship between im-lncRNAs and GBM immunophenotyping, we transformed the individual expression values of the im-lncRNAs into a score (im-lncScore) by applying a principal component analysis (PCA). We assessed the potential of the im-lncScore to predict GBM immunophenotypes in 144 cancer samples. Compared with the immune expression signatures, the im-lncScore also showed a promising performance. An AUC of 0.928 (95% CI, 0.87–0.97) suggested a predictive value for the im-lncScore (**Figure 3A**). Moreover, the optimal cutoff point determined by the ROC curve analysis was 0.0 (95% CI, 0.83–0.90). We also found that the im-lncScores of IH samples were usually greater than the optimal cutoff, while the opposite was observed for the IL samples (**Figure 3B**).

Besides, we also validated the immunophenotyping ability of im-lncScore in an independent dataset [GSE121810 (Cloughesy et al., 2019)]. The dataset included 29 GBM samples. We first calculated the im-lncScore to subclassify the GBM samples into IH/IL subtype. 13 IH and 16 IL samples were identified in the dataset (**Figure 3C**). Next, we also evaluated the ESs of five immune expression signatures and infiltration levels of tumor immune cells. As described above, there were higher ESs of IFN-γ, leukocyte infiltration and macrophages signatures, and higher levels of tumor-infiltrating immune cells in IH than IL samples (**Figure 2C**). These results suggested that the im-lncScore does not require a complex algorithm to effectively subclassify the GBM tumor immunophenotypes, which also further indicated the important role of im-lncRNAs in GBM tumor immunity.

## Im-lncRNAs Are Associated With the Tumor Immune Cell Infiltration

To evaluate whether the im-lncRNAs associated with the levels of tumor immune cell infiltration, we first subclassified the cancer samples into high and low immune infiltration groups by comparing the sample immune infiltration levels to the median immune infiltration level of each immune cell. And then, the univariate logistic regression was performed based on the six im-lncRNAs expression value and im-lncScore. We found that the im-lncRNAs significantly correlated with the infiltration level of multiple immune cells (**Figures 4A–F**). Notably, the im-lncScore

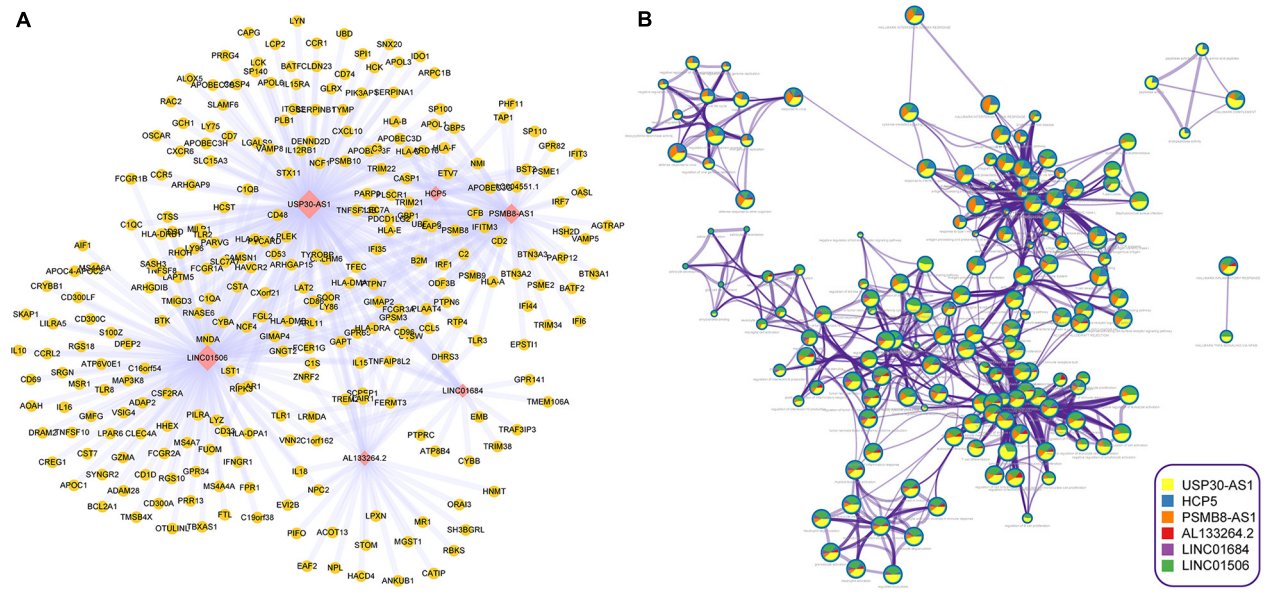



**FIGURE 5 |** Inferring the biological functions of im-lncRNAs. **(A)** The co-expressed im-lncRNAs-genes network. **(B)** Network of enriched terms represented as pie charts, where pies are color-coded based on the im-lncRNAs.

also showed the significantly correlation with multiple immune cell infiltration levels (except for T cell CD8). Besides, HCP5 and PSMB8-AS1 have been demonstrated could be the tumor-infiltrating immune-related lncRNA signature of non-small cell lung cancer and closely associated with outcome and immune cell infiltrates (Sun et al., 2020). These results suggested that the im-lncRNAs played crucial roles in the tumor immune infiltration.

### The Functional Enrichment Analysis of im-lncRNAs

To further explore the biological functions of im-lncRNAs, we identified the co-expressed genes with the im-lncRNAs using the Spearman's correlation test. The *P-values* were adjusted by multiple hypotheses. A total of 459 co-expressed im-lncRNA-gene pairs were identified (**Figure 5A**). Furthermore, we performed functional enrichment analysis on the co-expressed genes using the Metascape (Zhou et al., 2019). The result showed that the functions of co-expressed genes of each im-lncRNAs were all significantly enriched in the immune-related terms, such as hallmark interferon-gamma response (M5913), myeloid leukocyte activation (GO:0002274), tumor necrosis factor superfamily cytokine production (GO:0071706), ER-Phagosome pathway (R-HSA-1236974), etc. (**Figure 5B**). Moreover, we also found the im-lncRNAs were closely correlated with the GBM-related immune pathways (Li et al., 2020). For instance, the HCP5 and PSMB8-AS1 were related to the Antigen Processing and Presentation, Antimicrobials, and Natural Killer Cell Cytotoxicity; AL133264.2 was related to Interleukins; the LINC01684 and USP30-AS1 were related to Antimicrobials. These results further validated the important role of im-lncRNAs in the GBM immune regulation.

## DISCUSSION

Accumulating evidence suggests that lncRNA serves as a specific molecular marker for tumor immunoreactivity (Wang et al., 2019; Sun et al., 2020). In this study, we analyze the role of lncRNAs in IH and IL tumor immunophenotypes. Moreover, we identify im-lncRNAs based on the machine learning method. Furthermore, we construct an im-lncScore using the expression value of im-lncRNAs. The im-lncScore shows a good performance for distinguishing the GBM tumor immunophenotypes (AUC = 0.928, 95%CI: 0.885–0.970). The im-lncScore does not need a complex algorithm to effectively reflect the patient tumor immunoreactivity. Furthermore, these im-lncRNAs are also closely associated with the levels of tumor immune cell infiltration. This evidence indicates that the im-lncRNAs have the potential to be an important indicator for future clinical diagnosis of GBM immunophenotypes. However, these results are still at the level of the initial calculation, so to ensure accuracy the biology experiments are required to further validate the role of im-lncRNAs. Besides, due to the limited scale, we only use the TCGA data to train our models. Therefore, as the scale of data increases, we will continue to validate the efficiency of im-lncRNAs in GBM.

In summary, im-lncRNAs play an important role in tumor immunophenotyping. Identification of GBM immunophenotypes will provide us a novel insight to improve the therapeutic strategy of GBM. Therefore, the im-lncRNAs has the promising potential for clinical diagnosis of GBM immunophenotypes in the future.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The Cancer Genome Atlas (TCGA) project (https://www.cancer.gov/tcga), GEO (GSE121810, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121810).

## AUTHOR CONTRIBUTIONS

MG, XC, and SZ conceived and designed the experiments. MG, XW, DH, and EL analyzed the data. JZ, LW, and QY collected the data. QJ and JW validated the method and data. CZ checked the writing standard of the manuscript. SZ, XC, and MG wrote this manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China (81972363) and Research Fund for the Postdoctoral Science Foundation of China (2017M620119).

## ACKNOWLEDGMENTS

The authors gratefully thank the TCGA Research Network for providing data for this work.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.604655/full#supplementary-material

## REFERENCES

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108–112. doi: 10.1038/nature08460

Beck, A. H., Espinosa, I., Edris, B., Li, R., Montgomery, K., Zhu, S., et al. (2009). The macrophage colony-stimulating factor 1 response signature in breast carcinoma. *Clin. Cancer Res.* 15, 778–787. doi: 10.1158/1078-0432.CCR-08-1283

Calabro, A., Beissbarth, T., Kuner, R., Stojanov, M., Benner, A., Asslaber, M., et al. (2009). Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Res. Treat.* 116, 69–77. doi: 10.1007/s10549-008-0105-3

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

Chang, H. Y., Sneddon, J. B., Alizadeh, A. A., Sood, R., West, R. B., Montgomery, K., et al. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* 2:E7. doi: 10.1371/journal.pbio.0020007

Cloughesy, T. F., Mochizuki, A. Y., Orpilla, J. R., Hugo, W., Lee, A. H., Davidson, T. B., et al. (2019). Neoadjuvant anti-PD-1 immunotherapy promotes a survival benefit with intratumoral and systemic immune responses in recurrent glioblastoma. *Nat. Med.* 25, 477–486. doi: 10.1038/s41591-018-0337-7

Daud, A. I., Wolchok, J. D., Robert, C., Hwu, W. J., Weber, J. S., Ribas, A., et al. (2016). Programmed death-ligand 1 expression and response to the anti-programmed death 1 antibody pembrolizumab in melanoma. *J. Clin. Oncol.* 34, 4102–4109. doi: 10.1200/JCO.2016.67.2477

Doucette, T., Rao, G., Rao, A., Shen, L., Aldape, K., Wei, J., et al. (2013). Immune heterogeneity of glioblastoma subtypes: extrapolation from the cancer genome atlas. *Cancer Immunol. Res.* 1, 112–122. doi: 10.1158/2326-6066.CIR-13-0028

Engler, J. R., Robinson, A. E., Smirnov, I., Hodgson, J. G., Berger, M. S., Gupta, N., et al. (2012). Increased microglia/macrophage gene expression in a subset of adult and pediatric astrocytomas. *PLoS One* 7:e43339. doi: 10.1371/journal.pone.0043339

Feng, X., Szulzewsky, F., Yerevanian, A., Chen, Z., Heinzmann, D., Rasmussen, R. D., et al. (2015). Loss of CX3CR1 increases accumulation of inflammatory monocytes and promotes gliomagenesis. *Oncotarget* 6, 15077–15094. doi: 10.18632/oncotarget.3730

Hanchuan, P., Fuhui, L., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intellig.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Hu, F., Dzaye, O., Hahn, A., Yu, Y., Scavetta, R. J., Dittmar, G., et al. (2015). Glioma-derived versican promotes tumor expansion via glioma-associated microglial/macrophages Toll-like receptor 2 signaling. *Neuro Oncol.* 17, 200–210. doi: 10.1093/neuonc/nou324

Jeschke, J., Bizet, M., Desmedt, C., Calonne, E., Dedeurwaerder, S., Garaud, S., et al. (2017). DNA methylation-based immune response signature improves patient diagnosis in multiple cancers. *J. Clin. Invest.* 127, 3090–3102. doi: 10.1172/JCI91095

Jin, K. T., Yao, J. Y., Fang, X. L., Di, H., and Ma, Y. Y. (2020). Roles of lncRNAs in cancer: focusing on angiogenesis. *Life Sci.* 252:117647. doi: 10.1016/j.lfs.2020.117647

Kaffes, I., Szulzewsky, F., Chen, Z., Herting, C. J., Gabanic, B., Velazquez Vega, J. E., et al. (2019). Human Mesenchymal glioblastomas are characterized by an increased immune cell presence compared to proneural and classical tumors. *Oncoimmunology* 8:e1655360. doi: 10.1080/2162402X.2019.1655360

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

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

Li, Y., Jiang, T., Zhou, W., Li, J., Li, X., Wang, Q., et al. (2020). Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat. Commun.* 11:1000. doi: 10.1038/s41467-020-14802-2

Mandal, R., Senbabaoglu, Y., Desrichard, A., Havel, J. J., Dalin, M. G., Riaz, N., et al. (2016). The head and neck cancer immune landscape and its immunotherapeutic implications. *JCI Insight* 1:e89829. doi: 10.1172/jci.insight.89829

Martens-Uzunova, E. S., Bottcher, R., Croce, C. M., Jenster, G., Visakorpi, T., and Calin, G. A. (2014). Long noncoding RNA in prostate, bladder, and kidney cancer. *Eur. Urol.* 65, 1140–1151. doi: 10.1016/j.eururo.2013.12.003

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* 52, 91–118. doi: 10.1023/A:1023949509487

Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* 12, 252–264. doi: 10.1038/nrc3239

Pei, X., Wang, X., and Li, H. (2018). LncRNA SNHG1 regulates the differentiation of Treg cells and affects the immune escape of breast cancer via regulating miR-448/IDO. *Int. J. Biol. Macromol.* 118(Pt A), 24–30. doi: 10.1016/j.ijbiomac.2018.06.033

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

Sun, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110. doi: 10.1136/jitc-2019-000110

Teschendorff, A. E., Gomez, S., Arenas, A., El-Ashry, D., Schmidt, M., Gehrmann, M., et al. (2010). Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. *BMC Cancer* 10:604. doi: 10.1186/1471-2407-10-604

Urup, T., Staunstrup, L. M., Michaelsen, S. R., Vitting-Seerup, K., Bennedbaek, M., Toft, A., et al. (2017). Transcriptional changes induced by bevacizumab combination therapy in responding and non-responding recurrent glioblastoma patients. *BMC Cancer* 17:278. doi: 10.1186/s12885-017-3251-3

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, C. J., Zhu, C. C., Xu, J., Wang, M., Zhao, W. Y., Liu, Q., et al. (2019). The lncRNA UCA1 promotes proliferation, migration, immune escape and inhibits apoptosis in gastric cancer by sponging anti-tumor miRNAs. *Mol. Cancer* 18:115. doi: 10.1186/s12943-019-1032-0

Wolf, D. M., Lenburg, M. E., Yau, C., Boudreau, A., and van 't Veer, L. J. (2014). Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PLoS One* 9:e88309. doi: 10.1371/journal.pone.0088309

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

Zhai, X., Zhao, J., Wang, Y., Wei, X., Li, G., Yang, Y., et al. (2018). Bibliometric analysis of global scientific research on lncRNA: a swiftly expanding trend. *Biomed. Res. Int.* 2018:7625078. doi: 10.1155/2018/7625078

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. *Mol. Neurobiol.* 55, 3684–3697. doi: 10.1007/s12035-017-0572-9

Zhou, Q., Chung, A. C., Huang, X. R., Dong, Y., Yu, X., and Lan, H. Y. (2014). Identification of novel long noncoding RNAs associated with TGF-beta/Smad3-mediated renal inflammation and fibrosis by RNA sequencing. *Am. J. Pathol.* 184, 409–417. doi: 10.1016/j.ajpath.2013.10.007

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

**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: 03 February 2021
doi: 10.3389/fgene.2021.610087

# Stratification of Estrogen Receptor-Negative Breast Cancer Patients by Integrating the Somatic Mutations and Transcriptomic Data

*Jie Hou, Xiufen Ye*, Yixing Wang and Chuanlong Li*

*College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China*

Patients with estrogen receptor-negative breast cancer generally have a worse prognosis than estrogen receptor-positive patients. Nevertheless, a significant proportion of the estrogen receptor-negative cases have favorable outcomes. Identifying patients with a good prognosis, however, remains difficult, as recent studies are quite limited. The identification of molecular biomarkers is needed to better stratify patients. The significantly mutated genes may be potentially used as biomarkers to identify the subtype and to predict outcomes. To identify the biomarkers of receptor-negative breast cancer among the significantly mutated genes, we developed a workflow to screen significantly mutated genes associated with the estrogen receptor in breast cancer by a gene coexpression module. The similarity matrix was calculated with distance correlation to obtain gene modules through a weighted gene coexpression network analysis. The modules highly associated with the estrogen receptor, called important modules, were enriched for breast cancer-related pathways or disease. To screen significantly mutated genes, a new gene list was obtained through the overlap of the important module genes and the significantly mutated genes. The genes on this list can be used as biomarkers to predict survival of estrogen receptor-negative breast cancer patients. Furthermore, we selected six hub significantly mutated genes in the gene list which were also able to separate these patients. Our method provides a new and alternative method for integrating somatic gene mutations and expression data for patient stratification of estrogen receptor-negative breast cancers.

**Keywords: breast cancer patient stratification, estrogen receptor-negative, distance correlation, significantly mutated gene, gene coexpression network**

OPEN ACCESS

***Edited by:***
*Rosalba Giugno,*
*University of Verona, Italy*

***Reviewed by:***
*Yuqi Zhao,*
*Jackson Laboratory, United States*
*Ugur Sezerman,*
*Sabancı University, Turkey*

****Correspondence:***
*Xiufen Ye*
*yexiufen@hrbeu.edu.cn*

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

***Received:*** *25 September 2020*
***Accepted:*** *04 January 2021*
***Published:*** *03 February 2021*

***Citation:***
*Hou J, Ye X, Wang Y and Li C (2021) Stratification of Estrogen Receptor-Negative Breast Cancer Patients by Integrating the Somatic Mutations and Transcriptomic Data. Front. Genet. 12:610087. doi: 10.3389/fgene.2021.610087*

## 1. INTRODUCTION

Breast cancer is a heterogeneous disease with many subtypes which exhibits significant differences in response to therapy and patient outcomes (Jonasson et al., 2019). Breast cancer has been known to be an endocrine-related cancer (Wu et al., 2020), and the majority of breast cancer subtypes are hormone-associated (DeSantis et al., 2017; Xu et al., 2019). The expression of the estrogen receptor (ER), progesterone receptor (PR), and human epithelial growth factor receptor 2 (HER2) as predictive and/or prognostic markers has been well established in multiple studies (Francis et al., 2019). Endocrine therapies that target the ER have long been the cornerstone of therapy

approaches for the majority of breast cancer patients. However, 20–30% of breast tumors do not express ER and are not responsive to existing endocrine therapies (Ni et al., 2011). The prognosis of estrogen receptor-negative (ER$^-$) breast cancer is worse than estrogen receptor-positive (ER$^+$) breast cancer in most situations, but ER$^-$ breast cancer patients do not always have a poor clinical outcome. Due to the lack of reliable biomarkers, it is impossible to identify ER$^-$ tumors with a good prognosis (Teschendorff et al., 2007; Zhang et al., 2016). Several studies have revealed that different chromosomal and gene expression patterns are present in patients with different estrogen receptor statuses (Zhang et al., 2009; Fohlin et al., 2020). Thus, an accurate grouping of ER$^-$ breast cancer into clinically relevant subtypes is of particular importance for therapeutic decision making.

Cancer is often driven by the accumulation of genetic alterations. Until now, the somatic mutation landscapes and signatures of more than a dozen major cancer types have been reported. However, pinpointing the driver mutations and cancer genes from millions of available cancer somatic mutations remains a significant challenge (Cheng et al., 2016). In The Cancer Genome Atlas (TCGA) database, a phenomenon can be observed that the position and nature of somatic mutations can often be translated to changes of protein structures or functions of the genes. The affected gene is designated as a significantly mutated gene (SMG). SMGs are the result of splice-site change, nonsense, nonstop, or frame-shift mutations (Zhang et al., 2016). The prevalence of SMGs in almost all cancer types has allowed for postulation that they may be act potentially as biomarkers for subtyping and testing for use in cancer patient outcome predictions, or a starting point of clarifying the tumorigenesis process (Cancer Genome Atlas Network, 2012).

Network approaches have provided the means to bridge the gap between individual genes and systems oncology (Ghazalpour et al., 2006). Weighted gene coexpression network analysis (WGCNA) is a systems biology method used to analyze gene expression profiling data which is widely used in bioinformatics (Zhang and Horvath, 2005). WGCNA can help researchers analyze the relationships between genes and pathogenic mechanisms. Instead of linking thousands of genes to the disease, this method focuses on the relationship between gene modules and disease traits (Huang et al., 2020). Many studies that constructed the coexpression networks in breast cancer used WGCNA. Coexpression networks were used to screen hub genes from the co-expression module using the relationship between genes and traits, together with the core position of genes in the module (Tang et al., 2019; Jia et al., 2020). A coexpression network analysis can also identify the prognostic lncRNAs (Liu et al., 2019; Li et al., 2020). However, these studies did not consider the information of genetic mutations in breast cancer.

SMGs are not always concentrated in specific genomic loci, which suggests that instead of common genes, mutations may affect some pathways or gene interaction networks (Zhang et al., 2016). So, in this work, we propose a method to screen SMGs using gene coexpression networks to identify the SMGs that highly relate to ER_Status. We show the development of a workflow for screening SMGs associated with clinical data of the estrogen receptor in breast cancer by a gene coexpression module. The new gene list was designated as important-SMGs. The identified genes, which were used to stratify patients with different subtypes of cancers, were suggested to represent biomarkers. Our method provides a new alternative method for cancer patient stratification by integrating transcriptomic, variants, and clinic data.

## 2. METHODS

In this work, we propose a method for screening SMGs by a gene coexpression module associated with clinical data of breast cancer and the estrogen receptor; the workflow is summarized in **Figure 1**. We calculated the similarity coexpression matrix by distance correlation for WGCNA to construct a gene coexpression network and to obtain the gene modules. Distance correlation has a perfect theoretical system and works for both linear and nonlinear dependence between any two vectors (Székely et al., 2007). WGCNA is a method used to identify clusters (modules) of highly correlated genes (Zhang and Horvath, 2005). We identified some important modules that were significantly associated with the measured clinical estrogen receptor data. SMGs were then selected from the TCGA tumor somatic mutation data and the important-SMGs were obtained through the overlap between the important module genes and the SMGs. Furthermore, we respectively chose the hub SMGs in the important modules and acquired six genes which can be used as the biomarkers for stratification and prediction of patient survival of ER$^-$ breast cancer.

### 2.1. Datasets

The TCGA datasets used in this study can be found in the Data Portal for TCGA-Breast Cancer (Weinstein et al., 2013), For the construction of the gene coexpression and the SMGs selection, we used the TCGA dataset. The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform with $log_2(x + 1)$ transformed RSEM normalized count (Cancer Genome Atlas Network, 2012). The samples were screened based on RNA-seq data and clinical data, after which we selected genes with a variable coefficient of more than 0.2 and a mean >1. Ultimately, we obtained 5,076 genes.

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset from the cBioportal website (Cerami et al., 2012) contains cDNA microarray performed on the Illumina HT-12 platform (Curtis et al., 2012; Pereira et al., 2016). The details of data normalization can be found in Margolin et al. (2013). For validation, both datasets containing gene expression data and matching survival time (months) were used for survival analysis. Samples in the METABRIC were screened based on the clinical data (contain ER_Status, Days, Vital_Status). The sample numbers used in the two datasets are shown in **Table 1**.

### 2.2. Distance Correlation

In 2007, distance correlation was proposed by Szekely, Rizzo, and Bakirov in the paper titled *Measuring and Testing*

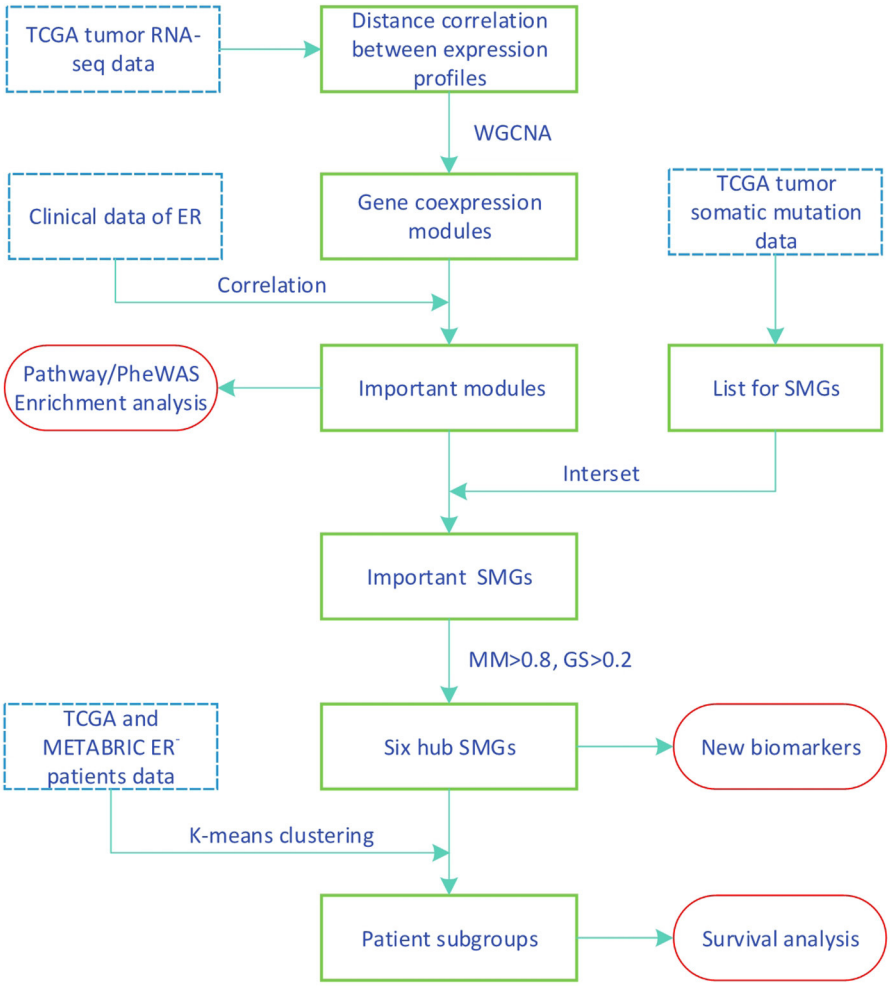



**FIGURE 1 |** Workflow of identifying new biomarkers using transcriptomic and variants data.

**TABLE 1 |** Sample numbers in two datsets.

| Dataset | Total | SMGs | ER$^+$ | ER$^-$ | Deceased/Living (ER$^-$) |
|---|---|---|---|---|---|
| TCGA | 637 | 383 | 499 | 133 | 23/110 ≈ 0.209 |
| METABRIC | 1,888 | – | 1,435 | 424 | 240/184 ≈ 1.304 |

*There are more samples in METABRIC and longer clinical follow-up time.*

*Dependence by Correlation of Distances* published in the Annals of Statistics (Székely et al., 2007). In this work, the similarity coexpression matrix was calculated with distance correlation for WGCNA to perform a gene coexpression network analysis. Unlike the Pearson correlation, distance correlation works for both linear and nonlinear dependence between two gene expression profiles. However, distance correlation is still a relatively expensive computation. The time complexity of distance correlation was $O(n^2)$. Distance correlation was calculated using the energy package in R (see the references in the manual for more package details).

## 2.3. WGCNA

WGCNA (Zhang and Horvath, 2005) can be used to identify clusters (modules) of highly correlated genes. This method summarizes such clusters using the module eigengene or an intramodular hub gene. Alternatively, it relates modules to one another and to external sample traits and calculating module membership measures using the eigengene network methodology (Langfelder and Horvath, 2008; Luo et al., 2018). The functions of WGCNA are plentiful, and only some of them were used in this study. We mainly used the process of module division of WGCNA. First, the correlation for all genes was calculated using correlation methods, and a similarity coexpression matrix was obtained. The similarity coexpression matrix was transformed to an adjacency matrix using the soft-thresholding power which was chosen based on the criteria of approximating the scale-free topology (SFT) of the network.

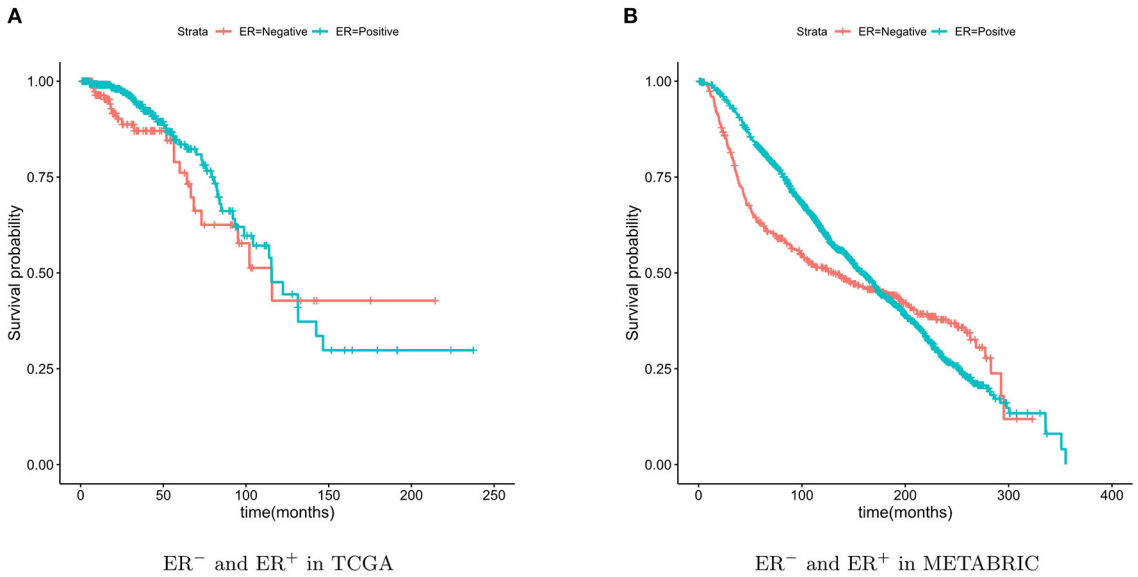



**FIGURE 2 |** Kaplan-Meier survival curves of ER⁻ and ER⁺. The ER⁻ breast cancer patient have a poor prognosis in the short term and a relatively good prognosis in the longer term.

Next, a topological overlap matrix was computed from the adjacency matrix. Finally, a tree (dendrogram) was produced from the dissimilarity topological overlap matrix by hierarchical clustering. The clusters (modules) were obtained using dynamic tree cutting. For functions of WGCNA, we refer to the corresponding tutorials package. The WGCNA package is now available from the *Comprehensive R Archive Network*(CRAN).

## 2.4. Enrichment Analysis

Enrichr (Chen et al., 2013; Kuleshov et al., 2016) was used to analyze the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2019) pathways and the phenome-wide association studies (PheWAS) (Denny et al., 2010) of diseases identified in the important modules. Enrichr is open source and freely available online.

## 2.5. SMGs and Important SMGs

The SMGs were obtained by screening the somatic mutations derived from the TCGA breast cancer patients. The SMGs are genes with frame-shift indels, splice-site changes, nonstop mutations, or nonsense mutations (Zhang et al., 2016). Mismatch, silent, RNA, and in-frame indel mutations did not belong to the SMGs. Among the samples we selected, the mutation types of 1920 SMGs and 383 samples are listed in **Supplementary Table 1**.

To obtain ER-related SMG, we acquired some SMGs contained in the important modules by taking the intersection of genes in important modules and SMGs, and we named them important SMGs.

## 2.6. Gene Significance and Module Membership

To find genes associated with clinical ER_Status, we defined a measure of gene significance (GS) between the $i$-th gene profile $x_i$ and the ER_Status as

$$\mathrm{GS}_i = cor(x_i, \mathrm{ER\_Status}), \tag{1}$$

where $cor(\cdot, \cdot)$ denotes the correlation coefficients. ER_Status can be mapped to a binary indicator variable where 1 is positive and 0 is negative. The higher the absolute value of $\mathrm{GS}_i$ of the gene, the more closely relevant it is to ER.

To measure the relationship between the $i$-th gene and the module to which it belongs, we introduced the module membership (MM) (Langfelder and Horvath, 2008; Wei et al., 2020) which was defined by calculating the correlation coefficient between the gene expression profile and the module eigengene.

## 2.7. Survival Analysis

Some subtypes of breast cancer have a poor prognosis in the short term and a relatively good prognosis in the longer term. This particular characteristic of ER⁻ breast cancer can be observed from **Figure 2**. Due to this characteristic, the two survival curves may cross. This made the log-rank test $P$-value large, although the two curves were obviously separate. The two-stage hypothesis test was developed for handling the crossing hazard rates problem. We evaluated the $P$-values of both the log-rank and the two-stage hypothesis tests.

For validation, the TCGA breast cancer dataset (containing 133 ER⁻ patients) and the METABRIC dataset (containing 424 ER⁻ patients) were used. The breast cancer characteristic led to the crossing of the two survival curves, so the two-stage hypothesis test was developed for handling the crossing

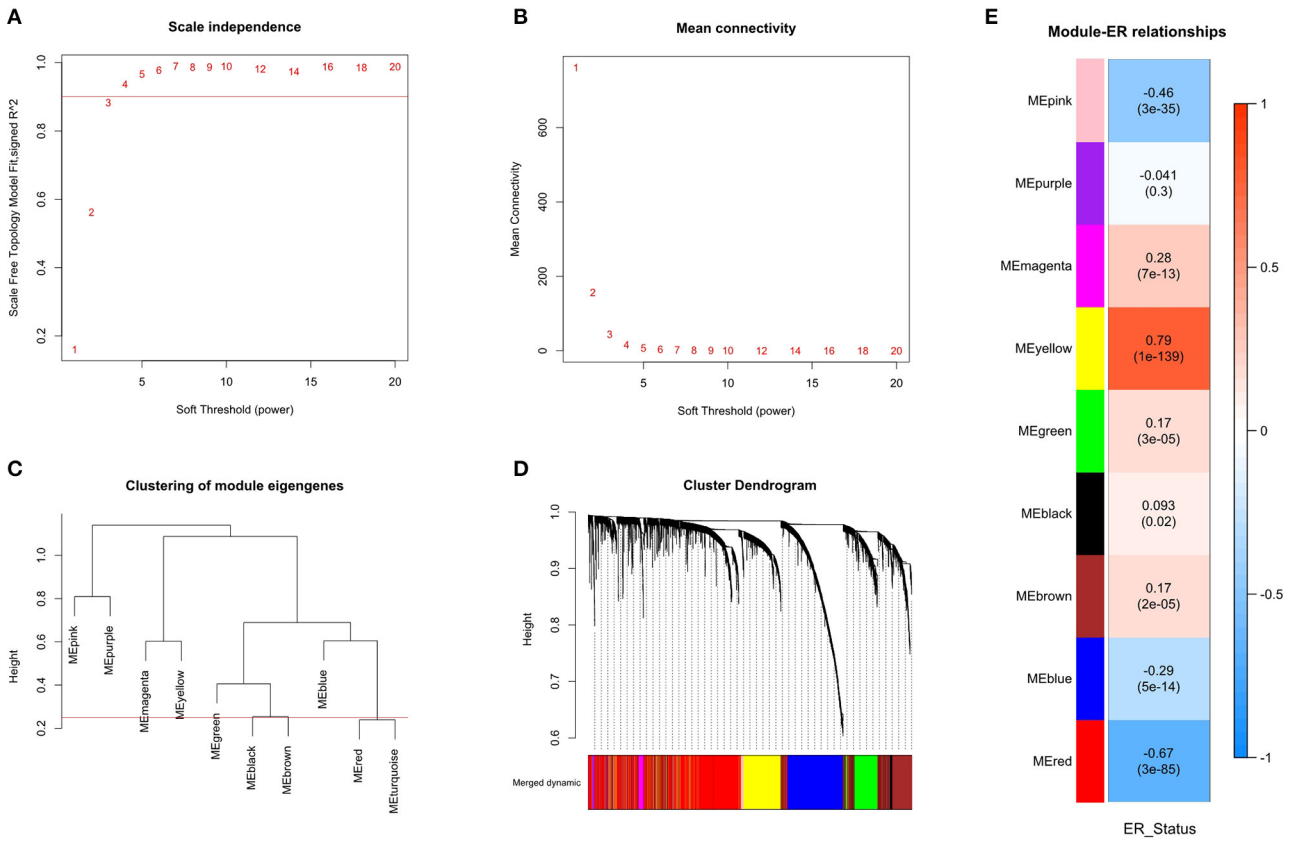



**FIGURE 3 |** Identification of modules associated with the ER_Status of breast cancer. **(A)** The scale-free fit index for various soft-thresholding power. Scale-free topology (SFT) was achieved when the recommended soft-thresholding power was 3. **(B)** The mean connectivity for various soft-thresholding powers. **(C)** The cluster dendrogram of module eigengenes. **(D)** The cluster dendrogram of all genes with corresponding color assignments. Nine colors present nine modules. **(E)** Module-ER_Status relationship heatmap. The values above the brackets represent the correlation coefficients between modules and ER_Status. The values in brackets are the *P*-values for the association test. The red and yellow modules were significantly related to the ER_Status and selected as the important modules.

hazard rates problem (Qiu and Sheng, 2008). To predicate the significance of the difference in the survival time between the two patient groups, we performed the Log-rank and two-stage tests.

# 3. RESULTS

## 3.1. Gene Co-expression Module Associated With Estrogen Receptor

The similarity coexpression matrix was calculated with distance correlation. When we chose 3 as the recommended soft-thresholding power, the SFT was achieved. The scale-free fit index is shown in **Figure 3A**, and the mean connectivity for various soft-thresholding powers is shown in **Figure 3B**. The modules were obtained by hierarchical clustering based on the minimum module size of 30. The modules were then merged if the similarity between module eigengenes were >0.75. The cluster trees (dendrograms) of the module eigengenes are shown in **Figure 3C** and the cluster dendrograms of the genes that were assigned module colors after the merge is shown in the **Figure 3D**. Finally, nine coexpression modules were constructed.

To find modules related to clinical ER_Status, the correlation between modules eigengenes and ER_Status was calculated and shown in **Figure 3E**. The modules eigengenes were associated with ER_Status when $p < 0.01$. There were four modules positively associated with ER_Status, and three modules that were negatively associated. The yellow and red modules, where the absolute value of the correlation coefficient was >0.6, had the highest correlations with ER_Status. This means that these modules have great biological significance related to the ER_Status, so these two modules were selected as the important modules.

## 3.2. Enrichment Analysis of the Important Modules

We analyzed the KEGG and PheWAS enrichments for the two important modules to associate each module with biological pathways and diseases (see **Table 2**). Enrichment results of all modules are available in **Supplementary Table 2**.

Several KEGG enriched terms related to cardiac diseases were enriched in the yellow module. Approximately 59% of cancer patients in the dataset used in this study received radiation therapy. What is more, hormonal therapy plays an important role in breast cancer treatments (Jones and Buzdar, 2004). Some reports showed that one of the side effects of

**TABLE 2 |** KEGG and PheWAS enrichment analysis by Enrichr of the important modules identified by WGCNA.

| Module | No. | KEGG | *P*-value | PheWeb | *P*-value |
|---|---|---|---|---|---|
| Yellow | 677 | Dilated cardiomyopathy (DCM) | 3.67E-03 | Cancer of stomach | 2.18E-03 |
| | | Adrenergic signaling in cardiomyocytes | 3.86E-03 | Pelvic peritoneal adhesions,-female (postoperative) (postinfection) | 5.20E-03 |
| | | Cardiac muscle contraction | 4.83E-03 | Cholecystitis without cholelithiasis | 5.85E-03 |
| | | Glutamatergic synapse | 5.35E-03 | Cancer of eye | 8.57E-03 |
| | | Hypertrophic cardiomyopathy (HCM) | 8.07E-03 | Elevated cancer antigen 125 [CA 125] | 8.57E-03 |
| Red | 1819 | Metabolism of xenobiotics-by cytochrome P450 | 2.80E-04 | Genital prolapse | 6.45E-04 |
| | | Chemical carcinogenesis | 3.42E-04 | Breast cancer | 2.55E-03 |
| | | Neuroactive ligand-receptor interaction | 1.31E-03 | Osteoarthrosis, localized, primary | 2.73E-03 |
| | | Caffeine metabolism | 6.53E-03 | Heart failure with preserved EF [Diastolic heart failure] | 2.88E-03 |
| | | Protein digestion and absorption | 6.83E-03 | Other venous embolism and thrombosis | 4.09E-03 |

*All the important modules were highly enriched with PheWAS in breast cancer, cancer or female-related diseases.*

breast cancer treatments (radiation therapy, hormonal therapy) is cardiotoxicity (Bird and Swain, 2008; Demissei et al., 2020). This may be the cause of the enrichment of the cardiac disease pathway in the yellow module. The yellow modules were highly enriched in cancer (For instance, cancer of stomach, cancer of eye, and elevated cancer antigen) or female-related diseases with PheWAS. With the KEGG pathway enrichment analysis, the red modules were enriched in the metabolism and chemical carcinogenesis pathways. This is consistent with the conclusion that the ER is a modulator in metabolic disorders (Mauvais-Jarvis et al., 2013). With PheWAS diseases enrichment analysis, the top two significant terms were breast cancer and female-related diseases. The results of the enrichment analysis confirmed the biological significance of the important modules related to breast cancer or other cancers.

## 3.3. Survival Analysis by Important-SMGs and RNA-Seq Data

The new gene list, designated as the important-SMGs, was obtained through overlapping the important module genes and the SMGs. The list contains 227 SMGs and is shown in **Supplementary Table 3**.

In Zhang et al. (2016), the ER⁻ samples were also separated into two groups. The authors developed an approach for stratifying cancer patients into groups with different clinical outcomes. They focused on this specific Group 1 with a significantly higher proportion of ER-negative patients. Thirteen SMGs among the 201 SMGs in Group 1 are identical to the important-SMGs obtained by our approach. The TCGA breast cancer dataset (containing 133 ER⁻ patients) and the METABRIC dataset (containing 424 ER⁻ patients) were used in this test. The important-SMGs in this work were compared with the Group 1-specific genes in Zhang et al. (2016). For survival analysis, the ER⁻ samples were separated into two groups based on the K-means algorithm with K = 2, using the two gene lists and the RNA-seq data. The results are shown in **Figure 4**.

From the two-stage *P*-value, the two gene lists in our test on the TCGA ER⁻ data were able to separate the patients into two groups with a significant survival time difference. The survival curves in **Figures 4A,B** were clearly separated, but the two curves obtained by the important-SMGs in **Figure 4B** were further apart than that obtained by the gene list of Group 1 in Zhang et al. (2016) in **Figure 4A**. Therefore, on the TCGA ER⁻ data, the important-SMGs were able to separate the patients into two more significant groups.

The test on METABRIC data shown in **Figure 4D** suggested that the important-SMGs were able to separate the patients into two groups with a significant survival time difference (the *P*-value of the two tests are 0.00853). However, the gene list of Group 1 in Zhang et al. (2016) shown in **Figure 4C** could effectively separate the ER⁻ patients with the bigger *P*-value (the *P*-values of the two tests larger than 0.01). The survival curves of the two groups obtained by the important-SMGs were also further apart. Therefore, on the METABRIC data, the important-SMGs were able to separate the patients into two more significant groups.

## 3.4. Survival Analysis by Six Hub SMGs and RNA-Seq Data

As discussed in the previous section, the 227 important-SMGs were able to more significantly separate the ER⁻ patients into two groups. As biomarkers, it is best to keep the number of genes as small as possible. Gene co-expression modules were composed of highly correlated genes, we just have to choose a few representative genes from 227 SMGs. The most representative genes are the hub genes within important modules.

We chose the GS $>$ 0.2 and MM $>$ 0.8 in the two important modules and obtain 29 hub genes. The six genes (FOXA1, GABRP, BCL11A, DNALI1, STAC, and ESR1) obtained by overlapping the 29 hub genes and the SMGs were called the hub-SMGs. The ER⁻ samples were separated into two groups based on the K-means algorithm with $K = 2$, using the hub-SMGs and the RNA-seq data. The results in the TCGA and the METABRIC datasets of survival analysis are shown in **Figure 5**. From the value of the two-stage *P*-value, the hub-SMGs can significantly separate the ER⁻ breast cancer patients

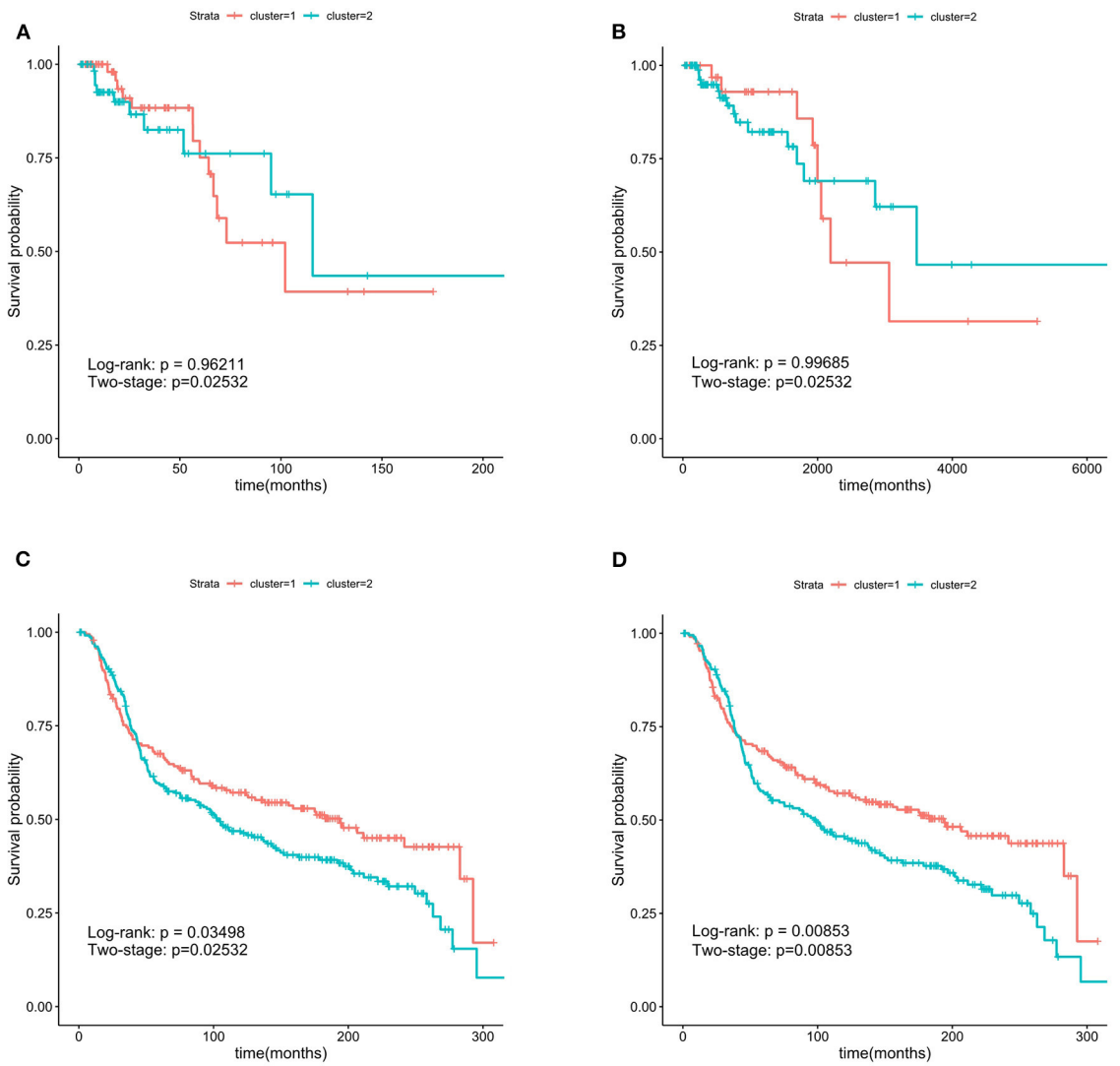



**FIGURE 4 |** Kaplan-Meier survival curves. The 227 important-SMGs were able to separate the patients into two groups more significantly. The *P*-values were smaller in METABRIC dataset. **(A)** Group 1 in TCGA, **(B)** important-SMGs in TCGA, **(C)** group 1 METABRIC, **(D)** important-SMGs in METABRIC.

into two groups. Patients in different groups have different survival times. From **Figure 5B**, the P-value in the METABRIC dataset is 0.00554 which is smaller than the *P*-value of the Important-SMGs 0.00853 (see **Figure 4D**). This suggests that a few genes can represent the important-SMGs and separate the ER$^-$ patients.

## 4. CONCLUSION

With rapid developments in massively parallel sequencing and computing capacity, a rich resource of data in different modalities for cancer specimens have been generated in public databases at an amazing speed. Therefore, integrating and mining the tremendous volume of data has become an important subject in the field of bioinformatics. In our study, we show the development of a new workflow to integrate somatic mutations, gene expression, and clinical data. We constructed a gene co-expression network and obtained nine coexpression modules. The yellow and red modules were selected as the important modules, because these two modules have the most significant correlation with ER. We obtained the important-SMGs list through the overlap between the important module genes and the SMGs. In the TCGA and METABRIC datasets, we verified that the important-SMGs were able to separate the ER$^-$ patients more significantly than other methods.

Furthermore, we selected the six hub SMGs as potential biomarkers which are also able to separate these patients. The genes ESR1, DNALI1, and FOXA1 belong to the yellow module, the genes GABRP, STAC, and BCL11A belong to the

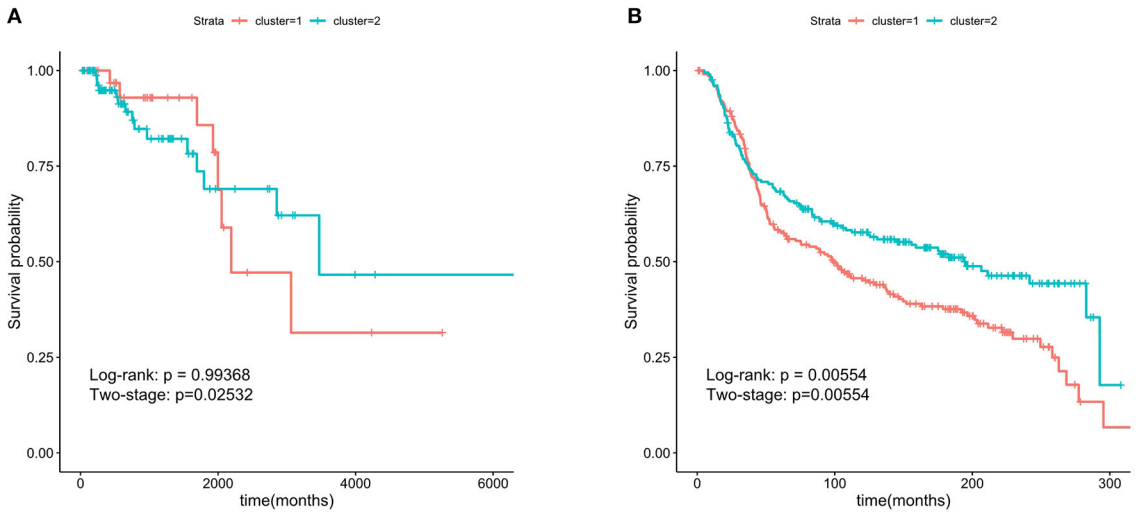



**FIGURE 5 |** Kaplan-Meier survival curves using the six hub SMGs. A few hub SMGs can represent the 227 important-SMGs and were able to separate the patients into two groups more significantly. **(A)** Six hub SMGs in TCGA. **(B)** Six hub SMGs in METABRIC.

red module. These six genes have been reported to be related to cancer or breast cancer in the literature. In particular, two genes in the yellow module are directly related to estrogen receptors. ESR1 (estrogen receptor 1, also known as ER) is a gene that encodes the estrogen receptor protein (Holst et al., 2007). FOXA1 is a key determinant of estrogen receptor function and endocrine response (Hurtado et al., 2011). The conclusion of the relevant literature verified the correctness of our algorithm flow.

Our work provided a novel workflow for identifying new biomarkers using transcriptomic and variants data. In future research, we will use the same workflow for other complex diseases to further test its effectiveness and to find a new gene list to stratify patients.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://portal.gdc.cancer.gov/, https://www.cbioportal.org/.

## AUTHOR CONTRIBUTIONS

XY supervised this work, made critical revisions, and approved final version. JH designed the study, analyzed the data, and wrote the original draft of the manuscript. YW and CL analyzed the data and revised the manuscript. All authors read and approved the final manuscript.

## FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 41876100), and the Development Project of Applied Technology in Harbin (Grant No. 2016RAXXJ071).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.610087/full#supplementary-material

## REFERENCES

Bird, B. R. H., and Swain, S. M. (2008). Cardiac toxicity in breast cancer survivors: review of potential cardiac problems. *Clin. Cancer Res.* 14, 14–24. doi: 10.1158/1078-0432.CCR-07-1033

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128. doi: 10.1186/1471-2105-14-128

Cheng, F., Zhao, J., and Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinformatics* 17, 642–656. doi: 10.1093/bib/bbv068

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Demissei, B. G., Hubbard, R. A., Zhang, L., Smith, A. M., Sheline, K., McDonald, C., et al. (2020). Changes in cardiovascular biomarkers with breast cancer therapy and associations with cardiac dysfunction. *J. Am. Heart Assoc.* 9:e014708. doi: 10.1161/JAHA.119.014708

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126

DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A., and Jemal, A. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. *Ca-Cancer J. Clin*. 67, 439–448. doi: 10.3322/caac.21412

Fohlin, H., Bekkhus, T., Sandström, J., Fornander, T., Nordenskjöld, B., Carstensen, J., et al. (2020). Rab6c is an independent prognostic factor of estrogen receptor-positive/progesterone receptor-negative breast cancer. *Oncol. Lett*. 19, 52–60. doi: 10.3892/mco.2020.2014

Francis, I. M., Altemaimi, R. A., Al-Ayadhy, B., Alath, P., Jaragh, M., Mothafar, F. J., and Kapila, K. (2019). Hormone receptors and human epidermal growth factor (her2) expression in fine-needle aspirates from metastatic breast carcinoma-role in patient management. *J. Cytol*. 36, 94–100. doi: 10.4103/JOC.JOC_117_18

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*. 2:e20130. doi: 10.1371/journal.pgen.0020130

Holst, F., Stahl, P. R., Ruiz, C., Hellwinkel, O., Jehan, Z., Wendland, M., et al. (2007). Estrogen receptor alpha (esr1) gene amplification is frequent in breast cancer. *Nat. Genet*. 39, 655–660. doi: 10.1038/ng2006

Huang, Y., Liu, H., Zuo, L., and Tao, A. (2020). Key genes and co-expression modules involved in asthma pathogenesis. *PeerJ*. 8:e8456. doi: 10.7717/peerj.8456

Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D., and Carroll, J. S. (2011). Foxa1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet*. 43, 27–33. doi: 10.1038/ng.730

Jia, R., Zhao, H., and Jia, M. (2020). Identification of co-expression modules and potential biomarkers of breast cancer by WGCNA. *Gene* 750:144757. doi: 10.1016/j.gene.2020.144757

Jonasson, E., Ghannoum, S., Persson, E., Karlsson, J., Kroneis, T., Larsson, E., et al. (2019). Identification of breast cancer stem cell related genes using functional cellular assays combined with single-cell RNA sequencing in MDA-MB-231 cells. *Front. Genet*. 10:500. doi: 10.3389/fgene.2019.00500

Jones, K. L., and Buzdar, A. U. (2004). A review of adjuvant hormonal therapy in breast cancer. *Endocr. Relat. Cancer* 11, 391–406. doi: 10.1677/erc.1.00594

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 47, D590–D595. doi: 10.1093/nar/gky962

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 44, W90–W97. doi: 10.1093/nar/gkw377

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, Z., Li, Y., Wang, X., and Yang, Q. (2020). Identification of a six-immune-related long non-coding RNA signature for predicting survival and immune infiltrating status in breast cancer. *Front. Genet*. 11:680. doi: 10.3389/fgene.2020.00680

Liu, Z., Li, M., Hua, Q., Li, Y., and Wang, G. (2019). Identification of an eight-lncRNA prognostic model for breast cancer using WGCNA network analysis and a cox-proportional hazards model based on l1-penalized estimation. *Int. J. Mol. Med*. 44, 1333–1343. doi: 10.3892/ijmm.2019.4303

Luo, M., Zhang, Q., Xia, M., Hu, F., Ma, Z., Chen, Z., et al. (2018). Differential co-expression and regulatory network analysis uncover the relapse factor and mechanism of T cell acute leukemia. *Mol. Ther. Nucleic Acids*. 12, 184–194. doi: 10.1016/j.omtn.2018.05.003

Margolin, A. A., Bilal, E., Huang, E., Norman, T. C., Ottestad, L., Mecham, B. H., et al. (2013). Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med*. 5:181re1. doi: 10.1126/scitranslmed.3006112

Mauvais-Jarvis, F., Clegg, D. J., and Hevener, A. L. (2013). The role of estrogens in control of energy balance and glucose homeostasis. *Endocr. Rev*. 34, 309–338. doi: 10.1210/er.2012-1055

Ni, M., Chen, Y., Lim, E., Wimberly, H., Bailey, S. T., Imai, Y., et al. (2011). Targeting androgen receptor in estrogen receptor-negative breast cancer. *Cancer Cell*. 20, 119–131. doi: 10.1016/j.ccr.2011.05.026

Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K. M., Provenzano, E., Bardwell, H. A., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun*. 7, 1–16. doi: 10.1038/ncomms11479

Qiu, P., and Sheng, J. (2008). A two-stage procedure for comparinghazard rate functions. *J. R. Stat. Soc. Ser. B Stat. Methodol*. 70, 191–208. doi: 10.1111/j.1467-9868.2007.00622.x

Székely, G. J., Rizzo, M. L., Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat*. 35, 2769–2794. doi: 10.1214/009053607000000505

Tang, J., Lu, M., Cui, Q., Zhang, D., Kong, D., Liao, X., et al. (2019). Overexpression of ASPM, CDC20, and TTK confer a poorer prognosis in breast cancer identified by gene co-expression network analysis. *Front. Oncol*. 9:310. doi: 10.3389/fonc.2019.00310

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., and Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol*. 8:R157. doi: 10.1186/gb-2007-8-8-r157

Wei, S., Chen, J., Huang, Y., Sun, Q., Wang, H., Liang, X., et al. (2020). Identification of hub genes and construction of transcriptional regulatory network for the progression of colon adenocarcinoma hub genes and TF regulatory network of colon adenocarcinoma. *J. Cell. Physiol*. 235, 2037–2048. doi: 10.1002/jcp.29067

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet*. 45:1113. doi: 10.1038/ng.2764

Wu, N., Fu, F., Chen, L., Lin, Y., Yang, P., and Wang, C. (2020). Single hormone receptor-positive breast cancer patients experienced poor survival outcomes: a systematic review and meta-analysis. *Clin. Transl. Oncol*. 22, 474–485. doi: 10.1007/s12094-019-02149-0

Xu, J., Bao, H., Wu, X., Wang, X., Shao, Y. W., and Sun, T. (2019). Elevated tumor mutation burden and immunogenic activity in patients with hormone receptor-negative or human epidermal growth factor receptor 2-positive breast cancer. *Oncol. Lett*. 18, 449–455. doi: 10.3892/ol.2019.10287

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. 4:17. doi: 10.2202/1544-6115.1128

Zhang, J., Abrams, Z., Parvin, J. D., and Huang, K. (2016). Integrative analysis of somatic mutations and transcriptomic data to functionally stratify breast cancer patients. *BMC Genomics* 17:513. doi: 10.1186/s12864-016-2902-0

Zhang, Y., Martens, J. W., Jack, X. Y., Jiang, J., Sieuwerts, A. M., Smid, M., et al. (2009). Copy number alterations that predict metastatic capability of human breast cancer. *Cancer Res*. 69, 3795–3801. doi: 10.1158/0008-5472.CAN-08-4596

**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 February 2021
doi: 10.3389/fgene.2021.632626

# Systemic Multi-Omics Analysis Reveals Amplified P4HA1 Gene Associated With Prognostic and Hypoxic Regulation in Breast Cancer

*Manikandan Murugesan and Kumpati Premkumar**

*Department of Biomedical Science, School of Biotechnology and Genetic Engineering, Bharathidasan University, Tiruchirappalli, India*

Breast cancer (BC) is a common malignant tumor in females around the world. While multimodality therapies exist, the mortality rate remains high. The hypoxic condition was one of the potent determinants in BC progression. The molecular mechanisms underpinning hypoxia and their association with BC can contribute to a better understanding of tailored therapies. In this study, two hypoxic induced BC transcriptomic cohorts (GSE27813 and GSE47533) were assessed from the GEO database. The P4HA1 gene was identified as a putative candidate and significantly regulated in hypoxic BC cells compared to normal BC cells at different time intervals (6 h, 9 h, 16 h, 32 h, and 48 h). In patients with Luminal ($p$ < 1E-12), triple-negative subclasses ($p$ = 1.35059E-10), Stage 1 ($p$ = 8.8817E-16), lymph node N1 ($p$ = 1.62436E-12), and in the 40–80 age group ($p$ = 1.62447E-12), the expression of P4HA1 was closely associated with the clinical subtypes of BC. Furthermore, at the 10q22.1 chromosomal band, the P4HA1 gene displayed a high copy number elevation and was associated with a poor clinical regimen with overall survival, relapse-free survival, and distant metastases-free survival in BC patients. In addition, using BioGRID, the protein–protein interaction (PPI) network was built and the cellular metabolic processes, and hedgehog pathways are functionally enriched with GO and KEGG terms. This tentative result provides insight into the molecular function of the P4HA1 gene, which is likely to promote hypoxic-mediated carcinogenesis, which may favor early detection of BC and therapeutic stratification.

**Keywords: breast cancer, hypoxia, prognosis, omics, computational biology**

OPEN ACCESS

***Edited by:***
*Suman Ghosal,*
*National Institutes of Health (NIH), United States*

***Reviewed by:***
*Vishal Midya,*
*Icahn School of Medicine at Mount Sinai, United States*
*Shankar Suman,*
*The Ohio State University, United States*
*Victor Banerjee,*
*University of Texas Health Science Center at Houston, United States*

***Correspondence:**
*Kumpati Premkumar*
*prems@bdu.ac.in*

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

***Received:*** *23 November 2020*
***Accepted:*** *29 January 2021*
***Published:*** *22 February 2021*

***Citation:***
*Murugesan M and Premkumar K (2021) Systemic Multi-Omics Analysis Reveals Amplified P4HA1 Gene Associated With Prognostic and Hypoxic Regulation in Breast Cancer. Front. Genet. 12:632626. doi: 10.3389/fgene.2021.632626*

## INTRODUCTION

The second leading cause of tumor-related death worldwide is breast cancer (BC) (WCRF, 2018). Poly-etiology and the constituent nature of BC threaten early diagnosis and treatment strategies (Feng et al., 2018). BC is divided into five prevailing subtypes based on molecular profiling techniques: luminal A/B, basal-like, HER2(+), and normal breast-like. Molecular heterogeneity in BC inter-/intra-tumor also increases tumor growth and becomes more complex in therapy (Koren and Bentires-Alj, 2015; Haynes et al., 2017)A common trait of cancer cells is that they quickly proliferate, consuming significant amounts of oxygen that hampers the low-level oxygen

state called hypoxia. The hypoxia-inducible factor 1 (HIF-1) regulator pathway gets activated once the cancer cell enters hypoxic conditions (1–5% $O_2$), contributing to the promotion of angiogenesis and metastatic tumor characteristics in BC (Murugesan and Premkumar, 2018; Al Tameemi et al., 2019; Dillekas et al., 2019). In invasive-BC tumors, about 50–60% with hypoxic regions and suggests a critical determinant of metastasis (Greer et al., 2012). Almost 90% of BC deaths are reported due to delayed late diagnosis (Dillekas et al., 2019). Clinical studies show that hypoxia is one of the primary drivers of epithelial-mesenchymal transformation (EMT) and metastatic cascade transition (Dillekas et al., 2019). In addition, HIF-1 was implicated in hematogenous breast metastases to lung cancer and was associated with low patient survival and resistance to chemotherapy in breast (Campbell et al., 2019), gastric (Cheng et al., 2012), and colorectal (Baba et al., 2010) cancer.

The accumulating knowledge in microarray databases (Oncomine, GEO, Array Express, and so on) using genome-wide technologies has played an essential role in exploring the cancer-related molecular pathogenesis portfolio (Siegel et al., 2018; Ha et al., 2019; Shou et al., 2020; Yang et al., 2020). In future contexts, the ability to dissect and incorporate cancer omics data opens the door to a new approach to the biomarker strategy for diagnosis and treatment. In the same way, TCGA provides a multi-cancer cohort of RNA-Seq transcriptomics, which has led to a significant increase in understanding the biology of malignancy. Its accessibility has led to a splendid opportunity to extend molecular tumors' fundamental mechanisms (Manzoni et al., 2018).

Prolyl collagen 4-hydroxylase (P4H) is a tetrameric α2β2 α-ketoglutarate (α-KG) –dioxygenase that is responsible for collagen folding and stabilization. Collagens, which are the most abundant proteins in humans, provide extracellular matrix (ECM) assembly scaffolding (rigidity and cell adhesion) (Koski et al., 2017) and are also associated with stabilizing tumor proliferation (Provenzano et al., 2006). Three P4HA isoforms in mammalian cells (P4HA1-3) were identified. Of the three isoforms, P4HA1 is the foremost isoform that contributes to the foremost peptide bond and protein scaffolding activity. P4HA2 is also involved in the collagen synthesis and folding of collagen chains. The P4HA1 is majorly expressed in the testis and placenta, P4HA2 in adipose tissue, and P4HA3 in the heart and placenta. Reports suggest P4HA1 and P4HA2 to be associated with cancer proliferation and hypoxic regulation (Weinschenk et al., 2002; Cioffi et al., 2003; Kukkola et al., 2003; Willam et al., 2006; Gorres and Raines, 2010). In addition, P4HA1 enhances EMT and stemness of malignant cells through the HIF-1 pathway (Kappler et al., 2017; D'Aniello et al., 2019). P4HA1 has recently been found to overexpress in gliomas and HNSCC; its expression associated with tumor microvessel density (Li et al., 2019). Recent studies have shown that increased production of collagen is linked to BC progression, adhesion, and invasion (Xiong et al., 2018; Wishart et al., 2020).

However, the potential effects of P4HA1 and their precise contribution to BC are not entirely explored. This research extensively examined the expression of P4HA1 in breast cancer cells and its therapeutic relevance in tumor-affected samples using integrative functional multi-omic approaches. In addition, the regulatory genes of P4HA1 and their molecular, pathological, and signaling predictive role in BC consented. In a diagnostic and treatment regimen to control BC malignancy, P4HA1 could be an effective target.

# MATERIALS AND METHODS

## Microarray Data

The GSE27813 and GSE47533 transcriptomic profiles of breast cancer cells subjected to hypoxia conditions (1% O2) were downloaded from the Gene Expression Omnibus (GEO) database[1] of the National Center for Biotechnology Information (NCBI) and explored in the current study. The studies were carried out on two different platforms GPL10558-Illumina Human HT-12 V4.0 bead chip expression and GPL6884-Illumina Human WG-6 v3.0 bead chip expression. The normalized data were downloaded, and probes were annotated with authentic gene symbols from each platform using the required Illumina annotation files. Integrative analysis of these BC mRNA transcriptomes with/without hypoxic exposure profiles was used to identify the potential genes at various time intervals. The full integrated analysis chart had shown in **Figure 1**.

## The Cancer Genome Atlas (TCGA) Data Validation

TCGA is a web-based platform that visualizes, integrates, and analyses malignancy genomics and associated clinical results. UALCAN[2] can be an intuitive, user-friendly, open-source web portal for an in-depth study of TCGA data (Chandrashekar et al., 2017). UALCAN uses RNA-Seq level 3 of TCGA and clinical data on 31 cancer types. The expression of the candidate gene in normal tissues was subsequently weighed against the corresponding BC tissues. Moreover, overall survival (OS)/recurrence-free survival (RFS) was assessed using Kaplan–Meier survival curves, and the hazard ratio (HR) was determined with 95% confidence intervals, and log-rank *p*-value was ascertained. Furthermore, assessment of mRNA expression of P4HA1 among different subtypes of breast tumors was achieved to explore the pathological characteristics of genes in tumor initiation or progression.

## Oncomine Database Analysis

The expression level of P4HA1 was derived from the oncomine database[3] in various BC transcriptomic profiles. The oncomine interface (Compendia biosciences, Ann Arbor, MI, United States) is an online archive of previously published, open-access microarray data widely distributed and freely accessible to cancer repositories (Rhodes et al., 2004). The differential expression of mRNA in cancer tissue relative to normal was achieved using the parameters of the *p*-value threshold of 0.01 and fold-change (FC) > 2.

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

[2]http://ualcan.path.uab.edu/index.html

[3]https://www.oncomine.org/resource/login.html

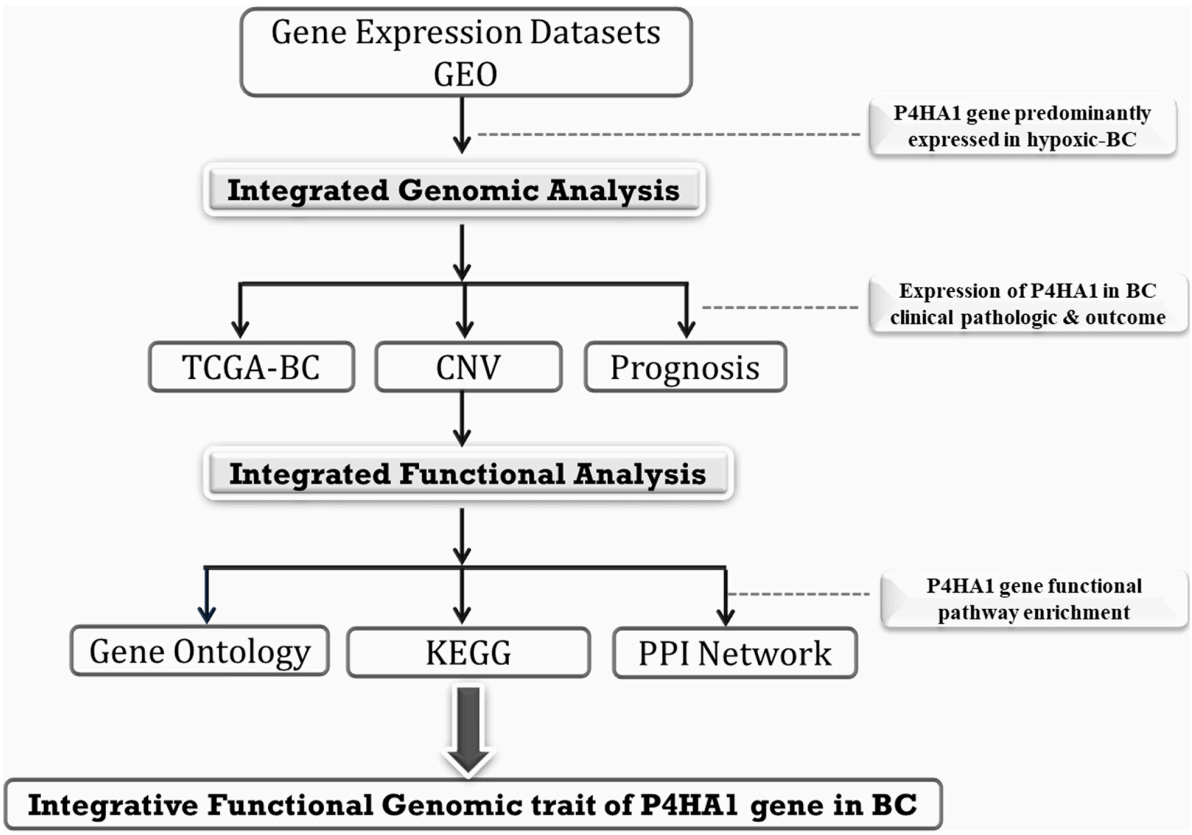



**FIGURE 1 |** A flow chart of integrated functional analysis.

## Copy Number Alteration Analysis

Using Progenetix (Progenetix, Stanford, CA, United States)[4], DNA copy number variations (CNVs), such as deletions and amplification in BC transcriptomic cohorts, were examined (Baudis and Cleary, 2001). It is an online repossession of cancer molecular-cytogenetic data that captures the robust, best-grained understanding of the absolute copy number aberration. The chromosomal variation features of the P4HA1 gene were analyzed in the TCGA-BC data to produce frequency gain/loss.

## Clinical Regimen Prognosis

Kaplan–Meier Plotter[5] is a data source that integrates gene expression and clinical data on about 21 cancer types, including breast cancer ($n$ = 6234) (Gyorffy and Schafer, 2009). KM Plotter was used to study the prognosis value for P4HA1 in BC. We centered our assessment on the overall patient survival (OS), distance metastasis-free survival (DMFS), and relapse-free survival (RFS). The log-rank $p$-value and hazard ratio with 95% confidence intervals additionally ascertained. The Cox proportional hazard regression model with microarray cohort GSE22133 was examined to verify the patient's overall survival between the expression of the P4HA1 gene and the BC's clinical characteristics. The median P4HA1 value is the threshold used to evaluate the prognostic score of each parameter.

## Protein–Protein Network

Protein–protein interaction networks provide information on the molecular framework of cellular processes and integral mobile activity. In the present research, a PPI network of P4HA1 regulatory genes built using an online database, the Biological General Repository for Interaction Datasets (BioGRID) v3.5.175[6], a database of already established networks; incorporates 1,728,498 protein and genetic interactions (Oughtred et al., 2019). In the BioGRID database, we have imported the lists of co-expressed P4HA1 genes. To create and visualize the PPI network for the P4HA1 protein, Cytoscape v3.5.1 was employed. The PPI network's primary nodes were then grouped according to the enrichment of the KEGG Pathway. Hub nodes with a higher degree would be in phase to delineate their significant role in the BC progression.

## Pathway Enrichment Analysis

We conducted pathway enrichment (GO and KEGG) using g:Profiler[7] to explore the function of P4HA1 gene sets with

[4]http:/www.progenetix.net
[5]http://kmplot.com/analysis/
[6]http://thebiogrid.org/
[7]http:/biit.cs.ut.ee/gprofiler/

biochemical, cellular, and molecular aspects (Raudvere et al., 2019). g:Profiler searches for a collection of the pathway, network, regulatory motif, and phenotype gene sets using a detailed set of accurate and concise annotation methods. The method also consolidates the exact Fisher test with an input gene list and *p*-value enrichment for each pathway. Using a threshold of 0.05, the g:Profiler computes the *p*-values from GO and KEGG route enrichment analysis.

## Statistical Analysis

The transcriptomic cohort analysis was performed using the R programming environment (version 3.2.5) with the criteria of *p*-value $< 0.05$. Survival analysis was conducted jointly with Kaplan–Meier plots and COX Proportional hazard model. The Kaplan–Meier curves were used to assess the overall survival, relapse-free survival, and distance metastasis-free survival associated with the P4HA1 gene expression. The univariate and multivariate Cox proportional model was carried to analyze the association of P4HA1 with the clinicopathologic variants of breast cancer and estimate the hazard ratio and 95% CIs. Logistic regression analysis was carried out in GSE22133 data to explore the association of P4HA1 gene expression with the clinicopathologic variants of breast cancer: ER, PR, and Grade. It estimates the breast cancer risk by examining the odds ratios (ORs) and 95% confident intervals (CIs), and *p*-value. The two-tailed *p*-values below 0.05 were considered statistically significant.

# RESULTS

## P4HA1 Expression in BC Under Hypoxic Condition

A detailed description of the transcriptomic data used in this study was given in **Table 1**. An integrative analysis of these cohorts identified a high-expression P4HA1 gene with a p-threshold criterion of $<0.05$ and FC $> 2$ in the two datasets. Moreover, P4HA1 was remarkably increased during the different time (6 h, 9 h, 16 h, 32 h, and 48 h) of the hypoxic state. The Violin Plot revealed the difference between with and without hypoxic exposure in breast cancer cells in the mRNA expression of P4HA1 (**Figures 2A,B**).

## Transcriptional Expression of P4HA1 in the Clinical Regimen of BC

A differential transcriptional level of P4HA1 between BC and paired normal breast tissue was evaluated by the UALCAN database to determine the mRNA expression of P4HA1 in BC patients. As illustrated in **Figure 2**, the transcriptional level of P4HA1 was substantially up-regulated in BC tissues (**Figure 2C**, $p \leq$ 1E-12) compared to normal tissues. Subsequently, P4HA1 differential transcriptional levels were compared for the molecular and histological subtypes, tumor grades, and other BC patient factors. Box plots were made to visualize the association between the expression levels of the clinicopathologic condition of BC. As shown in **Figure 2**, the level of P4HA1 was significantly associated with the intrinsic subclasses of the BC. Patients with Luminal ($p \leq$ 1E-12) and triple-negative subclasses ($p$ = 1.35059E-10) tend to express a higher P4HA1, than HER2-positive ($p$ = 1.9099E-05). The highest mRNA expressions of P4HA1 were found sequentially in the various stages of the BC, Stage 1 ($p$ = 8.8817E-16) <Stage 3 ($p$ = 1.670441E-12) <Stage 2 ($p$ = 1.62447E-12) <Stage 4 ($p$ = 1.31617E-03) (**Figure 2D**), and the highest mRNA expressions of P4HA1 were similarly found in-between the age group of 40–80 ($p$ = 1.62447E-12) and marginally lower in age <80 ($p$ = 3.9105E-08) than the >40 ($p$ = 6.3915E-04) age group (**Figure 2F**). Interestingly, P4HA1 expression was analyzed with the metastatic lymph node classification and elevated level of expression in N1 ($p$ = 1.62436E-12) than N0 ($p \leq$ 1E-12) <N2 ($p$ = 6.06390E-09) <N3 ($p$ = 2.31799E-07) (**Figure 2G**). Together, the results showed a positive association between P4HA1 transcriptional levels and typical subclasses in BC patients.

Oncomine analysis of malignant breast tissue relative to normal tissue analysis showed altered expression of P4HA1 in different transcriptomic profiles (**Figure 3**). In the Curtis data set, the P4HA1 mRNA rate was significantly higher in the breast tumor (FC = −1.570, $p$ = 4.72E-5). Furthermore, in invasive breast carcinoma, there was a substantial rise in mRNA levels of P4HA1 (FC = 1.219, $p$ = 5.25E-6). Moreover, P4HA1 was up-regulated in the Gluck (FC = 1.641, $p$ = 0.015) and Zhao (FC = 1.598, $p$ = 0.048) datasets.

## P4HA1 Genomic Alteration in BC

With genome-wide copy number profiles in the Progenetix database, we investigated the prevalent genomic amplification of the P4HA1 chromosomal region in BC. We focused on the use of the TCGA-BC cohort and obtained a recurring functional copy number gain for chromosome 10q22.1 (location P4HA1) (**Figure 3E**). Since this is the typical copy number peaks in cancers, it can aid BC's development and metastatic niche.

## Prognostic Significance of P4HA1 in BC

To evaluate the clinical significance of P4HA1 with BC, we analyzed the patient's survival index through the Kaplan–Meier plotter and UALCAN (**Figure 4**). The regulation of P4HA1 significantly contributes to the worst prognostic in BC patients. OS was significantly shorter in patients with elevated P4HA1 (HR = 1.35; 95% CI: 1.09–1.67; $p < 0.0059$) (**Figures 4A**, **5B**) compared to low P4HA1 expression. Moreover, the higher

**TABLE 1 |** Characteristics of transcriptomic data from Gene Expression Omnibus.

| GEO ID | Platform Acc. | Platform | Cell line | Time period of hypoxia (1% $O_2$) | Year |
|---|---|---|---|---|---|
| *GSE27813* | GPL10558 | Illumina Human HT-12 v4.0 Expression BeadChip | MCF-7 | 6 h, 9 h | 2011 |
| *GSE47533* | GPL6884 | Illumina HumanWG-6 v3.0 Expression BeadChip | MCF-7 | 16 h, 32 h, 48 h | 2014 |

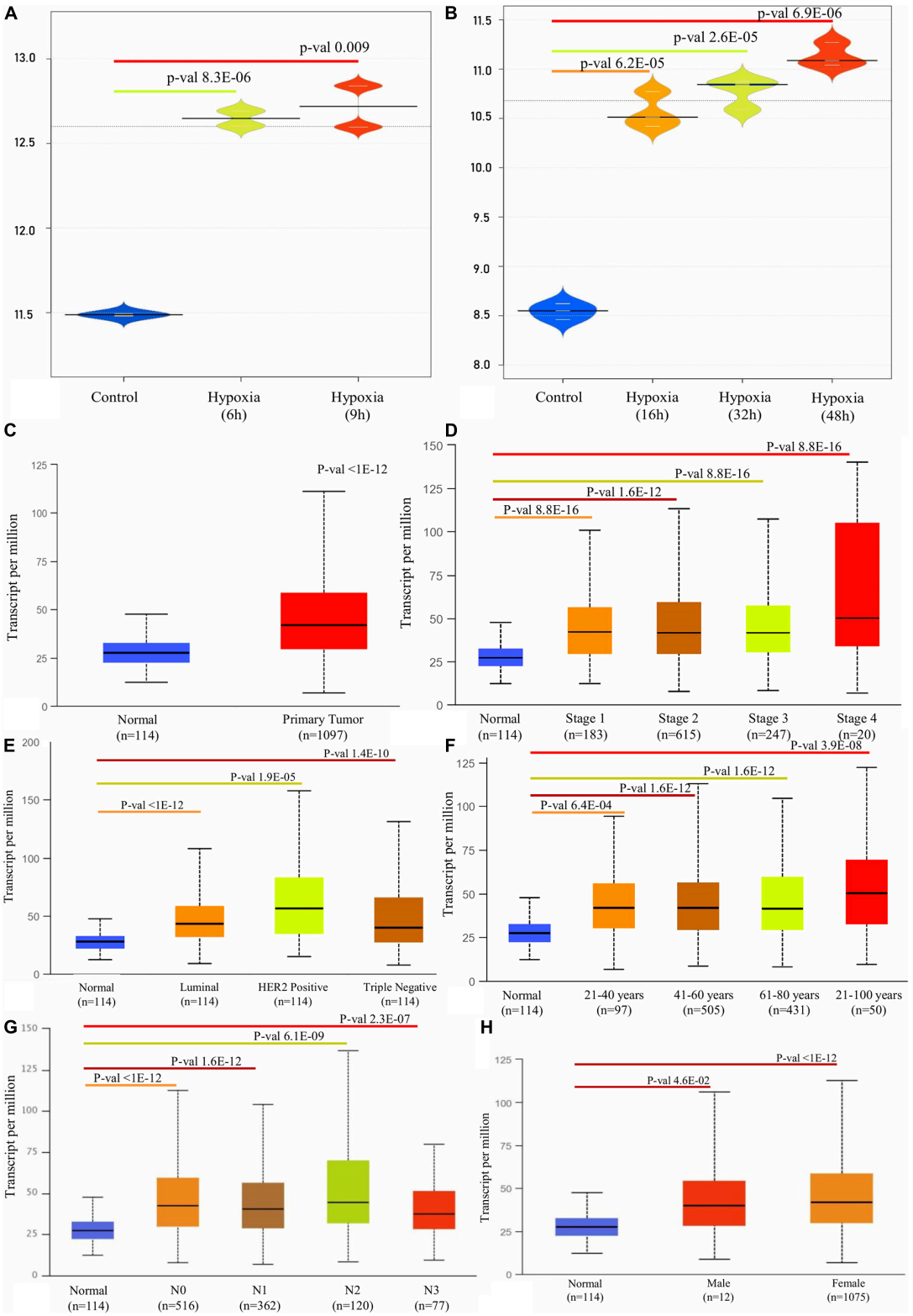



**FIGURE 2 |** Box plot representation of P4HA1 gene expression compared with a normal and different time period of hypoxic exposure in BC cells: cut-off *p*-value < 0.05. **(A)** GSE27813 and **(B)** GSE47533. **(C–H)** Box plot showing relative expression of P4HA1 in clinicopathologic of Breast Cancer, **(C)** Normal and Primary Tumor samples, **(D)** Normal and patients in Stages 1, 2, 3, and 4, **(E)** Normal and Subclass, **(F)** Normal and Age group, **(G)** Normal and Nodal subclass, and **(H)** Normal and Gender.

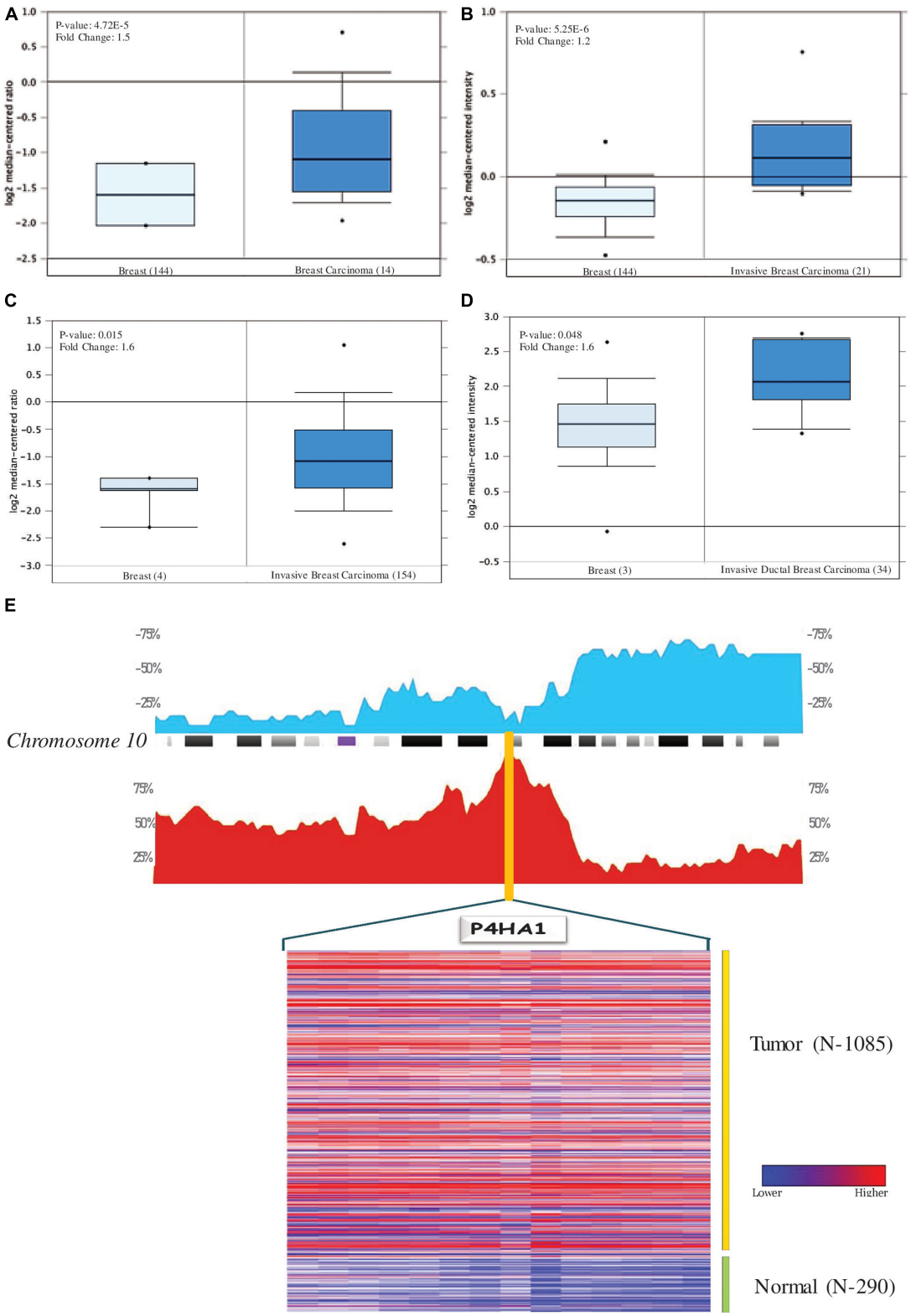



**FIGURE 3 |** Levels of P4HA1 mRNA expression in BC compared to normal cells. Figures generated based on Oncomine analysis with criteria fold-change and *p*-values. **(A)** Zhao Breast, **(B)** Curtis Breast, **(C)** Gluck Breast, and **(D)** Curtis Breast Dataset. **(E)** The distribution of Copy number variation of schematic physical map of Chromosome 10 (human genome 19 assembly (GRCh37) for TCGA Breast carcinoma generated from Progenetix tool. Heat map representation of P4HA1 between the normal and breast cancer patients – TCGA data. The color ratio red to green represents the change from high to low.

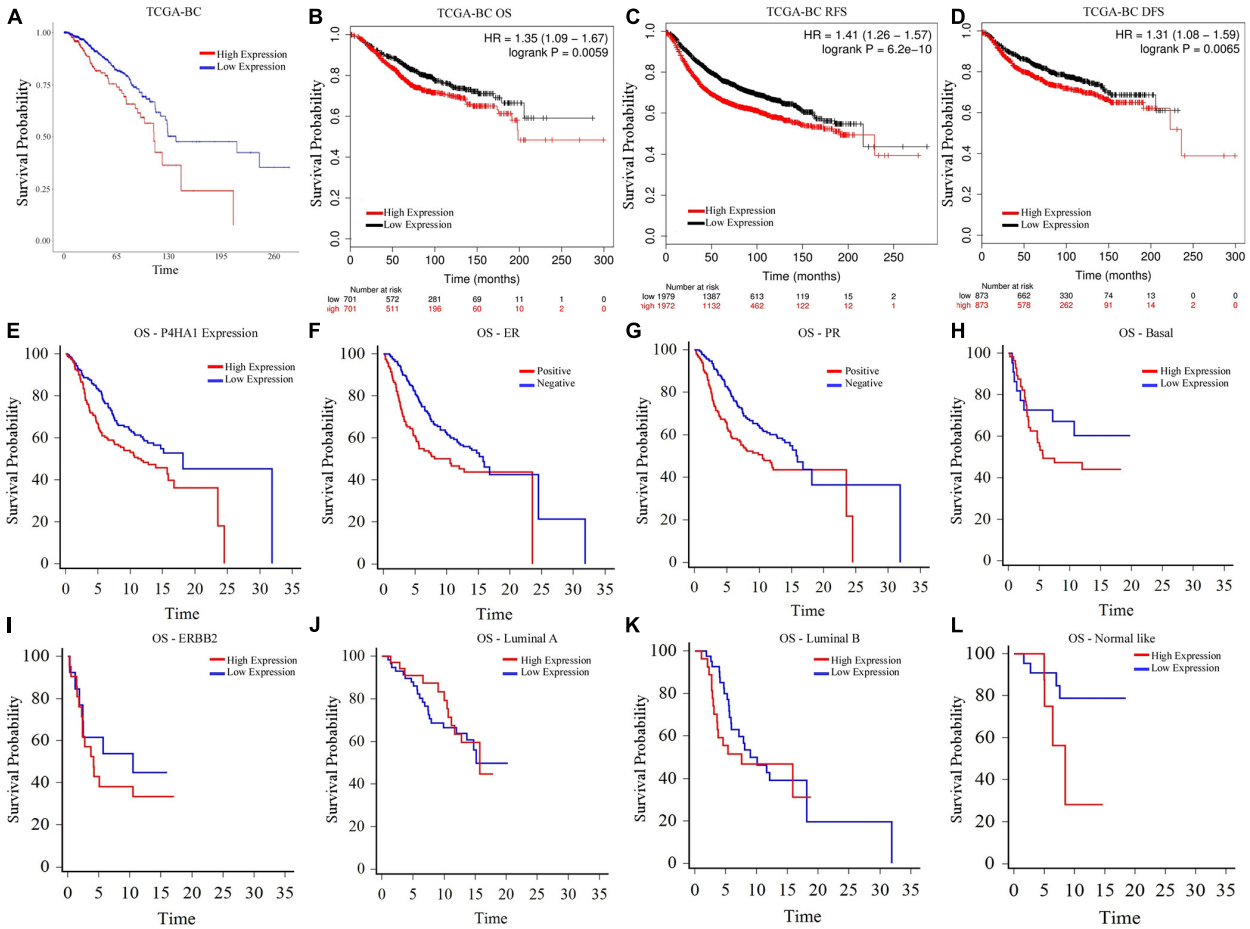



**FIGURE 4 |** Prognostic Index of Breast cancer as determined by Kaplan–Meier estimates **(A)** Overall Survival (UALCAN), **(B)** Overall Survival (Kaplan–Meier Plotter), **(C)** Relapse-free survival (Kaplan–Meier Plotter), **(D)** Distant free metastasis survival (Kaplan–Meier Plotter), **(E)** Overall Survival-P4HA1 Expression (GSE22133), **(F)** Overall Survival-Estrogen Receptor (GSE22133), **(G)** Overall Survival-Progesterone Receptor (GSE22133), **(H)** Overall Survival-Basal (GSE22133), **(I)** Overall Survival-ERBB2 (GSE22133), **(J)** Overall Survival-Luminal A (GSE22133), **(K)** Overall Survival-Luminal B (GSE22133), and **(L)** Overall Survival-Normal-like (GSE22133).

expression of P4HA1 indicated poor RFS (HR = 1.41; 95% CI: 1.26–1.57; $p$ < 6.2E-10) (**Figure 4C**) and DMFS (HR = 1.31; 95% CI: 1.08–1.59; $p$ < 0.0065) (**Figure 4D**). These findings show that P4HA1 is critically associated with a poorer clinical regimen in BC patients.

A univariate and multivariate regression analysis of Cox hazard regression using GSE22133 data was explored to verify the prognostic index of P4HA1. The association risk was estimated with the clinicopathologic covariates, including estrogen receptor (ER), progesterone receptor (PR), histological subtypes, and grades. **Table 2** shows how the P4HA1 gene is associated with clinical factors. Univariate Cox regression analysis indicated a significant association with hormonal receptor ER ($p$ = 0.0042, HR = 0.62, 95% CI = 0.46-0.86), PR ($p$ = 0.0043, HR = 0.63, 95% CI = 0.46-0.86), and Grade ($p$ = 0.051, HR = 1.21, 95% CI = 0.99-1.48) in GSE22133 data. In addition, multivariate Cox analysis found no strong association between histological subtypes and hormone receptors. Each clinical factor's survival plot was depicted in **Figures 4E–L**. These results indicate that the P4HA1 expression strongly attributes to the hormonal receptor, ER, and PR.

**Table 3** shows the logistic regression analysis of the association between the P4HA1 expression and clinicopathologic variants of breast cancer (ER, PR, and Grade). The expression of P4HA1 was significantly associated with the ER status group of breast cancer (OR = 0.38; 95% CI: 0.79–0.80, $P$ = 0.011) but less significantly associated with the PR status group cancer (OR = 1.47; 95% CI: 0.71–3.03, $P$ = 0.29). We assessed the association of P4HA1 expression with breast cancer grade through combing grade 1 and grade 2 vs. grade 3 and results revealed no significance associated with grades (OR = 1.40; 95% CI: 0.76–2.57, $P$ = 0.27). In addition, this analysis also revealed a strong association of P4HA1 gene expression with the ER of breast cancer.

## Biological Interaction of P4HA1

Gene Ontology (GO) analysis was carried out against using P4HA1 and its associated genes generated from the BioGRID source. We applied a hypergeometric test for each enriched

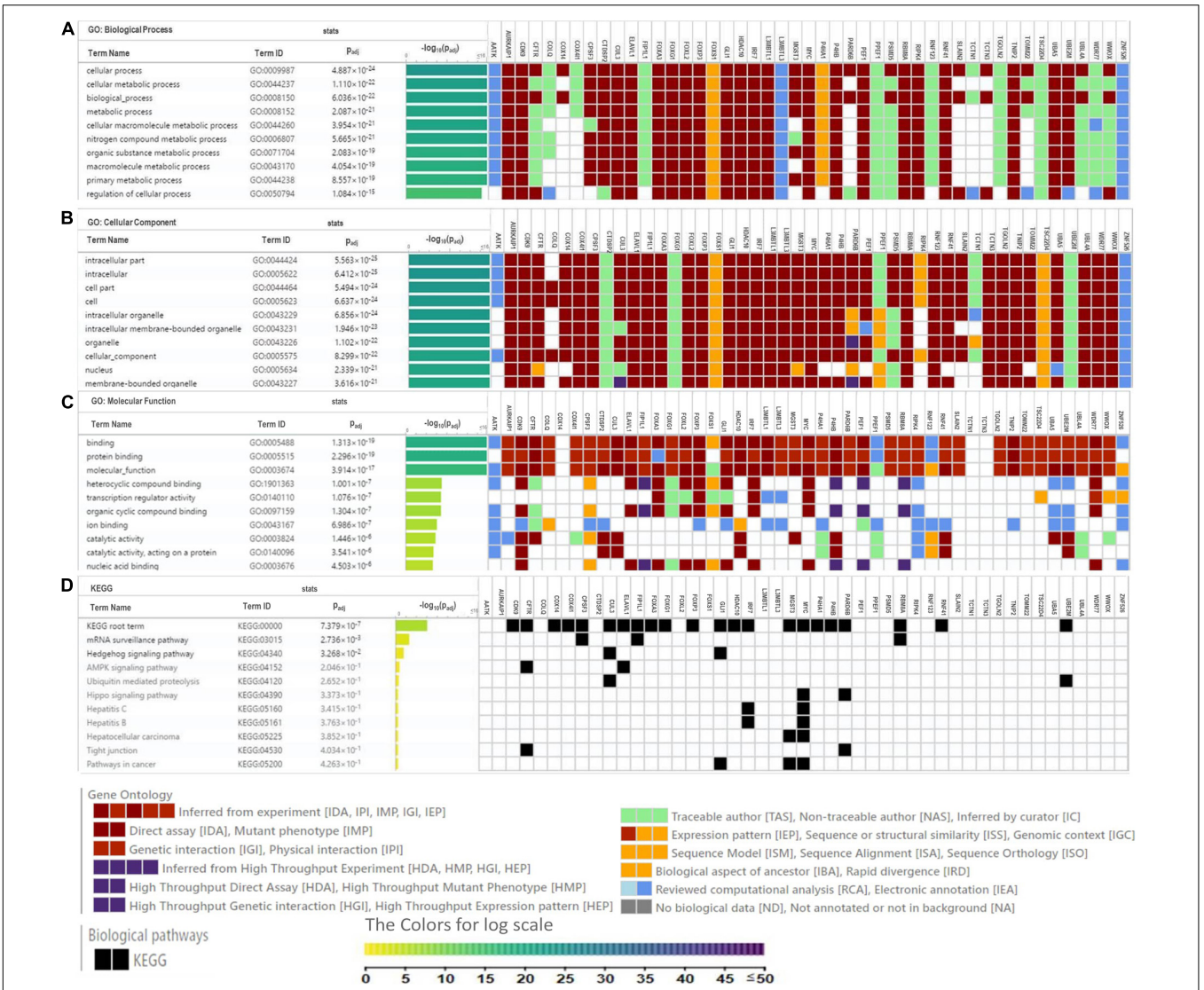



**FIGURE 5 |** Heat map depicts the associations of P4HA1 signature with GO term and KEGG generated by g:Profiler. **(A)** GO-Biological Process, **(B)** GO-Cellular Component, **(C)** GO-Molecular Function, and **(D)** KEGG pathway.

GO term, with a threshold lower than 0.05 in the g:Profiler tool: (**Figure 5**). Under the GO hierarchy, the ontology of highly enriched biological processes was "Cellular Process" (GO:0009987), "Cellular Metabolic Process" (GO:0044237). In cellular ontology, the enriched terms were "intracellular part" (GO:0044424) and, similarly, with the ontological molecular function "Binding" (GO:0005488), were highly enriched (**Table 4**). Apart from the significant enrichment of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway terms were the "mRNA surveillance pathway," "Hedgehog signaling pathway," and "AMPK signaling pathway." The full enrichment analysis output is listed in **Supplementary Table 1** (GO) and **Supplementary Table 2** (KEGG). Most critically, many of these genes are associated with cellular metabolic shift and oncogenic signaling pathways, a process intimately linked with invasion and proliferation.

## Protein Interaction Network of P4HA1

We constructed a P4HA1 mRNA interaction network generated from the BioGRID database. The final PPI network generated by Cytoscape consisted of 59 nodes and 382 interactions (**Supplementary Table 3**). Each signaling pathway's proteins were colored based on the KEGG enrichment (**Figure 6**).

## DISCUSSION

Breast cancer heterogeneity is still one of the most frequent causes of cancer mortality (Lin L. F. et al., 2019). Despite multimodal care for patients, the hypoxic condition is a critical factor that affects the treatment strategy and the clinical regimen (Tong et al., 2018). The knowledge in genotypical and their profound mechanisms will also advance the effective

**TABLE 2 |** Univariate and multivariate analysis of clinicopathological factors associated with the prognostic significance of P4HA1 expression in breast cancer.

| Clinical factors | GSE22133 | | | | | |
|---|---|---|---|---|---|---|
| | Univariate | | | Multivariate | | |
| | *p*-value | HR | CI (95%) | *p*-value | HR | CI (95%) |
| **P4HA1 expression** | 0.0097 | 1.51 | 1.10–2.07 | 0.1693 | 1.34 | 0.88–2.04 |
| **ER status** Positive vs. Negative | 0.0042 | 0.62 | 0.45–0.86 | 0.8055 | 0.92 | 0.51–1.66 |
| **PR status** Positive vs. Negative | 0.0043 | 0.63 | 0.46–0.86 | 0.5944 | 0.86 | 0.49–1.48 |
| **Grade** 1 and 2 vs. 3 | 0.0531 | 1.21 | 0.99–1.48 | 0.3255 | 1.12 | 0.88–1.43 |
| **Histological subtypes** | | | | | NA | |
| Basal | 0.2847 | 1.53 | 0.70–3.34 | | | |
| ERBB2 | 0.4776 | 1.39 | 0.56–3.45 | | | |
| Luminal A | 0.7437 | 0.88 | 0.43–1.79 | | | |
| Luminal B | 0.4840 | 1.26 | 0.65–2.44 | | | |
| Normal-Like | 0.0880 | 3.41 | 0.83–14.00 | | | |

*ER, Estrogen Receptor; PR, Progesterone Receptor; CI, Confidence Interval.*

therapeutic stratification of BC. Microarray and next-generation (NGS) sequencing methods have recently been used for early detection and personalized treatment (Wang et al., 2009; Marzancola et al., 2016). Such diverse data offers an outstanding opportunity to discuss more concerns relevant to tumor heterogeneity. A compendium of an integrative functional approach was systematically proposed to explore the P4HA1 gene fundamentally associated with hypoxia-induced BC. To delineate the processes involved in carcinogenesis, the reliability of this analysis was validated in terms of expression, clinical subtypes, copy number variation, and altered pathways in the clinical TCGA-BC cohort. Therefore this analysis merged transcriptional activities with molecular signaling pathways to underpin the proliferation of hypoxic-mediated BC.

Our findings revealed that P4HA1 gene expression is reliably expressed in breast cancer vs. normal cells. It was consistently noted in BC subclasses, that in patients with Luminal, triple-negative, and lymph node (N1), P4HA1 was overexpressed but comparatively lower in the positive HER2 group and P4HA1 was prominent in Stage I compared to the other BC stages. Overexpression of P4HA1 has previously been seen in TNBC-BC (Xiong et al., 2018), head and neck squamous cell carcinomas (HNSCC) (Li et al., 2019), prostate (Wolf et al., 2004), melanoma (Atkinson et al., 2019), and gastric cancer (Cheng et al., 2012). Importantly, our study showed that overexpression of P4HA1 could be associated with tumor progression, invasion and thus act as a diagnostic biomarker of BC.

A distinctive molecular mechanism explains the strong association between CNV and differential expression of P4HA1. We observed that the P4HA1-10q22.1 copy number showed a high-level positive amplification in the patient data for

**TABLE 3 |** Logistic regression analysis of associations between P4HA1 expression and the clinicopathologic variants of breast cancer.

| Variable | | Size | *P*-value | Odds ratio | 95% CI |
|---|---|---|---|---|---|
| ER | Pos (173) vs. Neg (173) | 346 | 0.0118 | 0.3811 | 0.1799 to 0.8076 |
| PR | Pos (172) vs. Neg (171) | 343 | 0.2964 | 1.4718 | 0.7126 to 3.0399 |
| Grade | 1 and 2 (116) vs. 3 (116) | 232 | 0.2730 | 1.4035 | 0.7656 to 2.5731 |

*ER, Estrogen Receptor; PR, Progesterone Receptor; CI, Confidence Interval.*

**TABLE 4 |** Functional enrichment pathway analysis: Top enriched terms of gene ontology-biological process, cellular component, molecular function, and KEGG pathways.

| Source | Term Id | Term name | *p*-value |
|---|---|---|---|
| ***Gene ontology-biological process*** | | | |
| *GO:BP* | GO:0009987 | cellular process | 4.89E-24 |
| *GO:BP* | GO:0044237 | cellular metabolic process | 1.11E-22 |
| *GO:BP* | GO:0008150 | biological_process | 6.04E-22 |
| *GO:BP* | GO:0008152 | metabolic process | 2.09E-21 |
| *GO:BP* | GO:0044260 | cellular macromolecule metabolic process | 3.95E-21 |
| ***Gene ontology-cellular component*** | | | |
| *GO:CC* | GO:0044424 | intracellular part | 5.56E-25 |
| *GO:CC* | GO:0005622 | intracellular | 6.41E-25 |
| *GO:CC* | GO:0044464 | cell part | 5.49E-24 |
| *GO:CC* | GO:0005623 | cell | 6.64E-24 |
| *GO:CC* | GO:0043229 | intracellular organelle | 6.86E-24 |
| ***Gene ontology-molecular function*** | | | |
| *GO:MF* | GO:0005488 | binding | 1.31E-19 |
| *GO:MF* | GO:0005515 | protein binding | 2.30E-19 |
| *GO:MF* | GO:0003674 | molecular_function | 3.91E-17 |
| *GO:MF* | GO:1901363 | heterocyclic compound binding | 1.00E-07 |
| *GO:MF* | GO:0140110 | transcription regulator activity | 1.08E-07 |
| ***KEGG*** | | | |
| *KEGG* | KEGG:03015 | mRNA surveillance pathway | 0.002436223 |
| *KEGG* | KEGG:04340 | Hedgehog signaling pathway | 0.002867797 |
| *KEGG* | KEGG:04152 | AMPK signaling pathway | 0.002946176 |
| *KEGG* | KEGG:04120 | Ubiquitin mediated proteolysis | 0.003652434 |

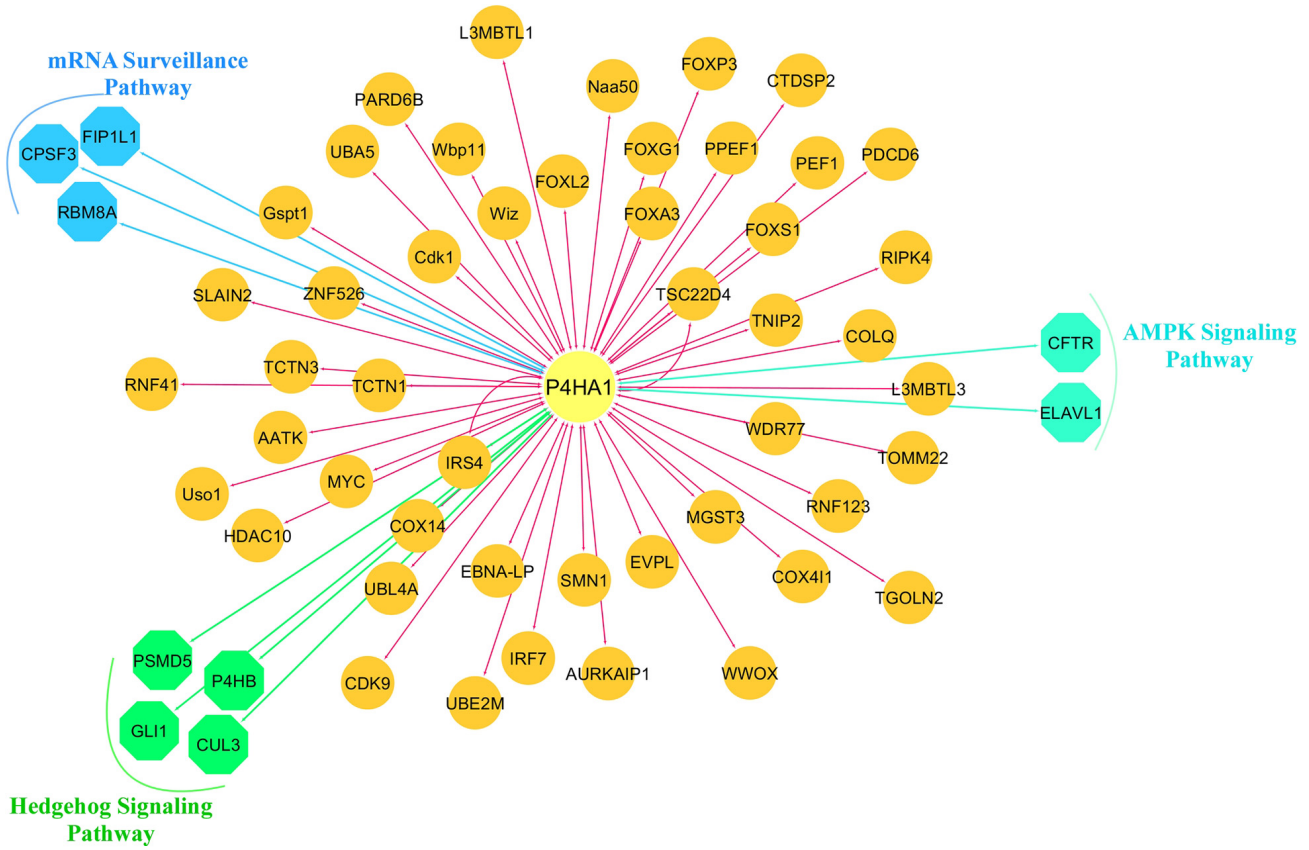



**FIGURE 6 |** Protein-protein interaction network of P4HA1. Highly interacted protein network generated from the BioGRID source. Different colors of the network edge indicate functional enrichment with the KEGG database.

TCGA-BC, suggesting its effect on the high mRNA transcription level. Moreover, the association in the elevated amplicon 10q22 was reported to have a remarkable role in tumorigenesis and weak prognostic significance in patients with prostate cancer (Wolf et al., 2004), gastric cancer (Cheng et al., 2012), glioma (Hu et al., 2017), melanoma (Atkinson et al., 2019), oral squamous cell carcinoma (Kappler et al., 2017), and HNSCC (Li et al., 2019). In line with previous studies, higher P4HA1 expression was also directly related to BC patients' poor survival and could be accomplished as a prognostic predictor.

Functional enrichment analysis of gene ontology revealed that genes were mostly involved in different cellular metabolic processes. Most frequently, by increased glycolytic flux and suppressed oxidative phosphorylation (Warburg effect), tumor cells adapt their resources to cope with high energy demands. Thus, the hypoxic state acquires energy via the hypoxic receptive elements (HRE) through the metabolic shift and tumor microenvironment (Dillekas et al., 2019). Under physiological oxygen concentrations, Prolyl hydroxylase (PHD1-3) enzymes strengthen the stability of HIF1 and HREs. Previous studies have shown that PHD enzymes involved in HRE's regulatory network in gastric cancer and PHD inhibition contribute to reduced tumor development under hypoxic conditions (Cheng et al., 2012). Interestingly, the presence of PHDs is closely related to tumor angiogenesis and metastasis during hypoxic cell proliferation.

We observed that the F gene and the RBM8A gene were closely associated with an mRNA surveillance pathway in the KEGG pathway enrichment. The Cleavage polyadenylation specificity factor (CPSF) is a multi-subunit that actively participates through the cleavage and polyadenylation of mRNA activation in the eukaryotic pre-messenger (m)RNA 3′-end process (Casanal et al., 2017). Importantly, these CPSF factors lead to the growth of human cancer, such as breast (Erson-Bensan and Can, 2016), ovarian cancer (Zhang et al., 2017), and even the inhibition of CPSF3 actuates apoptosis in prostate cancer cells (Van Etten et al., 2017). Interestingly, CPCF3 and CPCF4 were a major component of the OS and RFS based CPSF complex in non-small lung cancer (Ning et al., 2019).

RNA binding motif protein 8A (RBM8A), also known as Y14, is an essential factor in exon junction complex (EJC), translation, chromatin remodeling, damage checkpoints, regulation of apoptosis (Gerstberger et al., 2014), and deregulation contribute to cancer pathologies and cardiovascular diseases (Wurth and Gebauer, 2015). RBM8A up-regulation is critically involved in modulating apoptosis, and tumor proliferation and metastasis (Lu et al., 2017). Cell growth was blocked in RBM8A knockout cells, and the G2/M step of the cell cycle was arrested in

lung adenocarcinoma cells (Ishigaki et al., 2013). In addition, for individuals with hepatocellular carcinoma, elevated RBM8A expression was associated with poor prognosis and progression-free survival. RBM8A tends to be active in the EMT transition, an important occurrence in the metastatic niche (Lin Y. et al., 2019).

Hedgehog signaling (Hh) plays a vital role in embryonic cellular differentiation, and its alteration has oncogenic functions in initiating and progression (Sari et al., 2018; Chang and Lai, 2019). One of the downstream regulators of the Hh route was the Cullin gene. Cullin 3 proteins are active in cell cycle regulation and redox homeostasis, protein trafficking, and stress responses (Chen and Chen, 2016). Interestingly, CUL3 up-regulation is associated with an acquired carcinogenic state and oxidative stress in BC (Loignon et al., 2009). Recent evidence indicates that Cullin-dependent ubiquitin ligases play a crucial role in breast carcinogenesis and squamous cell carcinoma of the esophagus (Hu et al., 2018).

Glioma-associated oncogene transcription factors (GLI) is a Zinc finger protein and downstream regulator of the Hh pathway (Pietrobono et al., 2019). In early embryonic development, GLI members play a major role in the central nervous system; however, it is also involved in carcinogenesis and metastatic cascade niche (Niewiadomski et al., 2019). Since amplified GLI was first observed in glioblastoma, it has now been commonly detected in the breast (Song et al., 2016), lung (Panneerselvam et al., 2019), pancreatic (Kowolik et al., 2019), colorectal (Park et al., 2019), leukemia (Jetten, 2019), and renal cell carcinoma (Kotulak-Chrzaszcz et al., 2019). It was also stated that high-expression GLI prevails tumor suppression mediated by p53 (Abe et al., 2008). Silencing GLI decreases cancer cell proliferation and invasive potency (Mishra et al., 2019). These results indicate a mechanism of Hh signaling to stimulate malignant stemming and facilitate the growth of tumors.

## CONCLUSION

This study used robust multiple transcriptomic cohorts with an integrated omic analysis and found that P4HA1 may be a potential oncogenic biomarker in BC. Moreover, this gene showed a copy number gain, reliably more explicit in high-grade metastatic breast tumors with poor clinical patient results. Besides, we speculate the implication of the hedgehog signaling pathway and metabolic reprogramming during high cell proliferation in hypoxic breast tumors. Our studies have provided useful insights into the P4HA1; it can be a novel biomarker for the diagnosis and progression of BC therapy.

## DATA AVAILABILITY STATEMENT

This study was carried out on publicly available data on Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with accession numbers: GSE27813, GSE47533, and GSE22133.

## AUTHOR CONTRIBUTIONS

MM and KP conceived and designed the study. MM performed the integrated analysis, acquired the data, and drafted the manuscript. KP assisted with reviewing and editing the manuscript. Both authors approved the final manuscript for publication.

## ACKNOWLEDGMENTS

MM gratefully acknowledges the Indian Council of Medical Research, New Delhi, for sanctioning Senior Research Fellowship (ICMR SRF: 5/3/8/26/ITR-F/2018-ITR).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.632626/full#supplementary-material

## REFERENCES

Abe, Y., Oda-Sato, E., Tobiume, K., Kawauchi, K., Taya, Y., Okamoto, K., et al. (2008). Hedgehog signaling overrides p53-mediated tumor suppression by activating mdm2. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4838–4843. doi: 10.1073/pnas.0712216105

Al Tameemi, W., Dale, T. P., Al-Jumaily, R. M. K., and Forsyth, N. R. (2019). Hypoxia-modified cancer cell metabolism. *Front. Cell Dev. Biol.* 7:4. doi: 10.3389/fcell.2019.00004

Atkinson, A., Renziehausen, A., Wang, H. X., Lo Nigro, C., Lattanzio, L., Merlano, M., et al. (2019). Collagen prolyl hydroxylases are bifunctional growth regulators in melanoma. *J. Invest. Dermatol.* 139, 1118–1126. doi: 10.1016/j.jid.2018.10.038

Baba, Y., Nosho, K., Shima, K., Irahara, N., Chan, A. T., Meyerhardt, J. A., et al. (2010). HIF1A overexpression is associated with poor prognosis in a cohort of 731 colorectal cancers. *Am. J. Pathol.* 176, 2292–2301. doi: 10.2353/ajpath.2010.090972

Baudis, M., and Cleary, M. L. (2001). Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics* 17, 1228–1229. doi: 10.1093/bioinformatics/17.12.1228

Campbell, E. J., Dachs, G. U., Morrin, H. R., Davey, V. C., Robinson, B. A., and Vissers, M. C. M. (2019). Activation of the hypoxia pathway in breast cancer tissue and patient survival are inversely associated with tumor ascorbate levels. *BMC Cancer* 19:307. doi: 10.1186/s12885-019-5503-x

Casanal, A., Kumar, A., Hill, C. H., Easter, A. D., Emsley, P., Degliesposti, G., et al. (2017). Architecture of eukaryotic mRNA 3'-end processing machinery. *Science* 358, 1056–1059. doi: 10.1126/science.aao6535

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Chang, W. H., and Lai, A. G. (2019). Aberrations in notch-hedgehog signalling reveal cancer stem cells harbouring conserved oncogenic properties associated with hypoxia and immunoevasion. *Br. J. Cancer* 121, 666–678. doi: 10.1038/s41416-019-0572-9

Chen, H. Y., and Chen, R. H. (2016). Cullin 3 ubiquitin ligases in cancer biology: functions and therapeutic implications. *Front. Oncol.* 6:113. doi: 10.3389/fonc.2016.00113

Cheng, L., Wang, P., Yang, S., Yang, Y. Q., Zhang, Q., Zhang, W., et al. (2012). Identification of genes with a correlation between copy number and

expression in gastric cancer. *BMC Med. Genomics* 5:14. doi: 10.1186/1755-8794-5-14

Cioffi, C. L., Liu, X. Q., Kosinski, P. A., Garay, M., and Bowen, B. R. (2003). Differential regulation of HIF-1 alpha prolyl-4-hydroxylase genes by hypoxia in human cardiovascular cells. *Biochem. Biophys. Res. Commun.* 303, 947–953. doi: 10.1016/s0006-291x(03)00453-4

D'Aniello, C., Cermola, F., Palamidessi, A., Wanderlingh, L. G., Gagliardi, M., Migliaccio, A., et al. (2019). Collagen prolyl hydroxylation-dependent metabolic perturbation governs epigenetic remodeling and mesenchymal transition in pluripotent and cancer cells. *Cancer Res.* 79, 3235–3250. doi: 10.1158/0008-5472.Can-18-2070

Dillekas, H., Rogers, M. S., and Straume, O. (2019). Are 90% of deaths from cancer caused by metastases? *Cancer Med.* 8, 5574–5576. doi: 10.1002/cam4.2474

Erson-Bensan, A. E., and Can, T. (2016). Alternative polyadenylation: another foe in cancer. *Mol. Cancer Res.* 14, 507–517. doi: 10.1158/1541-7786.Mcr-15-0489

Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., et al. (2018). Breast cancer development and progression: risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes Dis.* 5, 77–106. doi: 10.1016/j.gendis.2018.05.001

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat. Rev. Genet.* 15, 829–845. doi: 10.1038/nrg3813

Gorres, K. L., and Raines, R. T. (2010). Prolyl 4-hydroxylase. *Crit. Rev. Biochem. Mol. Biol.* 45, 106–124. doi: 10.3109/10409231003627991

Greer, S. N., Metcalf, J. L., Wang, Y., and Ohh, M. (2012). The updated biology of hypoxia-inducible factor. *EMBO J.* 31, 2448–2460. doi: 10.1038/emboj.2012.125

Gyorffy, B., and Schafer, R. (2009). Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. *Breast Cancer Res. Treat.* 118, 433–441. doi: 10.1007/s10549-008-0242-8

Ha, M., Moon, H., Choi, D., Kang, W., Kim, J. H., Lee, K. J., et al. (2019). Prognostic role of TMED3 in clear cell renal cell carcinoma: a retrospective multi-cohort analysis. *Front. Genet.* 10:355. doi: 10.3389/fgene.2019.00355

Haynes, B., Sarma, A., Nangia-Makker, P., and Shekhar, M. P. (2017). Breast cancer complexity: implications of intratumoral heterogeneity in clinical management. *Cancer Metastasis Rev.* 36, 547–555. doi: 10.1007/s10555-017-9684-y

Hu, J. L., Hu, X. L., Lu, C. X., Chen, X. J., Fu, L., Han, Q., et al. (2018). Variants in the 3 '-untranslated region of CUL3 is associated with risk of esophageal squamous cell carcinoma. *J. Cancer* 9, 3647–3650. doi: 10.7150/jca.27052

Hu, W. M., Zhang, J., Sun, S. X., Xi, S. Y., Chen, Z. J., Jiang, X. B., et al. (2017). Identification of P4HA1 as a prognostic biomarker for high-grade gliomas. *Pathol. Res. Pract.* 213, 1365–1369. doi: 10.1016/j.prp.2017.09.017

Ishigaki, Y., Nakamura, Y., Tatsuno, T., Hashimoto, M., Shimasaki, T., Iwabuchi, K., et al. (2013). Depletion of RNA-binding protein RBM8A (Y14) causes cell cycle deficiency and apoptosis in human cells. *Exp. Biol. Med.* 238, 889–897. doi: 10.1177/1535370213494646

Jetten, A. M. (2019). Emerging roles of GLI-similar kruppel-like zinc finger transcription factors in leukemia and other cancers. *Trends Cancer* 5, 547–557. doi: 10.1016/j.trecan.2019.07.005

Kappler, M., Kotrba, J., Kaune, T., Bache, M., Rot, S., Bethmann, D., et al. (2017). P4HA1: a single-gene surrogate of hypoxia signatures in oral squamous cell carcinoma patients. *Strahlenther. Onkol.* 193:S84.

Koren, S., and Bentires-Alj, M. (2015). Breast tumor heterogeneity: source of fitness, hurdle for therapy. *Mol. Cell* 60, 537–546. doi: 10.1016/j.molcel.2015.10.031

Koski, M. K., Anantharajan, J., Kursula, P., Dhavala, P., Murthy, A. V., Bergmann, U., et al. (2017). Assembly of the elongated collagen prolyl 4-hydroxylase alpha(2)beta(2) heterotetramer around a central alpha(2) dimer. *Biochem. J.* 474, 751–769. doi: 10.1042/Bcj20161000

Kotulak-Chrzaszcz, A., Klacz, J., Matuszewski, M., Kmiec, Z., and Wierzbicki, P. M. (2019). Expression of the sonic hedgehog pathway components in clear cell renal cell carcinoma. *Oncol. Lett.* 18, 5801–5810. doi: 10.3892/ol.2019.10919

Kowolik, C. M., Lin, M., Xie, J., Overman, L. E., and Horne, D. A. (2019). Attenuation of hedgehog/GLI signaling by NT1721 extends survival in pancreatic cancer. *J. Exp. Clin. Cancer Res.* 38:431. doi: 10.1186/s13046-019-1445-z

Kukkola, L., Hieta, R., Kivirikko, K. I., and Myllyharju, J. (2003). Identification and characterization of a third human, rat, and mouse collagen prolyl 4-hydroxylase isoenzyme. *J. Biol. Chem.* 278, 47685–47693. doi: 10.1074/jbc.M306806200

Li, Q., Shen, Z. S., Wu, Z. H., Shen, Y., Deng, H. X., Zhou, C. C., et al. (2019). High P4HA1 expression is an independent prognostic factor for poor overall survival and recurrent-free survival in head and neck squamous cell carcinoma. *J. Clin. Lab. Anal.* 34:e23107. doi: 10.1002/jcla.23107

Lin, L. F., Yan, L., Liu, Y. L., Yuan, F., Li, H., and Ni, J. (2019). Incidence and death in 29 cancer groups in 2017 and trend analysis from 1990 to 2017 from the global burden of disease study. *J. Hematol. Oncol.* 12:96. doi: 10.1186/s13045-019-0783-9

Lin, Y., Liang, R., Qiu, Y. F., Lv, Y. F., Zhang, J. Y., Qin, G., et al. (2019). Expression and gene regulation network of RBM8A in hepatocellular carcinoma based on data mining. *Aging* 11, 423–447. doi: 10.18632/aging.101749

Loignon, M., Miao, W. M., Hu, L. G., Bier, A., Bismar, T. A., Scrivens, P. J., et al. (2009). Cul3 overexpression depletes Nrf2 in breast cancer and is associated with sensitivity to carcinogens, to oxidative stress, and to chemotherapy. *Mo.l Cancer Ther.* 8, 2432–2440. doi: 10.1158/1535-7163.Mct-08-1186

Lu, C. C., Lee, C. C., Tseng, C. T., and Tarn, W. Y. (2017). Y14 governs p53 expression and modulates DNA damage sensitivity. *Sci. Rep.* 7:45558. doi: 10.1038/srep45558

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., et al. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* 19, 286–302. doi: 10.1093/bib/bbw114

Marzancola, M. G., Sedighi, A., and Li, P. C. H. (2016). DNA microarray-based diagnostics. *Methods Mol. Biol.* 1368, 161–178. doi: 10.1007/978-1-4939-3136-1_12

Mishra, S., Bernal, C., Silvano, M., Anand, S., and Ruiz, I. A. A. (2019). The protein secretion modulator TMED9 drives CNIH4/TGFalpha/GLI signaling opposing TMED3-WNT-TCF to promote colon cancer metastases. *Oncogene* 38, 5817–5837. doi: 10.1038/s41388-019-0845-z

Murugesan, M., and Premkumar, K. (2018). Hypoxia stimulates microenvironment in human embryonic stem cell through inflammatory signalling: an integrative analysis. *Biochem. Bioph. Res. Commun.* 498, 437–444. doi: 10.1016/j.bbrc.2018.02.194

Niewiadomski, P., Niedziolka, S. M., Markiewicz, L., Uspienski, T., Baran, B., and Chojnowska, K. (2019). Gli proteins: regulation in development and cancer. *Cells* 8:147. doi: 10.3390/cells8020147

Ning, Y., Liu, W. X., Guan, X. Y., Xie, X. B., and Zhang, Y. J. (2019). CPSF3 is a promising prognostic biomarker and predicts recurrence of non-small cell lung cancer. *Oncol. Lett.* 18, 2835–2844. doi: 10.3892/ol.2019.10659

Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47, D529–D541. doi: 10.1093/nar/gky1079

Panneerselvam, J., Srivastava, A., Mehta, M., Chen, A., Zhao, Y. D., Munshi, A., et al. (2019). IL-24 inhibits lung cancer growth by suppressing GLI1 and inducing DNA damage. *Cancers (Basel)* 11:1879. doi: 10.3390/cancers11121879

Park, S. H., Jeong, S., Kim, B. R., Jeong, Y. A., Kim, J. L., Na, Y. J., et al. (2019). Activating CCT2 triggers Gli-1 activation during hypoxic condition in colorectal cancer. *Oncogene* 39, 136–150. doi: 10.1038/s41388-019-0972-6

Pietrobono, S., Gagliardi, S., and Stecca, B. (2019). Non-canonical hedgehog signaling pathway in cancer: activation of GLI transcription factors beyond smoothened. *Front. Genet.* 10:556. doi: 10.3389/fgene.2019.00556

Provenzano, P. P., Eliceiri, K. W., Campbell, J. M., Inman, D. R., White, J. G., and Keely, P. J. (2006). Collagen reorganization at the tumor-stromal interface facilitates local invasion. *BMC Med.* 4:38. doi: 10.1186/1741-7015-4-38

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. doi: 10.1093/nar/gkz369

Rhodes, D. R., Yu, J. J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6, 1–6. doi: 10.1016/S1476-5586(04)80047-2

Sari, I. N., Phi, L. T. H., Jun, N. Y., Wijaya, Y. T., Lee, S., and Kwon, H. Y. (2018). Hedgehog signaling in cancer: a prospective therapeutic target for eradicating cancer stem cells. *Cells* 7:208. doi: 10.3390/cells7110208

Shou, Y., Yang, L., Yang, Y., Zhu, X., Li, F., and Xu, J. (2020). Identification of signatures of prognosis prediction for melanoma using a hypoxia score. *Front. Genet.* 11:570530. doi: 10.3389/fgene.2020.570530

Siegel, M. B., He, X. P., Hoadley, K. A., Hoyle, A., Pearce, J. B., Garrett, A. L., et al. (2018). Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *J. Clin. Invest.* 128, 1371–1383. doi: 10.1172/Jci96153

Song, L., Wang, W., Liu, D., Zhao, Y., He, J., Wang, X., et al. (2016). Targeting of sonic hedgehog-Gli signaling: a potential therapeutic target for patients with breast cancer. *Oncol. Lett.* 12, 1027–1033. doi: 10.3892/ol.2016.4722

Tong, C. W. S., Wu, M. X., Cho, W. C. S., and To, K. K. W. (2018). Recent advances in the treatment of breast cancer. *Front. Oncol.* 8:227.

Van Etten, J. L., Nyquist, M., Li, Y. M., Yang, R. D., Ho, Y., Johnson, R., et al. (2017). Targeting a single alternative polyadenylation site coordinately blocks expression of androgen receptor mRNA splice variants in prostate cancer. *Cancer Res.* 77, 5228–5235. doi: 10.1158/0008-5472.Can-17-0320

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

Weinschenk, T., Gouttefangeas, C., Schirle, M., Obermayr, F., Walter, S., Schoor, O., et al. (2002). Integrated functional genomics approach for the design of patient-individual antitumor vaccines. *Cancer Res.* 62, 5818–5827.

Willam, C., Maxwell, P. H., Nichols, L., Lygate, C., Tian, Y. M., Bernhardt, W., et al. (2006). HIF prolyl hydroxylases in the rat; organ distribution and changes in expression following hypoxia and coronary artery ligation. *J. Mol. Cell Cardiol.* 41, 68–77. doi: 10.1016/j.yjmcc.2006.04.009

Wishart, A. L., Conner, S. J., Guarin, J. R., Fatherree, J. P., Peng, Y., McGinn, R. A., et al. (2020). Decellularized extracellular matrix scaffolds identify full-length collagen VI as a driver of breast cancer cell invasion in obesity and metastasis. *Sci. Adv.* 6:eabc3175. doi: 10.1126/sciadv.abc3175

Wolf, M., Mousses, S., Hautaniemi, S., Karhu, R., Huusko, P., Allinen, M., et al. (2004). High-resolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. *Neoplasia* 6, 240–247. doi: 10.1593/neo.03439

Wurth, L., and Gebauer, F. (2015). RNA-binding proteins, multifaceted translational regulators in cancer. *Biochim. Biophys. Acta* 1849, 881–886. doi: 10.1016/j.bbagrm.2014.10.001

Xiong, G., Stewart, R. L., Chen, J., Gao, T., Scott, T. L., Samayoa, L. M., et al. (2018). Collagen prolyl 4-hydroxylase 1 is essential for HIF-1 alpha stabilization and TNBC chemoresistance. *Nat. Commun.* 9:4456. doi: 10.1038/s41467-018-06893-9

Yang, H., Wang, Y., Zhang, Z., and Li, H. (2020). Identification of KIF18B as a hub candidate gene in the metastasis of clear cell renal cell carcinoma by weighted gene co-expression network analysis. *Front. Genet.* 11:905. doi: 10.3389/fgene.2020.00905

Zhang, B. G., Liu, Y., Liu, D. H., and Yang, L. (2017). Targeting cleavage and polyadenylation specific factor 1 via shRNA inhibits cell proliferation in human ovarian cancer. *J. Biosci.* 42, 417–425. doi: 10.1007/s12038-017-9701-x

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 May 2021
doi: 10.3389/fgene.2021.669928

# Elevated Expression of PDZD11 Is Associated With Poor Prognosis and Immune Infiltrates in Hepatocellular Carcinoma

*Yao Chen[1†], Haifeng Xie[2†], Ting Xie[3], Xunjun Yang[3,4], Yilin Pang[3*] and SongDao Ye[4*]*

*[1] Department of Pathology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China, [2] Hangzhou Traditional Chinese Medicine (TCM) Hospital Affiliated to Zhejiang Chinese Medical University, Hangzhou, China, [3] Zhejiang Provincial Key Laboratory of Medical Genetics, Key Laboratory of Laboratory Medicine, Ministry of Education, School of Laboratory Medicine and Life Sciences, Wenzhou Medical University, Wenzhou, China, [4] Department of Laboratory Medicine, The Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University, Wenzhou, China*

OPEN ACCESS

**Edited by:**
*Shaoli Das, National Institutes of Health (NIH), United States*

**Reviewed by:**
*Vishal Midya, Icahn School of Medicine at Mount Sinai, United States*
*Michael Poidinger, Murdoch Childrens Research Institute, Australia*

***Correspondence:**
*Yilin Pang*
*ylpang2010@126.com*
*SongDao Ye*
*yesd955022@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:** *19 February 2021*
**Accepted:** *30 April 2021*
**Published:** *21 May 2021*

**Citation:**
*Chen Y, Xie H, Xie T, Yang X, Pang Y and Ye S (2021) Elevated Expression of PDZD11 Is Associated With Poor Prognosis and Immune Infiltrates in Hepatocellular Carcinoma. Front. Genet. 12:669928. doi: 10.3389/fgene.2021.669928*

Epithelial cells are held together by tight and adherent junctions, which are destroyed by the activation of epithelial-to-mesenchymal transition (EMT). The PLEKHA7-PDZD11 complex has been reported to be important for epithelial cell adhesion and connecting tissues. However, there is no research regarding the expression and role of PDZD11 in liver hepatocellular carcinoma (LIHC) progression. Here, we analyzed *PDZD11* mRNA expression and its clinical results in LIHC patient RNA sequencing data based on different open databases. Furthermore, we examined differences in PDZD11 expression in LIHC tissues and cell lines using western blotting and real-time qPCR. These results are the first to report that the mRNA and protein levels of PDZD11 are significantly overexpressed in LIHC. Moreover, high expression of *PDZD11* was correlated with poor overall survival in patients with LIHC. Gene regulatory network analysis suggested that PDZD11 is mainly involved in copper ion homeostasis, proteasome, and oxidative phosphorylation pathways. Interestingly, we found that PDZD11 levels were positively correlated with the abundance of immune infiltrates. In particular, higher infiltration levels of $CD4^+$ T cells and macrophage subsets significantly affected LIHC patient prognosis. Taken together, these results demonstrate that PDZD11 could be a potential diagnostic and prognostic biomarker in LIHC.

**Keywords: PDZD11, hepatocellular carcinoma, prognostic biomarker, immune infiltrates, functional network analysis**

## INTRODUCTION

Liver hepatocellular carcinoma (LIHC) accounts for the most common form of primary liver cancers (Villanueva, 2019), with an increasing incidence, particularly in East Asia (Bray et al., 2018; Siegel et al., 2019). LIHC is currently the third leading cause of cancer-related death worldwide (Jiang et al., 2019). LIHC likely occurs in patients with underlying liver diseases since infection with the hepatitis B or C virus (HBV or HCV) and long-term intoxication with alcohol or aflatoxin are the leading risk factors for developing LIHC (Jemal et al., 2017; Villanueva, 2019). Due to the high rate of recurrence and metastasis, the 5-year overall rate of survival for LIHC is only 18%, making liver cancer the second-leading cause

of cancer deaths, after pancreatic cancer (Jemal et al., 2017). Although operative treatment may be effective in the early stage of LIHC, the 5-year survival rate after developing to later stage is only 50–70% [European Association for the Study of the Liver (EASL), 2018]. Therefore, it is important to further screen LIHC oncogenes to help identifying novel biomarkers, therapeutic targets and immune-related biomarkers, and ultimately contribute to better diagnosis and prognosis of LIHC.

The human *PDZD11* gene is located at chromosome Xq13.1 and is 3.92 kb long with 6 exons. The PDZD11 protein (140 aa) is a ubiquitously expressed small protein and mainly composed of a single PDZ domain (Stephenson et al., 2005). Previous studies have shown that diseases associated with PDZD11 include Purulent Acute Otitis Media and Middle Ear Disease (GeneCards database). PDZD11 was previously known as PISP, based on its interaction with the plasma membrane calcium ATPase (PMCA) b-splice variants, which may play a role in their sorting to or from the plasma membrane (Goellner et al., 2003). PDZD11 is also known as AIPP1, because it interacts with Menkes copper ATPase (ATP7A), which involves in maintaining copper homeostasis (Stephenson et al., 2005). Nabokina et al. (2011) demonstrated that PDZD11 interacted with human sodium-dependent multivitamin transporter (hSMVT) in intestinal epithelial cells and that this interaction affected biotin uptake process. Shah et al. (2016) reported that the interaction of the N-terminal region of PDZD11 with the WW1 domain of pleckstrin homology domain-containing A7 (PLEKHA7) was essential to stabilize junctional nectins at adherens junctions (AJ), and promote efficient junction assembly. Recent work has also shown that cooperative binding of the tandem WW domains (e.g., WW1 and WW2) of PLEKHA7 to PDZD11 promoted the binding of the C-terminus of Tspan33 to PLEKHA7. Furthermore, the complex formation of PLEKHA7, PDZD11, ADAM10 and its molecular chaperone Tspan33 through promoting the junctional clustering of the α-toxin receptor ADAM10 makes cells more sensitive to the cytotoxic effects of *Staphylococcus aureus* α-toxin (Vasileva et al., 2017; Rouaud et al., 2020).

Epithelial-to-mesenchymal transition (EMT) is a reversible cellular procedure that can transiently dedifferentiate epithelial cells into a mesenchymal phenotype (Dongre and Weinberg, 2019). Epithelial cells build strong connections with their neighbors and an apical-to-basal polarity via the sequential arrangement of adherens junctions, desmosomes, and tight junctions (Thiery et al., 2009). Conversely, EMT confers cells with invasive and metastatic potential, induces stem cell properties, inhibits apoptosis and senescence, and contributes to immunosuppression (Thiery et al., 2009). Therefore, EMT plays a crucial role in embryogenesis, wound-healing, organ fibrosis, tumor invasion and metastasis (Yan et al., 2018). In particular, about 90% of cancer-associated mortality is attributed to metastasis (Chaffer and Weinberg, 2011). Previous studies have shown that the combination of metastasis-related gene signatures and serum alpha-fetoprotein can be used as a good predictor of LIHC prognosis regardless of etiology and race (Yan et al., 2018).

The tumor microenvironment is composed of infiltrating inflammatory cells, stromal cells, and inflammatory mediators (Yan et al., 2018). Undoubtedly, the inflammatory microenvironment associated with hepatitis virus infection is an important factor influencing the invasion and metastasis of LIHC (Yan et al., 2018). Lara-Pezzi et al. (2001) also reported that hepatitis B virus HBx protein was able to induce adherens junction disruption in a src-dependent manner, which might contribute to the development of LIHC.

In this study, we first performed a bioinformatics analysis using different open databases to acquire detailed information about potential functions and prognostic value of PDZD11 in LIHC, and to explore whether the abnormal expression of PDZD11 is closely related to immune infiltrates of LIHC. Further, we verified the expression of PDZD11 in LIHC tissues, various human liver cancer cell lines and matched normal hepatocytes. The findings of this study may help us to understand the role of PDZD11 in the development of LIHC.

# MATERIALS AND METHODS

## Patients and LIHC Tissue Specimens

Seven pairs of matched LIHC tumor tissues and adjacent normal tissues of each pair of patients were immediately quenched in liquid nitrogen after surgical removal in the First Affiliated Hospital of Wenzhou Medical University. All the patients were clinically and pathologically confirmed as liver cancer. Informed consent was approved by the board of directors and the ethics committee of the First Affiliated Hospital of Wenzhou Medical University. Written informed consent was obtained from all subjects.

## Cell Culture

HCCLM3, MHCC97H, HepG2, and L02 cells were cultured in high-glucose DMEM (GIBCO, Waltham, MA, United States) containing 10% FBS (fetal bovine serum) (GIBCO, United States) and antibiotics (100 U/ml penicillin and 100 μg/ml streptomycin) (GIBCO, United States), and incubated in an incubator containing 5% $CO_2$ at 37°C.

## Western Blot Analysis

Proteins in clinical tissues and whole-cells were extracted with 1%Triton X-100 lysis buffer supplemented with protease and phosphatase inhibitors (Sigma-Aldrich). Protein concentrations of the extracts were determined by the BCA assay kit (Thermo Fisher Scientific, Waltham, MA, United States). 40 μg of total protein in each sample was separated by a 12% SDS-PAGE gels and transferred onto PVDF membrane (Bio-Rad, Hercules, CA) with a wet transfer system (Bio-Rad, United States). Block the blot in blocking buffer (5% skim milk in TBST) on a shaker at room temperature for 1 h, and then incubated with primary antibodies specific for PDZD11 (ab121210) (Abcam, Cambridge, MA) (1:2,000) and β-actin (Beyotime Biotechnology Co., Ltd., Shanghai, China) (1:5,000) overnight at 4°C. The membrane was washed in TBST for 3 × 15 min and then incubated with horseradish peroxidase (HRP)-conjugated anti-rabbit (1:5,000)

and anti-mouse (1:20,000) immunoglobulin G on a shaker at room temperature for 1.5 h. Immunoreactive proteins were visualized using ECL reagent according to the manufacturer's protocol (Thermo Fisher Scientific, Rockford, IL). The optical density was quantified by executing ImageJ software.

## Quantitative RT-PCR

LIHC cell lines and hepatocytes L-02 were seeded in 10 cm culture dish at a density of $2 \times 10^6$ cells per culture dish. After 36 h of incubation at 37°C, cells were harvested and washed once with ice-cold PBS. The mRNA expression levels of genes were tested by SYBR green-based real-time quantitative PCR. Total RNA was extracted from all the cells using TRIzol reagent (Thermo Fisher Scientific, Waltham, MA, United States) according to the manufacturer's instructions. Total RNA (1 μg) was reverse-transcribed into cDNA (+ gDNA wiper) HiScript II Q Select RT SuperMix (Vazyme Biotech Co., Ltd., Nanjing, China) according to the manufacturer's instructions. The RT reaction was subsequently used as a template for real-time PCR. The reactions were performed on a CFX Connect™ Real-Time PCR Detection System (Bio-Rad, Hercules, CA) using ChamQ Universal SYBR qPCR Master Mix (Vazyme Biotech Co., Ltd., Nanjing, China). Primer sequences were as follows: PDZD11 5′CGGTGGTTTTCTTGCCTGCC3′ (forward), 5′-TCAGTGTGATGGTTCGGGGC-3′ (reverse) and β-actin 5′-AGCACAGAGCCTCGCCTTTG-3′ (forward), 5′-AAGCCGGCCTTGCACATG-3′ (reverse). The PCR amplification procedures were as follows: pre-denaturation at 95℃ for 3 min, followed by 40 cycles of (95°C for 10 s, 60°C for 30 s). Record the threshold cycle number (Ct) for each reaction. The Ct values of target genes were normalized to that of β-actin. Each sample was analyzed in triplicate and repeated 3 times.

## GEPIA2 Database

The expression of *PDZD11* mRNA in LIHC was analyzed using the GEPIA2 database[1], which was developed by Peking University, China, and is based on The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) databases, including RNA sequencing and expression data from 33 malignant tumors, 8,587 normal tissues, and 9,736 tumor samples (Tang et al., 2017).

## Oncomine Analysis

The Oncomine v.4.5 database[2] is a comprehensive and user-friendly online cancer microarray database for DNA and RNA sequence analysis (Rhodes et al., 2007). In our study, mRNA expression levels and DNA copy number of *PDZD11* in normal controls and cancer specimens were obtained from the Oncomine database. The retrieval conditions were as follows: analysis type/cancer vs. normal analysis, cancer type/liver cancer, dataset filters/data type/mRNA or DNA, and sample filters/sample type/clinical specification. The significance threshold was designed using the following specific parameters: $p$-value of 1E-4, -fold change of 2, and gene rank in the top 10%. Student's $t$-test was used to analyze differences in the expression of *PDZD11* between normal controls and cancer specimens.

## DriverDBV3 Database

DriverDBV3[3] uses a variety of -omics techniques to identify cancer driver genes and to present them with different molecular features, including somatic mutations, RNA expression, miRNA expression, DNA methylation, copy number variation, and clinical data, in addition to annotation of bases (Liu S.H. et al., 2020). The Gene Summary of *PDZD11* in various cancer tissues and mRNA expression of *PDZD11* in LIHC was analyzed using the DriverDBV3 database. Survival with a log-rank $p < 0.05$, was considered statistically significant.

## UALCAN Database Analysis

The UALCAN database[4] is a website for online analysis based on level 3 RNA-seq and clinical data of 31 cancer types from TCGA datasets (Chandrashekar et al., 2017). We used this database to analyze the differential expression and promoter methylation profile of *PDZD11* in primary LIHC tissues and their association with clinicopathological parameters. Student's $t$-test was used to generate $p$-values; after Bonferroni correction for multiple measures, $p$ was still $< 0.05$, which was statistically significant.

## cBioPortal Analysis

The cBioPortal[5] is an open-access web resource that provides visualization and analysis of multidimensional cancer genomics data (Gao et al., 2013). In this study, genetic alterations to *PDZD11* in LIHC patients (TCGA, Firehose Legacy, 360 patients/samples) were investigated using the cBioPortal database.

## Protein-Protein Interaction (PPI) Network Analysis

PPI network analysis of PDZD11 was conducted using the STRING[6] (von Mering et al., 2003) and GeneMANIA[7] (Warde-Farley et al., 2010) online databases. We also used GeneMANIA to construct gene networks and predict the biological functions of gene sets in which Gene Set Enrichment Analysis (GSEA) was identified as being enriched in LIHC.

## LinkedOmics Database Analysis

The LinkedOmics database[8] (Vasaikar et al., 2018) is an online open-access powerful bioinformatics platform, which includes multi-omics information and clinical data involving 11,158 patients and 32 cancer types in the TCGA project. LinkedOmics was used to study genes differentially expressed in correlation with *PDZD11* in LIHC. Pearson's correlation coefficient was applied to statistical analysis of the results produced by

[1]http://gepia2.cancer-pku.cn/#index

[2]https://www.oncomine.org/resource/login.html

[3]http://driverdb.tms.cmu.edu.tw/

[4]http://ualcan.path.uab.edu/

[5]https://www.cbioportal.org/

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

[7]http://genemania.org/

[8]http://www.linkedomics.org/login.php

LinkedOmics. Then, genes positively and negatively correlated with *PDZD11* in LIHC were selected based on the criteria of coefficient $> 0.3$ and $< -0.3$. Finally, we enriched these gene sets by Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using the DAVID database[9] (Huang da et al., 2009), and the results were visualized using an online platform[10]. Moreover, GSEA was utilized to perform various enrichment analyses, including for kinase targets, miRNA targets, and transcription factor targets. Ranking was based on the criteria of false discovery rate (FDR) $<0.05$, and 500 simulations were performed.

## TIMER Analysis

Tumor Immune Estimation Resource (TIMER)[11] is a comprehensive website for the systematic analysis of tumor-infiltrating immune cells (Li et al., 2017). TIMER2.0[12] is the latest version of TIMER. We first analyzed the expression of *PDZD11* in various tumors using the TIMER database, and the results were analyzed statistically using Wilcoxon rank sum test. Then, correlations between the expression of PDZD11 and the abundance of the six immune cell types (B cells, $CD8^+$ T cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells) in LIHC were analyzed using Spearman tests (tumor purity adjusted). Finally, the survival module was used to draw Kaplan-Meier plots for immune infiltrates and *PDZD11* to determine survival differences. Statistical significance was set at $p < 0.05$.

## Statistical Analysis

GraphPad Prism (v.5.0) for Windows was used for statistical analysis, and $p < 0.05$ was considered statistically significant. The log-rank test was used in Kaplan-Meier survival analysis. Student's *t*-test and Wilcoxon rank sum test were employed in two-group comparisons. Moreover, we conducted Bonferroni's correction for multiple measurements to ensure the credibility of multiple group comparisons. After Bonferroni correction, *p* was still less than 0.05, which represents a statistically significant difference.

# RESULTS

## Elevated Expression of PDZD11 in LIHC

To determine the differential expression of PDZD11 in diverse cancer types, *PDZD11* mRNA expression was analyzed using the TIMER database. It was shown that the mRNA level of *PDZD11* was significantly upregulated in bladder, breast, gallbladder, esophagus, kidney, liver, lung, gastric, thyroid, and uterine corpus endometrial carcinoma (**Figure 1A**). Further analysis showed that *PDZD11* was overexpressed in LIHC patients in the GEPIA2 database (**Figure 1B**). In the ONCOMINE database, *PDZD11* was also identified with significantly higher levels in LIHC in multiple datasets. In the Chen Liver dataset, *PDZD11* overexpression was found in LIHC tissues compared with normal tissues with a -fold change of 1.812 ($p$ = 3.44E-22), while Wurmbach observed a 1.651-fold increase in *PDZD11* mRNA expression in LIHC samples ($p$ = 8.16E-5) (**Figures 1C,D**). In addition, we analyzed *PDZD11* expression using the DriverDBV3 database. We found that the results were largely consistent with those in the ONCOMINE database. However, there was no statistically significant difference in the mRNA expression of *PDZD11* observed in recurrent solid tumors compared to adjacent normal liver tissues (**Figure 1E**). Consistently, protein analysis involving eight patients (including eight tumor tissue and eight matched adjacent normal tissues) diagnosed with liver cancer confirmed that *PDZD11* abundance was elevated in LIHC tissues (**Figure 1F**). Additionally, we found increased levels of *PDZD11* in LIHC cell lines at both the mRNA and protein levels (**Figures 1G,H**) compared to that found in normal human L-02 hepatocytes. However, the protein expression of PDZD11 in HepG2 cells was significantly lower than that in L-02 cells.

## Relationship Between *PDZD11* mRNA Levels and Clinicopathological Parameters in LIHC Patients

Next, relationships between *PDZD11* mRNA expression and clinicopathological parameters of LIHC patients were analyzed using the UALCAN database. The results showed that *PDZD11* was upregulated in primary LIHC tissues compared to adjacent normal tissues (**Figure 2A**, $p < 0.001$). As shown in **Figures 2B–I**, according to subgroup analysis based on race, gender, age, weight, and lymph node metastasis status, the mRNA expression of *PDZD11* in LIHC patients was evidently higher than that in healthy individuals. In particular, the expression of *PDZD11* mRNA was clearly correlated with more advanced and less-differentiated tumors in LIHC patients, who tended to express higher *PDZD11* mRNA levels. The highest mRNA expression of *PDZD11* was found in stage 3 and/or tumor grade 3 cases (**Figures 2B,G**). The reason why mRNA expression of *PDZD11* in stage 3 and/or tumor grade 3 seemed to be higher than that in stage 4 and tumor grade 4 may be due to the small number of samples. In addition, *PDZD11* mRNA expression was positively correlated with *TP53* mutation status, and was also significantly elevated in LIHC patients with *TP53* mutations (**Figure 2I**).

## Frequency and Types of *PDZD11* Alterations in LIHC

Genetic alterations to *PDZD11* in LIHC were evaluated using the cBioPortal database. As shown in **Figure 3A**, among the 360 LIHC patients sequenced, 27 showed genetic alterations, with a mutation rate of 8%. Moreover, we observed that mRNA upregulation was the only aberrant type of genetic alteration involving *PDZD11* in LIHC.

## Decreased *PDZD11* Promoter Methylation Levels in LIHC

To analyze why expression of *PDZD11* mRNA was significantly higher in LIHC tissues than in adjacent normal liver tissues, we

[9]https://david.ncifcrf.gov/home.jsp
[10]http://www.bioinformatics.com.cn/
[11]https://cistrome.shinyapps.io/timer/
[12]http://timer.cistrome.org/

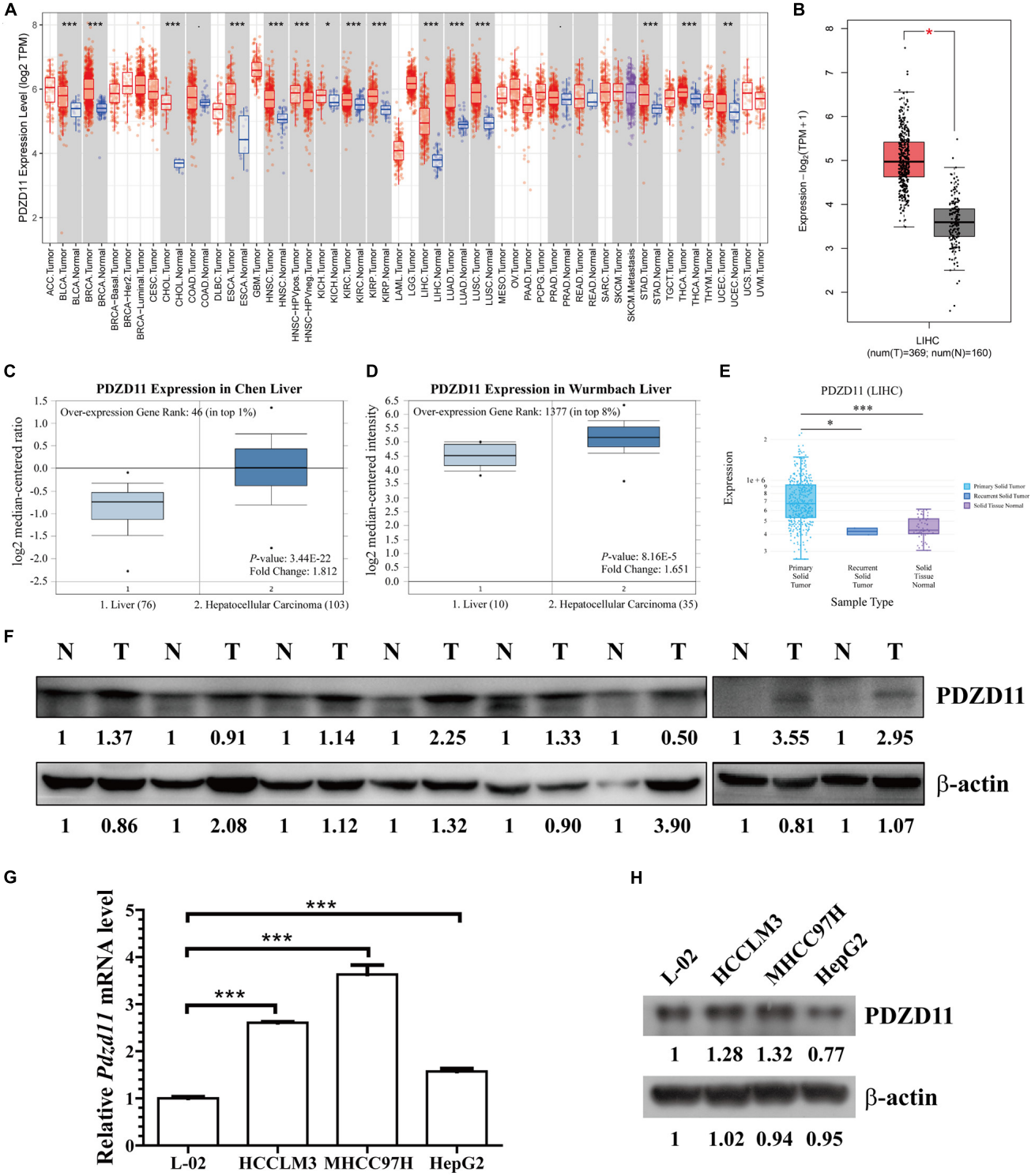



**FIGURE 1 |** PDZD11 expression levels in LIHC. **(A)** Transcription levels of *PDZD11* in different types of cancers (TIMER database). **(B)** *PDZD11* mRNA expression levels in LIHC tissues and adjacent normal liver tissues from GEPIA 2 database. **(C,D)** Box plots show *PDZD11* mRNA expression in liver (left plot) and hepatocellular carcinoma tissue (right plot) of the Chen Liver **(C)** and Wurmbach Liver **(D)** datasets. The fold-change of *PDZD11* expression in LIHC was determined using the Oncomine database. The threshold was designed using the following specific parameters: $p$ = 1E-4, fold change = 2, and gene rank 10%. **(E)** mRNA expression of *PDZD11* in primary solid tumors, recurrent solid tumors, and adjacent normal liver tissues (DriverDBV3 database). **(F)** A representative western blot showing PDZD11 protein is expressed in LIHC tissues (T) and matched normal liver tissues (N) ($n$ = 8). **(G,H)** Real-time qPCR and Western blotting analysis of *PDZD11* mRNA **(G)** and protein expression **(H)** in human hepatocytes and LIHC cell lines. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

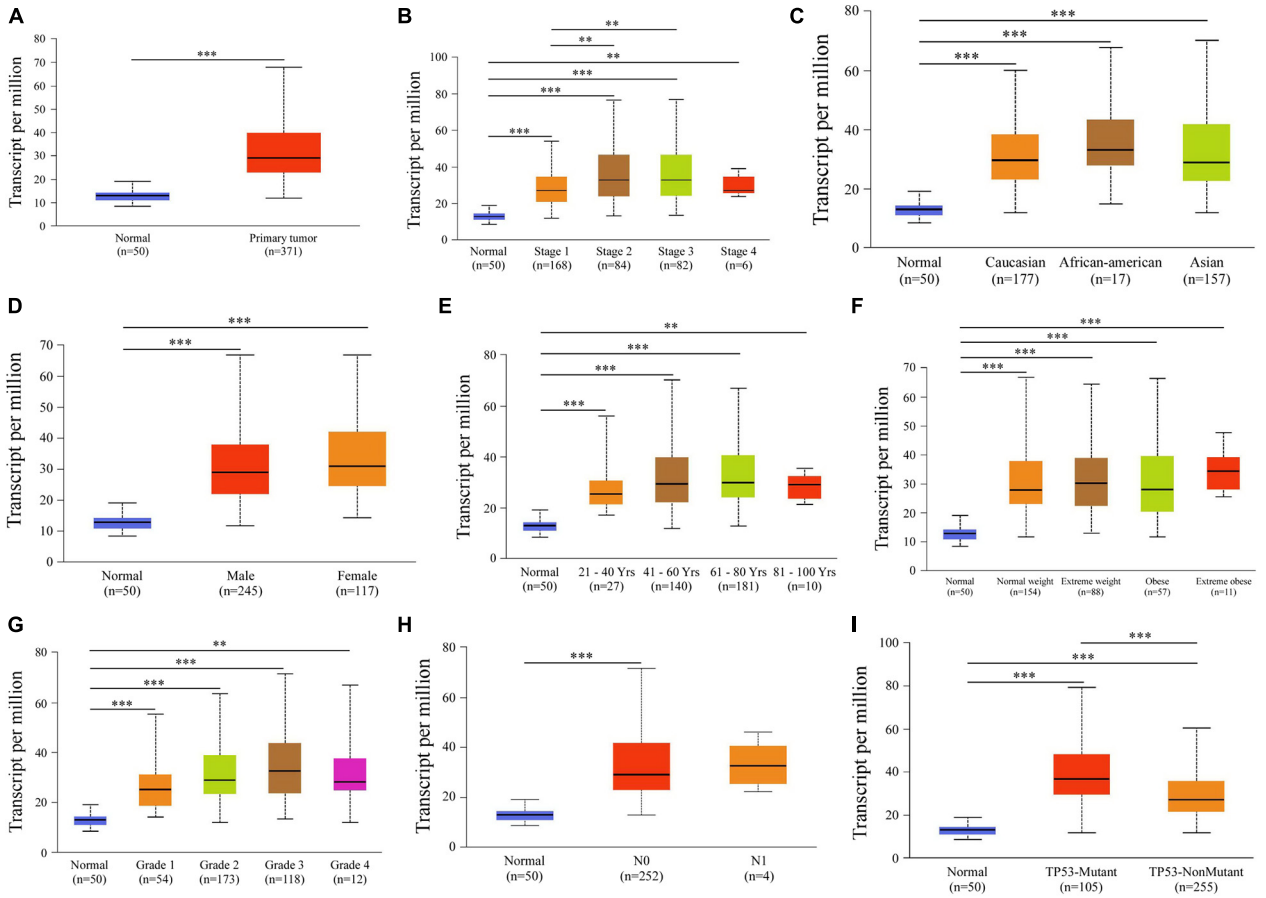



**FIGURE 2 |** Analysis of subgroup expression of *PDZD11* in LIHC (UALCAN database). **(A)** *PDZD11* mRNA expression in LIHC tissue and adjacent normal liver tissue. **(B–I)** Box plot shows the *PDZD11* mRNA expression of LIHC patients in the subgroups of different cancer stages **(B)**, race **(C)**, gender **(D)**, age **(E)**, weight **(F)**, tumor grade **(G)**, nodal metastasis status **(H),** and TP53 mutation status **(I)**. The data are shown as mean ± SE. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. The asterisk indicates a significant difference between the two sets of data.

used the UALCAN database to evaluate the extent of *PDZD11* promoter methylation in LIHC samples and investigated the association between promoter DNA methylation and *PDZD11* expression levels. The results indicated that *PDZD11* promoter methylation levels were lower in LIHC cases than in normal control samples (**Figure 3B**). To explore the factors that affect levels of *PDZD11* promoter methylation, we further analyzed promoter DNA methylation of *PDZD11* in different subgroups according to different clinicopathological parameters. The subgroup analysis results showed that promoter methylation of *PDZD11* was possibly affected by individual cancer stages, race, gender, age, weight, tumor grade, nodal metastasis status, and *TP53* mutation status in LIHC (**Figures 3C–J**).

## Prognostic Value of *PDZD11* mRNA Expression in LIHC Patients

To explore whether high expression levels of *PDZD11* are associated with cancer-promoting or tumor suppressor genes, we evaluated the prognostic value of *PDZD11* mRNA expression in patients with LIHC using the DriverDBV3 database. As shown in **Figures 4A,B**, *PDZD11* overexpression was associated with unfavorable 5-year survival [hazard ratio (HR) = 1.69, log-rank $p = 0.0036$] and overall survival (OS, HR = 1.53, log-rank $p = 0.0153$) in LIHC patients.

## Biological Interaction Network of PDZD11

Using the STRING and GeneMANIA databases, a functional protein interaction network of PDZD11 was constructed to enrich for possible PDZD11-mediated signaling pathways (**Figures 4C,D**). ATP7A, a transmembrane protein that functions in copper transport across cell membranes, was the only gene that intersected two protein-protein interaction (PPI) networks (Schmidt et al., 2018). Furthermore, STRING was used to perform GO and KEGG analyses to determine the functional enrichment of these 29 interactors. The results indicated that biological processes included copper ion homeostasis and copper ion transmembrane transport; Cellular components analysis found that these proteins are localized mainly in endosomes and early endosomes. Molecular function analysis indicated that these

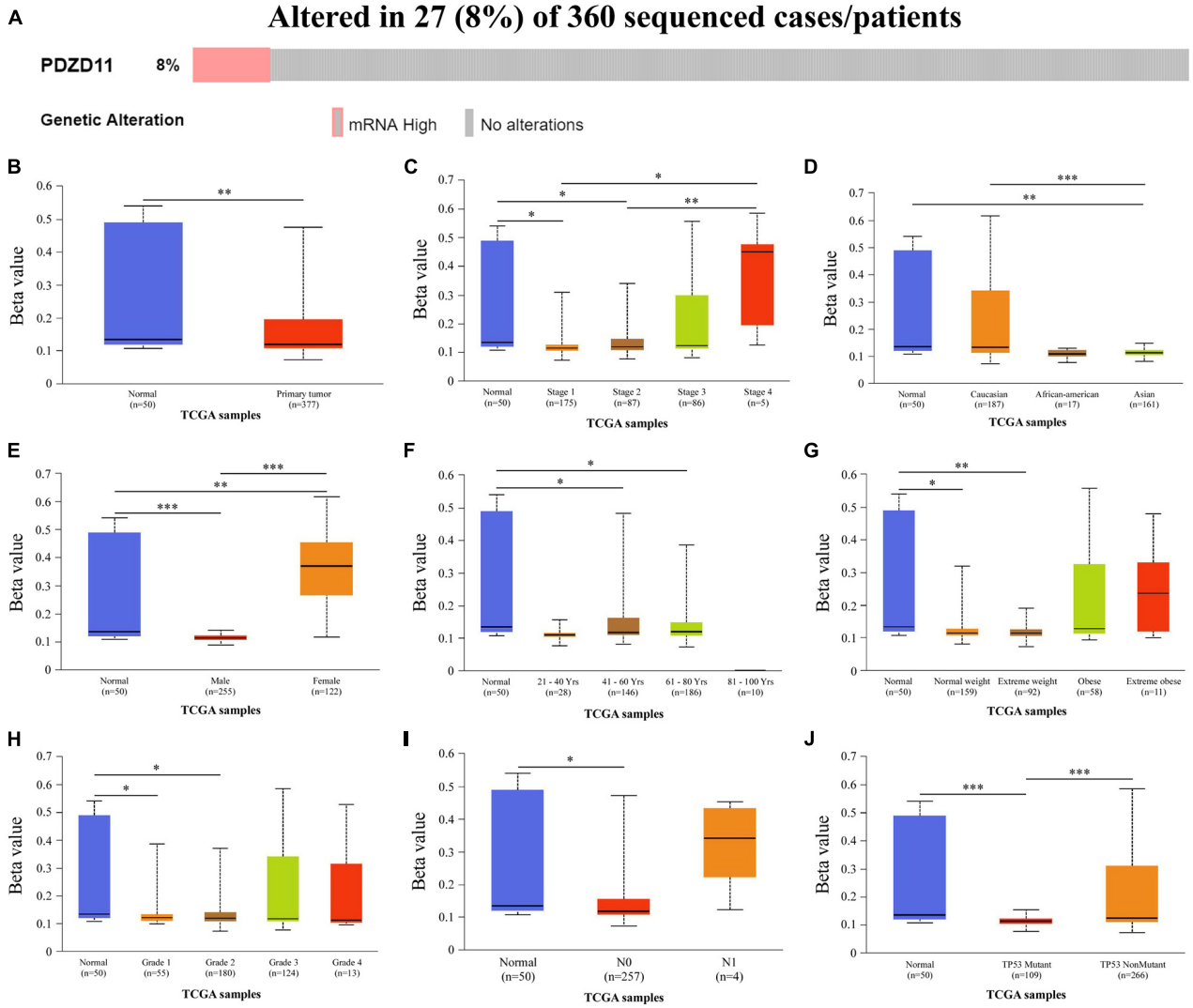



**FIGURE 3 |** Genetic alternations and promoter methylation levels of *PDZD11* in LIHC. **(A)** Oncoprint in cBioPortal showed the distribution and proportion of samples with alternations in *PDZD11*. **(B)** The promoter methylation levels of *PDZD11* in LIHC. **(C–J)** Box plots show promoter methylation level of *PDZD11* in normal vs. LIHC tissues and different individual cancer stages **(C)**, race **(D)**, gender **(E)**, age **(F)**, weight **(G)**, tumor grade **(H)**, nodal metastasis status **(I)**, and TP53 mutation status **(J)** (UALCAN database). Data are shown as mean ± SE. $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

proteins are primarily involved in copper-dependent protein binding, copper ion transmembrane transporter activity, copper ion binding, copper chaperone activity and phosphatidylinositol-3,5-bisphosphate binding (data not shown).

## Enrichment Analysis of PDZD11 Functional Networks in LIHC

### Predicted Functions and Pathways of Co-expressed Genes Correlate With *PDZD11* in LIHC

LinkedOmics was used to analyze TCGA mRNA sequencing data from 371 LIHC patients. Pearson's test was used to analyze the co-expression of genes correlated with *PDZD11* levels in LIHC. As shown in the volcano plot (**Figure 5A**), 2,960 genes (dark red dots) showed significant positive correlation with *PDZD11*, whereas 3,234 genes (dark green dots) showed opposite correlations (false discovery rate (FDR) < 0.01). The top 50 significant genes were positively and negatively associated with *PDZD11*, as shown in the heat map (**Figures 5B,C**). As shown in **Figures 5D–F**, the mRNA expression of *PDZD11* showed the strongest positive association with expression of *FAM50A* (Pearson correlation = 0.62, $p$ = 7.71e-41), *NDUFA1* (Pearson correlation = 0.60, $p$ = 9.78e-38), and *LAGE3* (Pearson correlation = 0.60, $p$ = 1.04e-37), which reflect changes in the spliceosome complex (Lee et al., 2020), mitochondrial respiratory chain complex I (Fernandez-Moreira et al., 2007), and the KEOPS/EKC complex (tRNA modification complex) (Wan et al., 2017).

Furthermore, based on the results of the Pearson test (**Figures 5A–C**), we selected positively and negatively correlated

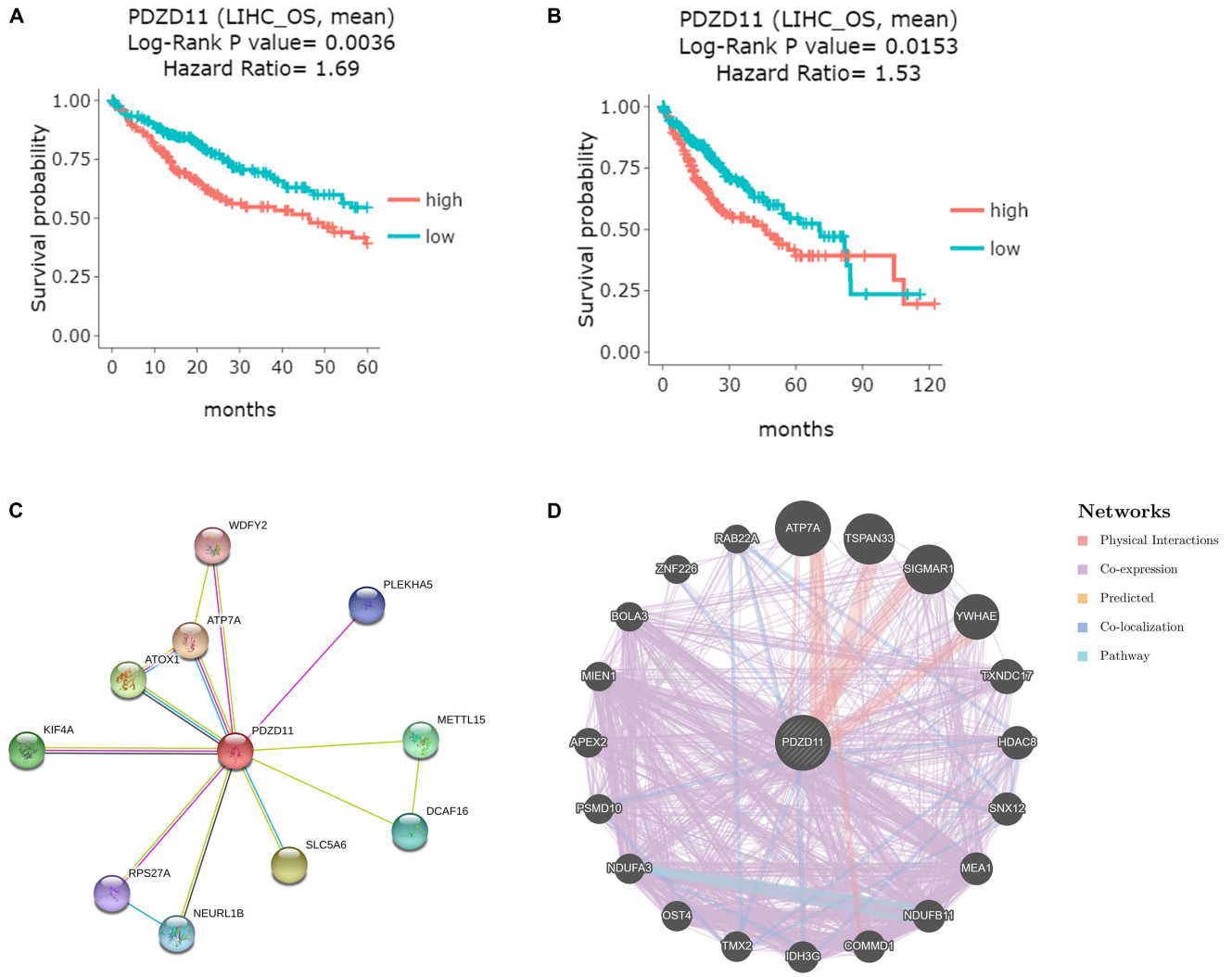



**FIGURE 4 |** Visual summary of the prognostic value and biological interaction network of PDZD11 in LIHC. **(A,B)** Kaplan-Meier plot of the relationship of *PDZD11* gene expression and survival in LIHC patients (DriverDBV3 database). The LIHC patient samples are stratified into 2 groups using the mean expression value as the cut-off point: 149 samples with highly expressed *PDZD11* mRNA (red) and 216 samples with lowly expressed *PDZD11* mRNA (green). *Y*-axis is survival probability. The left figure is the 5-year survival, and the right figure shows overall survival (OS). The *X*-axis is the months of survival period. Log-Rank *P*-value and Hazard Ration (HR) are provided on the top of plots. **(C)** Protein-protein interaction (PPI) network of PDZD11 (Top 10) (STRING database). **(D)** PPI network (Top 20) of PDZD11 from GeneMANIA database.

genes with coefficients $> 0.3$ and $< -0.3$. Finally, 617 genes positively correlated with *PDZD11* and 411 genes negatively correlated with *PDZD11* were selected (FDR $< 0.001$). Moreover, these genes were used for GO and KEGG enrichment analyses using the DAVID database. The cutoff criterion was set at FDR $< 0.01$. As shown in **Figures 6A–D**, cellular component analysis indicated that these proteins were mainly located in the nucleoplasm, proteasome complex, and mitochondrial inner membrane. Biological processes were primarily enriched in NIK/NF-κB signaling, regulation of cellular amino acid metabolic processes, and anaphase-promoting complex-dependent catabolic processes. Molecular function analysis revealed that these proteins were mostly involved in protein binding, poly(A) RNA binding, and threonine-type endopeptidase activity. KEGG pathway results showed that the co-expressed genes for the most part participated in proteasomes, oxidative phosphorylation (OXPHOS) (**Figure 6E**), and Alzheimer's disease.

## PDZD11 Networks of Kinase, MicroRNA or Transcription Factor Targets in LIHC

To further explore the gene regulatory network of *PDZD11* in LIHC, we also analyzed the important kinase, miRNA, and transcription factor target networks that were connected to *PDZD11* in LIHC via gene set enrichment analysis (GSEA). The results showed that the most frequent kinase targets, miRNA targets, and transcription factor targets were kinase CDK5, three miR-200 family members (miR-200b, miR-200c, and miR-429),

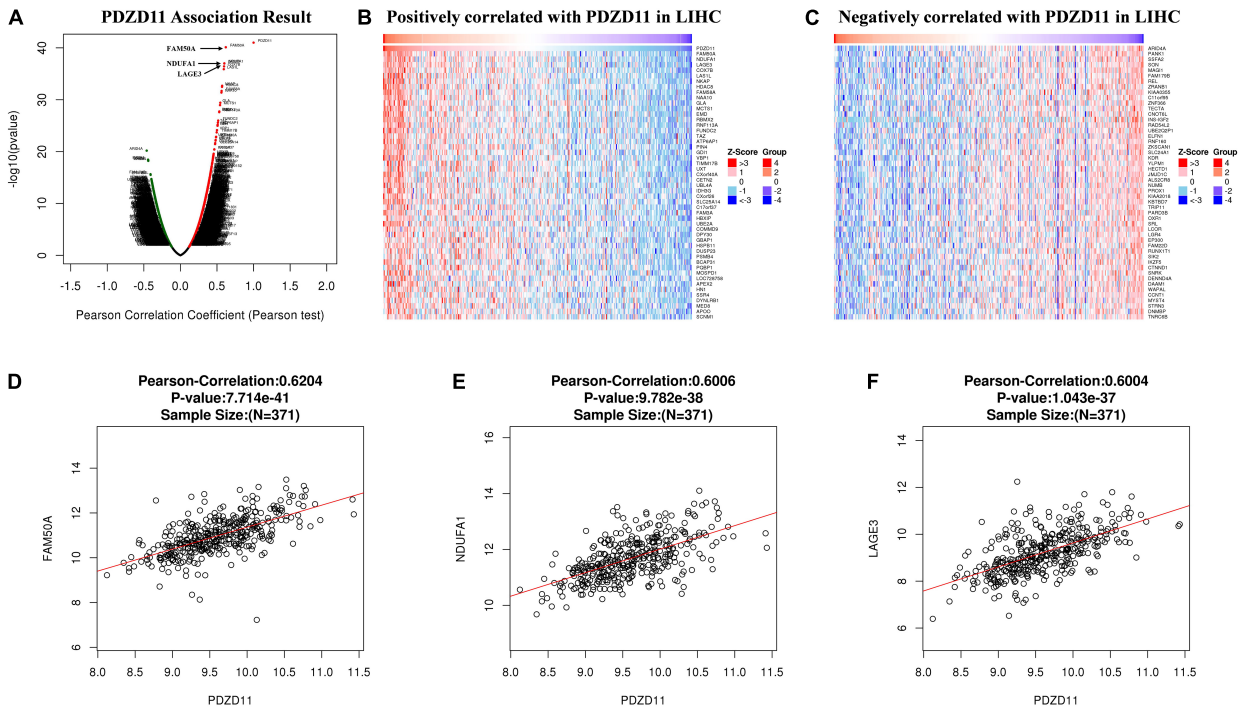



**FIGURE 5 |** Differentially expressed genes that correlated with *PDZD11* in LIHC (LinkedOmics database). **(A)** A Pearson test was used to determine the correlations between *PDZD11* and differently expressed genes in LIHC. **(B,C)** Heat maps are showing genes (Top 50) positively or negatively correlated with *PDZD11* in LIHC. Red indicates positively correlated genes and blue indicates negatively correlated genes. **(D–F)** The scatter plots mean that *PDZD11* expression is positively correlated with the expression of *FAM50A* **(D)**, *NDUFA1* **(E)**, and *LAGE3* **(F)**.

and V$SOX9_B1, respectively (**Table 1** and **Supplementary Tables 5–7**). Furthermore, PPI networks were constructed by STRING, and biological enrichment was performed using the DAVID database, indicating that all three gene sets were mainly involved in the KEGG pathway of prostate cancer, MAPK signaling pathway, and transcriptional dysregulation in cancer (**Supplementary Figures 1–3**).

## Association of PDZD11 Expression and Immune Infiltration in LIHC

LIHC is one of the most common malignant tumors (Okajima et al., 2017). Because *PDZD11* overexpression is associated with poor prognosis in LIHC patients (**Figure 4**), we explored whether the expression of *PDZD11* was correlated with levels of immune infiltration in LIHC from the TIMER database and/or TIMER2.0 database. As shown in **Figure 7**, there was a positive correlation between *PDZD11* expression and infiltration by B cells, $CD8^+$ T cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells. Furthermore, under the premise of high expression of *PDZD11* mRNA in LIHC, we found that higher infiltration levels of two immune cells (T cell CD4 + memory resting-CIBERSORT, and Macrophage-EPIC) were associated with better survival outcomes in LIHC patients (**Figures 7B,C**). In contrast, higher infiltrating levels of the macrophage M2 subset was a risk factor for disease prognosis in LIHC patients (**Figure 7D**).

# DISCUSSION

EMT-induced changes in epithelial cell plasticity are evidenced by the loss of epithelial markers, such as the adherence junction component E-cadherin and cytokeratins of the intermediate filament system (K8, K18, K19). Conversely, the expression of mesenchymal proteins such as N-cadherin, α-SMA, FSP-1, and the EMT transcription factors Snail (SNA1), Slug (SNA2), Twist, and ZEB are increased (Giannelli et al., 2016). Konopka et al. (2007) have also reported that junctional adhesion molecule-A (JAM-A) is critical for the formation of pseudocanaliculi and regulates E-cadherin expression through feedback signaling pathways in hepatic cells. However, the present study lacked a well-defined consensus on EMT-MET (mesenchymal-epithelial transition) biomarkers, which hinders definitive conclusions on how EMT affects clinical outcomes in LIHC patients (Giannelli et al., 2016). Therefore, there is an urgent need to identify biomarkers or therapeutic targets related to EMT for early diagnosis and for predicting the progression and recurrence of LIHC.

Current research reports that the interaction of PDZD11 with PLEKHA7 is significantly associated with tight and adherens junctions (Guerrera et al., 2016; Vasileva et al., 2017). However, to the best of our knowledge, no study has investigated the role of PDZD11 in liver cancer. In this study, we provide the first evidence that *PDZD11* mRNA expression is significantly upregulated in LIHC and is associated with poor prognosis

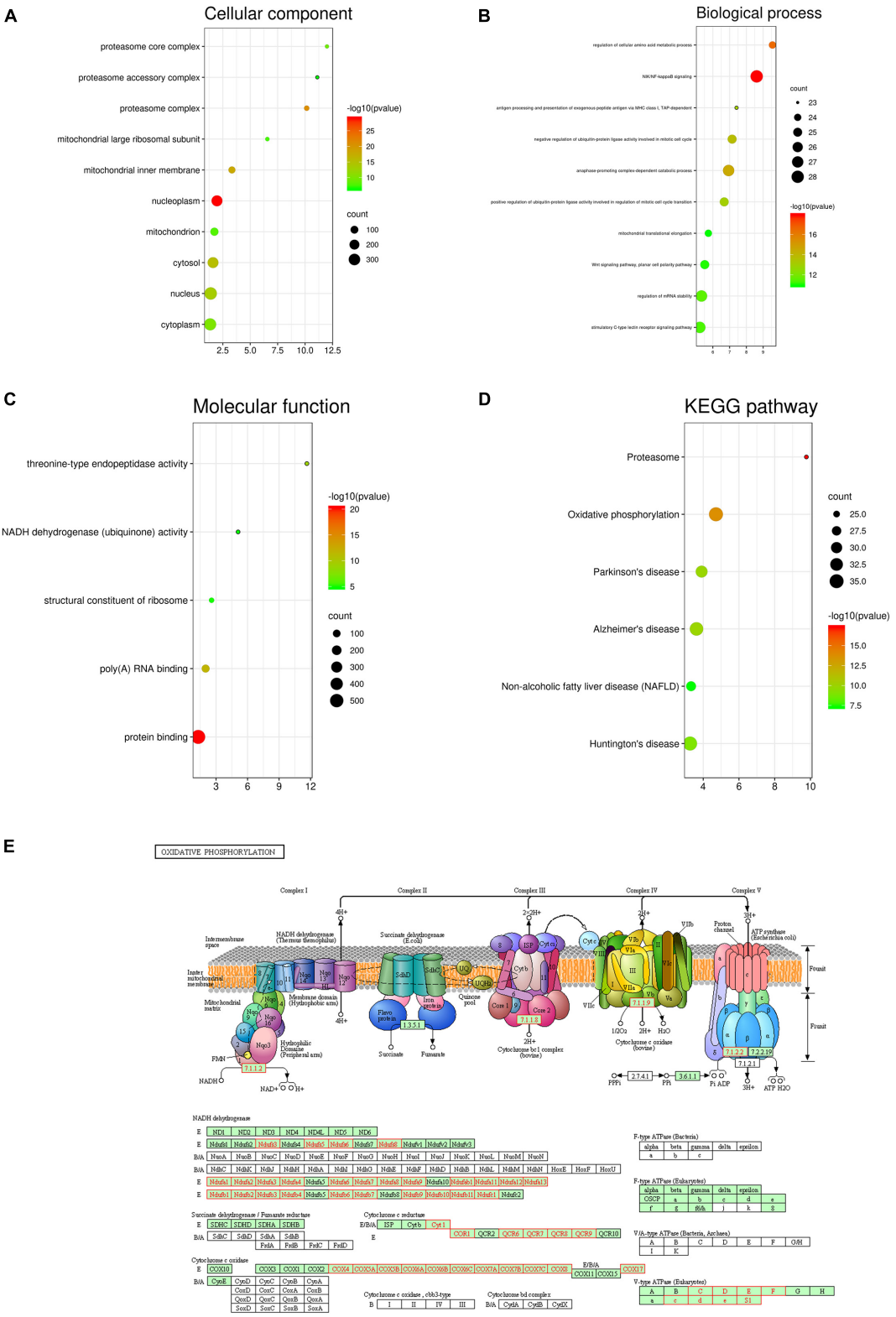



**FIGURE 6 |** Enrichment analysis of *PDZD11* co-expression genes in LIHC. **(A–D)** The significantly enriched GO annotations and KEGG pathways of *PDZD11* co-expression genes in LIHC are analyzed using DAVID. Based on the Pearson test (**Figures 5A–C**), we selected the positively and negatively correlated genes with coefficient > 0.3 and < −0.3 (LinkedOmics and bioinformatics databases). **(E)** KEGG pathway annotations of the oxidative phosphorylation pathway. Red marked nodes are associated with the Leading Edge Genes. FDR < 0.05 was considered statistically significant.

**TABLE 1 |** Kinase, miRNA and transcription factor-target networks of *PDZD11* in LIHC (LinkedOmics).

| Enriched category | Geneset | Leading edge number | FDR | *P*-value |
|---|---|---|---|---|
| Kinase target | Kinase_CDK5 | 26 | 0.0066696 | 0 |
| | Kinase_NLK | 5 | 0.010671 | 0 |
| | Kinase_MAPK7 | 14 | 0.034682 | 0 |
| | Kinase_DYRK1A | 7 | 0.045353 | 0.0040650 |
| miRNA target | GTGTTGA,MIR-505 | 46 | 0 | 0 |
| | CAGTATT,MIR-200B,MIR-200C,MIR-429 | 155 | 0 | 0 |
| | ACTGAAA,MIR-30A-3P,MIR-30E-3P | 81 | 0 | 0 |
| | AAAGGGA,MIR-204,MIR-211 | 101 | 0 | 0 |
| | TACTTGA,MIR-26A,MIR-26B | 131 | 0 | 0 |
| Transcription factor target | GGAANCGGAANY_UNKNOWN | 35 | 0 | 0 |
| | V$FREAC4_01 | 49 | 0 | 0 |
| | V$HOX13_01 | 15 | 0 | 0 |
| | V$SOX9_B1 | 80 | 0 | 0 |
| | V$STAT5A_02 | 50 | 0 | 0 |

*FDR, false discovery rate from Benjamini and Hochberg from gene set enrichment analysis (GSEA). V$, the annotation found in Molecular Signatures Database (MSigDB) for transcription factors (TF).*

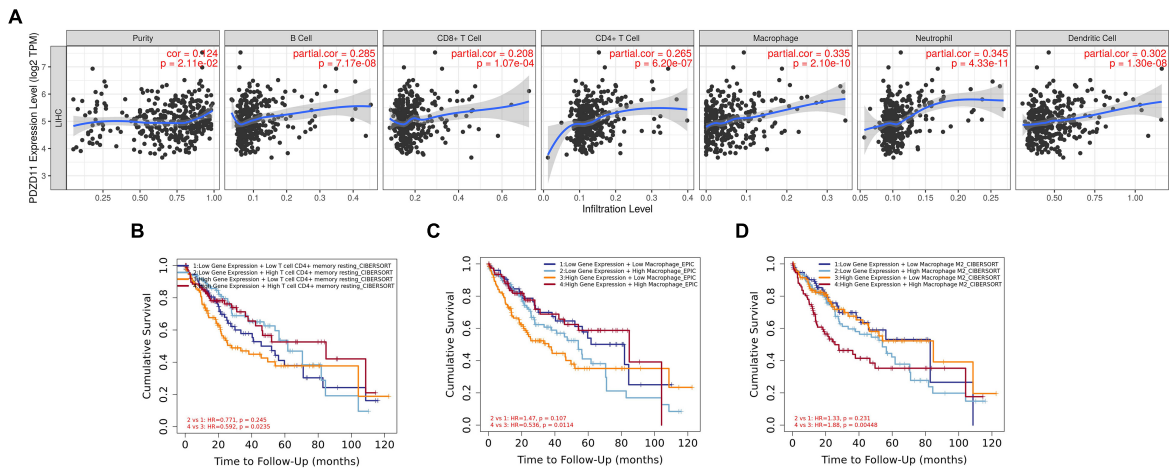



**FIGURE 7 |** Associations between mRNA expression of *PDZD11* and immune infiltration in LIHC (TIMER2.0 database). **(A)** Association of *PDZD11* expression with abundance of immune infiltrates (B cells, $CD8^+$ T cells, $CD4^+$ T cells, macrophages, neutrophils and dendritic cells). **(B–D)** The effect of *PDZD11* expression in correlation with infiltration levels of immune cells on the prognosis of LIHC.

(**Figures 1A–E**, **4A,B**). In particular, we demonstrated that *PDZD11* is aberrantly expressed in human liver cancer tissues and cell lines (**Figures 1F–H**). Moreover, subgroup analysis showed that the mRNA expression of *PDZD11* was also upregulated in different subgroups of LIHC (**Figure 2**). In particular, the mRNA expression of *PDZD11* increased as tumors progressed (**Figures 2B,G**). Additionally, we found that the expression of *PDZD11* may be negatively regulated by wild-type p53 at the transcriptional level (**Figure 2I**). Similarly, previous studies have shown that E-cadherin, the most reliable and closely investigated marker in a large number of LIHC patients, was directly correlated with poorer prognosis and shorter survival (Yamada et al., 2014). Consequently, these results suggest that PDZD11 and the EMT marker E-cadherin could serve as potential diagnostic and prognostic biomarkers in LIHC patients.

mRNA upregulation is the most aberrant type of genetic alteration involving *PDZD11* in LIHC (**Figure 3A**). We further analyzed *PDZD11* promoter DNA methylation levels and found that the higher expression of *PDZD11* in LIHC may be negatively correlated with the extent of promoter methylation (**Figure 3B**). Subgroup analysis showed that *PDZD11* promoter methylation level was also downregulated in different subgroups of LIHC (**Figures 3C–J**). Cano et al. (2000) reported that the expression of the EMT marker E-cadherin is negatively regulated by the transcription factor Snail. These results suggest that the mechanism of high expression of *PDZD11* mRNA in LIHC could be different from the classical Snail/E-cadherin axis.

Further analysis of the gene regulatory network of PDZD11 in LIHC suggested that the functions of these genes were primarily related to copper ion homeostasis, proteasome, and OXPHOS pathway. As shown in **Figures 4C,D**, *ATP7A* is the

only gene that intersects the two PPI networks. A previous study demonstrated that ATP7A is a transmembrane protein that functions in copper transport across cell membranes (Schmidt et al., 2018). Bortezomib is a first-in-class proteasome inhibitor that has been repeatedly demonstrated to exert anti-proliferative, anti-metastatic, and pro-apoptotic effects in LIHC (Yang et al., 2016; Huang et al., 2019). This study showed that the protein expression level of PDZD11 was irreconcilable with its mRNA transcription level. However, this proteasome-mediated PDZD11 protein degradation pathway requires further research. A recent study has also reported that induced E-cadherin expression and subsequent induction of NF-κB signaling increases OXPHOS, glycolysis, and cell proliferation in human gastric adenocarcinomacells (Park et al., 2017). Therefore, further research is needed to determine how abnormal expression of PDZD11 affects OXPHOS in LIHC and its role in LIHC metastasis.

We also sought important networks of target kinases, miRNAs, and transcription factors of the differentially expressed PDZD11 in LIHC. We found that PDZD11 in LIHC was linked to a network of kinases, including CDK5, NLK, and MAPK7. Previous studies have reported that levels of these kinases are significantly higher in human LIHC tissue than in normal liver tissue. Moreover, the downregulated expression of these kinases significantly inhibits the development and growth of LIHC *in vitro* and *in vivo* (Jung et al., 2010; Ehrlich et al., 2015; Lu et al., 2016). The probable miRNAs involved in the regulation of PDZD11 expression in LIHC included miR-505, three miR-200 family members (miR-200b, miR-200c, and miR-429), and two miR-30 family members (miR-30a-3p and miR-30e-3p). Lu et al. (2016) found that miR-505 regulates proliferation, invasion, and EMT in MHCC97 hepatoma cells by targeting high-mobility group box 1 (HMGB1). Ding et al. (2012) showed that the combination of a DNA methyltransferase (DNMT) inhibitor and upregulation of miR-200b could block lung metastasis of mesenchymal-phenotype hepatocellular carcinoma. Wang et al. (2014) indicated that miR-30a-3p inhibits tumor proliferation, invasion, and migration, and is downregulated in LIHC. Our data indicated that V$FREAC4_01, V$HOX13_01, and V$SOX9_B1 may be key transcription factors in the regulation of PDZD11. Liu et al. (2016) demonstrated that Sox9 regulates self-renewal and tumorigenicity by promoting symmetrical cell division of cancer stem cells in LIHC. Taken together, abnormal expression of PDZD11 may modulate tumor cell proliferation, invasion, metastasis, and the development of LIHC by regulating these targets. Further studies are required to verify this hypothesis.

The emergence and development of LIHC are accompanied by a persistent inflammatory reaction. Inflammatory cells in the tumor microenvironment of LIHC mainly include macrophages, infiltrating lymphocytes, neutrophils, mast cells, dendritic cells, and eosinophils (Kim and Bae, 2016; Yan et al., 2018). In particular, Liu W.R. et al. (2020) reported that among these tumor-related regulatory T cells (Tregs), macrophages, and neutrophils are strongly correlated with OS and relapse-free survival (RFS) in LIHC patients. Here, we found that PDZD11 expression in LIHC was positively correlated with infiltrating levels of six immune cell types (i.e., B cells, $CD4^+$ T cells, CD8 + T cells, macrophages, neutrophils, and dendritic cells). Moreover, under the premise of high expression of *PDZD11* mRNA in LIHC, the higher infiltration levels of the $CD4^+$ memory resting T cell subset were favorable factors for prognosis in LIHC patients. In contrast, the higher infiltration levels of the macrophage M2 subset had an unfavorable prognosis in LIHC (**Figure 7**). Previous studies have shown that $CD4^+$ T cells and tumor-associated macrophages (TAMs) play a central role in pro-tumor immunity; their interactions with tumor cells can directly promote tumor growth, progression, invasion, and metastasis. Conversely, $CD8^+$ T cells are responsible for anti-tumor responses, and increased $CD8^+$ T cell infiltration usually indicates a better prognosis in LIHC (Yan et al., 2018; Ansari et al., 2020; Li et al., 2020). In summary, these data indicate that PDZD11 is not only a prognostic biomarker, but may also reflect the immune status of LIHC patients.

In summary, these findings highlight the critical role of PDZD11 in the development and progression of LIHC. However, immunohistochemistry and functional analysis are needed in future studies to verify the relationship between PDZD11 and EMT in LIHC at the clinical and cellular levels. In particular, compared with normal human hepatocytes, the overexpression level of *PDZD11* mRNA was significantly higher than its protein level. In addition, the protein expression level of PDZD11 in HepG2 cells was significantly reduced. Therefore, overexpression of *PDZD11* in LIHC could not be ruled out, which is a self-protective feedback regulation mechanism that inhibits tumor metastasis. Further studies are needed to determine whether the aberrant expression of PDZD11 is detrimental or beneficial to patients with LIHC, and further studies are needed to explore how the aberrant expression of PDZD11 regulates the onset and progression of LIHC via EMT and OXPHOS pathways.

## 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/s.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the board of directors and the ethics committee of the First Affiliated Hospital of Wenzhou Medical University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

SY and YP designed and supervised the study. YC and XY collected the patient samples and performed the study. YC, HX, TX, YP, and SY participated in data analysis and figure preparation. YP, SY, and TX revised the article was written.

SY and YP reviewed the manuscript. All authors read and approved the final manuscript.

## FUNDING

This study was supported by the Zhejiang Medical and Health Science and Technology Plan Project, No. 2016KYB191 and Key Discipline of Zhejiang Province in Medical Technology (first class, category A).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.669928/full#supplementary-material

**Supplementary Figure 1 |** Protein-protein interaction (PPI) network of CDK5 Kinase-target networks (STRING).

**Supplementary Figure 2 |** PPI network of MIR-505 (STRING).

**Supplementary Figure 3 |** PPI network of V$FREAC4_01 transcription factor -target networks (STRING).

**Supplementary Table 1 |** Significantly enriched GO annotations (biological processes) of *PDZD11* in LIHC (LinkedOmics).

**Supplementary Table 2 |** Significantly enriched GO annotations (cellular components) of *PDZD11* in LIHC (LinkedOmics).

**Supplementary Table 3 |** Significantly enriched GO annotations (molecular functions) of *PDZD11* in LIHC (LinkedOmics).

**Supplementary Table 4 |** Significantly enriched KEGG pathway annotations of *PDZD11* in LIHC (LinkedOmics).

**Supplementary Table 5 |** Significantly enriched kinase-target networks of *PDZD11* in LIHC (LinkedOmics).

**Supplementary Table 6 |** Significantly enriched miRNA-target networks of *PDZD11* in LIHC (LinkedOmics).

**Supplementary Table 7 |** Significantly enriched transcription factor-target networks of *PDZD11* in LIHC (LinkedOmics).

## REFERENCES

Ansari, R. E., Craze, M. L., Althobiti, M., Alfarsi, L., Ellis, I. O., Rakha, E. A., et al. (2020). Enhanced glutamine uptake influences composition of immune cell infiltrates in breast cancer. *Br. J. Cancer* 122, 94–101. doi: 10.1038/s41416-019-0626-z

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

Cano, A., Pérez-Moreno, M. A., Rodrigo, I., Locascio, A., Blanco, M. J., del Barrio, M. G., et al. (2000). The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat. Cell Biol.* 2, 76–83. doi: 10.1038/35000025

Chaffer, C. L., and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science* 331, 1559–1564. doi: 10.1126/science.1203543

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Ding, W., Dang, H., You, H., Steinway, S., Takahashi, Y., Wang, H. G., et al. (2012). miR-200b restoration and DNA methyltransferase inhibitor block lung metastasis of mesenchymal-phenotype hepatocellular carcinoma. *Oncogenesis* 1:e15. doi: 10.1038/oncsis.2012.15

Dongre, A., and Weinberg, R. A. (2019). New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* 20, 69–84. doi: 10.1038/s41580-018-0080-4

Ehrlich, S. M., Liebl, J., Ardelt, M. A., Lehr, T., De Toni, E. N., Mayr, D., et al. (2015). Targeting cyclin dependent kinase 5 in hepatocellular carcinoma–a novel therapeutic approach. *J. Hepatol.* 63, 102–113. doi: 10.1016/j.jhep.2015.01.031

European Association for the Study of the Liver (EASL), (2018). EASL clinical practice guidelines: management of hepatocellular carcinoma. *J. Hepatol.* 69, 182–236. doi: 10.1016/j.jhep.2018.03.019

Fernandez-Moreira, D., Ugalde, C., Smeets, R., Rodenburg, R. J., Lopez-Laso, E., Ruiz-Falco, M. L., et al. (2007). X-linked NDUFA1 gene mutations associated with mitochondrial encephalomyopathy. *Ann. Neurol.* 61, 73–83.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:2004088.

Giannelli, G., Koudelkova, P., Dituri, F., and Mikulits, W. (2016). Role of epithelial to mesenchymal transition in hepatocellular carcinoma. *J. Hepatol.* 65, 798–808.

Goellner, G. M., DeMarco, S. J., and Strehler, E. E. (2003). Characterization of PISP, a novel single-PDZ protein that binds to all plasma membrane Ca2+-ATPase b-splice variants. *Ann. N. Y. Acad. Sci.* 986, 461–471. doi: 10.1111/j.1749-6632.2003.tb07230.x

Guerrera, D., Shah, J., Vasileva, E., Sluysmans, S., Méan, I., Jond, L., et al. (2016). PLEKHA7 Recruits PDZD11 to adherens junctions to stabilize nectins. *J. Biol. Chem.* 291, 11016–11029. doi: 10.1074/jbc.M115.712935

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, I. T., Dhungel, B., Shrestha, R., Bridle, K. R., Crawford, D. H. G., Jayachandran, A., et al. (2019). Spotlight on bortezomib: potential in the treatment of hepatocellular carcinoma. *Expert Opin. Investig. Drugs* 28, 7–18. doi: 10.1080/13543784.2019.1551359

Jemal, A., Ward, E. M., Johnson, C. J., Cronin, K. A., Ma, J., Ryerson, B., et al. (2017). Annual report to the nation on the status of cancer, 1975-2014, featuring survival. *J. Natl. Cancer Inst.* 109:djx030.

Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., et al. (2019). Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 567, 257–261.

Jung, K. H., Kim, J. K., Noh, J. H., Eun, J. W., Bae, H. J., Xie, H. J., et al. (2010). Targeted disruption of Nemo-like kinase inhibits tumor cell growth by simultaneous suppression of cyclin D1 and CDK2 in human hepatocellular carcinoma. *J. Cell Biochem.* 110, 687–696. doi: 10.1002/jcb.22579

Kim, J., and Bae, J. S. (2016). Tumor-associated macrophages and neutrophils in tumor microenvironment. *Mediators Inflamm.* 2016:6058147. doi: 10.1155/2016/6058147

Konopka, G., Tekiela, J., Iverson, M., Wells, C., and Duncan, S. A. (2007). Junctional adhesion molecule-A is critical for the formation of pseudocanaliculi and modulates E-cadherin expression in hepatic cells. *J. Biol. Chem.* 282, 28137–28148. doi: 10.1074/jbc.M703592200

Lara-Pezzi, E., Roche, S., Andrisani, O. M., Sánchez-Madrid, F., and López-Cabrera, M. (2001). The hepatitis B virus HBx protein induces adherens junction disruption in a src-dependent manner. *Oncogene* 20, 3323–3331. doi: 10.1038/sj.onc.1204451

Lee, Y. R., Khan, K., Armfield-Uhas, K., Srikanth, S., Thompson, N. A., Pardo, M., et al. (2020). Mutations in FAM50A suggest that Armfield XLID syndrome is a spliceosomopathy. *Nat. Commun.* 11:3698. doi: 10.1038/s41467-020-17452-6

Li, R., Liu, H., Cao, Y., Wang, J., Chen, Y., Qi, Y., et al. (2020). Identification and validation of an immunogenic subtype of gastric cancer with abundant intratumoural CD103(+)CD8(+) T cells conferring favourable prognosis. *Br. J. Cancer* 122, 1525–1534. doi: 10.1038/s41416-020-0813-y

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.

Liu, C., Liu, L., Chen, X., Cheng, J., Zhang, H., Shen, J., et al. (2016). Sox9 regulates self-renewal and tumorigenicity by promoting symmetrical cell division of cancer stem cells in hepatocellular carcinoma. *Hepatology* 64, 117–129. doi: 10.1002/hep.28509

Liu, S. H., Shen, P. C., Chen, C. Y., Hsu, A. N., Cho, Y. C., Lai, Y. L., et al. (2020). DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.* 48, D863–D870.

Liu, W. R., Tian, M. X., Tang, Z., Fang, Y., Zhou, Y. F., Song, S. S., et al. (2020). Nine-factor-based immunohistochemistry classifier predicts recurrence for early-stage hepatocellular carcinoma after curative resection. *Br. J. Cancer* 123, 92–100. doi: 10.1038/s41416-020-0864-0

Lu, L., Qiu, C., Li, D., Bai, G., Liang, J., and Yang, Q. (2016). MicroRNA-505 suppresses proliferation and invasion in hepatoma cells by directly targeting high-mobility group box 1. *Life Sci.* 157, 12–18. doi: 10.1016/j.lfs.2016.05.039

Nabokina, S. M., Subramanian, V. S., and Said, H. M. (2011). Association of PDZ-containing protein PDZD11 with the human sodium-dependent multivitamin transporter. *Am. J. Physiol. Gastrointest Liver Physiol.* 300:23.

Okajima, W., Komatsu, S., Ichikawa, D., Miyamae, M., Ohashi, T., Imamura, T., et al. (2017). Liquid biopsy in patients with hepatocellular carcinoma: circulating tumor cells and cell-free nucleic acids. *World J. Gastroenterol.* 23, 5650–5668. doi: 10.3748/wjg.v23.i31.5650

Park, S. Y., Shin, J. H., and Kee, S. H. (2017). E-cadherin expression increases cell proliferation by regulating energy metabolism through nuclear factor-κB in AGS cells. *Cancer Sci.* 108, 1769–1777. doi: 10.1111/cas.13321

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9, 166–180. doi: 10.1593/neo.07112

Rouaud, F., Tessaro, F., Aimaretti, L., Scapozza, L., and Citi, S. (2020). Cooperative binding of the tandem WW domains of PLEKHA7 to PDZD11 promotes conformation-dependent interaction with tetraspanin 33. *J. Biol. Chem.* 295, 9299–9312. doi: 10.1074/jbc.RA120.012987

Schmidt, K., Ralle, M., Schaffer, T., Jayakanthan, S., Bari, B., Muchenditsi, A., et al. (2018). ATP7A and ATP7B copper transporters have distinct functions in the regulation of neuronal dopamine-β-hydroxylase. *J. Biol. Chem.* 293, 20085–20098. doi: 10.1074/jbc.ra118.004889

Shah, J., Guerrera, D., Vasileva, E., Sluysmans, S., Bertels, E., and Citi, S. (2016). PLEKHA7: cytoskeletal adaptor protein at center stage in junctional organization and signaling. *Int. J. Biochem. Cell Biol.* 75, 112–116. doi: 10.1016/j.biocel.2016.04.001

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

Stephenson, S. E., Dubach, D., Lim, C. M., Mercer, J. F., and La Fontaine, S. (2005). A single PDZ domain protein interacts with the Menkes copper ATPase, ATP7A. a new protein implicated in copper homeostasis. *J. Biol. Chem.* 280, 33270–33279. doi: 10.1074/jbc.m505889200

Tang, Z., Li, C., Kang, B., Gao, G., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102.

Thiery, J. P., Acloque, H., Huang, R. Y., and Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *Cell* 139, 871–890. doi: 10.1016/j.cell.2009.11.007

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963.

Vasileva, E., Sluysmans, S., Bochaton-Piallat, M. L., and Citi, S. (2017). Cell-specific diversity in the expression and organization of cytoplasmic plaque proteins of apical junctions. *Ann. N. Y. Acad. Sci.* 1405, 160–176. doi: 10.1111/nyas.13391

Villanueva, A. (2019). Hepatocellular carcinoma. *N. Engl. J. Med.* 380, 1450–1462.

von Mering, C., 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

Wan, L. C., Maisonneuve, P., Szilard, R. K., Lambert, J. P., Ng, T. F., Manczyk, N., et al. (2017). Proteomic analysis of the human KEOPS complex identifies C14ORF142 as a core subunit homologous to yeast Gon7. *Nucleic Acids Res.* 45, 805–817. doi: 10.1093/nar/gkw1181

Wang, W., Lin, H., Zhou, L., Zhu, Q., Gao, S., Xie, H., et al. (2014). MicroRNA-30a-3p inhibits tumor proliferation, invasiveness and metastasis and is downregulated in hepatocellular carcinoma. *Eur. J. Surg. Oncol.* 40, 1586–1594. doi: 10.1016/j.ejso.2013.11.008

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220.

Yamada, S., Okumura, N., Wei, L., Fuchs, B. C., Fujii, T., Sugimoto, H., et al. (2014). Epithelial to mesenchymal transition is associated with shorter disease-free survival in hepatocellular carcinoma. *Ann. Surg. Oncol.* 21, 3882–3890. doi: 10.1245/s10434-014-3779-2

Yan, L., Xu, F., and Dai, C. L. (2018). Relationship between epithelial-to-mesenchymal transition and the inflammatory microenvironment of hepatocellular carcinoma. *J. Exp. Clin. Cancer Res.* 37:203.

Yang, Z., Liu, S., Zhu, M., Zhang, H., Wang, J., Xu, Q., et al. (2016). PS341 inhibits hepatocellular and colorectal cancer cells through the FOXO3/CTNNB1 signaling pathway. *Sci. Rep.* 6:22090. doi: 10.1038/srep22090

**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: 03 June 2021
doi: 10.3389/fgene.2021.671585

# Upregulation of LIMK1 Is Correlated With Poor Prognosis and Immune Infiltrates in Lung Adenocarcinoma

*Guojun Lu[1†], Ying Zhou[2†], Chenxi Zhang[2] and Yu Zhang[1*]*

[1] Department of Respiratory Medicine, Nanjing Chest Hospital, Affiliated Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China, [2] Central Laboratory, Nanjing Chest Hospital, Affiliated Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China

**Background:** Protein-coding gene LIM Domain Kinase 1 (*LIMK1*) is upregulated in various tumors and reported to promote tumor invasion and metastasis. However, the prognostic values of *LIMK1* and correlation with immune infiltrates in lung adenocarcinoma are still not understood. Therefore, we evaluated the prognostic role of *LIMK1* and its correlation with immune infiltrates in lung adenocarcinoma.

**Methods:** Transcriptional expression profiles of *LIMK1* between lung adenocarcinoma tissues and normal tissues were downloaded from the Cancer Genome Atlas (TCGA). The LIMK1 protein expression was assessed by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and the Human Protein Atlas. Receiver operating characteristic (ROC) curve was used to differentiate lung adenocarcinoma from adjacent normal tissues. Kaplan-Meier method was conducted to assess the effect of *LIMK1* on survival. Protein-protein interaction (PPI) networks were constructed by the STRING. Functional enrichment analyses were performed using the "ClusterProfiler" package. The relationship between *LIMK1* mRNA expression and immune infiltrates was determined by tumor immune estimation resource (TIMER) and tumor-immune system interaction database (TISIDB).

**Results:** The expression of *LIMK1* in lung adenocarcinoma tissues was significantly upregulated than those in adjacent normal tissues. Increased *LIMK1* mRNA expression was associated with lymph node metastases and high TNM stage. The ROC curve analysis showed that with a cutoff level of 4.908, the accuracy, sensitivity, and specificity for *LIMK1* differentiate lung adenocarcinoma from adjacent controls were 69.5, 93.2, and 71.9%, respectively. Kaplan-Meier survival analysis showed lung adenocarcinoma patients with high- *LIMK1* had a worse prognosis than those with low- *LIMK1* (43.1 vs. 55.1 months, $P = 0.028$). Correlation analysis indicated *LIMK1* mRNA expression was correlated with tumor purity and immune infiltrates.

**Conclusion:** Upregulated *LIMK1* is significantly correlated with poor survival and immune infiltrates in lung adenocarcinoma. Our study suggests that *LIMK1* can be used as a biomarker of poor prognosis and potential immune therapy target in lung adenocarcinoma.

**Keywords: lung adenocarcinoma, *LIMK1*, LIM domain kinase1, biomarker, prognosis, immune infiltrates**

OPEN ACCESS

***Edited by:***
*Suman Ghosal, National Institutes of Health (NIH), United States*

***Reviewed by:***
*Enrique Medina-Acosta, State University of the North Fluminense Darcy Ribeiro, Brazil*
*Vijaykumar Muley, Universidad Nacional Autónoma de México, Mexico*

***Correspondence:**
*Yu Zhang*
*zhangyu2113_nj@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:*** *24 February 2021*
***Accepted:*** *06 May 2021*
***Published:*** *03 June 2021*

***Citation:***
*Lu G, Zhou Y, Zhang C and Zhang Y (2021) Upregulation of LIMK1 Is Correlated With Poor Prognosis and Immune Infiltrates in Lung Adenocarcinoma. Front. Genet. 12:671585. doi: 10.3389/fgene.2021.671585*

## INTRODUCTION

Lung cancer is one of the most common malignant tumors around the world and the leading cause for cancer-related death (Jemal et al., 2011). The incidence of lung cancer has steadily increased over recent years. Lung cancer remains refractory and the 5-year survival rate continues to be the lowest among the major cancers. It is speculated that numerous people will be diagnosed with lung cancer in the future, which bring a heavy economic burden to our society (Torre et al., 2016; Albaba et al., 2017). In the subtypes of lung cancer, lung adenocarcinoma accounts for about 50% (Brustugun et al., 2018). Despite many therapeutic endeavors has been made in lung adenocarcinoma, such as targeted therapy and immunotherapy, the survival rate remains bleak and staggers at about 20% 5 years after treatment (Hirsch et al., 2017). Thus, it is imperative to search novel biomarkers for advancing the prognosis of lung adenocarcinoma.

LIM Domain Kinase 1 (*LIMK1*) is a protein known as a member of the LIM kinase protein family. *LIMK1* is consisted of gene spans 39,499 base pairs with 16 exons and encoded by a gene located on human chromosome 7q11.23 (Scott and Olson, 2007). Through phosphorylation and inactivation to its downstream effector of cofilin, *LIMK1* has been shown to be important in regulating the polymerization of actin (Liu et al., 2019). When *LIMK1* is phosphorylated, cofilin loses the ability to bind to actin, leading to the accumulation of actin polymers dysregulation of actin-mediated cytoskeletal changes (Nishimura et al., 2006). The phosphorylation of *LIMK1* has been implicated with many cellular functions including angiogenesis, proliferation, cell cycle, and metastasis progression (Foletta et al., 2004; Nishimura et al., 2006). Previous studies have confirmed that ectopic expression of *LIMK1* was associated with the progression of several tumor types, such as colorectal cancer, gastric cancer, prostate cancer, and breast cancer (Davila et al., 2003; McConnell et al., 2011; You et al., 2015; Liao et al., 2017). A paper from Huang et al. (2020) indicated that the upregulation of *LIMK1* is correlated with lymph node metastasis and poor biochemical-free survival in prostate cancer. In pancreatic cancer, Vlecken and Bagowski (2009) reported that knockdown of *LIMK1* can lead to an inhibition of invasion and metastatic behavior, as well as suppression of pancreatic cancer cell-induced angiogenesis. Moreover, some recent findings suggested that downregulation of *LIMK1* can inhibit lung cancer cell migration (Chen et al., 2013; Wan et al., 2014; Zhang et al., 2020). Thus, *LIMK1* has great potential to be a biomarker of poor prognosis and therapeutic target for lung cancer.

The prognostic values and correlation with immune infiltrates of *LIMK1* in lung adenocarcinoma are still not fully understood. Given the overexpression of *LIMK1* in lung cancer and the downregulation of *LIMK1* can inhibit lung cancer cell migration, we hypothesized that the level of *LIMK1* is associated with survival in lung adenocarcinoma. To test this hypothesis, we evaluated the prognostic role of *LIMK1* in lung adenocarcinoma based on data from The Cancer Genome Atlas (TCGA). In this study, we found that *LIMK1* is upregulated in lung adenocarcinoma. Significantly, the upregulation of *LIMK1* is correlated with poor clinical characteristics and risk factors. We further evaluated the diagnostic and prognostic values, the correlation with immune infiltrates of *LIMK1* for lung adenocarcinoma. Our study links the overexpression of *LIMK1* and poor survival in lung adenocarcinoma.

## MATERIALS AND METHODS

### TCGA Datasets

Transcriptional expression data of *LIMK1* and corresponding clinical information were downloaded from TCGA official website[1] (Tomczak et al., 2015). The 18 enrolled cancer types contained at least 5 samples in the normal group. Finally, the RNA-Seq gene expression data with workflow type of FPKM was transformed into TPM format and log2 conversion for further study. Since all the data were downloaded from TCGA, this study did not need approval from the Ethics Committee.

### RNA-Sequencing Data of *LIMK1* in Lung Adenocarcinoma

The RNA-Seq expression data of *LIMK1* in lung adenocarcinoma was also downloaded from TCGA. Therefore, 535 lung adenocarcinoma and 59 adjacent normal tissue data were retained. The samples selected contained *LIMK1* gene expression data and associated clinical information, including age, gender, smoker condition, T stage, N stage, M stage, and tumor location. The mRNA expression data were characterized by mean ± *SD*.

### Clinical Proteomic Tumor Analysis Consortium (CPTAC) and UALCAN

With the application of proteomic technologies, CPTAC[2] analyzes tumor biospecimens using mass spectrometry, quantifying and identifying the constituent proteins and characterizing proteome of each tumor sample (Edwards et al., 2015). UALCAN[3] is a user-friendly online web resource for analyzing publicly available cancer data (Chandrashekar et al., 2017). In this study, we performed UALCAN to present a throughout analysis of LIMK1 protein expression from CPTAC.

### The Human Protein Atlas (HPA)

HPA[4] contains normal tissues and tumor tissues information regarding the expression profiles of human genes on protein level (Uhlen et al., 2015, 2017). In this study, we conducted HPA to compare the protein expression of LIMK1 between normal lung tissue and lung adenocarcinoma tissue.

### Protein-Protein Interaction (PPI) Networks and Functional Enrichment Analysis

STRING is an online database for the retrieval of interacting genes (version 11.0[5]; Szklarczyk et al., 2011). In this study,

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

[2]https://proteomics.cancer.gov/programs/cptac

[3]http://ualcan.path.uab.edu/

[4]https://proteinatlas.org/

[5]http://string-db.org

we conducted STRING to search co-expression genes and construct PPI networks with an interaction score >0.4. Gene ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses of co-expression genes were performed by the "ClusterProfiler" package and visualized by the "ggplot2" package (Wickham, 2016; Yu et al., 2012).

## Tumor Immune Estimation Resource (TIMER) Database

TIMER is a comprehensive online resource for systematic analysis of immune infiltrates across various cancer types[6] (Li et al., 2017). In this study, we performed TIMER to determine the relationship between *LIMK1* expression in lung adenocarcinoma and six immune infiltrates (B cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils, macrophages, and dendritic cells).

## Tumor-Immune System Interaction Database (TISIDB)

TISIDB[7] is an online web integrated repository portal for tumor-immune system interaction (Ru et al., 2019). In this study, we performed TISIDB to determine the expression of *LIMK1* and tumor-infiltrating lymphocytes (TILs) across human cancers. Based on the gene expression profile, the relative abundance of TILs was inferred by using gene set variation analysis. The correlations between *LIMK1* and TILs were measured by Spearman's test.

## PrognoScan Database

PrognoScan database[8] is a powerful online platform to evaluate the correlation between gene expression and survival across various types of cancers (Mizuno et al., 2009). In this study, we performed PrognoScan database to analyze the correlation between *LIMK1* expression and overall survival in lung adenocarcinoma with two different datasets (jacob-00182-CANDF, jacob-00182-MSK).

## Statistical Analyses

All statistical analyses were performed with R (V 3.6.3)[9] and R package ggplot2 was used to visualize expression differences. Paired *t*-test and Mann-Whitney *U*-test were used to determine the differences between lung adenocarcinoma tissues and adjacent normal tissues. ROC curve was performed to detect the cutoff value of *LIMK1* using the pROC package (Robin et al., 2011). Kaplan-Meier and log-rank tests were conducted with the survminer package[10] to assess the effect of *LIMK1* on survival.

[6]https://cistrome.shinyapps.io/timer/

[7]http://cis.hku.hk/TISIDB/

[8]http://dna00.bio.kyutech.ac.jp/PrognoScan/index.html

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

[10]https://CRAN.R-project.org/package=survminer

# RESULTS

## Expression Pattern of *LIMK1* in Pan-Cancer Perspective

To evaluate the mRNA expression pattern of *LIMK1* across different cancer types, we excluded from the analysis the datasets from 15 cancer types that contained less than five samples in the normal group. The final working set refers to 18 cancer types. As shown in **Figure 1**, compared with normal tissues, *LIMK1* was significantly upregulated in 16 of all 18 cancer types. This data indicated the mRNA expression of *LIMK1* was abnormally expressed across different cancer types.

## Upregulated mRNA and Protein Expression of LIMK1 in Patients With Lung Adenocarcinoma

To determine the mRNA and protein expression of LIMK1 in lung adenocarcinoma, the LIMK1 expression data from TCGA and HPA were analyzed. The baseline characteristics of lung adenocarcinoma patients from TCGA were listed in **Supplementary Table 1**. As shown in **Figure 2A**, paired data analysis showed that the mRNA expression levels of *LIMK1* in lung adenocarcinoma tissues ($n$ = 57) were significantly higher than those in adjacent normal tissues ($n$ = 57) (**Figure 2A**, 5.584 ± 0.747 vs. 4.320 ± 0.442, $P < 0.001$). Unpaired data analyses also showed that the mRNA expression levels of *LIMK1* in lung adenocarcinoma tissues ($n$ = 535) were significantly higher than those in adjacent normal tissues ($n$ = 59) (**Figure 2B**, 5.314 ± 0.847 vs. 4.324 ± 0.437, Mann-Whitney *U*-test, $P < 0.001$). To present a throughout analysis of LIMK1 protein expression, we performed analysis on CPTAC with UALCAN. The result showed that the protein expression of LIMK1 in lung adenocarcinoma was significantly higher than those in normal tissues (**Figure 2C**). As shown in **Figure 2D**, immunohistochemical staining from HPA also revealed LIMK1 protein was upregulated in lung adenocarcinoma tissue. These results indicated that both mRNA and protein expression of LIMK1 are upregulated in lung adenocarcinoma tissues.

## Relationships Between *LIMK1* mRNA Levels and Clinical Pathological Characteristics of Lung Adenocarcinoma Patients

To evaluate the association between the mRNA expression of *LIMK1* and clinical pathological characteristics of lung adenocarcinoma samples, we performed Mann-Whitney *U*-test and logistic regression analysis. As shown in **Table 1** and **Figures 3A–I**, higher expression levels of *LIMK1* were observed in male patients ($P$ = 0.004), patients with lymph node metastases ($P$ = 0.022), and patients with high TNM stage ($P$ = 0.048). However, no statistically significant correlation were found between the expression levels of *LIMK1* and other clinical pathological characteristics, such as age ($P$ = 0.113), smoker ($P$ = 0.270), T stage ($P$ = 0.129), M stage ($P$ = 0.921), and anatomic subdivision (right vs. left, $P$ = 0.959; peripheral vs.

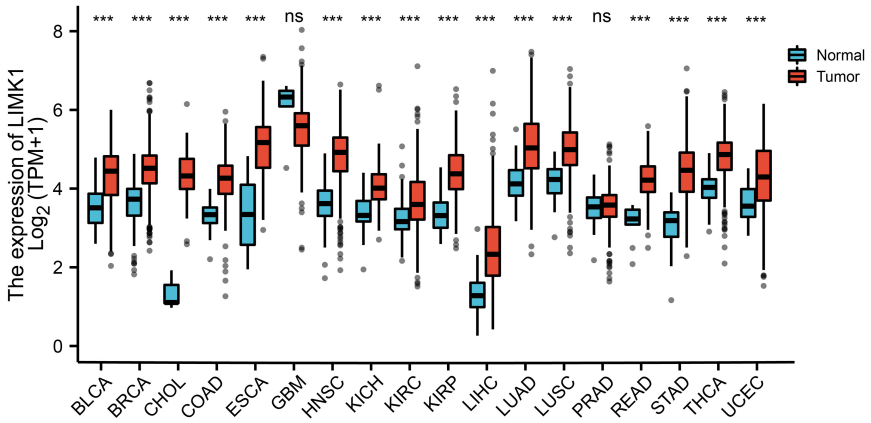



**FIGURE 1 |** Expression pattern of *LIMK1* in Pan-cancer perspective. The mRNA expression of *LIMK1* was upregulated in 16 of 18 cancer types compared with normal tissues. (***$P < 0.001$). ns, no significance; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma mutiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma.

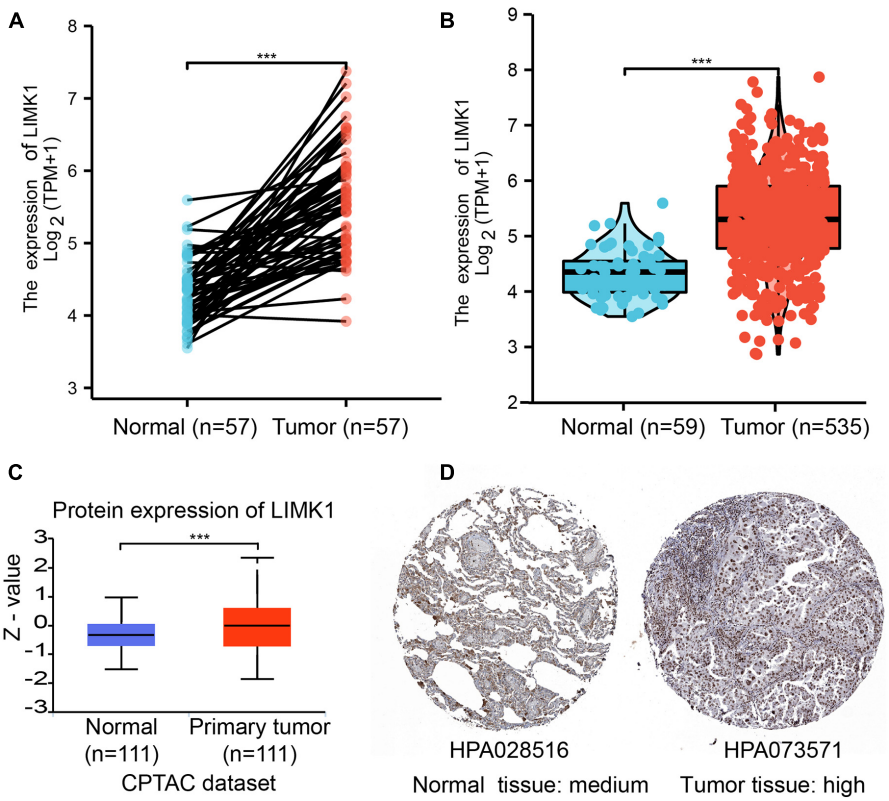



**FIGURE 2 |** The mRNA and protein expression of LIMK1 in lung adenocarcinoma. **(A)** The mRNA expression levels of *LIMK1* in 57 lung adenocarcinoma and matched-adjacent normal samples. **(B)** The mRNA expression levels of *LIMK1* in 535 lung adenocarcinoma samples and 59 normal samples. **(C)** The protein expression levels of LIMK1 based on CPTAC. **(D)** The protein levels of LIMK1 based on Human Protein Atlas. Normal tissue, https://www.proteinatlas.org/ENSG00000106683-LIMK1/tissue/lung#img; Tumor tissue, https://www.proteinatlas.org/ENSG00000106683-LIMK1/pathology/lung+cancer#img (***$P < 0.001$).

central, $P = 0.562$). Taken together, these results suggested that *LIMK1* is correlated with lymph node metastases and high TNM stage, further suggesting *LIMK1* may act as a biomarker of poor prognosis for lung adenocarcinoma.

**TABLE 1 |** Clinical characteristics of the lung adenocarcinoma patients (TCGA).

| Characteristics | Total | Low expression | High expression | *P*-value |
|---|---|---|---|---|
| | *N* (%) | *N* (%) | *N* (%) | |
| **T stage** | | | | 0.199 |
| T1 | 175 (32.9%) | 97 (18.2%) | 78 (14.7%) | |
| T2 | 289 (54.3%) | 140 (26.3%) | 149 (28%) | |
| T3 | 49 (9.2%) | 21 (3.9%) | 28 (5.3%) | |
| T4 | 19 (3.6%) | 7 (1.3%) | 12 (2.3%) | |
| **N stage** | | | | 0.006** |
| N0 | 348 (67.0%) | 188 (36.2%) | 160 (30.8%) | |
| N1 | 95 (18.3%) | 39 (7.5%) | 56 (10.8%) | |
| N2 | 74 (14.3%) | 28 (5.4%) | 46 (8.9%) | |
| N3 | 2 (0.4%) | 0 (0%) | 2 (0.4%) | |
| **M stage** | | | | 1.000 |
| M0 | 361 (93.5%) | 184 (47.7%) | 177 (45.9%) | |
| M1 | 25 (6.5%) | 13 (3.4%) | 12 (3.1%) | |
| **Pathologic stage** | | | | 0.002** |
| Stage I | 294 (55.8%) | 165 (31.3%) | 129 (24.5%) | |
| Stage II | 123 (23.3%) | 50 (9.5%) | 73 (13.9%) | |
| Stage III | 84 (16.0%) | 31 (5.9%) | 53 (10.1%) | |
| Stage IV | 26 (4.9%) | 14 (2.7%) | 12 (2.3%) | |
| **Gender** | | | | 0.005** |
| Female | 286 (53.5) | 126 (23.6%) | 160 (29.9%) | |
| Male | 249 (46.5) | 141 (26.4%) | 108 (20.2%) | |
| **Age** | | | | 0.134 |
| < = 65 | 255(49.4) | 118 (22.9%) | 137 (26.6%) | |
| >65 | 261(50.6) | 139 (26.9%) | 122 (23.6%) | |
| **Smoker** | | | | 0.327 |
| No | 75(14.4) | 33 (6.3%) | 42 (8.1%) | |
| Yes | 446(85.6) | 227 (43.6%) | 219 (42%) | |
| **Anatomic neoplasm subdivision** | | | | 1.000 |
| Left | 205(39.4) | 102 (19.6%) | 103 (19.8%) | |
| Right | 315 (60.6) | 156 (30%) | 159 (30.6%) | |
| **Anatomic neoplasm subdivision 2** | | | | 0.671 |
| Central lung | 62 (32.8) | 27 (14.3%) | 35 (18.5%) | |
| Peripheral lung | 127 (67.2) | 61 (32.3%) | 66 (34.9%) | |

*$^{**}P < 0.01$.*

## Differential RNA-Seq Levels of *LIMK1* as a Prospective Biomarker to Distinguish Lung Adenocarcinoma Samples From Normal Samples

To investigate the value for *LIMK1* to distinguish lung adenocarcinoma samples from normal smples, we performed a ROC curve analysis. As showed in **Figure 4A**, the ROC curve analysis showed *LIMK1* had an AUC value of 0.851 (95% CI: 0.813–0.888). At a cutoff of 4.908, *LIMK1* had a sensitivity, specificity, and accuracy of 69.5, 93.2, and 71.9%, respectively. The positive predictive value was 98.9% and the negative predictive value was 25.2%. These findings indicated that *LIMK1* could be a promising biomarker to differentiate lung adenocarcinoma tissues from normal tissues.

## High mRNA Expression of *LIMK1* Is Associated With Short OS

To explore the relationship between *LIMK1* mRNA expression and OS in lung adenocarcinoma patients, Kaplan-Meier curves and PrognoScan database were performed. As shown in **Figure 4B**, the OS of lung adenocarcinoma patients with high-level of *LIMK1* was significantly shorter than those with low-level of *LIMK1* (43.1 vs. 55.1 months, $P = 0.028$). PrognoScan result with two different datasets (**Supplementary Figure 1**) also indicated that high expression of *LIMK1* was correlated with poor overall survival in lung adenocarcinoma. These data indicated that high mRNA expression of *LIMK1* is a biomarker of poor prognosis in lung adenocarcinoma.

## PPI Networks and Functional Annotations

To construct PPI networks and functional annotations, we conducted STRING database, GO, and KEGG analyses. **Figure 5A** showed a network of *LIMK1* and its 10 co-expression genes. As shown in **Figure 5B**, changes in the biological process of *LIMK1* were associated with actin filament organization, regulation of actin filament-based process, and actin cytoskeleton organization. Functional annotations indicated that these genes were involved in purine ribonucleoside binding, GTP Binding, and GTPase activity. The correlation analyses between the expression of *LIMK1* and co-expressed genes in lung adenocarcinoma from TCGA were shown in **Figures 5C–I**.

## Correlation Analysis Between *LIMK1* Expression and Immune Cell Infiltration in Lung Adenocarcinoma

We analyzed the correlation between *LIMK1* expression and the six types of tumor infiltrating immune cells in the TIMER database. As shown in **Figure 6A**, *LIMK1* expression had correlations with tumor purity ($r = -0.189$, $P = 2.37e\text{-}05$), $CD4^+$ T cell ($r = 0.285$, $P = 1.65e\text{-}10$), macrophage ($r = 0.143$, $P = 1.64e\text{-}03$), neutrophil ($r = 0.263$, $P = 4.53e\text{-}09$), dendritic cell ($r = 0.363$, $P = 1.20e\text{-}16$). We also evaluated the correlation between *LIMK1* expression and 28 types of TILs in the TISIDB database. **Figure 6B** shown the relations between expression of *LIMK1* and 28 types of TILs across human cancers. As shown in **Figure 6C**, the expression of *LIMK1* was correlated with abundance of $CD8^+$ T cells ($r = 0.401$, $P = 2.2e\text{-}16$), $CD4^+$ T cells ($r = 0.317$, $P = 1.92e\text{-}16$), monocyte cells ($r = 0.289$, $P = 2.71e\text{-}11$), treg cells ($r = 0.289$, $P = 4.41e\text{-}11$), CD56dim cells ($r = 0.275$, $P = 2.31e\text{-}10$), and myeloid derived suppressor cells (MDSC, $r = 0.275$, $P = 2.41e\text{-}10$). These data indicated that *LIMK1* may play a specific role in immune infiltration in lung adenocarcinoma.

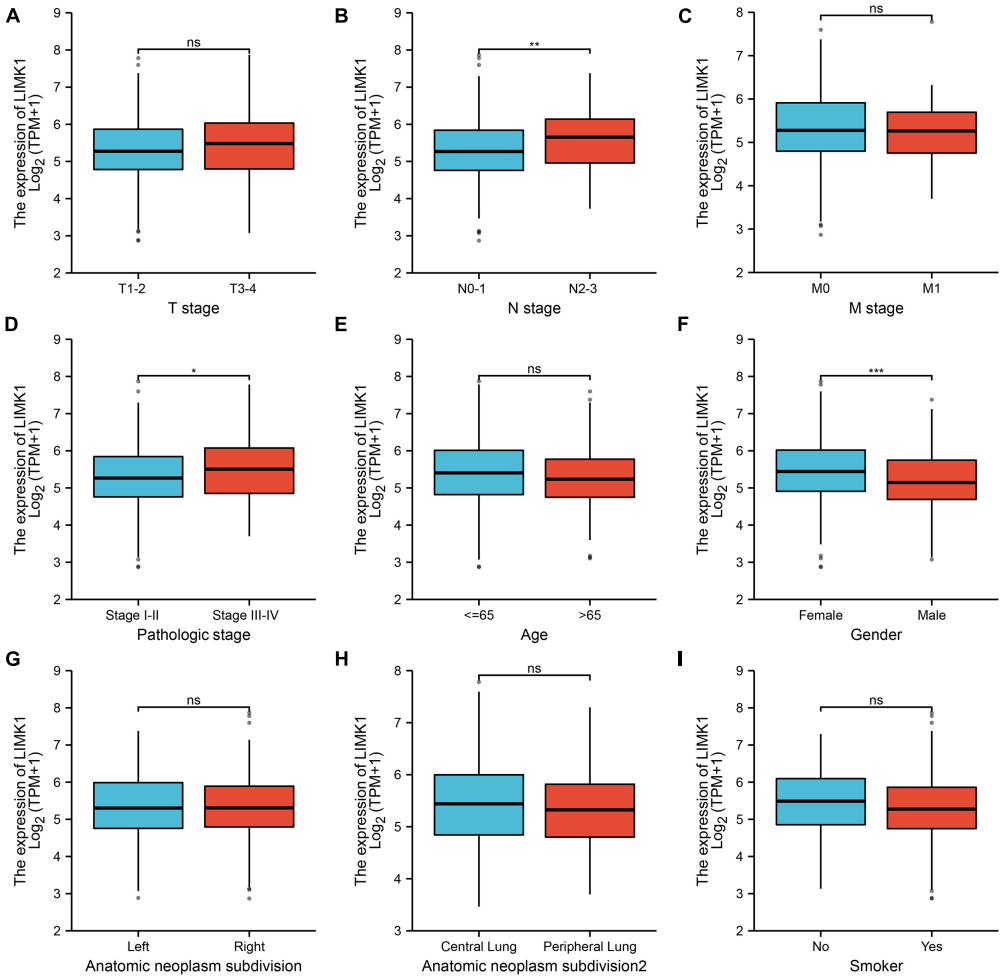



**FIGURE 3 |** Relationships between *LIMK1* mRNA levels and clinical pathological characteristics. *LIMK1* mRNA expression was significantly correlated with lymph node metastases **(B)**, high TNM stage **(D)** and gender was male **(F)**. However, no statistically significant correlation were found between the expression levels of *LIMK1* and T stage **(A)**, M stage **(C)**, age **(E)**, anatomic neoplasm subdivision **(G,H)** and smoke condition **(I)** (ns, no significance, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$).

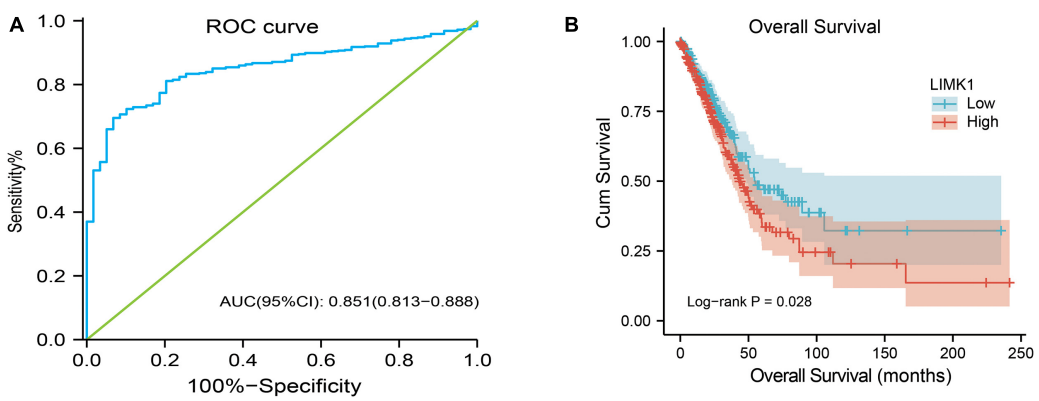



**FIGURE 4 |** ROC and Kaplan-Meier curves for *LIMK1*. **(A)** ROC curve showed that *LIMK1* had an AUC value of 0.851 to discriminate lung adenocarcinoma tissues from healthy controls. With a cutoff of 4.908, the sensitivity, specificity and accuracy were 93.2, 71.9, and 69.5%, respectively. **(B)** Kaplan-Meier survival curves indicated that lung adenocarcinoma patients with high *LIMK1* mRNA expression had a shorter OS than those with low-level of *LIMK1* (43.1 vs. 55.1 months, $P = 0.028$).

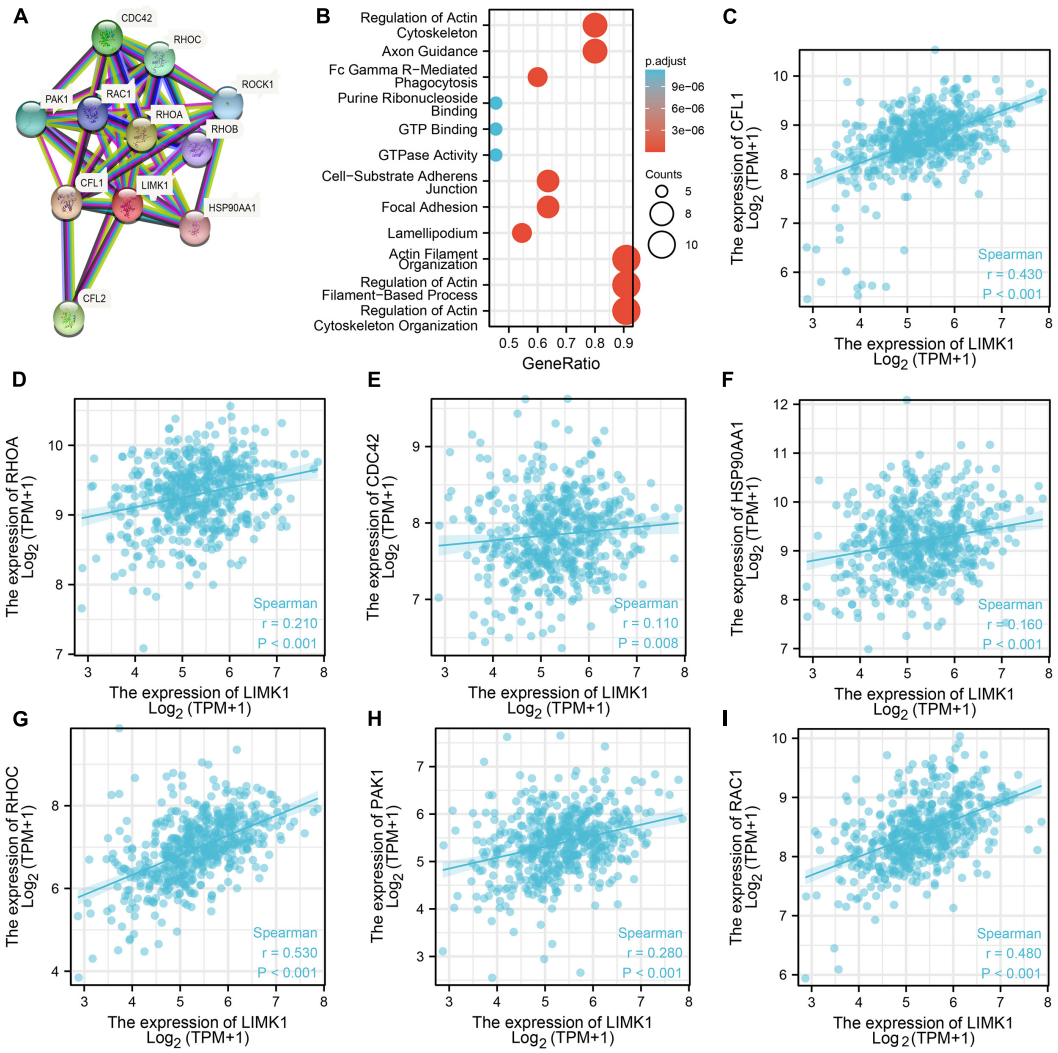



**FIGURE 5 |** PPI networks and functional enrichment analyses. **(A)** A network of *LIMK1* and its co-expression genes. **(B)** Functional enrichment analyses of 11 involved genes. *LIMK1* was associated with actin filament organization, regulation of actin filament-based process, and actin cytoskeleton organization. These genes were involved in purine ribonucleoside binding, GTP Binding, and GTPase Activity. **(C–I)** The correlation analyses between the expression of *LIMK1* and co-expressed genes in lung adenocarcinoma.CFL1, cofilin-1; RHOA, transforming protein RhoA; CFL2, cofilin-2; CDC42, cell division control protein 42 homolog; RHOC, Rho-related GTP-binding protein RhoC; PAK1, serine/threonine-protein kinase PAK 1; ROCK1, Rho-associated protein kinase 1; RAC1, Ras-related C3 botulinum toxin substrate 1; RHOB, Rho-related GTP-binding protein RhoB; HSP90AA1, heat shock protein HSP 90-alpha.

## DISCUSSION

In this study, we found that the mRNA expression of *LIMK1* is upregulated in lung adenocarcinoma tissues. The upregulated mRNA expression of *LIMK1* is positively correlated with lymph node metastases and high TNM stage. ROC curve analysis indicated that *LIMK1* could be a promising diagnostic biomarker to differentiate lung adenocarcinoma from normal tissues. In light of Kaplan-Meier curves and univariate analysis, we confirmed that high mRNA expression of *LIMK1* is associated with short OS and *LIMK1* can be used as a potential biomarker of poor prognosis for lung adenocarcinoma. Moreover, *LIMK1* may play a specific role in immune infiltration in lung adenocarcinoma.

*LIMK1* is one of the members of the LIM kinase family and has been reported to play a significant role in promoting cell invasion and metastasis (Scott and Olson, 2007). Many studies about the oncogenic role of *LIMK1* in several human cancers have been emerged in recent years, including gastric cancer, pancreatic cancer, as well as lung cancer (McConnell et al., 2011; Chen et al., 2013). Furthermore, it is reported that *LIMK1* is upregulated in various cancers and associates with an unfavorable prognosis (Huang et al., 2020). However, the expression of *LIMK1* and its prognostic value has not been fully investigated in lung adenocarcinoma. Here, in this study, based on pan-cancer analysis, our results are consistent with those reports that *LIMK1* mRNA is abnormally expressed in various cancers. We also confirmed that *LIMK1* is significantly upregulated in lung

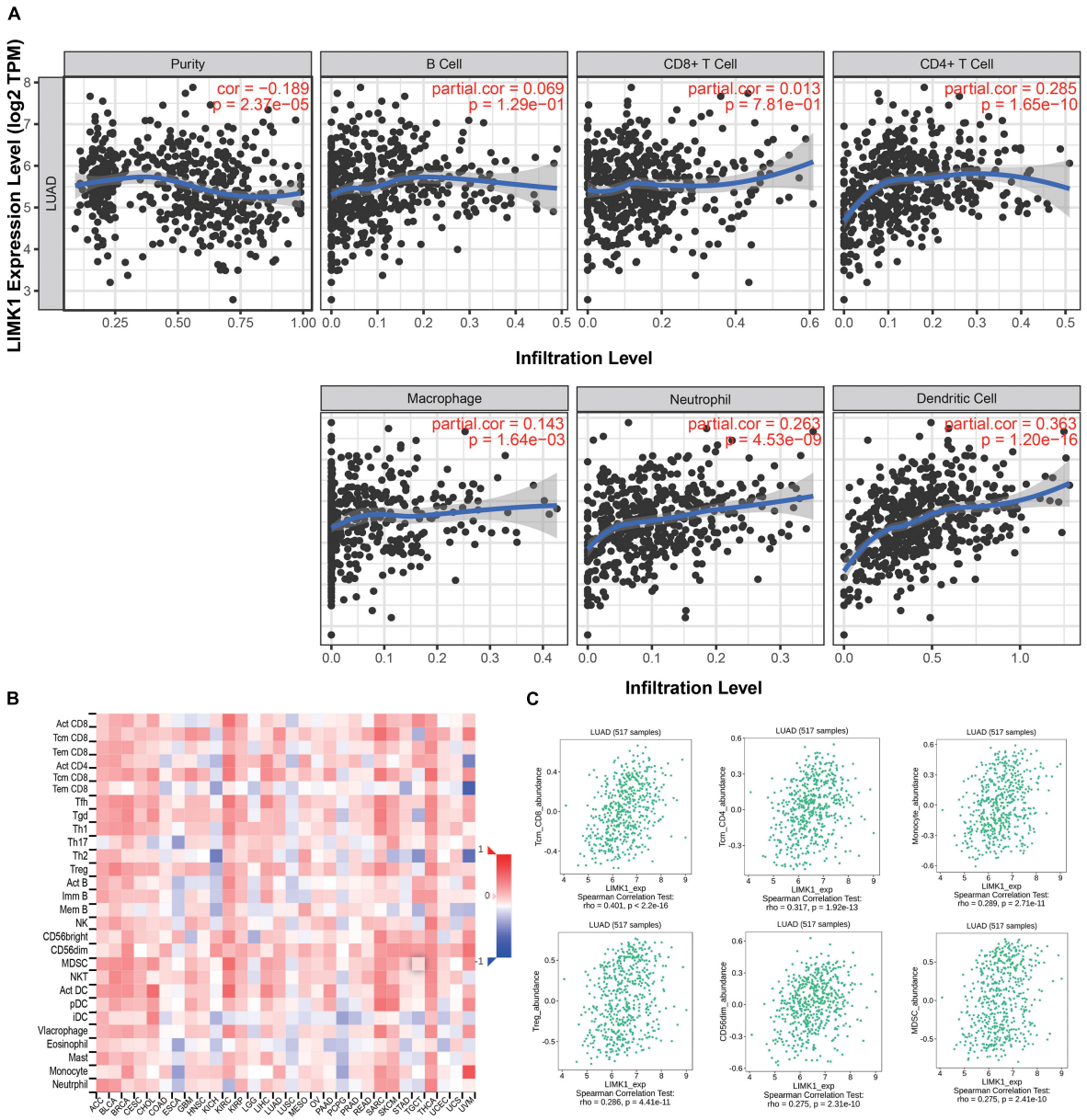



**FIGURE 6 |** Correlations of *LIMK1* expression with immune infiltration level. **(A)** *LIMK1* expression is negatively related to tumor purity and has correlations with dendritic cell, CD4$^+$ T cell, neutrophil, and macrophage in lung adenocarcinoma. **(B)** Relations between the expression of *LIMK1* and 28 types of TILs across human cancers. **(C)** *LIMK1* was correlated with abundance of CD8$^+$ T cells, CD4$^+$ T cells, monocyte cells, Treg cells, CD56dim cells, and MDSC cells.

adenocarcinoma. High mRNA expression of *LIMK1* is positively associated with lymph node metastases and high TNM stage, our finding agrees with the previous report by Chen et al. (2013). These findings suggest that *LIMK1* might act as a potential biomarker of poor prognosis to identify lung adenocarcinoma with poor clinical outcome.

Currently, the function of *LIMK1* in tumors had not been fully reported. Previous trials suggest that *LIMK1* may be a target of dasatinib which can inhibit *LIMK1* to suppress lung cancer cell proliferation and growth (Zhang et al., 2020). Other studies have shown *LIMK1* acts as a direct target of miRNA-27-3p and miRNA-128-3p (Chen et al., 2017; Zhao et al., 2019), both miRNA-27-3p and miRNA-128-3p can suppress cancer cell proliferation, migration, and invasion. The underlying mechanism analysis showed that the *LIMK1*-cofilin signaling pathway plays an important role in tumor progression (Nishimura et al., 2006). All these results suggest that *LIMK1* could be regarded as a promising biomarker or emerging target for cancer therapy. Given the condition that mRNA expression of *LIMK1* is significantly higher in lung adenocarcinoma than in normal lung tissues, we speculate *LIMK1* can act as a biomarker to differentiate lung adenocarcinoma from normal controls. In

order to validate the clinical value of *LIMK1* in the diagnosis of lung adenocarcinoma, we conducted ROC curve analysis. Our results showed that *LIMK1* had a significantly high AUC value in the detection of lung adenocarcinoma, with 69.5% in sensitivity, 93.2% in specificity, and 71.9% in accuracy. On the basis of our finding, we conclude that *LIMK1* might act as a potential diagnostic biomarker to differentiate lung adenocarcinoma from normal controls.

Recent studies have characterized *LIMK1* as an important biomarker for poor prognosis and associated upregulated mRNA expression of *LIMK1* with poor overall survival in many cancers. In prostate cancer, it is reported that elevated *LIMK1* is positively associated with higher Gleason Scores and incidence of metastasis, as well as poor clinical outcome and reduced survival (Davila et al., 2007; Mardilovich et al., 2015; Huang et al., 2020). In ovarian cancer, Zhang et al. (2012) demonstrated that overexpression of *LIMK1* is significant correlated with severity and poor differentiation level of ovarian cancer. A paper from Zhang et al. (2011) suggested that upregulation of *LIMK1* can promote the invasion and metastasis in drug-resistant osteosarcoma and in turn *LIMK1* can act as a potential novel therapeutic target. In glioblastoma, Chen et al. (2020) reported that *LIMK1* is increased and the overexpression of *LIMK1* is associated with high grade and poor prognosis. In contrast, suppression of *LIMK1* can prolong survival time. However, the prognostic value of *LIMK1* has not been investigated in lung adenocarcinoma. Given the upregulation of *LIMK1* is positively correlated with lymph node metastases and high TNM stage, we speculated *LIMK1* is involved in the development of lung adenocarcinoma. Moreover, since lymph node metastases and high TNM stage are correlated with poor survival, we speculated that the upregulation of *LIMK1* is a biomarker of poor prognosis. Furthermore, in light of Kaplan-Meier curves and log-rank test, lung adenocarcinoma patients with high mRNA expression of *LIMK1* are associated with a decreased survival rate than those with low *LIMK1* levels. On the basis of our data, we concluded that *LIMK1* can be used as a biomarker of poor prognosis for determining prognosis in lung adenocarcinoma.

*LIMK1* is a crucial component of Rac1/PAK1/LIMK1/cofilin signaling pathway, which is involved in several cancers. For example, in cervical cancer, miR-509-3p can regulate this pathway to enhance the apoptosis and chemo-sensitivity of cervical cancer cells (Xu et al., 2012). In gastric cancer, the inhibition of Rho GDP dissociation inhibitor 2 can suppress tumor cell migration and invasion via signaling pathway (Zeng et al., 2020). In this study, co-expression analyses indicated that the expression of *LIMK1* is significantly correlated to that of *Rac1*, *PAK1*, and *CLF1*. On the basis of our finding, we speculate that the upregulation of LIMK1 expression would affect the entire pathway. However, this should be tested in other experiments.

Many studies about the possible role of *LIMK1* in human TILs have emerged in recent years. Xu et al. (2012) reported that *LIMK1* may be involved in spontaneous actin polarization in transformed CD4 T cells. However, the correlation analysis between *LIMK1* expression and immune cell infiltration in lung adenocarcinoma has not been investigated. In this study, we found that several tumor infiltrating immune cells ($CD4^+$ T cell, macrophage, neutrophil, dendritic cell) were correlated with the expression of *LIMK1* in lung adenocarcinoma by using TIMER. We also found that positive correlation were indicated between *LIMK1* expression and $CD8^+$ T cells, $CD4^+$ T cells, monocyte cells, treg cells, CD56dim cells, and myeloid derived suppressor cells. These findings suggest that there is a potential correlation between *LIMK1* and immune infiltration in lung adenocarcinoma. However, further research should be designed to confirm this correlation.

There are several limitations in this study. First, the expression and prognostic implication of *LIMK1* were conducted with online public databases, further study with clinical samples is required to validate these results. Second, to further examine the detailed mechanism of the impact of *LIMK1* on immune infiltration in lung adenocarcinoma, *in vivo*/*vitro* experiments should be designed.

In conclusion, in this study, we showed for the first time that mRNA expression of *LIMK1* is upregulated in lung adenocarcinoma and positively correlated with lymph node metastases and high TNM stage. Our research suggests that *LIMK1* could be regarded as a potential biomarker of poor prognosis to identify lung adenocarcinoma patients with poor clinical outcomes and may play a specific role in immune infiltration.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## AUTHOR CONTRIBUTIONS

YuZ conceived and designed the study. GL performed data analysis and wrote the manuscript. YiZ and CZ contributed analysis tools. All authors reviewed the manuscript.

## FUNDING

This study was supported by the Science and Technology Development Fund of Nanjing Medical University (NMUB2019112), "The 13TH Five-Year Plan" Major Program of Nanjing Medical Science and Technique Development Foundation (ZDX16012).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.671585/full#supplementary-material

**Supplementary Figure 1 |** The correlation between *LIMK1* and overall survival in two different datasets analyzed with PrognoScan. High expression of *LIMK1* was correlated with poor overall survival in dataset jacob-00182-CANDF **(A)** and jacob-00182-MSK **(B)**.

## REFERENCES


Albaba, H., Lim, C., and Leighl, N. B. (2017). Economic Considerations in the Use of Novel Targeted Therapies for Lung Cancer: review of Current Literature. *Pharmacoeconomics* 35, 1195–1209. doi: 10.1007/s40273-017-0563-8

Brustugun, O. T., Gronberg, B. H., Fjellbirkeland, L., Helbekkmo, N., Aanerud, M., Grimsrud, T. K., et al. (2018). Substantial nation-wide improvement in lung cancer relative survival in Norway from 2000 to 2016. *Lung Cancer* 122, 138–145. doi: 10.1016/j.lungcan.2018.06.003

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B., et al. (2017). UALCAN: a Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Chen, J., Ananthanarayanan, B., Springer, K. S., Wolf, K. J., Sheyman, S. M., Tran, V. D., et al. (2020). Suppression of LIM Kinase 1 and LIM Kinase 2 Limits Glioblastoma Invasion. *Cancer Res.* 80, 69–78. doi: 10.1158/0008-5472.CAN-19-1237

Chen, Q., Jiao, D., Hu, H., Song, J., Yan, J., Wu, L., et al. (2013). Downregulation of LIMK1 level inhibits migration of lung cancer cells and enhances sensitivity to chemotherapy drugs. *Oncol. Res.* 20, 491–498. doi: 10.3727/096504013x13657689382699

Chen, Y., Chen, G., Zhang, B., Liu, C., Yu, Y., and Jin, Y. (2017). miR-27b-3p suppresses cell proliferation, migration and invasion by targeting LIMK1 in colorectal cancer. *Int. J. Clin. Exp. Pathol.* 10, 9251–9261.

Davila, M., Frost, A. R., Grizzle, W. E., and Chakrabarti, R. (2003). LIM kinase 1 is essential for the invasive growth of prostate epithelial cells: implications in prostate cancer. *J. Biol. Chem.* 278, 36868–36875. doi: 10.1074/jbc.M306196200

Davila, M., Jhala, D., Ghosh, D., Grizzle, W. E., and Chakrabarti, R. (2007). Expression of LIM kinase 1 is associated with reversible G1/S phase arrest, chromosomal instability and prostate cancer. *Mol. Cancer* 6:40. doi: 10.1186/1476-4598-6-40

Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., et al. (2015). The CPTAC Data Portal: a Resource for Cancer Proteomics Research. *J. Proteome. Res.* 14, 2707–2713. doi: 10.1021/pr501254j

Foletta, V. C., Moussi, N., Sarmiere, P. D., Bamburg, J. R., and Bernard, O. (2004). LIM kinase 1, a key regulator of actin dynamics, is widely expressed in embryonic and adult tissues. *Exp. Cell. Res.* 294, 392–405. doi: 10.1016/j.yexcr.2003.11.024

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Wu, Y. L., et al. (2017). Lung cancer: current therapies and new targeted treatments. *Lancet* 389, 299–311. doi: 10.1016/s0140-6736(16)30958-8

Huang, J. B., Wu, Y. P., Lin, Y. Z., Cai, H., Chen, S. H., Sun, X. L., et al. (2020). Up-regulation of LIMK1 expression in prostate cancer is correlated with poor pathological features, lymph node metastases and biochemical recurrence. *J. Cell. Mol. Med.* 24, 4698–4706. doi: 10.1111/jcmm.15138

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

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

Liao, Q., Li, R., Zhou, R., Pan, Z., Xu, L., Ding, Y., et al. (2017). LIM kinase 1 interacts with myosin-9 and alpha-actinin-4 and promotes colorectal cancer progression. *Br. J. Cancer* 117, 563–571. doi: 10.1038/bjc.2017.193

Liu, J., Zhang, Z., Liu, J., and Wang, D. (2019). LIM Kinase 1 Mediates Estradiol Effects on the Phosphorylation of Cofilin1 in Eutopic Endometrial Stromal Cells During the Invasion and Proliferation of Endometriosis. *Reprod. Sci.* 26, 1499–1505. doi: 10.1177/1933719119828076

Mardilovich, K., Gabrielsen, M., McGarry, L., Orange, C., Patel, R., Shanks, E., et al. (2015). Elevated LIM kinase 1 in nonmetastatic prostate cancer reflects its role in facilitating androgen receptor nuclear translocation. *Mol. Cancer Ther.* 14, 246–258. doi: 10.1158/1535-7163.MCT-14-0447

McConnell, B. V., Koto, K., and Gutierrez-Hartmann, A. (2011). Nuclear and cytoplasmic LIMK1 enhances human breast cancer progression. *Mol. Cancer* 10:75. doi: 10.1186/1476-4598-10-75

Mizuno, H., Kitada, K., Nakai, K., and Sarai, A. (2009). PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med. Genom.* 2:18. doi: 10.1186/1755-8794-2-18

Nishimura, Y., Yoshioka, K., Bernard, O., Bereczky, B., and Itoh, K. (2006). A role of LIM kinase 1/cofilin pathway in regulating endocytic trafficking of EGF receptor in human breast cancer cells. *Histochem. Cell. Biol.* 126, 627–638. doi: 10.1007/s00418-006-0198-x

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* 12:77. doi: 10.1186/1471-2105-12-77

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an integrated repository portal for tumor-immune system interactions. *Bioinformatics* 35, 4200–4202. doi: 10.1093/bioinformatics/btz210

Scott, R. W., and Olson, M. F. (2007). LIM kinases: function, regulation and association with human disease. *J. Mol. Med.* 85, 555–568. doi: 10.1007/s00109-007-0165-6

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–8. doi: 10.1093/nar/gkq973

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Torre, L. A., Siegel, R. L., and Jemal, A. (2016). Lung Cancer Statistics. *Adv. Exp. Med Biol.* 893, 1–19. doi: 10.1007/978-3-319-24223-1_1

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. *Science* 347:1260419. doi: 10.1126/science.1260419

Uhlen, M., Zhang, C., Lee, S., Sjostedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357:eaan2507. doi: 10.1126/science.aan2507

Vlecken, D. H., and Bagowski, C. P. (2009). LIMK1 and LIMK2 are important for metastatic behavior and tumor cell-induced angiogenesis of pancreatic cancer cells. *Zebrafish* 6, 433–439. doi: 10.1089/zeb.2009.0602

Wan, L., Zhang, L., Fan, K., and Wang, J. (2014). MiR-27b targets LIMK1 to inhibit growth and invasion of NSCLC cells. *Mol. Cell. Biochem.* 390, 85–91. doi: 10.1007/s11010-013-1959-1

Wickham, H. (2016). *ggplot2 Elegant Graphics for Data Analysis*. Germany: Springer International Publishing. doi: 10.1007/978-3-319-24277-4

Xu, X., Guo, J., Vorster, P., and Wu, Y. (2012). Involvement of LIM kinase 1 in actin polarization in human CD4 T cells. *Commun. Integr. Biol.* 5, 381–383. doi: 10.4161/cib.20165

You, T., Gao, W., Wei, J., Jin, X., Zhao, Z., Wang, C., et al. (2015). Overexpression of LIMK1 promotes tumor growth and metastasis in gastric cancer. *Biomed. Pharmacother.* 69, 96–101. doi: 10.1016/j.biopha.2014.11.011

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

Zeng, Y., Ren, M., Li, Y., Liu, Y., Chen, C., Su, J., et al. (2020). Knockdown of RhoGDI2 represses human gastric cancer cell proliferation, invasion and drug resistance via the Rac1/Pak1/LIMK1 pathway. *Cancer Lett.* 492, 136–146. doi: 10.1016/j.canlet.2020.07.013

Zhang, H., Wang, Y., Xing, F., Wang, J., Wang, Y., Wang, H., et al. (2011). Overexpression of LIMK1 promotes migration ability of multidrug-resistant osteosarcoma cells. *Oncol. Res.* 19, 501–509. doi: 10.3727/096504012x13286534482511

Zhang, M., Tian, J., Wang, R., Song, M., Zhao, R., Chen, H., et al. (2020). Dasatinib Inhibits Lung Cancer Cell Growth and Patient Derived Tumor Growth in Mice

by Targeting LIMK1. *Front. Cell. Dev. Biol.* 8:556532. doi: 10.3389/fcell.2020.556532

Zhang, W., Gan, N., and Zhou, J. (2012). Immunohistochemical Investigation of the Correlation between LIM Kinase 1 Expression and Development and Progression of Human Ovarian Carcinoma. *J. Int. Med. Res.* 40, 1067–1073. doi: 10.1177/14732300120400 0325

Zhao, J., Li, D., and Fang, L. (2019). MiR-128-3p suppresses breast cancer cellular progression via targeting LIMK1. *Biomed. Pharmacother.* 115:108947. doi: 10.1016/j.biopha.2019.108947

**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: 03 June 2021
doi: 10.3389/fgene.2021.670749

# Multi-Omics Marker Analysis Enables Early Prediction of Breast Tumor Progression

*Haifeng Xu[1,2], Tonje Lien[1], Helga Bergholtz[1], Thomas Fleischer[1], Lounes Djerroudi[3], Anne Vincent-Salomon[3], Therese Sørlie[1]* and Tero Aittokallio[1,2,4]**

[1] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, [2] Oslo Centre for Biostatistics and Epidemiology (OCBE), University of Oslo, Oslo, Norway, [3] Institut Curie, Ensemble Hospitalier, Pôle de Médecine Diagnostique et Théranostique, Département de Pathologie, Paris, France, [4] Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

OPEN ACCESS

**Edited by:**
Rosalba Giugno,
University of Verona, Italy

**Reviewed by:**
Biju Issac,
Leidos Biomedical Research, Inc., United States
Rodrigo Gularte Mérida,
Memorial Sloan Kettering Cancer Center, United States

***Correspondence:**
Therese Sørlie
therese.sorlie@medisin.uio.no
Tero Aittokallio
t.a.aittokallio@medisin.uio.no

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

**Received:** 22 February 2021
**Accepted:** 26 April 2021
**Published:** 03 June 2021

**Citation:**
Xu H, Lien T, Bergholtz H, Fleischer T, Djerroudi L, Vincent-Salomon A, Sørlie T and Aittokallio T (2021) Multi-Omics Marker Analysis Enables Early Prediction of Breast Tumor Progression. Front. Genet. 12:670749. doi: 10.3389/fgene.2021.670749

Ductal carcinoma *in situ* (DCIS) is a preinvasive form of breast cancer with a highly variable potential of becoming invasive and affecting mortality of the patients. Due to the lack of accurate markers of disease progression, many women with detected DCIS are currently overtreated. To distinguish those DCIS cases who are likely to require therapy from those who should be left untreated, there is a need for robust and predictive biomarkers extracted from molecular or genetic profiles. We developed a supervised machine learning approach that implements multi-omics feature selection and model regularization for the identification of biomarker combinations that could be used to distinguish low-risk DCIS lesions from those with a higher likelihood of progression. To investigate the genetic heterogeneity of disease progression, we applied this approach to 40 pure DCIS and 259 invasive breast cancer (IBC) samples profiled with genome-wide transcriptomics, DNA methylation, and DNA copy number variation. Feature selection using the multi-omics Lasso-regularized algorithm identified both known genes involved in breast cancer development, as well as novel markers for early detection. Even though the gene expression-based model features led to the highest classification accuracy alone, methylation data provided a complementary source of features and improved especially the sensitivity of correctly classifying DCIS cases. We also identified a number of repeatedly misclassified DCIS cases when using either the expression or methylation markers. A small panel of 10 gene markers was able to distinguish DCIS and IBC cases with high accuracy in nested cross-validation (AU-ROC = 0.99). The marker panel was not specific to any of the established breast cancer subtypes, suggesting that the 10-gene signature may provide a subtype-agnostic and cost-effective approach for breast cancer detection and patient stratification. We further confirmed high accuracy of the 10-gene signature in an external validation cohort (AU-ROC = 0.95), profiled using distinct transcriptomic assay, hence demonstrating robustness of the risk signature.

**Keywords: risk signature, breast cancer, disease progression, early detection, machine learning**

## INTRODUCTION

Ductal carcinoma *in situ* (DCIS) is a non-invasive precursor to invasive breast cancer (IBC) with low risk of progression (Cowell et al., 2013). Recent advances in breast cancer screening have resulted in an increasing number of women with detected DCIS lesions (Virnig et al., 2010; Seely and Alhassan, 2018; van Seijen et al., 2019), many of which actually will never progress to invasive disease (Page et al., 1982, 1995; Nielsen et al., 1984; Collins et al., 2005; Sanders et al., 2005). To distinguish the DCIS lesions with invasive potential from those that may be left untreated, there is need for robust biomarkers (or risk signatures) for accurate classification between high-risk and low-risk DCIS cases. However, DCIS lesions exhibit heterogeneous clinical, histopathological, and molecular characteristics that may vary considerably between the lesions and as a function of time (Vincent-Salomon et al., 2008). Furthermore, the underlying mechanisms of progression from DCIS to IBC are still poorly understood. The diagnostic classification has therefore considerable uncertainty, and the DCIS lesions may vary from indolent lesions to tumors on the verge of becoming invasive (Gorringe and Fox, 2017). Due to this uncertainty, treatment for DCIS is often extensive, resulting in substantial overtreatment (Esserman et al., 2014; Groen et al., 2017).

Even though histological grade and growth pattern provide some information on disease risk, there is a need for more precise risk prediction methods (Wang et al., 2011; Wallis et al., 2012; Onega et al., 2017). It has been shown that the "intrinsic" breast cancer subtypes (luminal A, luminal B, HER2-enriched, and basal-like) have prognostic significance, and a supervised risk predictor was developed based on the intrinsic subtypes and clinical information (Parker et al., 2009). We have also previously performed comparative analyses across the breast cancer subtypes and identified molecular differences between DCIS and IBC for subtype-specific disease progression (Bergholtz et al., 2020). In these subtype-stratified analyses, prominent molecular differences were identified especially for the basal-like DCIS, which was found to be less proliferative and showed a higher degree of differentiation than the basal-like IBC. However, for clinical use of the risk signatures, there is a need for cost-effective and subtype-agnostic biomarker panels that are widely applicable among diagnosed women regardless of their breast cancer subtype or other risk classifications that would require extensive clinical, histopathological, or molecular information.

In this study, we developed a supervised machine learning approach that implements multi-omics feature selection for the identification of biomarker combinations to distinguish DCIS and IBC cases. As a secondary objective, we identified a robust marker panel to identify those DCIS cases that may have a higher risk of progression (i.e., DCIS cases susceptible to invasion). To investigate the molecular, genetic, and epigenetic heterogeneity of disease progression, we applied the regularized approach to 40 DCIS and 259 IBC samples, profiled with genome-wide transcriptomics, DNA methylation, and DNA copy number variation. For economic clinical implementation, we further investigated the effect of the number of model features on the classification accuracy with each omics measurements. In doing so, we identified a minimal risk signature of 10 highly predictive and subtype-agnostic transcriptomic markers, originating from a single omics platform (microarrays), which could be developed as a decision support tool in clinical practice. We further validated our minimal risk signature in an independent validation cohort (with RNA-seq data) and studied how the signature predicted also lesions between DCIS and IBC classes, as well as relapsing DCIS cases.

## MATERIALS AND METHODS

### Training Material

As a model training data, we used multi-omics molecular and genomic profiles combined from three patient cohorts, Oslo2, Uppsala, and Milan (Muggerud et al., 2010; Fleischer et al., 2014; Lesurf et al., 2016; Aure et al., 2017; Bergholtz et al., 2020). Each patient cohort contains three levels of omics data from gene expression microarrays, DNA methylation, and DNA copy number. Gene expression was measured with Agilent Sureprint G3 Human Gene Expression 8 × 60 K microarrays (G4851A) (Agilent Technologies, Santa Clare, United States), with Low Input Quick Amp Labeling protocol. The DNA methylation was profiled using the Illumina Infinium Human Methylation 450K microarray (Illumina, CA, United States), following the manufacturer's instructions, and preprocessed with subset quantile normalization (Touleimat and Tost, 2012). The DNA copy number changes were profiled using Affymetrix SNP 6.0 arrays (Affymetrix, Santa Clara, United States) at Aros Applied Biotechnology (Aarhus, Denmark), following the manufacturer's instructions. In total, there were 370 patients included in these three cohorts. We included only patients with all three omics data levels, resulting in 299 patients as our training material, including 40 DCIS cases and 259 IBC cases (**Supplementary Figure 1** and **Supplementary Data 1**).

The gene expression and DNA copy number changes were mapped to protein-coding genes to make it easier to interpret the results and integrate across the omics data. To investigate the effect of DNA methylation data processing on predictive modeling, we considered two versions of the DNA methylation data. The first option was to use directly the original CpG level methylation data as model features, and therefore we performed feature preselection using only CpGs thought to be involved in important biological variation between breast cancer samples ($N$ = 44,263 CpGs) (Fleischer et al., 2017). These CpGs were thought to be involved in one of four breast cancer biological properties, namely, regulation of estrogen signaling, regulation of non-estrogen-related proliferation, fibroblast infiltration, or immune infiltration. The CpGs were located both inside and outside CpG islands and were enriched in both enhancers and promoters. The second option used gene-level processing, where we calculated a methylation score to represent each protein-coding gene using a principal component analysis (PCA), taking into account the variation of all CpGs mapped to a gene, similarly as before (Bergholtz et al., 2020). The second option leads to gene-level features, whereas in the first option, each gene can be associated with hundreds of CpGs.

## Validation Material

The validation data set was collected at Institute Curie, France (referred to as Curie Cohort), where the gene expression was profiled using RNA sequencing with the Illumina HiSeq2500 sequencer. The read counts were normalized with the rlog and cpm options in edgeR (v3.1.2) and DESeq2 (v1.4.5) R-packages, respectively (Robinson et al., 2009; Love et al., 2017). Pseudocount data were calculated as $\log_{10}$(RNAseq count + 1), and it was centered for each gene around the mean of the pseudocounts. The validation cohort included 18 pure DCIS cases and 20 IBC cases, as well as 16 micro-invasive (MI) DCIS cases, which are DCIS lesions with invasive foci of maximum 1 mm.

## Classification Models

Our main objective was to identify the most discriminating molecular and genetic differences between DCIS and IBC, regardless of their intrinsic subtype and the nuclear grade. We initially constructed Lasso, Support Vector Machine (SVM), and Random Forest (RF) models based on each type of omics data (gene expression, DNA methylation, and DNA copy number). We used the R-package "glmnet" to build Lasso models, R-package "e1071" to build SVM models, and R-package "randomForest" to build RF models (Liaw and Wiener, 2002; Friedman et al., 2010; Meyer et al., 2019). To assess the classification accuracy, we used 10-fold cross validation (CV), where the training dataset was divided into 10-fold, testing on each fold at a time, while the remaining ninefold were used for the model estimation (sub-training set). Stratified CV was used to make sure each CV fold had the same proportion of breast cancer subtypes. To test the generalizability of the Lasso models, and to avoid selection bias, we used nested cross-validation, where another 10-fold CV was applied within each sub-training set to determine the optimal model regularization parameters, e.g., the lambda and beta values of the Lasso model. The other model parameters were set to their default values. When training the SVM models, we used Recursive Feature Elimination (RFE) implemented in the R-package "caret" to select the model features (Kuhn, 2008). The size parameter of RFE was set to a vector (2, 5, 10, and 20), the parameter "number" of the rfeControl function was set to 5, and the kernel parameter was set to svmRadial to use the radial kernel. We used 10-fold CV for the SVM models, and in each fold, RFE was run to select the model features using nested CV.

## Evaluation Metrics

To evaluate the predictive accuracy, we used Area Under the ROC Curve (AU-ROC) and Area under the Precision-Recall Curve (AU-PRC) (**Supplementary Figure 2**). Moreover, classification cutoff-specific evaluation metrics, such as sensitivity and specificity, were also recorded to evaluate the trade-off between correctly classifying either DCIS or IBC cases. For avoiding overtreatment, it is especially important to correctly predict true DCIS cases, and therefore we labeled DCIS as positive and IBC as negative cases. Accordingly, sensitivity TP/(TP + FN) refers to the rate of how many DCIS cases are correctly classified, while specificity TN/(TN + FP) refers to the percentage of correctly classified IBC cases. Balanced accuracy is defined as the average of sensitivity and specificity. Precision–Recall analysis provides an alternative evaluation metric for the unbalanced classification problem. The AU-ROC and the AU-PRC were plotted and calculated with the R-packages "PRROC" (Grau et al., 2015) and "pROC" (Robin et al., 2011), respectively. As a continuous evaluation metric, we used Mean Squared Error (MSE), where MSE values close to zero indicate more accurate models.

## Multi-Omics Classifiers

To test whether integrating the three types of omics data improved the prediction accuracies, we combined gene expression data, DNA methylation, and DNA copy number data together in a single Lasso model. The CpG-level and gene-level methylation data were combined separately with the other data types to investigate their respective predictive contribution. To unify the scales between the different data types, we applied z-score scaling over each feature (gene or CpG) and then combined the z-scored features into a single model.

## Limiting Model Complexity

To test the effect of limiting the maximum allowed number of model features on the prediction accuracy, we adjusted the parameter "dfmax" of the glmnet function, which limits the maximum number of variables in the Lasso-regularized model (Friedman et al., 2010). We varied the dfmax parameter from 2 to 51 with each separate omics data and their combination using nested CV to explore the most predictive feature subsets and to construct a maximally sparse, cost-effective, and transparent models for economic clinical implementation.

## Robust Gene Selection

We considered the common features identified by the two classification models, SVM and Lasso, as robust biological signatures. To further improve the reliability of these signatures, and to avoid reporting unstable features, we considered only those features that were returned more than five times during the 10-fold CV (i.e., >50% of the folds), where each feature can be selected up to 10 times. This analysis was limited to the gene expression data only (without using z-scoring), since gene expression data was found generally most predictive.

## Model Validation

In the validation phase, we trained a new Lasso model using the subset of 10 most robust genes on the entire training set and tested its predictive power on the validation set (the Curie Cohort). Only RNA-seq transcriptomics data were available in the validation set. We used z-score scaling over each gene separately in the training and validation sets to normalize their scales between the microarray and RNA sequencing data. The model outcome was the predicted class probability in DCIS vs. IBC classification for each validation case separately.

## Existing Risk Scores

We compared the 10-gene signature against three existing risk scores. The first was ROR, risk of recurrence after surgical treatment for IBC, calculated based on expression of the PAM50 genes (Parker et al., 2009). Firstly, the correlation to the four breast cancer subtypes (Basal-like, Her2-enriched, Luminal A, and Luminal B) was calculated, and the ROR score was then defined as a weighted sum of the four correlations. We also calculated an invasiveness score based on a previously proposed 64-gene signature (Anastassiou et al., 2011). We summarized the 64-signature using z-score to obtain an invasiveness score for each sample and then used the mean value of each case as the final invasive score. As the third comparison score, we used the Oncotype DX® DCIS Score that has been suggested to quantify the risk of developing an ipsilateral breast event (i.e., local recurrence of DCIS or invasive carcinoma) (Solin et al., 2013). The original DCIS score was calculated using qPCR expression values from 12 genes. However, since our training cohort included normalized microarray expression data, we did not perform the first step of the DCIS Score calculation, i.e., normalizing seven signature genes relative to the expression of five housekeeping genes. The ROR, invasiveness, and DX® DCIS scores were included in the simple linear model using function "glm" from basic R, where only the score was used when building these models using 10-fold CV.

## Identification of Misclassified DCIS Cases

Some DCIS cases may never progress to IBC and will remain intraductal, while other DCIS lesions may have future invasive potential but were discovered while still intraductal. We hypothesized that even though some lesions are discovered while still intraductal, they may carry molecular or genomic changes that distinguish them from the low-risk DCIS cases that will never progress. To address the secondary questions of whether we can divide DCIS samples into two groups, low- and high-risk DCIS, and how accurately we can find those higher-risk DCIS cases that might carry the potential for future invasion, we built additional machine learning models based on gene expression and DNA methylation data, and the cases incorrectly classified by more than one model-data combinations were considered for further scrutiny. Next, we used so-called pseudo labeling, where the repeatedly misclassified DCIS cases were relabeled as IBC, then retrained a Lasso model with 10-fold nested CV and checked whether or not its classification accuracy increased, compared to the original Lasso model with the original class labeling.

# RESULTS

## Predictive Model Development in Multi-Omics Data

We started by testing various prediction algorithms, including Lasso, SVM, and RF, to classify the patient samples of the training cohort into two groups, DCIS and IBC. These algorithms were evaluated in terms of their classification accuracy and robustness in the heterogeneous omics data (gene expression, DNA copy number, and DNA methylation). In the initial runs, the classifiers were allowed to freely make use of an unlimited number of the omics features (genes and CpGs), and nested CV was then used to evaluate the predictive power of the models and the selected feature panels. In this section, we show the results of the Lasso model that performed generally the best, while the results of RF and SVM models are provided in **Supplementary Tables 2, 3**, respectively, showing similar performance trends with slightly decreased accuracies.

Notably, gene expression features provided the best overall accuracy among the single omics datasets when using summary metrics AU-ROC and AU-PRC (**Figure 1**). Interestingly, the CpG-level methylation data provided almost as high AU-ROC levels, but the Lasso model selected more than three times the CpG features compared to expression features, and the CpG model had much a lower AU-PRC value (**Figure 1**). DNA copy number variation profiles showed the poorest performance among the three omics datasets, even though the Lasso model selected the largest number of copy number features, suggesting that copy number changes do not contain a sufficient predictive signal for the classification between DCIS and IBC cases. All the omics profiles resulted in close to perfect specificity (**Figure 1**).

The combined use of the three omics features in a single Lasso model using z-score scaling resulted in similar AU-ROC and AU-PRC values when using the gene expression features alone (**Figure 1**). However, the sensitivity of correctly classified DCIS cases increased when using all the omics data together. In clinical practice, sensitivity is more important for avoiding overtreatment. Omics data integration also led to higher levels of balanced accuracy, while the specificity of correctly classifying IBC cases remained perfect, similar to that when using the gene expression data only. The two versions of the DNA methylation data provided a similar contribution to the multi-omics Lasso model; however, the gene-level methylation features led to slightly increased performance, especially in terms of MSE, whereas CpG-level data required less features (**Supplementary Table 1**).

## The Effect of Limiting the Number of Features

We next studied the effect of limiting the number of features of the Lasso model on its predictive accuracy, with the aim to investigate what are the minimal panels of biomarkers that could cost-effectively distinguish DCIS cases from IBC. A feature number limit from 2 to 50 was imposed on each data type separately and in combination, and for each limit, 10-fold nested CV was applied to investigate the classification accuracy of the Lasso models with limited number of features. Notably, already two gene features provided an almost perfect AU-ROC of 0.95 when using the expression data only (**Figure 2**), indicating that sparse models enable accurate classification. However, the variability of the AU-ROC decreased when using the feature limit higher than 12 (**Supplementary Figure 4**), suggesting that the additional gene features make the classifier more stable.

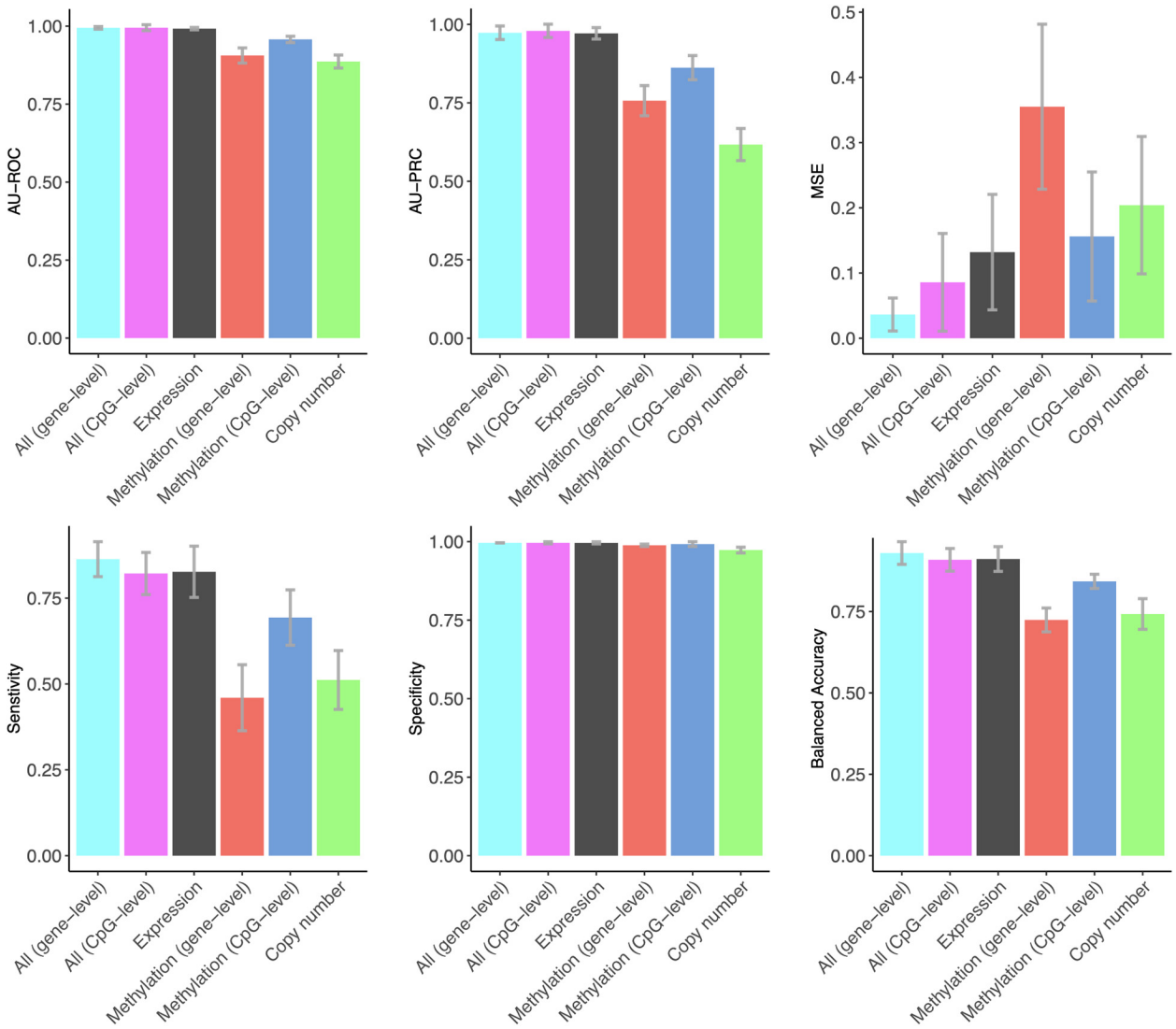



**FIGURE 1 |** The predictive contribution of omics profiles and their combinations to DCIS vs. IBC classification. The bars show the average accuracy over 10 CV folds, and the error bars indicate the standard error of the mean (SEM). These data correspond to those shown in **Supplementary Table 1**. ROC and PRCs of CV folds using various omics data combinations are shown in **Supplementary Figure 3**.

When considering AU-ROC, the CpG methylation model performed initially worse, when compared to the gene-level methylation model, but after 30 CpGs its classification accuracy increased (**Figure 2**). The variability of the classification accuracy was also lower with the CpG-level model compared to the gene-level methylation model. These results suggest that when the variance of individual CpGs is large, the model cannot make reliable classification using only a small number of CpG features. Since the gene-level methylation signature consists of many CpGs collapsed to single genes, its variance tends to be smaller due to measurement noise being canceled out in the collapsing process. When considering AU-PRC, the gene-level methylation model remained slightly better than the CpG model across all the feature numbers (**Supplementary Figure 4**), and it also led to increased sensitivity of the multi-omics model, comparable to that of the gene expression only model (**Supplementary Figure 5**).

Since the features were selected in 10-fold nested CV at each feature number limit, the model may identify in total more features than the limit, since the different CV folds may select different features. **Figure 2** lists as examples features that were selected in all the 10 CV folds, suggesting they are robust to training data subsampling and therefore likely to present robust classifying features. Such robust features could not be identified from the copy number data. DNA methylation profiles identified genes that are distinct from those identified using the gene expression data, both when using the gene-level or CpG-level methylation data (and the corresponding genes). However, a total of four genes (MMP11, RUFY3, UNCX, and MAMDC2) were selected using both versions of the integrated

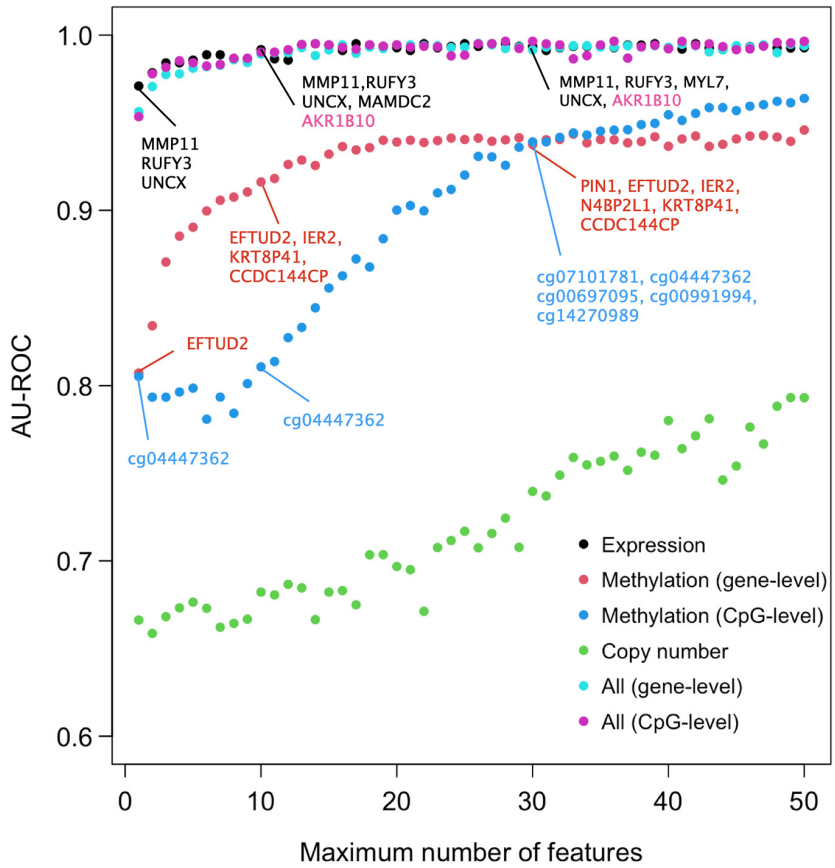



**FIGURE 2 |** Predictive accuracy of the omics profiles and their combinations when the maximum number of features was limited. The points are average AU-ROC values over the 10 CV rounds in nested CV. Example feature sets from omics data are shown at limits $x = 2$, 10, and 30. The gene lists contain the features that were selected by Lasso in all the 10-fold at that limit. Expression data and integrated data share most genes in common. Black color indicates genes that were selected by both gene expression and integrated data, the pink color those genes that were selected by integrated data only. Note that the top genes of the two types of integrated data are the same. No copy number features were selected in all the 10 CV folds (no robust CNV features). See **Supplementary Figure 4** for the version of ROC and PRCs with SEMs included.

data; these are exactly the same genes Lasso model identified when using the gene expression data only and the feature limit of 2, further suggesting that transcriptomics alone leads to sparse and accurate signatures.

## Identification of Repeatedly Misclassified DCIS Cases

We next investigated whether the multi-omics data and the classification models could identify those DCIS-labeled samples with a potentially higher likelihood for progressing to an invasive state. Even if these DCIS samples have been originally labeled as DCIS in the diagnostic classification, they may still possess molecular changes that promote invasion later in time. In this analysis, we used Lasso and RF models, together with gene expression and CpG methylation profiles, due to their overall good performance. We considered for further investigation those DCIS cases in the training cohort that were repeatedly misclassified by these model–data combinations more than once (**Table 1**). Misclassification by one model–data combination may represent merely technical noise.

Out of the 40 DCIS cases, there were 19 lesions that were always correctly classified, and 11 DCIS cases were misclassified once, whereas eight and 2 DCIS cases were misclassified two or three times, respectively. We next applied so-called pseudo-labeling, where the repeatedly misclassified DCIS cases were relabeled as IBC, and then trained a new Lasso model with nested CV. Notably, such pseudo-labeling slightly increased the AU-PRC levels in the training cohort, while the AU-ROC levels remained similar to those with the original class labels (**Supplementary Table 4**). The multi-omics patterns provide evidence that these DCIS cases have molecular signatures more similar to the IBC cases and may have an increased likelihood to progress to an invasive disease stage.

## The Most Robust Genes for Classification

Since gene expression was found to be generally the most predictive among the single omics data, we next identified the set of common genes selected by both the Lasso and SVM models using the gene expression features alone. We further required that

**TABLE 1 |** Misclassified training samples when using various classification models and omics data.

| Lasso expression | Lasso methylation (cpg) | RF expression |
|---|---|---|
| DCIS033 | DCIS029 | **DCIS026** |
| DCIS038 | DCIS031 | **DCIS053** |
| **DCIS026** | DCIS032 | DCIS051 |
| DCIS051 | DCIS024 | DCIS056 |
| DCIS052 | **DCIS026** | DCIS033 |
| **DCIS053** | **DCIS053** | DCIS031 |
| IBC301 | DCIS056 | DCIS052 |
| | DCIS035 | DCIS022 |
| | DCIS022 | DCIS032 |
| | DCIS017 | DCIS017 |
| | | DCIS001 |
| | | DCIS037 |
| | | DCIS008 |
| | | DCIS004 |
| | | DCIS030 |
| | | DCIS034 |
| | | DCIS013 |

*Color coding indicates the number of times any of the 40 DCIS cases were misclassified as IBC in the training data; green, once; brown, twice; red, three times. CpG methylation data with the RF model was not used in these analyses since it misclassified a total of 34 DCIS cases, which was considered too many. We considered for further investigation only those DCIS cases in the training cohort that were repeatedly misclassified by these model–data combinations more than once.*

a gene needs to be selected in more than 50% of the CV folds (i.e., more than five out of 10-folds), with the aim to guarantee robust and stable feature selection. In total, we found 10 such common and robust genes identified as robust risk signature. Notably, each of the 10 genes had a similar direction of differential expression between the DCIS and IBC classes across the established breast cancer subtypes (**Figure 3**), suggesting that they provide subtype-agnostic markers for breast cancer risk prediction.

Interestingly, there were marked differences in the expression levels of the 10 genes across the DCIS cases misclassified as IBC (**Figure 4**). For instance, RUFY3, UNCX, PRSS33, and COL10A1 showed an increasing trend of absolute expression changes between the DCIS cases as a function of the number of times the DCIS samples were misclassified by the models. This further demonstrates the molecular information captured in the expression profiles. Furthermore, based on the expression levels of the 10-gene signature, most of the sure DCIS cases that were always correctly classified were clustered together, whereas the repeatedly misclassified DCIS cases were scattered around in the unsupervised hierarchical clustering dendrogram (**Supplementary Figure 6**).

We next compared the classification accuracy of the 10-gene Lasso model against three existing risk signatures relevant for breast cancer progression: ROR (risk of recurrence), the invasiveness score (64-gene signature) and seven-gene DX® DCIS score (see Methods). Our results showed that none of these risk scores was able to accurately distinguish between DCIS and IBC cases in our training cohort (**Figure 5**). In particular, using the default Lasso cutoff of 0.5, both the ROR and invasiveness score always classified all the DCIS lesions as IBC, whereas the DX® DCIS Score classified all the IBC cases as DCIS (**Supplementary Table 5**). There were three common genes between the 64-gene invasiveness signature and our 10-gene signature (COL1A1, COL10A1, and MMP11), hence explaining its higher classification accuracy compared to ROR.

## Validation Set Results

The final step was to validate the 10-gene signature on an external data set, the Curie Cohort, with the aim to investigate whether the DCIS classification model generalizes also beyond the training cohort to an independent validation dataset. The Lasso model of 10 genes estimated in the full training dataset was shown to provide highly accurate classification between the DCIS and IBC cases also in the validation dataset (**Figure 6**). Notably, both the AU-ROC and AU-PRC values dropped only slightly from the training to the validation cohort, further demonstrating the reliability and robustness of the classification model based on the 10-gene signature. However, we note that the default classification cutoff of 0.5 was not optimal in the validation data, but instead smaller thresholds led to better classification accuracy (**Supplementary Figure 7**). This is likely due to the differences between the microarray gene expression data (training cohort) and RNA-sequencing data (validation cohort). Although we performed z-scoring to unify the scales, it cannot correct for all the distributional differences between microarray and RNA-sequencing data.

We further tested how the model predicts the microinvasive (MI) DCIS cases in the validation cohort to explore whether the 10-gene signature could also distinguish the MI cases from pure DCIS and IBC cases. Interestingly, the classification probabilities of the MI DCIS cases were in between the pure DCIS and IBC classes but remained significantly closer to the pure DCIS cases (**Figure 7**, left). However, there was a relatively large variability in the distribution of the predicted probabilities also within the classes, showing individual variability in the risk scoring based on the 10-gene signature. This suggests that there are molecular-level changes in these genes between the classes of pure DCIS, DCIS-MI, and IBC lesions. Interestingly, there appeared to be three outlier cases in the DCIS-MI class with the classification probability comparable to that of the IBC cases. The six genes that were related to the microenvironment (COL10A1, COL1A1, MFAP2, PRSS33, PRSS53 and MMP11) showed higher prediction probability in the recurrent DCIS cases, compared to DCIS without recurrence, and these became close to those of the IBC cases (**Figure 7**, right).

To further investigate the features of the sparse Lasso model, we plotted the expression distributions of the 10 genes on both the training and validation cohorts (**Figure 8**). After z-scoring, most of the genes showed similar distributions, except for UNCX and PRSS33. In particular, for UNCX, there were only two distinct expression values in the test RNA-seq data, and 53 out of 55 cases (96%) corresponded to zero expression in the original expression data before z-scoring. There were also marked differences in the expression levels of the 10 genes across the three disease classes of the validation cohort (**Supplementary Figure 9**), mostly differentiating IBC cases from DCIS and DCIS-MI, even though the differences were not as clear as in the training cohort (**Figure 3**). However, regardless of these technical and biological differences between the training and validation

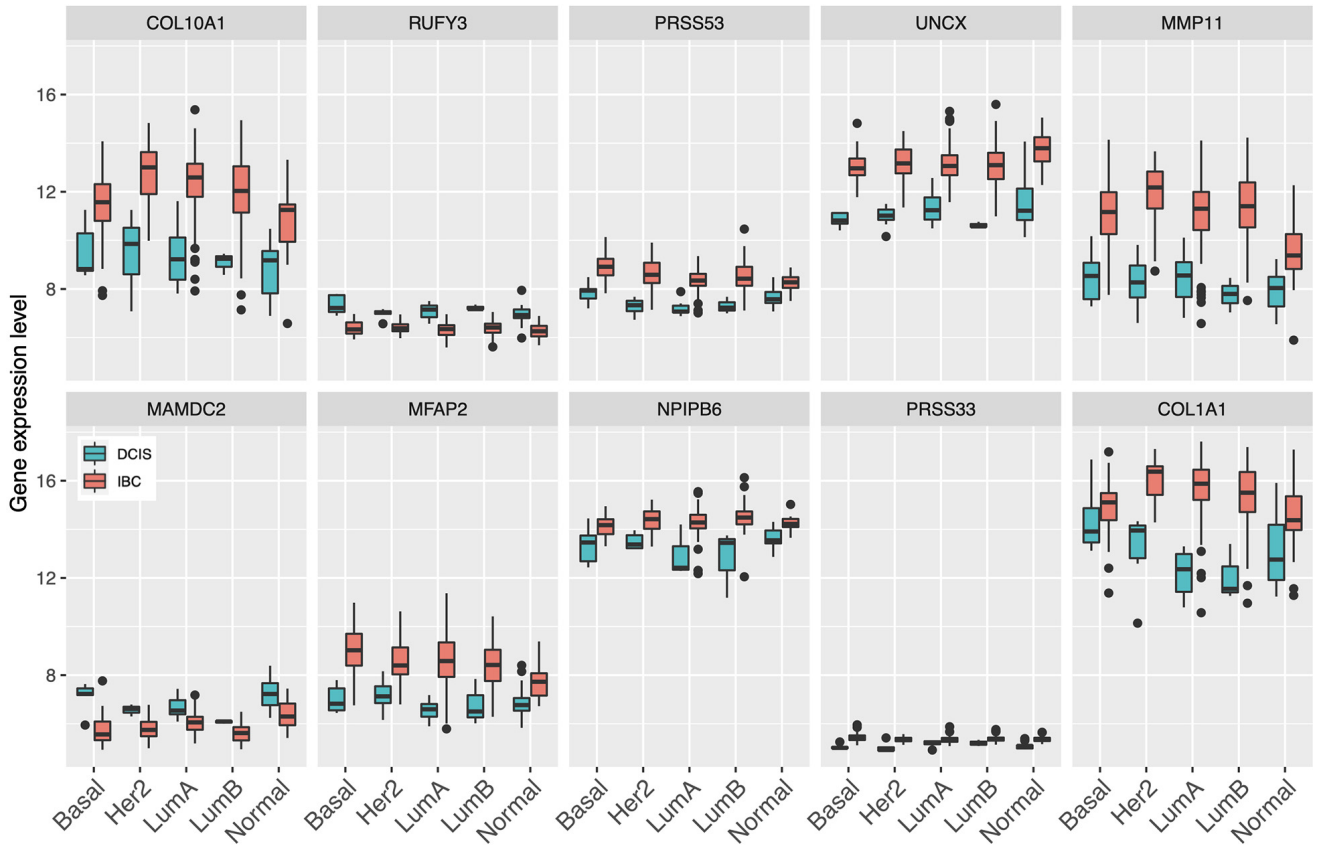



**FIGURE 3 |** Expression levels of the select 10 genes across the established breast cancer subtypes. Basal, basal-like; HER2, HER2-enriched; LumA, luminal A; LumB, luminal B; Normal, normal-like.

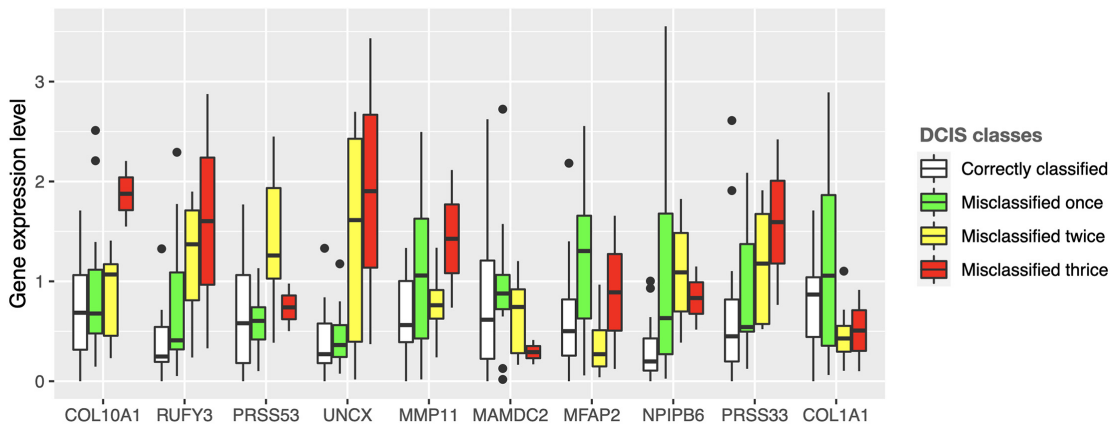



**FIGURE 4 |** Expression values of the 10 genes across the DCIS cases repeatedly misclassified as IBC. The bars show absolute deviance from the median expression level of the correctly classified DCIS class for each gene. The median level of the correctly classified DCIS class was subtracted from the other classes to better show differential expression levels. **Supplementary Figure 10** shows the expression differences before subtracting the median level of the correctly classified DCIS class.

cohorts, the 10-gene signature provided accurate classification performance in both of the datasets, further demonstrating its robust behavior. Taken together, these results indicate that the 10-gene signature can reliably identify those DCIS cases that are less likely to progress to invasive disease.

## DISCUSSION

In our multi-omics classification analysis between DCIS and IBC, we found that the gene expression-based model features led to the highest classification accuracy alone; however, methylation

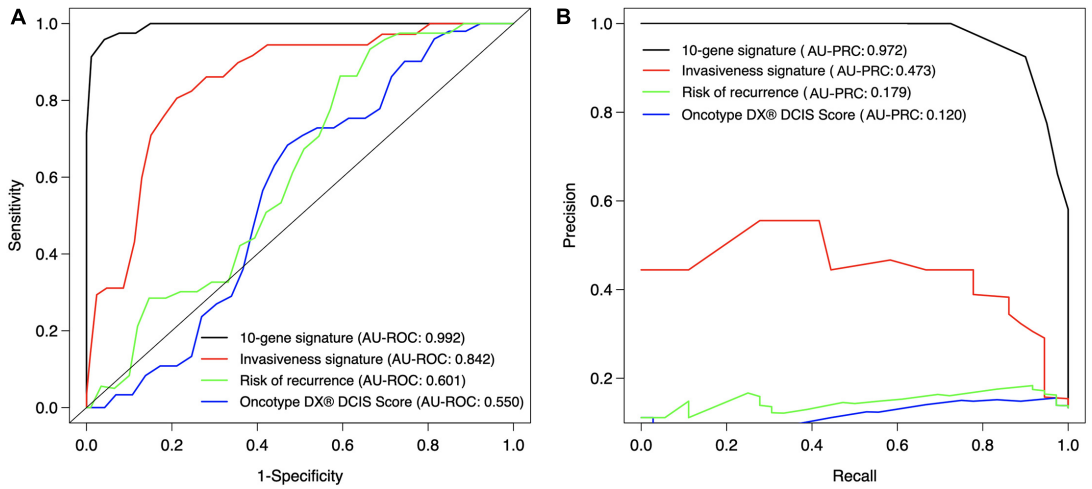



**FIGURE 5 |** Classification accuracy of the 10-gene signature against existing scores. **(A)** ROC, **(B)** PRC. Each signature was calculated based on the training patient cohort. The ROR score is based on the PAM50 genes (Parker et al., 2009), invasiveness score is based on 64 invasiveness related genes (Anastassiou et al., 2011), and DX® DCIS score based on seven genes (Solin et al., 2013). The expression values of the 64 genes were converted to z-score over each gene, and the average z-score was used as the invasiveness score for each sample. The original DX® DCIS score was based on qPCR data, but here it was applied to microarray gene expression data. The curves show the mean sensitivity and specificity over 10 CV folds in the training cohort. See **Supplementary Table 5** for the SD of the AU-ROC and AU-PRC values.

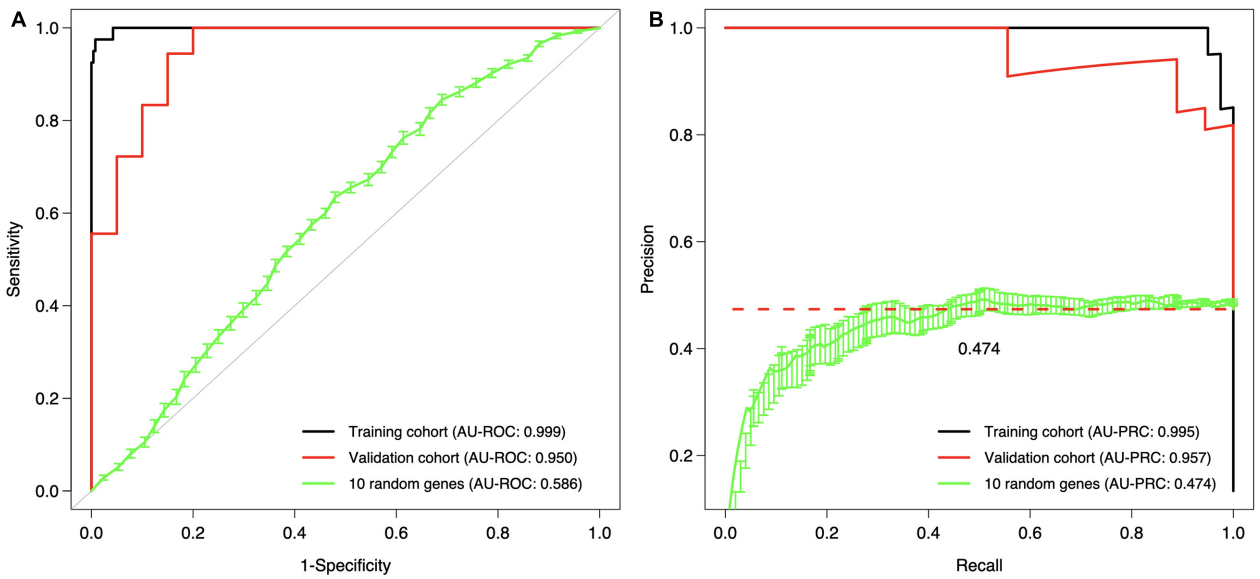



**FIGURE 6 |** Validation and training cohort accuracies of the 10-gene signature. **(A)** ROC, **(B)** PRC. The Lasso model was first estimated based on the full training dataset using the 10 genes as features, and then the estimated model was applied to the validation cohort. The training cohort model accuracy is overoptimistic as no cross validation was used and the training and test data are the same; see **Figure 5** for cross-validated training cohort model accuracy. For comparison, we randomly selected 10 genes 100 times, estimated 100 Lasso models in the training cohort, and then tested these random gene classifiers on the validation cohort. The 10 random gene curve shows the average performance of the random classifiers, and the error bars show the standard error of the mean (SEM). In panel **(B)**, the dashed horizontal line corresponds to a theoretical random classifier with AU-PRC = 0.473.

data provided a complementary source of predictive signal, and it improved especially the sensitivity of correctly classifying DCIS cases, which is important for clinical application of risk signatures. No better prediction results could be obtained with any of the two-data combinations, and the gene expression data was always required for the best prediction results, indicating its high predictive contribution. Due to the challenges of acquiring fresh frozen DCIS tissue, the number of DCIS cases was much smaller in the training cohort, compared to the IBC cases. We used several computational approaches to take into account such unbalanced classification setting: (i) we used several evaluation metrics to provide multiple views into the predictive

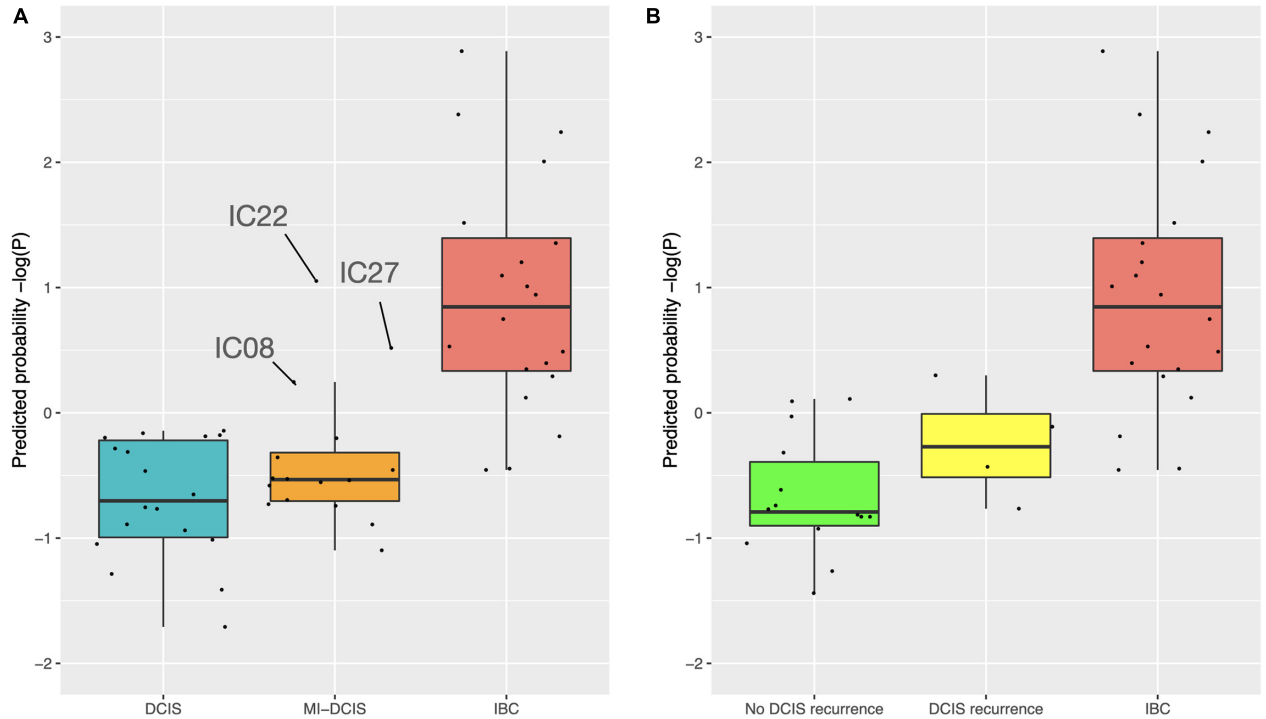



**FIGURE 7 |** Predicted class probabilities of the validation cohort samples. **(A)** The probabilities were calculated based on the 10-gene Lasso model estimated on the full training data. The validation cohort included 18 pure DCIS cases, 20 IBC cases, and 16 micro-invasive (MI) DCIS cases. Note: y-axis is log-scaled for better visualization. The non-logged version is shown in **Supplementary Figure 8**. **(B)** The probabilities were calculated based on the six microenvironment related genes. The validation cohort included 14 DCIS cases without recurrence, four DCIS recurrence cases, and 20 IBC cases. The y-axes are z-scaled for better comparability of the prediction probabilities. The horizontal lines in the boxplots indicate median values, the whiskers the interquartile range (IQR) of the cases in each class, and the error bars show cases withing 1.5xIQR, while the remaining cases are considered outliers.

performance of the models, including precision–recall analysis, which is often considered more suitable for the unbalanced classification problem; (ii) we included only those omics features in the signature that were robustly identified using multiple algorithms and across several CV rounds; (iii) we carried out the pseudo-labeling approach to investigate whether relabeling of some of the recurrently mis-classified cases could increase the predictive performance of the model and reveal potentially high-risk DCIS cases; and finally (iv) we validated the predictive power of the signature in an external validation cohort with more balanced classes.

Previous studies have found only moderate genomic and epigenomic differences between DCIS and IBC (Ma et al., 2003; Hannemann et al., 2006; Fleischer et al., 2014; Abba et al., 2015; Pang et al., 2017). In this study, we identified 10 genes using both the Lasso and SVM models that were selected in >50% of the CV rounds, indicating their robust behavior for classification between DCIS and IBC cases. We also found that these genes were differentially expressed between DCIS and IBC across all the breast cancer subtypes (**Figure 3**). One should interpret such gene lists with caution, however, as there may be other gene combinations with similar predictive power due to the correlated nature of the gene expression profiles among genes in the same pathways or biological processes. Nevertheless, the genes were selected by two independent methods, which increases the robustness of their biological signal. The 10-gene signature was also validated in independent test data (Curie cohort), where the transcriptional profiling was done with RNA-seq. The high classification accuracy observed for the 10 genes, originally identified using gene expression microarrays, further demonstrates the robustness of the signature, although there remained some variability that is beyond z-score normalization (**Figure 8**). We also note that the 10-gene signature was not able to predict recurrence in the validation cohort, as expected, since the genes were selected specifically for distinguishing between DCIS and IBC classes, not the progression of DCIS cases.

The comparison between our 10-gene signature and traditional breast cancer risk scores further demonstrated the added value of our 10-gene markers especially for the accurate DCIS classification (high sensitivity). We note that ROR is mostly affected by proliferation, and it is highly associated with breast cancer subtypes (Parker et al., 2009). Our results therefore indicate that proliferation may not be very important when distinguishing between DCIS and IBC cases. However, the invasiveness score has previously been found highly associated with cancer cell motility and invasiveness of several cancer types, including non-epithelial cancers such as neuroblastoma (Anastassiou et al., 2011). This should make it a competitive biological marker to classify DCIS and IBC. Our results showed that the invasiveness score achieved a relatively high

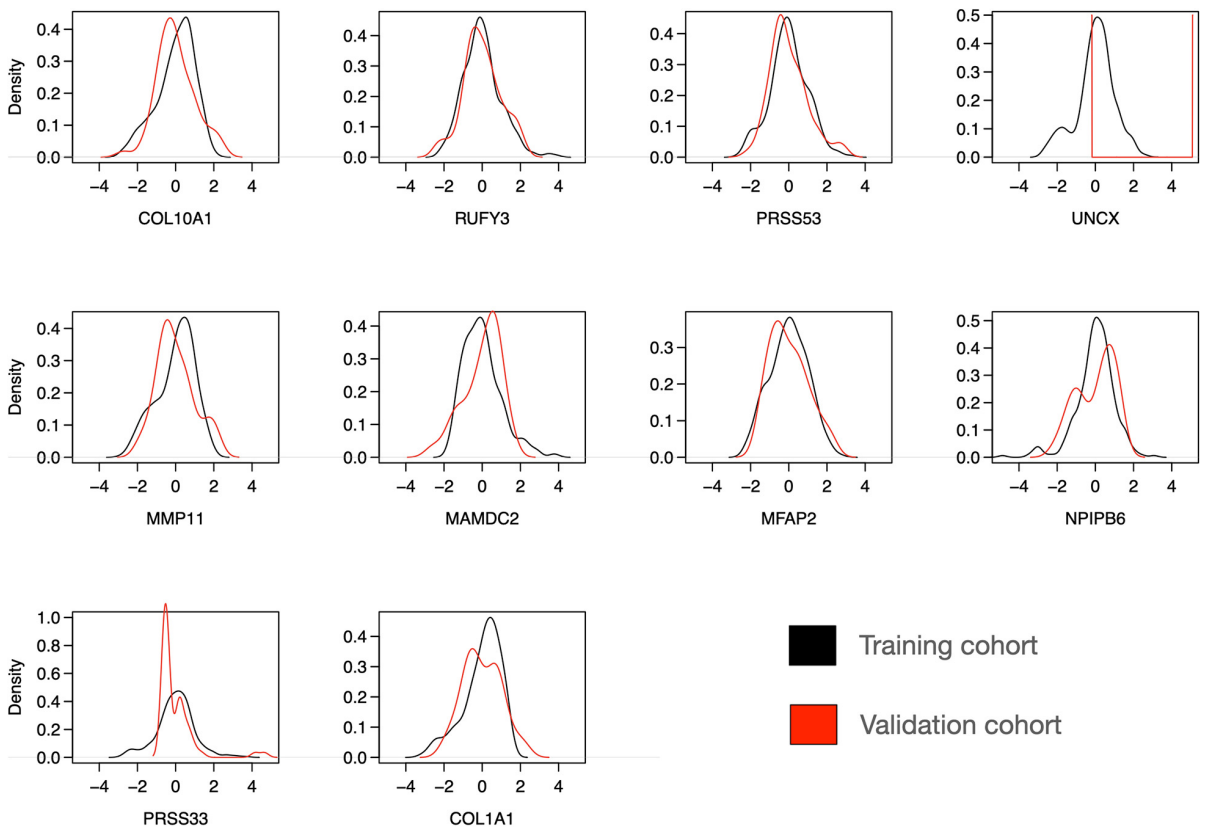



**FIGURE 8 |** Distribution of expression levels of 10 selected genes over all the individuals in the training and validation cohorts. The scales between the microarray (training cohort) and RNA-seq (validation cohort) data were harmonized using z-score normalization over each gene separately in the training and validation datasets. Note: the y-axis density ranges differ between the panels.

classification accuracy (AU-ROC 0.842), but not as high as our 10-gene signature (AU-ROC 0.992). The 64 genes included in the invasiveness signature had three genes in common with our 10-gene signature (COL1A1, COL10A1, and MMP11). Since we demonstrated that already two genes can give a relatively high AUC (**Figure 2**), and MMP11 is one of the selected genes when the feature limit was two, the higher performance of the invasiveness signature was as expected. However, the extended set of 10 genes provided increased performance especially for classification sensitivity. Furthermore, measuring 10 genes is more practical than measuring 64 genes using, for instance, qPCR-based clinical assays.

Many of the genes included in the model have previously been identified as differentially expressed between DCIS and IBC (Lesurf et al., 2016), but there are also some novel genes. Out of the 10 genes, six are related to the tumor microenvironment (COL10A1, COL1A1, MFAP2, PRSS33, PRSS53, and MMP11), and these genes showed predictive power for recurrent DCIS (**Figure 7**), although its added value for clinical practice remains to be investigated on a larger series. COL10A1, COL1A1, and MFAP2 are constituents of the extracellular matrix remodeling, which is an important process in breast tumor invasion and tumor cell dissemination (McSherry et al., 2007). Overexpression of the genes encoding these proteins is associated with poor breast cancer survival, and MFAP2 has been shown to promote cell proliferation, migration, invasion, and epithelial to mesenchymal transition (Wang et al., 2018; Liu et al., 2018; Zhang et al., 2020). MMP11 is a proteinase that is involved in extracellular matrix degradation directly by degrading collagen IV and indirectly by inhibiting the alpha1-proteinase inhibitor (Pan et al., 2003; Motrescu et al., 2008. MMP11 has been characterized extensively for its role in breast cancer and has been shown to be a predictive factor for tumor invasiveness, hence serving as positive control here (Ahmad et al., 1998; Zhang et al., 2016). In contrast, the roles of the serine proteases PRSS33 and PRSS53 have been less investigated in cancer progression, but there are indications that PRSS33 may play a role in tumor cell invasion (Jeong et al., 2016).

The remaining four genes in our gene list are not directly associated with the microenvironment. For instance, RUFY3 is involved in F-actin-enriched protrusions from the cell surface and it has been shown to be involved in gastric cancer cell migration and invasion (Wang et al., 2015). This gene, however, shows paradoxical expression in our training data with higher expression in DCIS than in invasive samples (**Figure 3**). In the validation cohort, however, the expression levels of RUFY3 were as expected in the DCIS and IBC classes (**Supplementary Figure 9**), especially when focusing the recurrent DCIS cases (**Supplementary Figure 11**). UNCX

was another gene with distinct expression distribution between training and validation data. It is a homeobox transcription factor that has been associated with acute myeloid leukemia (Daniele et al., 2017). MAMDC2 is a known tumor suppressor involved in glycosaminoglycan binding (Lee et al., 2020), whereas NPIPB6 has not previously been associated with cancers to the best of our knowledge. We note that the 64 genes included in the invasiveness signature are mainly related to epithelial–mesenchymal transition (EMT) (Anastassiou et al., 2011). The improved performance of the 10-gene signature indicates that the molecular changes from DCIS to IBC not only are related to the EMT process but also involve other biological processes captured by the 10-gene signature. To further study the biological processes, larger DCIS cohorts will need to be collected beyond those in the current training cohort (Sweden, Italy, and Norway).

Since our analyses were performed across the molecular intrinsic subtypes of breast cancer, the identified genes can detect DCIS cases, regardless of their subtype. The genes therefore represent general invasion processes, while the subtype-specific tumor progression processes may be obscured. A major proportion of breast cancer samples are Luminal A, and this is also the case in the training cohort. We have previously shown that Luminal A DCIS and IBC are highly similar at a molecular level, while basal-like DCIS differ substantially from basal-like IBC (Bergholtz et al., 2020). Stratification by subtype prior to creating the models could yield different results and identify genes and biological processes relevant within each subtype, but this approach would, in our high-dimensional analysis, be limited due to rather low sample size of the current cohorts. We believe that a subtype-agnostic model should become more practical for a clinical application of the signature, avoiding the need for subtype classification of each DCIS case. Additional genes would need to be included, such as those in the PAM50 signature, if one wants to construct risk signatures separately for the established subtypes. Furthermore, many studies have found stromal difference between DCIS and IBC (Dabiri et al., 2013; Toss et al., 2020), and it would be interesting to investigate how these 10 genes are expressed in stromal component vs. other components using spatial gene expression profiling.

Our results of the classification analyses using the two options to represent DNA methylation (preselected enhancer and promoter CpGs related to breast cancer biology or PCA-derived gene-level methylation) suggests that few individual CpGs cannot capture enough variation for accurate prediction and that a certain number of CpGs (>30 features) are needed to represent a meaningful information identifying DCIS from IBC. Moreover, we observed that CpG-level methylation features show higher sensitivity than gene-level methylation features using the Lasso model. This result highlights the importance of both enhancer and promoter methylation for gene regulation in breast cancer. On the other hand, the gene-level methylation represents many CpGs for each gene and thus it captures more variation, but some important CpGs may be masked by the PCA summarization approach. Furthermore, classification made using only a few individual CpGs may be vulnerable to measurement noise, and this can be overcome by increasing the number of CpGs in the classifier. Using all the 450,000 CpGs led to a poor class prediction performance, likely due to model overfitting (data not shown). Since the optimal processing of DNA methylation data is still poorly understood, we hope these results will provide guidance for the community on how to use methylation features in predictive modeling, either alone or combined with other omics features.

We initially tested several classification algorithms, Lasso, SVM, and RF, which all supported the importance of multi-omics profiles for increased DCIS detection sensitivity (**Supplementary Tables 1–3**). The lasso-regularized model generally showed the best performance and was therefore selected to showcase the classification results, for instance, when limiting the maximum number of features in sparse predictive modeling (**Figure 2**). Compared to genome-wide measurements, such minimal predictive signatures may lead to more practical prediction models for clinical decision tools in the form of cost-effective signatures for economic implementation. As observed before, nested CV was found important to avoid selection bias and reporting of overoptimistic results about the predictive power of classifier (Ambroise and McLachlan, 2002; Varma and Simon, 2006). As a future research direction, we plan to make use of pathway information for mapping the predictive genes that may potentially lead to even more robust and accurate models using pathway-level biomarkers (Ben-Hamo et al., 2020; Madani Tonekaboni et al., 2020). While the present work focused solely on protein-coding genes, since this enabled better interpretation of the model results and easier integration among the three data types, recent work has shown the influence of non-coding gene expression on cancer progression (Bhan et al., 2017; Chi et al., 2019; Zhang et al., 2021). As a future development, it would be interesting to use also non-coding DNA or RNA as additional source of features in the classification between DCIS and IBC cases.

In conclusion, our results support the use of the 10-gene signature to reliably identify those DCIS cases that are less likely to progress to invasive disease and may therefore have potential for reducing the current overtreatment in breast cancer. Longitudinal follow-up data of the DCIS cases will be needed for prognostic validation of the signature in terms of its accuracy at identifying high-risk vs. low-risk DCIS cases, and to explore how many of the initially DCIS diagnosed cases will eventually progress to an invasive disease or become invasive recurrent.

## 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/s.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by approval numbers: 2016/433 (Oslo, Norway), PG/U-25/01/2012-00001497 (Milan, Italy), and 2005/118 (Uppsala, Sweden). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

HX analyzed the data, implemented the predictive models, prepared the figures, and drafted the manuscript. TL contributed to the data analysis and predictive modeling. HB existed risk scores. HB and TS interpreted the biological results. TF provided the methylation data process. LD and AV-S provided the validation data. LD, AV-S, and TA interpreted the results. TS and TA co-supervised the work. TA designed the study. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

We thank Zhi Zhao (OUH/OCBE) for his help with the predictive modeling, Jørgen Ankill (OUH) for his help with the methylation data processing, and Arnoldo Frigessi (OCBE) for fruitful discussions about the modeling approaches. We are also indebted to Maria Grazia Daidone (Istituto Nazionale dei Tumori, Italy), Fredrik Wärnberg (Sahlgrenska University Hospital, Sweden), and to the Oslo Breast Cancer Consortium for the continuous support and access to tumor samples.

## FUNDING

This work was supported by Grants from Helse Sør-Øst (2020026 to TA, 2012056 to TS, and 2017065 to TF) and the Norwegian Cancer Society (216104 to TA and 420056 to TS).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.670749/full#supplementary-material

## REFERENCES

Abba, M. C., Gong, T., Lu, Y., Lee, J., Zhong, Y., Lacunza, E., et al. (2015). A molecular portrait of high-grade ductal carcinoma *in situ*. *Cancer Res.* 75, 3980–3990. doi: 10.1158/0008-5472.CAN-15-0506

Ahmad, A., Hanby, A., Dublin, E., Poulsom, R., Smith, P., Barnes, D., et al. (1998). Stromelysin 3: an independent prognostic factor for relapse-free survival in node-positive breast cancer and demonstration of novel breast carcinoma cell expression. *Am. J. Pathol.* 152, 721–728.

Ambroise, C., and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6562–6566. doi: 10.1073/pnas.102102699

Anastassiou, D., Rumjantseva, V., Cheng, W., Huang, J., Canoll, P. D., Yamashiro, D. J., et al. (2011). Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained *in vivo*. *BMC Cancer* 11:529. doi: 10.1186/1471-2407-11-529

Aure, M. R., Vitelli, V., Jernström, S., Kumar, S., Krohn, M., Due, E. U., et al. (2017). Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Res.* 19:18. doi: 10.1186/s13058-017-0812-y

Ben-Hamo, R., Jacob Berger, A., Gavert, N., Miller, M., Pines, G., Oren, R., et al. (2020). Predicting and affecting response to cancer therapy based on pathway-level biomarkers. *Nat. Commun.* 11:3296. doi: 10.1038/s41467-020-17090-y

Bergholtz, H., Lien, T. G., Swanson, D. M., Frigessi, A., Bathen, T. F., Borgen, E., et al. (2020). Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions. *npj Breast Cancer* 6:26. doi: 10.1038/s41523-020-0167-x

Bhan, A., Soleimani, M., and Mandal, S. S. (2017). Long noncoding RNA and cancer: a new paradigm. *Cancer Res.* 77, 3965–3981. doi: 10.1158/0008-5472.CAN-16-2634

Chi, Y., Wang, D., Wang, J., Yu, W., and Yang, J. (2019). Long non-coding RNA in the pathogenesis of cancers. *Cells* 8:1015. doi: 10.3390/cells8091015

Collins, L. C., Tamimi, R. M., Baer, H. J., Connolly, J. L., Colditz, G. A., and Schnitt, S. J. (2005). Outcome of patients with ductal carcinoma *in situ* untreated after diagnostic biopsy: results from the nurses' health study. *Cancer* 103, 1778–1784. doi: 10.1002/cncr.20979

Cowell, C. F., Weigelt, B., Sakr, R. A., Ng, C. K. Y., Hicks, J., King, T. A., et al. (2013). Progression from ductal carcinoma *in situ* to invasive breast cancer: revisited. *Mol. Oncol.* 7, 859–869. doi: 10.1016/j.molonc.2013.07.005

Dabiri, S., Talebi, A., Shahryari, J., Meymandi, M. S., and Safizadeh, H. (2013). Distribution of myofibroblast cells and microvessels around invasive ductal carcinoma of the breast and comparing with the adjacent range of their normal-to-DCIS zones. *Arch. Iran. Med.* 16, 93–99.

Daniele, G., Simonetti, G., Fusilli, C., Iacobucci, I., Lonoce, A., Palazzo, A., et al. (2017). Epigenetically induced ectopic expression of uncx impairs the proliferation and differentiation of myeloid cells. *Haematologica* 102, 1204–1214. doi: 10.3324/haematol.2016.163022

Esserman, L. J., Thompson, I. M., Reid, B., Nelson, P., Ransohoff, D. F., Welch, H. G., et al. (2014). Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol.* 15, e234–e242. doi: 10.1016/S1470-2045(13)70598-9

Fleischer, T., Frigessi, A., Johnson, K. C., Edvardsen, H., Touleimat, N., Klajic, J., et al. (2014). Genome-wide DNA methylation profiles in progression to *in situ* and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol.* 15:435. doi: 10.1186/PREACCEPT-2333349012841587

Fleischer, T., Tekpli, X., Mathelier, A., Wang, S., Nebdal, D., Dhakal, H. P., et al. (2017). DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* 8:1379. doi: 10.1038/s41467-017-00510-x

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

Gorringe, K. L., and Fox, S. B. (2017). Ductal carcinoma *in Situ* biology, biomarkers, and diagnosis. *Front. Oncol.* 7:248. doi: 10.3389/fonc.2017.00248

Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing Precision-recall and receiver operating characteristic curves in R. *Bioinformatics* 31, 2595–2597. doi: 10.1093/bioinformatics/btv153

Groen, E. J., Elshof, L. E., Visser, L. L., Rutgers, E. J. T., Winter-Warnars, H. A. O., Lips, E. H., et al. (2017). Finding the balance between over- and under-treatment of ductal carcinoma *in situ* (DCIS). *Breast* 31, 274–283. doi: 10.1016/j.breast.2016.09.001

Hannemann, J., Velds, A., Halfwerk, J. B., Kreike, B., Peterse, J. L., and Van de Vijver, M. J. (2006). Classification of ductal carcinoma *in situ* by gene expression profiling. *Breast Cancer Res.* 8:R61. doi: 10.1186/bcr1613

Jeong, D., Ban, S., Kim, H., Oh, S., Ji, S., Kim, H. J., et al. (2016). Abstract 709: *in vitro* functional study of novel oncogene serine protease 33 (PRSS33) and the clinical significance of PRSS33 expression in colorectal cancer patients. *Cancer Res.* 76, 709–709. doi: 10.1158/1538-7445.am2016-709

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Lee, H., Park, B. C., Soon Kang, J., Cheon, Y., Lee, S., and Jae Maeng, P. (2020). MAM domain containing 2 is a potential breast cancer biomarker that exhibits tumour-suppressive activity. *Cell Prolif.* 53:e12883. doi: 10.1111/cpr.12883

Lesurf, R., Aure, M. R., Mørk, H. H., Vitelli, V., Lundgren, S., Børresen-Dale, A. L., et al. (2016). Molecular features of subtype-specific progression from ductal carcinoma *in situ* to invasive breast cancer. *Cell Rep.* 16, 1166–1179. doi: 10.1016/j.celrep.2016.06.051

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2, 18–22.

Liu, J., Shen, J. X., Wu, H. T., Li, X. L., Wen, X. F., Du, C. W., et al. (2018). Collagen 1A1 (COL1A1) promotes metastasis of breast cancer and is a potential therapeutic target. *Discov. Med.* 25, 211–223.

Love, M., Anders, S., and Huber, W. (2017). Analyzing RNA-seq data with DESeq2. *Bioconductor* 2, 1–63.

Ma, X. J., Salunga, R., Tuggle, J. T., Gaudet, J., Enright, E., McQuary, P., et al. (2003). Gene expression profiles of human breast cancer progression. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5974–5979. doi: 10.1073/pnas.0931261100

Madani Tonekaboni, S. A., Beri, G., and Haibe-Kains, B. (2020). Pathway-based drug response prediction using similarity identification in gene expression. *Front. Genet.* 11:1016. doi: 10.3389/fgene.2020.01016

McSherry, E. A., Donatello, S., Hopkins, A. M., and McDonnell, S. (2007). Molecular basis of invasion in breast cancer. *Cell. Mol. Life Sci.* 64, 3201–3218. doi: 10.1007/s00018-007-7388-0

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., et al. (2019). *Package 'e1071': Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package version* 1.7-3, 1–63. Available online at: https://cran.r-project.org/web/packages/e1071/index.html (accessed November 26, 2019).

Motrescu, E. R., Blaise, S., Etique, N., Messaddeq, N., Chenard, M. P., Stoll, I., et al. (2008). Matrix metalloproteinase-11/stromelysin-3 exhibits collagenolytic function against collagen VI under normal and malignant conditions. *Oncogene* 27, 6347–6355. doi: 10.1038/onc.2008.218

Muggerud, A. A., Hallett, M., Johnsen, H., Kleivi, K., Zhou, W., Tahmasebpoor, S., et al. (2010). Molecular diversity in ductal carcinoma *in situ* (DCIS) and early invasive breast cancer. *Mol. Oncol.* 4, 357–368. doi: 10.1016/j.molonc.2010.06.007

Nielsen, M., Jensen, J., and Andersen, J. (1984). Precancerous and cancerous breast lesions during lifetime and at autopsy. A study of 83 women. *Cancer* 54, 612–615.

Onega, T., Weaver, D. L., Frederick, P. D., Allison, K. H., Tosteson, A. N. A., Carney, P. A., et al. (2017). The diagnostic challenge of low-grade ductal carcinoma *in situ*. *Eur. J. Cancer* 80, 39–47. doi: 10.1016/j.ejca.2017.04.013

Page, D. L., Dupont, W. D., Rogers, L. W., Jensen, R. A., and Schuyler, P. A. (1995). Continued local recurrence of carcinoma 15–25 years after a diagnosis of low grade ductal carcinoma *in situ* of the breast treated only by biopsy. *Cancer* 76, 1197–1200.

Page, D. L., Dupont, W. D., Rogers, L. W., and Landenberger, M. (1982). Intraductal carcinoma of the breast: follow-up after biopsy only. *Cancer* 49, 751–758.

Pan, W., Arnone, M., Kendall, M., Grafstrom, R. H., Seitz, S. P., Wasserman, Z. R., et al. (2003). Identification of peptide substrates for human MMP-11 (stromelysin-3) using phage display. *J. Biol. Chem.* 278, 27820–27827. doi: 10.1074/jbc.M304436200

Pang, J. M. B., Savas, P., Fellowes, A. P., Mir Arnau, G., Kader, T., Vedururu, R., et al. (2017). Breast ductal carcinoma *in situ* carry mutational driver events representative of invasive breast cancer. *Mod. Pathol.* 30, 952–963. doi: 10.1038/modpathol.2017.21

Parker, J. S., Bernard, P. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi: 10.1200/JCO.2008.18.1370

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Sanders, M. E., Schuyler, P. A., Dupont, W. D., and Page, D. L. (2005). The natural history of low-grade ductal carcinoma *in situ* of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* 103, 2481–2484. doi: 10.1002/cncr.21069

Seely, J. M., and Alhassan, T. (2018). Screening for breast cancer in 2018—what should we be doing today? *Curr. Oncol.* 25, S115–S124. doi: 10.3747/co.25.3770

Solin, L. J., Gray, R., Baehner, F. L., Butler, S. M., Hughes, L. L., Yoshizawa, C., et al. (2013). A multigene expression assay to predict local recurrence risk for ductal carcinoma *in situ* of the breast. *J. Natl. Cancer Inst.* 105, 701–710. doi: 10.1093/jnci/djt067

Toss, M. S., Abidi, A., Lesche, D., Joseph, C., Mahale, S., Saunders, H., et al. (2020). The prognostic significance of immune microenvironment in breast ductal carcinoma *in situ*. *Br. J. Cancer* 122, 1496–1506. doi: 10.1038/s41416-020-0797-7

Touleimat, N., and Tost, J. (2012). Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 4, 325–341. doi: 10.2217/epi.12.21

van Seijen, M., Lips, E. H., Thompson, A. M., Nik-Zainal, S., Futreal, A., Hwang, E. S., et al. (2019). Ductal carcinoma *in situ*: to treat or not to treat, that is the question. *Br. J. Cancer* 121, 285–292. doi: 10.1038/s41416-019-0478-6

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91. doi: 10.1186/1471-2105-7-91

Vincent-Salomon, A., Lucchesi, C., Gruel, N., Raynal, V., Pierron, G., Goudefroye, R., et al. (2008). Integrated genomic and transcriptomic analysis of ductal carcinoma *in situ* of the breast. *Clin. Cancer Res.* 14, 1956–1965. doi: 10.1158/1078-0432.CCR-07-1465

Virnig, B. A., Tuttle, T. M., Shamliyan, T., and Kane, R. L. (2010). Ductal carcinoma *in situ* of the breast: a systematic review of incidence, treatment, and outcomes. *J. Natl. Cancer Inst.* 102, 170–178. doi: 10.1093/jnci/djp482

Wallis, M. G., Clements, K., Kearins, O., Ball, G., MacArtney, J., and Lawrence, G. M. (2012). The effect of DCIS grade on rate, type and time to recurrence after 15 years of follow-up of screen-detected DCIS. *Br. J. Cancer* 106, 1611–1617. doi: 10.1038/bjc.2012.151

Wang, G., Zhang, Q., Song, Y., Wang, X., Guo, Q., Zhang, J., et al. (2015). PAK1 regulates RUFY3-mediated gastric cancer cell migration and invasion. *Cell Death Dis.* 6:e1682. doi: 10.1038/cddis.2015.50

Wang, J. K., Wang, W. J., Cai, H. Y., Du, B. B., Mai, P., Zhang, L. J., et al. (2018). MFAP2 promotes epithelial–mesenchymal transition in gastric cancer cells by activating TGF-β/SMAD2/3 signaling pathway. *Onco Targets Ther.* 11, 4001–4017. doi: 10.2147/OTT.S160831

Wang, S. Y., Shamliyan, T., Virnig, B. A., and Kane, R. (2011). Tumor characteristics as predictors of local recurrence after treatment of ductal carcinoma *in situ*: a meta-analysis. *Breast Cancer Res. Treat.* 127, 1–14. doi: 10.1007/s10549-011-1387-4

Zhang, M., Chen, H., Wang, M., Bai, F., and Wu, K. (2020). Bioinformatics analysis of prognostic significance of COL10A1 in breast cancer. *Biosci. Rep.* 40:BSR20193286. doi: 10.1042/BSR20193286

Zhang, R., Zhu, Q., Yin, D., Yang, Z., Guo, J., Zhang, J., et al. (2021). Identification and validation of an autophagy-related lncRNA signature for patients with breast cancer. *Front. Oncol.* 10:597569. doi: 10.3389/fonc.2020.597569

Zhang, X., Huang, S., Guo, J., Zhou, L., You, L., Zhang, T., et al. (2016). Insights into the distinct roles of MMP-11 in tumor biology and future therapeutics (Review). *Int. J. Oncol.* 48, 1783–1793. doi: 10.3892/ijo.2016.3400

**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: 09 June 2021
doi: 10.3389/fgene.2021.648800

# Whole Transcriptome Data Analysis Reveals Prognostic Signature Genes for Overall Survival Prediction in Diffuse Large B Cell Lymphoma

*Mengmeng Pan[1,2†], Pingping Yang[1†], Fangce Wang[1†], Xiu Luo[1], Bing Li[1], Yi Ding[1], Huina Lu[1], Yan Dong[1], Wenjun Zhang[1*], Bing Xiu[1*] and Aibin Liang[1*]*

[1] Department of Hematology, Tongji Hospital, Tongji University School of Medicine, Shanghai, China, [2] National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

OPEN ACCESS

**Edited by:**
Rosalba Giugno,
University of Verona, Italy

**Reviewed by:**
Deli Liu,
Weill Cornell Medicine, United States
Michael Poidinger,
Royal Children's Hospital, Australia

***Correspondence:**
Wenjun Zhang
zhangwenjun@tongji.edu.cn
Bing Xiu
xiubing1233@tongji.edu.cn
Aibin Liang
lab7182@tongji.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:** 02 January 2021
**Accepted:** 17 May 2021
**Published:** 09 June 2021

**Citation:**
Pan M, Yang P, Wang F, Luo X, Li B, Ding Y, Lu H, Dong Y, Zhang W, Xiu B and Liang A (2021) Whole Transcriptome Data Analysis Reveals Prognostic Signature Genes for Overall Survival Prediction in Diffuse Large B Cell Lymphoma. Front. Genet. 12:648800. doi: 10.3389/fgene.2021.648800

**Background:** With the improvement of clinical treatment outcomes in diffuse large B cell lymphoma (DLBCL), the high rate of relapse in DLBCL patients is still an established barrier, as the therapeutic strategy selection based on potential targets remains unsatisfactory. Therefore, there is an urgent need in further exploration of prognostic biomarkers so as to improve the prognosis of DLBCL.

**Methods:** The univariable and multivariable Cox regression models were employed to screen out gene signatures for DLBCL overall survival (OS) prediction. The differential expression analysis was used to identify representative genes in high-risk and low-risk groups, respectively, where student *t* test and fold change were implemented. The functional difference between the high-risk and low-risk groups was identified by the gene set enrichment analysis.

**Results:** We conducted a systematic data analysis to screen the candidate genes significantly associated with OS of DLBCL in three NCBI Gene Expression Omnibus (GEO) datasets. To construct a prognostic model, five genes (*CEBPA*, *CYP27A1*, *LST1*, *MREG*, and *TARP*) were then screened and tested using the multivariable Cox model and the stepwise regression method. Kaplan–Meier curve confirmed the good predictive performance of this five-gene Cox model. Thereafter, the prognostic model and the expression levels of the five genes were validated by means of an independent dataset. High expression levels of these five genes were significantly associated with favorable prognosis in DLBCL, both in training and validation datasets. Additionally, further analysis revealed the independent value and superiority of this prognostic model in risk prediction. Functional enrichment analysis revealed some vital pathways responsible for unfavorable outcome and potential therapeutic targets in DLBCL.

**Conclusion:** We developed a five-gene Cox model for the clinical outcome prediction of DLBCL patients. Meanwhile, potential drug selection using this model can help clinicians to improve the clinical practice for the benefit of patients.

**Keywords: diffuse large B cell lymphoma, overall survival, prognosis, biomarkers, risk score**

## INTRODUCTION

Diffuse large B cell lymphoma (DLBCL) is the most common type of aggressive non-Hodgkin lymphoma with an annual incidence of 1–5/10,000 (Li et al., 2018; Marangon et al., 2019). DLBCL is an aggressive and potentially curable hematological malignancy, which makes an early diagnosis and effective treatments essential for patients. R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) is currently the standard first line treatment of DLBCL (Coiffier et al., 2002). Despite the high rate of complete response (76%), approximately 40% of patients will relapse, and the molecular mechanism underlying recurrence remains largely unknown (Coiffier et al., 2010). DLBCL displays tremendous clinical, genetic and molecular heterogeneity. The International Prognostic Index (IPI) has been used to predict the prognosis of patients with DLBCL for nearly 30 years, yet there still exists a minority of patients whose clinical process were not in accord with the IPI stratification (International Non-Hodgkin's Lymphoma Prognostic Factors Project, 1993). Gene expression profiling has helped identify two major subtypes, known as germinal center B-cell-like (GCB) and activated B-cell-like (ABC), and patients with ABC DLBCL exhibit a generally worse prognosis (Lenz et al., 2008a). However, the high prices and strict requirements regarding tissue limit the routine use of this method. Therefore, efforts have been made to find novel biomarkers with prognostic values in order to improve therapeutic strategy selection based on potential targets (Cabanillas and Shah, 2017).

Currently, various markers are defined through immunophenotyping, such as CD5, CD30, BCL2, MYC, and TP53 (Pierce and Mehta, 2017; Zhao et al., 2019). CD5 promotes downstream B-cell receptor signaling, is associated with ABC subtype and more aggressive clinical traits. Patients with $CD30^{+}$ DLBCL, which leads to the downregulation of NF-κB and B-cell receptor signaling, tend to exhibit a better prognosis (Bhatt et al., 2016; Thakral et al., 2017). Meanwhile, in patients with the GCB subtype, BCL2 and MYC rearrangements would lead to worse prognosis (Visco et al., 2013). TP53 mutation also adversely affects patients' prognosis (Xu-Monette et al., 2012). Based on the new integrated genetic map, Chapuy et al. (2018) identified distinct subsets, including a previously unrecognized group of low-risk ABC-DLBCLs, two GCB-DLBCLs subsets with different prognoses and an ABC/GCB-independent group. In addition, Schmitz et al. (2018) uncovered some previously unknown subtypes of DLBCL by differences in gene-expression signatures and responses to immunochemotherapy. The subset of high-risk patients requires revolutionized therapeutics, and personalized therapy based on patient's histological and molecular-genetic characteristics will bring greater benefits to patients. Therefore, further exploration of prognostic indicators is still needed to distinguish DLBCL patients of varied prognosis.

**Abbreviations:** DLBCL, diffuse large B cell lymphoma; IPI, International Prognostic Index; GCB, germinal center B-cell-like; ABC, activated B-cell-like; GEO, Gene Expression Omnibus; LDH, serum lactate dehydrogenase; ECOG, Eastern Cooperative Oncology Group; CHOP, combine with intensive chemotherapy; circRNA, circular RNAs; HCC, hepatocellular carcinoma; ncRNA, non-coding RNAs; PVTT, portal vein tumor thrombosis; GSEA, gene set enrichment analysis.

## MATERIALS AND METHODS

### Data Collection

The gene expression data and corresponding clinical information were collected from NCBI Gene Expression Omnibus (GEO) database with accession numbers of GSE32918 (Barrans et al., 2012) ($n$ = 172), GSE4475 (Hummel et al., 2006) ($n$ = 166), GSE69051 (Sha et al., 2015) ($n$ = 149), TCGA (Schmitz et al., 2018) ($n$ = 43), GSE31312 (Visco et al., 2012) ($n$ = 470), GSE34171 (Monti et al., 2012) ($n$ = 68), GSE11318 (Lenz et al., 2008b) ($n$ = 203), and GSE10846 (Lenz et al., 2008a) ($n$ = 414). It should be noted that Burkitt lymphoma samples in GSE69051 and GSE4475 have been excluded in this study. Among these datasets, GSE32918, GSE4475, and GSE69051 were used for feature selection and model training, while the remaining datasets including TCGA, GSE31312, GSE34171, GSE11318, and GSE10846 were used as independent validation datasets. The expression values were normalized by the data submitters, and discretized by median values, which were used for downstream analysis.

### Cox Proportional Hazard Model

The univariable Cox proportional hazard model was used to screen prognostic genes in the first three datasets. To integrate the three datasets and remove batch effect, we converted the continuous expression values of the shared genes into two discrete expression levels, i.e., high and low expression, using the median expression as the threshold value. The principal component analysis based on the discretized expression levels revealed that no clear batch effect was observed between the three datasets (Kruskal–Wallis test for the top two principal components, $P$-value $> 0.05$, **Supplementary Figure 1**), suggesting that there was no significant transcriptional difference between the three datasets. The comparison of the clinical factors indicated that there were significant differences in age and proportion of deceased cases among the three datasets (**Supplementary Table 1**). Those three discretized datasets of the shared prognostic signatures were then merged and used as the training set for the multivariable Cox model, and the stepwise regression method was used to determine the best model based on the Akaike Information Criterion (AIC). The risk scores for the samples of training and validation sets were estimated using the multivariable Cox model based on the expression levels of those five genes. The high- and low-risk groups were stratified based on the median of the risk scores in the training set. The independent value of this risk stratification was also assessed by multivariable Cox model.

### Differential Gene Expression Analysis

The differential gene expression analysis was conducted to identify the genes that were upregulated or downregulated between specific risk groups. The Wilcoxon rank-sum test and fold change methods were employed, and the thresholds

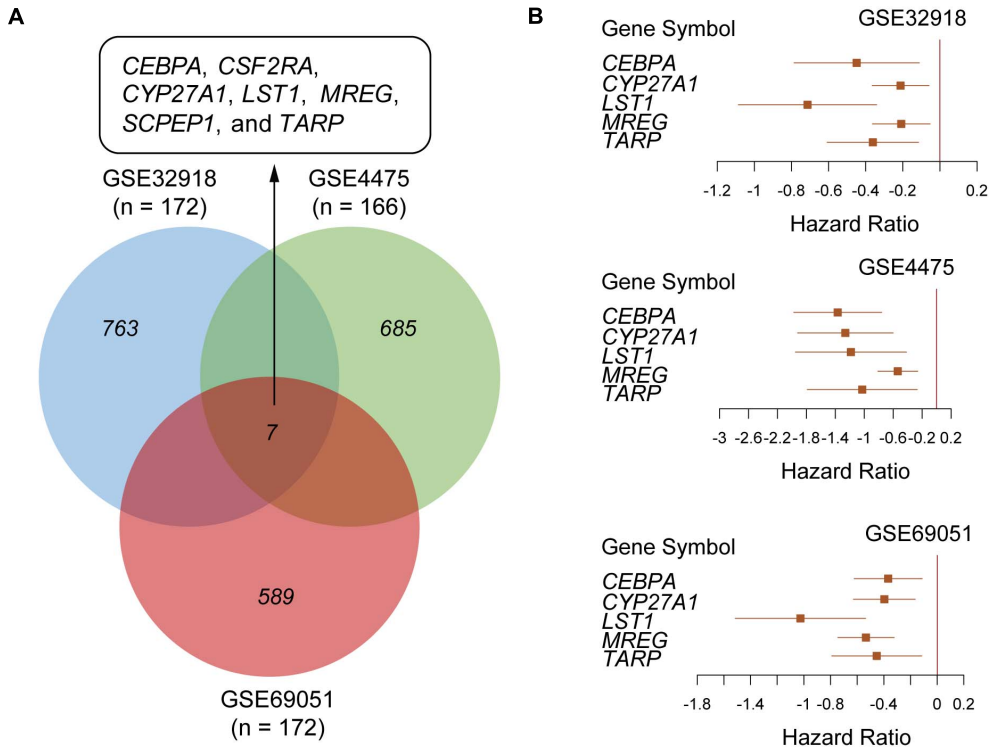



**FIGURE 1 |** Screening a five-gene Cox model in the public DLBCL datasets from Gene Expression Omnibus (GEO). **(A)** Venn diagram summarizing the overlap between the prognostic genes identified by univariable Cox regression analysis in three public DLBCL datasets with accession numbers of GSE32918 (*n* = 172), GSE4475 (*n* = 166) and GSE69051 (*n* = 172). **(B)** The forest plots represent the association of the five gene signatures with overall survival in the three public DLBCL datasets.

**TABLE 1 |** The statistics for the gene signatures in the multivariable Cox model.

| Gene | coef | exp (coef) | se(coef) | *Z* | Pr(> \| z\|) |
|---|---|---|---|---|---|
| *CEBPA* | −0.384 | 0.681 | 0.180 | −2.138 | 3.25E-02 |
| *CYP27A1* | −0.390 | 0.677 | 0.187 | −2.086 | 3.69E-02 |
| *LST1* | −0.468 | 0.626 | 0.178 | −2.631 | 8.50E-03 |
| *MREG* | −0.420 | 0.657 | 0.170 | −2.471 | 1.35E-02 |
| *TARP* | −0.292 | 0.746 | 0.156 | −1.873 | 6.11E-02 |

of adjusted *p*-value and log2-fold change were determined at 0.05 and 0.5.

## The Pathway Enrichment Analysis

The upregulated genes in each risk group were further investigated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, respectively. Hypergeometric test was applied to test the statistical significance of those identified pathways. The threshold for adjusted *P*-value was determined at 0.05.

## The Drug-Target Identification

The therapeutic targets were selected from the upregulated genes in each risk group. The drugs and upregulated genes were mapped by the R package maftools with *drugInteractions*.

# RESULTS

## Systematic Identification of Prognostic Gene Signatures for Overall Survival Prediction

To identify the prognostic gene signatures, we collected three public DLBCL datasets with accession numbers of GSE32918 ($n$ = 172), GSE4475 ($n$ = 166), and GSE69051 ($n$ = 149) from GEO database as depicted in the flow chart in **Supplementary Figure 1**. Subsequently, univariable Cox regression analysis was conducted, and a total of 763, 685, and 589 genes were identified to be associated with overall survival (OS) based on the gene expression profiles of these three datasets (**Figure 1A**, log-rank test, $P < 0.01$), respectively. Particularly, *CEBPA*, *CSF2RA*, *CYP27A1*, *LST1*, *MREG*, *SCPEP1*, and *TARP* were found to be significantly associated with OS in all the three datasets at the stringent threshold (**Figure 1A**). Furthermore, the three datasets were merged into one training set ($n$ = 487), and a multivariable Cox regression model was then built from gene expression profiles of the merged dataset. A stepwise method was used to select a subset of those gene signatures to construct a multivariable Cox regression model that could achieve the highest performance. Specifically, five genes including *CEBPA*, *CYP27A1*, *LST1*, *MREG*, and *TARP* were retained in the

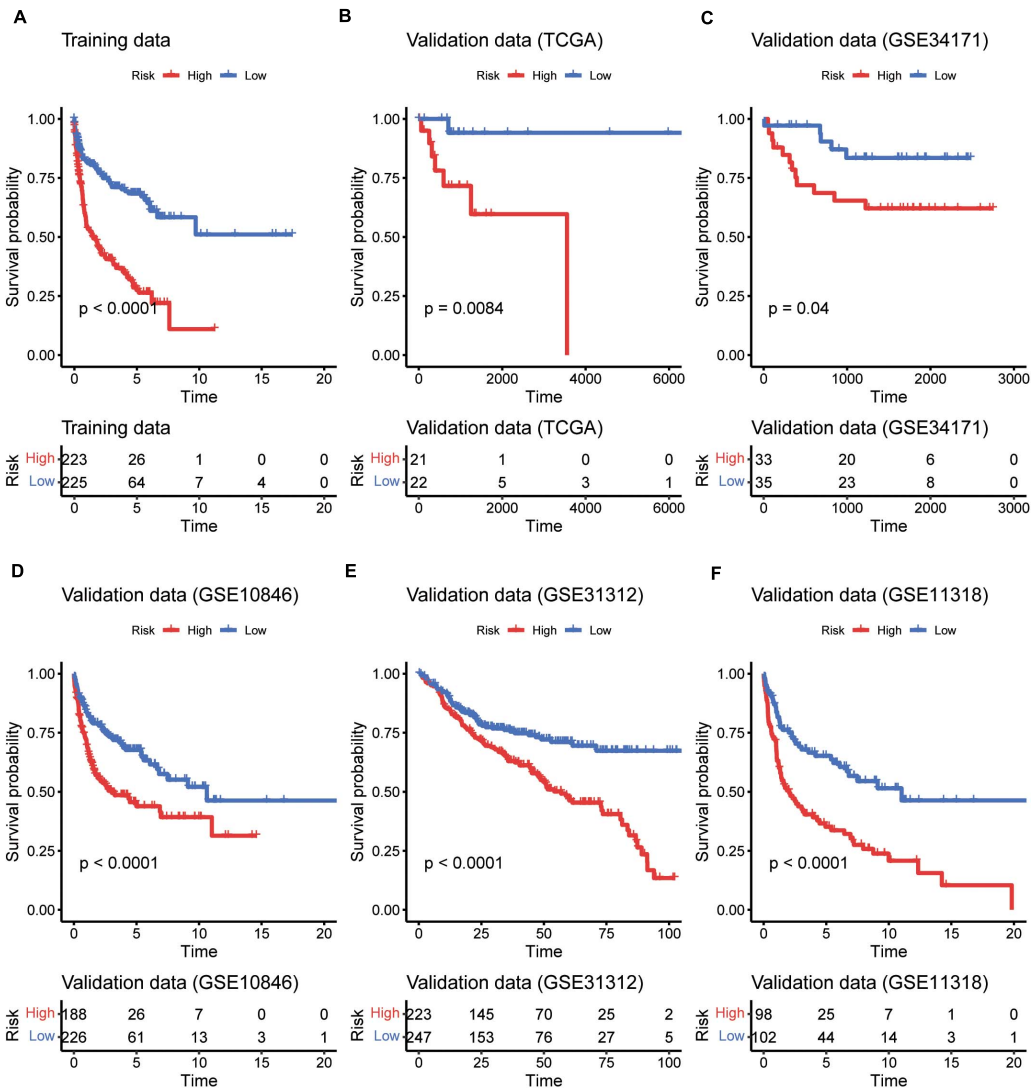



**FIGURE 2 |** The performance of the five gene signatures in predicting the patients' risk. K-M curves for the prognostic model in the training datasets **(A)** and the five validation datasets **(B–F)**. The red and blue lines represent the high- and low-risk groups, respectively. The numbers within risk tables on the bottom represent the number of survivors at that time point.

multivariable Cox model (**Table 1**), which was termed as the five-gene Cox model, and all of them were associated with favorable prognoses (**Figure 1B**).

## Performance Validation in an Independent Dataset

To evaluate the performance of the multivariable model in risk prediction, we first calculated the risk scores of the DLBCL samples in the training set, and stratified these samples into high- and low-risk groups by the median of risk scores. The high-risk group exhibited worse prognosis than the low-risk group (**Figure 2A**, $P < 0.0001$). Moreover, we also collected five independent gene expression datasets with long-term follow-up (TCGA, GSE31312, GSE34171, GSE11318, and GSE10846), predicted the risk scores and stratified the samples of those datasets into high- and low-risk groups. Consistently, these two groups also had significant difference in prognosis (**Figures 2B–F**, $P < 0.05$). Furthermore, the five gene signatures were found to be upregulated in low-risk group than high-risk group in both the training (**Figure 3A**) and validation sets (**Figures 3B–F**). These results indicated that these five gene signatures were robust and consistently associated with OS in both training and validation datasets.

## The Five-Gene Cox Model Is Superior to Other Gene Expression-Based Cox Models

To demonstrate the superiority of this five-gene Cox model based on the five gene signatures, we compared its performance

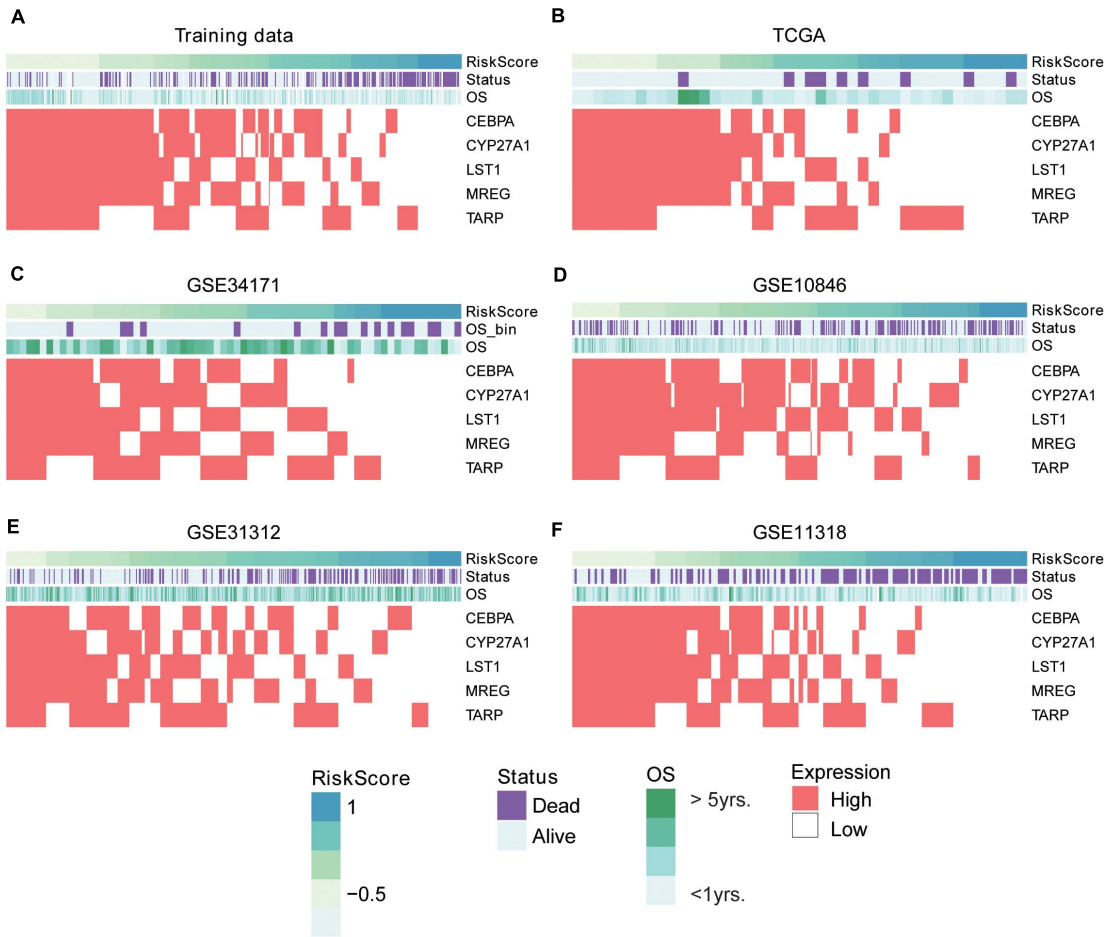



**FIGURE 3 |** The expression patterns of five prognostic gene signatures in the training and five validation sets. The expression patterns of the five prognostic genes in training **(A)** and validation **(B–F)** sets. The risk scores were estimated by the linear predictors of the Cox model. The samples were ordered by the risk scores.

with three sets of gene signatures (Rosenwald et al., 2002; Wright et al., 2003; Lossos, 2008) on the five validation datasets. Utilizing the trained models that were constructed from different gene signatures, the samples in the validation sets could also be stratified into high- and low-risk groups. The gene signatures proposed by Rosenwald et al. (2002) had the worst performance on almost all validation datasets (**Figure 4**). However, survival difference between samples stratified by our proposed five gene signatures was the most statistically significant across all the validation datasets (**Figure 4**), especially in the TCGA and GSE34171 cohorts with smaller sample size (**Figures 4A,B**), suggesting that the Cox model based on our five gene signatures was superior to other models.

## The Five-Gene-Based Risk Stratification Is a Prognostic Factor Independent of Clinical Factors

To further investigate the robustness of the five-gene Cox model, we tested whether the five-gene-based risk stratification was an independent predictor in the validation set. Since the IPI scoring system was a well-recognized factor for prognostic risk prediction and widely applied in clinical practice (Martelli et al., 2013), the samples were first divided into two groups of high ($\geq$3) or low ($<$3) IPI scores, considering age, serum lactate dehydrogenase (LDH), Eastern Cooperative Oncology Group (ECOG) Performance Status, Ann Arbor stage, and extranodal infiltration sites (International Non-Hodgkin's Lymphoma Prognostic Factors Project, 1993). As shown in **Figure 5A**, no significant difference was observed between the risk scores of the two groups, which were estimated using the five-gene Cox model (high vs. low IPI). Moreover, the differences were also not observed across the four stages. In contrast, the samples with high IPI had significantly higher risk scores when estimated with the three sets of gene signatures as mentioned above, than those with low IPI (**Supplementary Figure 2**). These results suggested that the risk scores were not only irrelevant to IPI scoring system and tumor stage, but also had a higher independent predictive values than those derived from previous gene signatures.

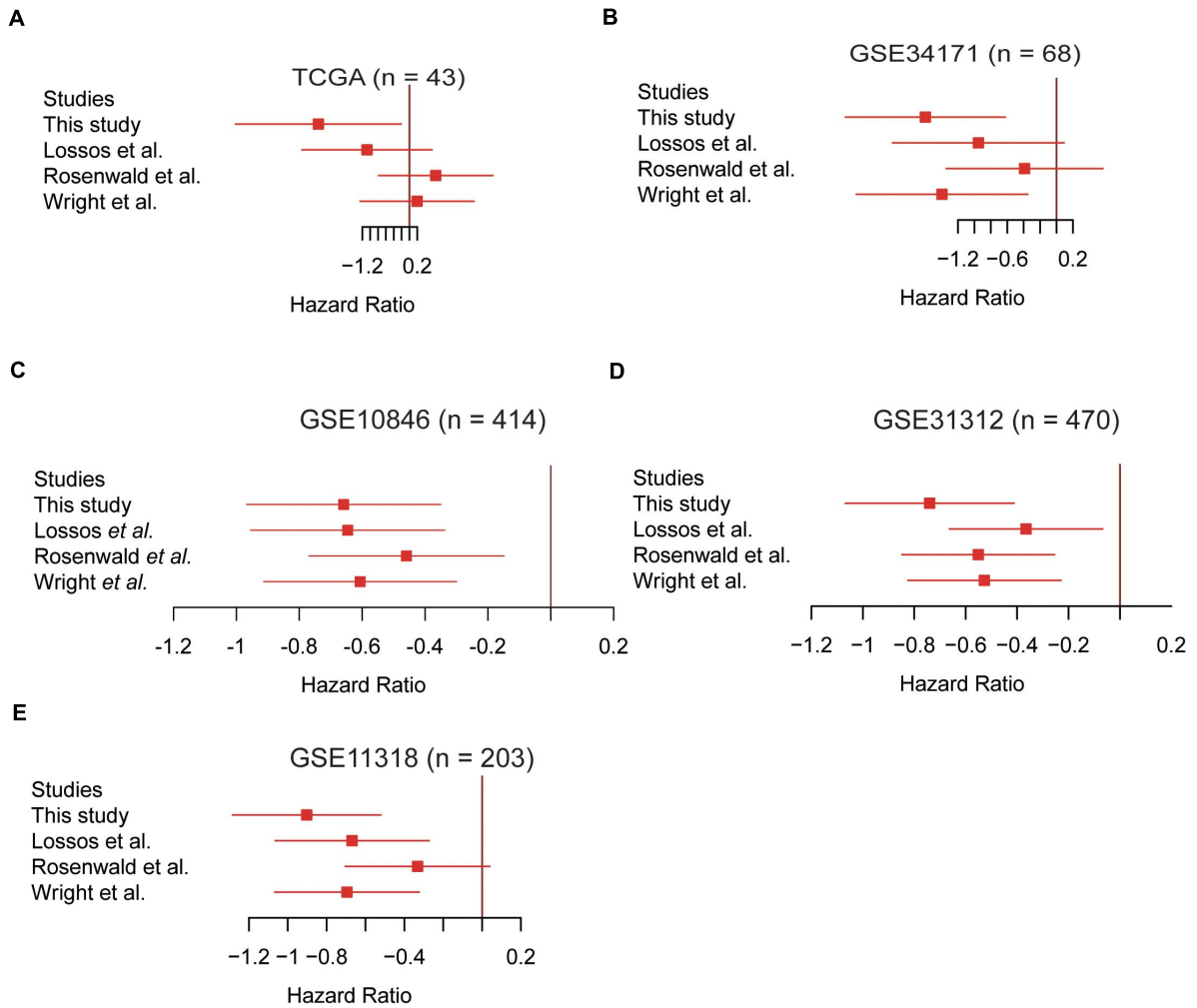



**FIGURE 4 |** The Cox model based on the five gene signatures was superior to other models. The performance of the four prognostic models in the validation datasets of TCGA ($n$ = 43), GSE34171 ($n$ = 68), GSE10846 ($n$ = 414), GSE31312 ($n$ = 470), and GSE11318 ($n$ = 203) are displayed in panels **(A–E)**. The log2-hazard ratios and 95% confidence intervals were denoted by the red boxes and lines.

Notably, the samples could be classified into four groups by combining the IPI scoring system and the five-gene-based risk stratification, and the four groups exhibited significantly prognostic difference (**Figure 5B**, $P < 0.0001$). It should be noted that the differences of OS were not observed between the two groups with the worse prognosis, but the samples with IPI $\geq 3$ in high-risk group still had shorter OS than samples with IPI $\geq 3$ in the low-risk group based on the KM curve.

Moreover, we also tested whether the risk stratification was independent of the DLBCL subtypes. Consistently, the three subtypes, including ABC, GCB and unclassified subtypes, could be further stratified into high- and low-risk groups. Except unclassified subtype, the ABC and GBC subtypes still maintained the statistical difference in OS between the high-risk and low-risk groups (**Figures 5C,D**, FDR $< 0.05$, and **Figure 5E**, FDR $> 0.05$). To test whether the chemotherapy treatment affects the performance of the gene signatures, we compared the two risk groups of patients treated with R-CHOP-like or CHOP-like regimens. Consistently, high-risk patients, who were treated with R-CHOP-like or CHOP-like regimens, still had shorter OS than the corresponding low-risk patients (**Figures 5F,G**), suggesting that the gene signatures were independent of the chemotherapy treatment. In addition, we also fitted the IPI scoring system, stage, subtype and risk stratification into a multivariable Cox model, and found that the risk stratification was still statistically significant with these prognostic factors as cofactors (**Table 2**). These results further demonstrated that the five-gene-based risk stratification was an independent prognostic factor for DLBCL risk prediction.

## The Molecular Characteristics and Potential Drugs for the Two Risk Groups

To reveal the molecular characteristics of the two risk groups, we compared the gene expression profiles of high-risk with those of low-risk group using the five validation datasets. A total

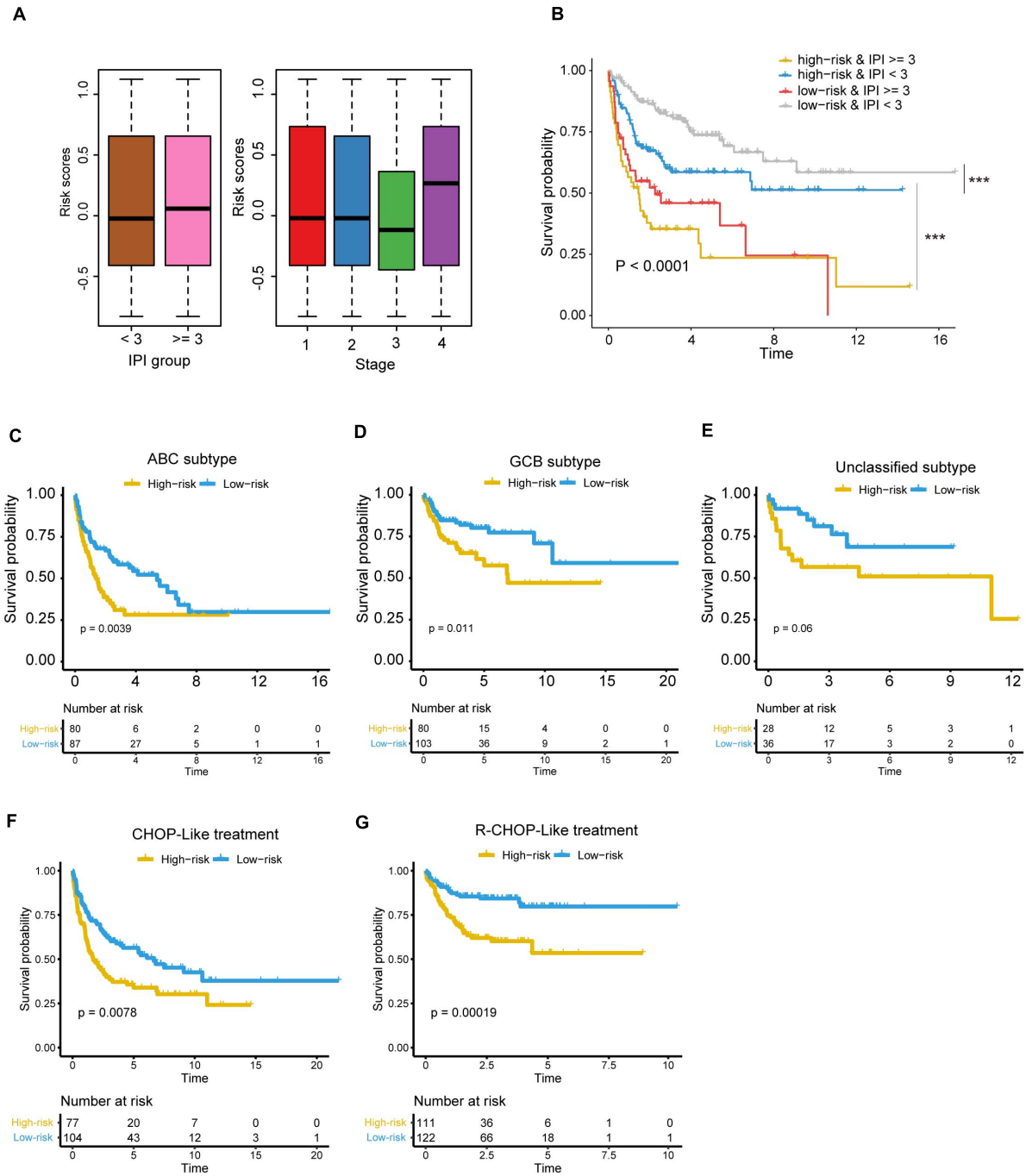



**FIGURE 5 |** The risk stratification based on the five prognostic genes is independent of clinical factors. **(A)** The risk scores in different IPI groups (left panel) and clinical stages (right panel). The boxes show the median and the interquartile range (IQR) of the risk scores grouped by the IPI scoring system and clinical stage in the validation dataset. There are no significant differences between those groups ($P > 0.05$). **(B)** Kaplan–Meier survival curves show the overall survival of samples grouped by combining the IPI scoring system and the five-gene-based risk stratification. ***$P < 0.0001$. The differences of overall survival between the high-risk and low-risk groups in specific subtype or with specific chemotherapy regiment [**(C)** ABC subtype; **(D)** GCB subtype; **(E)** unclassified subtype, **(F)** DLBCL treated with CHOP-Like regiment, **(G)** DLBCL treated with R-CHOP-Like regiment].

of 1,158 genes, jointly differentially expressed between high- and low-risk groups of the five validation datasets, were then selected by Wilcoxon rank-sum test and fold change (Adjusted *P*-value < 0.05 and log2-fold change > 0.5). Moreover, the overrepresentation enrichment analysis (ORA) was employed to identify the pathways potentially involved in the DLBCL

**TABLE 2 |** The statistics for the risk stratification and prognostically clinical factors in the multivariable Cox model.

| Variables | Log2 hazard ratio | Hazard ratio | Standard error | *Z* score | *P*-value |
|---|---|---|---|---|---|
| **Subtype** | | | | | |
| ABC | | | | | |
| GCB | −0.94 | 0.39 | 0.20 | −4.66 | 3.18E-06 |
| Unclassified | −0.79 | 0.45 | 0.27 | −2.94 | 3.26E-03 |
| **Stage** | | | | | |
| 1 | | | | | |
| 2 | 0.99 | 2.70 | 0.41 | 2.41 | 1.62E-02 |
| 3 | 0.64 | 1.89 | 0.44 | 1.45 | 1.47E-01 |
| 4 | 0.99 | 2.69 | 0.42 | 2.34 | 1.94E-02 |
| **Risk stratification** | | | | | |
| High-risk | | | | | |
| Low-risk | −0.59 | 0.55 | 0.18 | −3.34 | 8.46E-04 |
| **IPI** | | | | | |
| <3 | | | | | |
| ≥3 | 1.02 | 2.77 | 0.21 | 4.83 | 1.40E-06 |
| **Treatment** | | | | | |
| R-CHOP | | | | | |
| R-CHOP-like | −0.72 | 0.48 | 0.19 | −3.74 | 1.82E-04 |

progression (**Figure 6A**). Specifically, cell cycle-related pathway and those associated with genomic stability maintenance, such as mismatch repair, were highly upregulated in high-risk group (Adjusted *P*-value < 0.05). In contrast, immune-related pathways such as rheumatoid arthritis, antigen processing and presentation, hematopoietic cell lineage, and Th1 and Th2 cell differentiation were upregulated in low-risk group (Adjusted *P*-value < 0.05). Moreover, we also conducted correlation analysis between our signature genes and the DEGs in the five validation datasets. As high expression of the five signature genes indicates better prognosis, consistently, they are positively or conversely correlated with most of the upregulated genes in high-risk or low groups, respectively, indicating that those DEGs might also be associated with prognosis to a certain extent (**Figure 6B**).

For the low-risk group, some immune checkpoint proteins and inhibitors were identified, such as PDCD1 (PD-1), CD274 (PD-L1), CTLA4, and their corresponding drugs (**Figure 6C**), suggesting that the low-risk samples might benefit from inhibiting the immune checkpoint pathway. Besides, the cell cycle kinase, CDK1, was upregulated in high-risk group, and BARASERTIB and DINACICLIB might be the potential drugs for treating DLBCL classified as high-risk (**Figure 6D**). As we have known, CD20 (also termed *MS4A1*) is expressed on the surface of normal B lymphocytes and is detected in almost all DLBCL cases. At present, RITUXIMAB, a chimeric monoclonal antibody directed against the CD20, combined with intensive chemotherapy (CHOP) is the standard therapy for DLBCL (**Figure 6D**). These results indicated the stratification may contribute to the selection of targeted drugs for the DLBCL patients.

## DISCUSSION

Diffuse large B cell lymphoma is a remarkably heterogeneous disease, both histologically and genetically. Despite significant advances in subtype classification of DLBCL, accurate prediction of prognosis remains a challenge. With the development of high throughput sequencing technology, some potential prognostic genomic markers for DLBCL patients have been identified (Rosenwald et al., 2002; Wright et al., 2003; Lossos, 2008). However, the number of prognostic markers is still limited. There is an urgent need to screen out more biomarkers to improve the accuracy of prognostic prediction.

In the present study, we identified potential gene candidates through the univariable Cox regression analysis to examine associations between gene expression and patient prognosis of three DLBCL cohorts in GEO. To further narrow down the list of candidate gene signatures, multivariate Cox analysis was carried out on the merged datasets. A stepwise approach was used to select a subset of gene candidates to achieve the highest performance, and a risk model was established for predicting DLBCL prognosis based on the expression levels of five genes including *CEBPA*, *CYP27A1*, *LST1*, *MREG*, and *TARP*. We evaluated the model performance using an independent gene expression dataset and compared it with previously reported models. Our five-gene based risk model showed improved robustness, accuracy, and efficiency compared to those models and was demonstrated to be an independent prognostic factor for OS in patients with DLBCL. Subsequently, we compared the gene expression profiles of high-risk with those of low-risk group and performed ORA to identify pathways potentially involved in the DLBCL progression. Thus, we believe that our five-gene-based risk scoring model can be

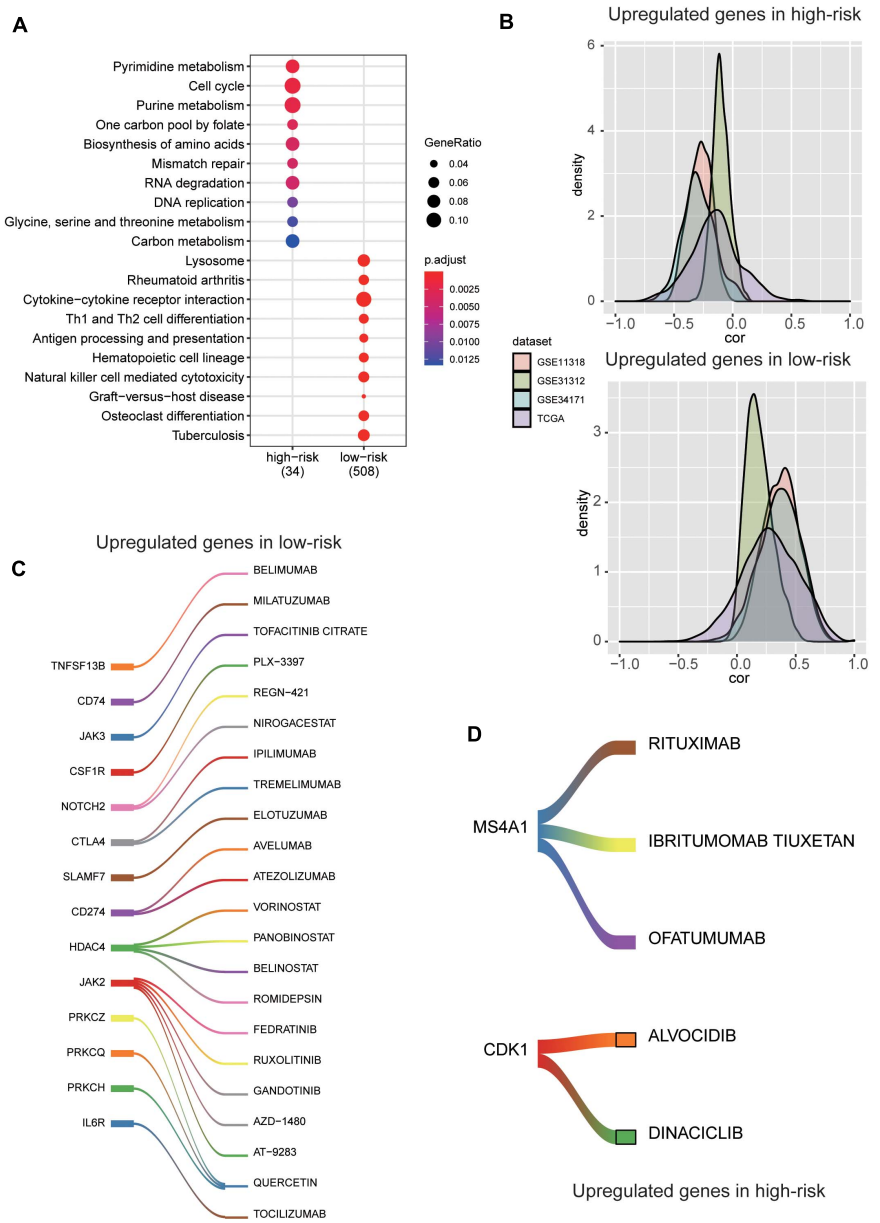



**FIGURE 6 |** The molecular characteristics and potential drugs for the two risk groups. **(A)** The top-ten GO terms enriched by the upregulated genes in high-risk and low-risk groups. The dots size and color represent the ratio of gene counts and statistical significance, respectively. **(B)** The probability density function of the Spearman's correlation between the five prognostic genes and the differentially expressed genes (DEGs). The colors represent the validation datasets. **(C)** The upregulated immune checkpoint proteins and the corresponding drugs in the low-risk group. **(D)** The upregulated cell cycle kinase and their potential drugs in high-risk group.

used for refining DLBCL subtypes and potentially improving patient therapy.

According to the multivariable Cox model, high expression of the five genes was all associated with a favorable survival outcome. CEBPA is a transcription factor playing roles in regulating proliferation and differentiation of many cell types (Gery et al., 2005). Within the hematopoietic system, inactivation mutation of CEBPA blocks the granulocytic differentiation in acute myeloid leukemia (AML) (Wang et al., 1999). In addition, it has been reported that CEBPA-regulated PER2 activation is a potential tumor suppressor pathway in diffuse large B-cell lymphoma (DLBCL) (Thoennissen et al., 2012). CYP27A1, a cytochrome P450 oxidase family member, is closely related to the proliferation of multiple tumor cells, such as prostate, breast and colon cancer (Ji et al., 2016; Alfaqih et al., 2017; Kimbung et al., 2017). LST1 is encoded within the TNF region of the human MHC which regulates lymphocyte proliferation (Rollinger-Holzinger et al., 2000). MREG is reported to suppress

thyroid cancer cell invasion and proliferation through PI3K/Akt-mTOR signaling pathway (Meng et al., 2017). The biological roles of these genes in DLBCL need to be further investigated.

The ORA of DEGs suggests that the abnormal cell cycle progression and increased genomic instability contribute to the rapid progression of DLBCL. Inhibitors of cell cycle kinase, such as BARASERTIB and DINACICLIB, may be effective in high-risk patients. On the contrary, genes related to immune-related pathways, such as antigen processing and presentation, Th1 and Th2 cell differentiation, were enriched in low-risk group, suggesting that activated host immune response may indicate favorable prognosis and response to therapy. These findings provide novel clues into the explanation of the mechanisms of DLBCL.

The prognostic model we proposed is helpful for further risk stratification at the genetic level on the basis of the present traditional subtyping, but this study still has some limitations. Some potential prognostic factors may be excluded in the model such as the racial factors and the roles that the five genes play in DLBCL requires further experimental validation. To sum up, our research indicates that the five-gene prognostic model is a reliable tool for predicting the OS of DLBCL patients and providing some hints on drug selection, which can assist clinicians in selecting personalized treatment, although specific drug selection requires further molecular biology research and clinical trials.

## 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/s.

## ETHICS STATEMENT

Participants gave their written informed consent for the materials to appear in publications without limit on the duration of publication.

## AUTHOR CONTRIBUTIONS

BX, WZ, and AL conceived and designed the experiments. MP, PY, and FW acquired data, related materials, and analysis tools. MP, XL, and BL analyzed the data. MP, PY, and FW wrote the manuscript. YDi, HL, and YDo revised the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was sponsored by Shanghai Sailing Program (No. 19YF1444300), Program of Outstanding Young Scientists of Tongji Hospital of Tongji University (No. HBRC1802), Youth Project of Scientific Research Project of Shanghai Health and Family Planning Commission (No. 20174Y0110), the Key Project of Natural Science Foundation of China (No. 81830004), and Clinical Research Plan of SHDC (No. SHDC2020CR6005).

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at Research Square (Pan mengmeng et al.).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.648800/full#supplementary-material

**Supplementary Figure 1 |** The principal component analysis (PCA) of the discretized expression profiles of the three cohorts used for model training.

**Supplementary Figure 2 |** The association of risk scores derived from the three previous gene signature sets with IPI scoring system and tumor stage.

**Supplementary Table 1 |** Clinical difference between the three cohorts used for model training.

## REFERENCES

Alfaqih, M. A., Nelson, E. R., Liu, W., Safi, R., Jasper, J. S., Macias, E., et al. (2017). CYP27A1 Loss Dysregulates Cholesterol Homeostasis in Prostate Cancer. *Cancer Res.* 77, 1662–1673. doi: 10.1158/0008-5472.CAN-16-2738

Barrans, S. L., Crouch, S., Care, M. A., Worrillow, L., Smith, A., Patmore, R., et al. (2012). Whole genome expression profiling based on paraffin embedded tissue can be used to classify diffuse large B-cell lymphoma and predict clinical outcome. *Br. J. Haematol.* 159, 441–453. doi: 10.1111/bjh.12045

Bhatt, G., Maddocks, K., and Christian, B. (2016). CD30 and CD30-Targeted Therapies in Hodgkin Lymphoma and Other B cell Lymphomas. *Curr. Hematol. Malign. Rep.* 11, 480–491. doi: 10.1007/s11899-016-0345-y

Cabanillas, F., and Shah, B. (2017). Advances in Diagnosis and Management of Diffuse Large B-cell Lymphoma. *Clin. Lymph. Myeloma Leukemia* 17, 783–796. doi: 10.1016/j.clml.2017.10.007

Chapuy, B., Stewart, C., Dunford, A. J., Kim, J., Kamburov, A., Redd, R. A., et al. (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* 24, 679–690. doi: 10.1038/s41591-018-0016-8

Coiffier, B., Lepage, E., Briere, J., Herbrecht, R., Tilly, H., Bouabdallah, R., et al. (2002). CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 346, 235–242. doi: 10.1056/nejmoa011795

Coiffier, B., Thieblemont, C., Van Den Neste, E., Lepeu, G., Plantier, I., Castaigne, S., et al. (2010). Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the Groupe d'Etudes des Lymphomes de l'Adulte. *Blood* 116, 2040–2045. doi: 10.1182/blood-2010-03-276246

Gery, S., Gombart, A. F., Yi, W. S., Koeffler, C., Hofmann, W.-K., and Koeffler, H. P. (2005). Transcription profiling of C/EBP targets identifies Per2 as a gene implicated in myeloid leukemia. *Blood* 106, 2827–2836. doi: 10.1182/blood-2005-01-0358

Hummel, M., Bentink, S., Berger, H., Klapper, W., Wessendorf, S., Barth, T. F., et al. (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* 354, 2419–2430. doi: 10.1056/NEJMoa055351

International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *N. Engl. J. Med.* 329, 987–994. doi: 10.1056/nejm199309303291402

Ji, Y.-C., Liu, C., Zhang, X., Zhang, C.-S., Wang, D., and Zhang, Y. (2016). Intestinal bacterium-derived cyp27a1 prevents colon cancer cell apoptosis. *Am. J. Res.* 8, 4434–4439.

Kimbung, S., Chang, C.-Y., Bendahl, P.-O., Dubois, L., Thompson, J. W., McDonnell, D. P., et al. (2017). Impact of 27-hydroxylase (CYP27A1) and 27-hydroxycholesterol in breast cancer. *Endocr. Relat. Cancer* 24, 339–349. doi: 10.1530/ERC-16-0533

Lenz, G., Wright, G. W., Emre, N. C., Kohlhammer, H., Dave, S. S., Davis, R. E., et al. (2008b). Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci U S A* 105, 13520–13525. doi: 10.1073/pnas.0804295105

Lenz, G., Wright, G., Dave, S. S., Xiao, W., Powell, J., Zhao, H., et al. (2008a). Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* 359, 2313–2323. doi: 10.1056/NEJMoa0802885

Li, S., Young, K. H., and Medeiros, L. J. (2018). Diffuse large B-cell lymphoma. *Pathology* 50, 74–87. doi: 10.1016/j.pathol.2017.09.006

Lossos, I. S. (2008). Diffuse large B cell lymphoma: from gene expression profiling to prediction of outcome. *Biol. Blood Marrow Transplant.* 14(1 Suppl. 1), 108–111. doi: 10.1016/j.bbmt.2007.10.020

Marangon, A. V., Colli, C. M., Cardozo, D. M., Visentainer, J. E. L., Sell, A. M., Guimaraes, F., et al. (2019). Impact of SNPs/Haplotypes of and on the Development of Diffuse Large B-Cell Lymphoma. *J. Immunol. Res.* 2019:2137538. doi: 10.1155/2019/2137538

Martelli, M., Ferreri, A. J. M., Agostinelli, C., Di Rocco, A., Pfreundschuh, M., and Pileri, S. A. (2013). Diffuse large B-cell lymphoma. *Crit. Rev. Oncol. Hematol.* 87, 146–171. doi: 10.1016/j.critrevonc.2012.12.009

Meng, X., Dong, Y., Yu, X., Wang, D., Wang, S., Chen, S., et al. (2017). MREG suppresses thyroid cancer cell invasion and proliferation by inhibiting Akt-mTOR signaling. *Biochem. Biophys. Res. Commun.* 491, 72–78. doi: 10.1016/j.bbrc.2017.07.044

Monti, S., Chapuy, B., Takeyama, K., Rodig, S. J., Hao, Y., Yeda, K. T., et al. (2012). Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* 22, 359–372. doi: 10.1016/j.ccr.2012.07.014

Pierce, J. M. R., and Mehta, A. (2017). Diagnostic, prognostic and therapeutic role of CD30 in lymphoma. *Expert Rev. Hematol.* 10, 29–37. doi: 10.1080/17474086.2017.1270202

Rollinger-Holzinger, I., Eibl, B., Pauly, M., Griesser, U., Hentges, F., Auer, B., et al. (2000). LST1: a gene with extensive alternative splicing and immunomodulatory function. *J. Immunol.* 164, 3169–3176. doi: 10.4049/jimmunol.164.6.3169

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England journal of medicine* 346, 1937–1947.

Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., et al. (2018). Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 378, 1396–1407. doi: 10.1056/NEJMoa1801445

Sha, C., Barrans, S., Care, M. A., Cunningham, D., Tooze, R. M., Jack, A., et al. (2015). Transferring genomics to the clinic: distinguishing Burkitt and diffuse large B cell lymphomas. *Genome Med* 7, 64. doi: 10.1186/s13073-015-0187-6

Thakral, B., Medeiros, L. J., Desai, P., Lin, P., Yin, C. C., Tang, G., et al. (2017). Prognostic impact of CD5 expression in diffuse large B-cell lymphoma in patients treated with rituximab-EPOCH. *Eur. J. Haematol.* 98, 415–421. doi: 10.1111/ejh.12847

Thoennissen, N. H., Thoennissen, G. B., Abbassi, S., Nabavi-Nouis, S., Sauer, T., Doan, N. B., et al. (2012). Transcription factor CCAAT/enhancer-binding protein alpha and critical circadian clock downstream target gene PER2 are highly deregulated in diffuse large B-cell lymphoma. *Leukemia Lymphoma* 53, 1577–1585. doi: 10.3109/10428194.2012.658792

Visco, C., Li, Y., Xu-Monette, Z. Y., Miranda, R. N., Green, T. M., Li, Y., et al. (2012). Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the International DLBCL Rituximab-CHOP Consortium Program Study. *Leukemia* 26, 2103–2113. doi: 10.1038/leu.2012.83

Visco, C., Tzankov, A., Xu-Monette, Z. Y., Miranda, R. N., Tai, Y. C., Li, Y., et al. (2013). Patients with diffuse large B-cell lymphoma of germinal center origin with BCL2 translocations have poor outcome, irrespective of MYC status: a report from an International DLBCL rituximab-CHOP Consortium Program Study. *Haematologica* 98, 255–263. doi: 10.3324/haematol.2012.066209

Wang, X., Scott, E., Sawyers, C. L., and Friedman, A. D. (1999). C/EBPalpha bypasses granulocyte colony-stimulating factor signals to rapidly induce PU.1 gene expression, stimulate granulocytic differentiation, and limit proliferation in 32D cl3 myeloblasts. *Blood* 94, 560–571. doi: 10.1182/blood.v94.2.560.414k41_560_571

Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., and Staudt, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci.* 100, 9991–9996. doi: 10.1073/pnas.1732008100

Xu-Monette, Z. Y., Wu, L., Visco, C., Tai, Y. C., Tzankov, A., Liu, W.-M., et al. (2012). Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an International DLBCL Rituximab-CHOP Consortium Program Study. *Blood* 120, 3986–3996. doi: 10.1182/blood-2012-05-433334

Zhao, P., Li, L., Zhou, S., Qiu, L., Qian, Z., Liu, X., et al. (2019). CD5 expression correlates with inferior survival and enhances the negative effect of p53 overexpression in diffuse large B-cell lymphoma. *Hematol Oncol.* 37, 360–367. doi: 10.1002/hon.2657

**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: 11 June 2021
doi: 10.3389/fgene.2021.656114

# Identification of the Signature Associated With m$^6$A RNA Methylation Regulators and m$^6$A-Related Genes and Construction of the Risk Score for Prognostication in Early-Stage Lung Adenocarcinoma

*Bingzhou Guo[1†], Hongliang Zhang[2†], Jinliang Wang[3], Rilige Wu[4,5], Junyan Zhang[4,5], Qiqin Zhang[6], Lu Xu[7], Ming Shen[7], Zhibo Zhang[8], Fangyan Gu[9], Weiliang Zeng[1]*, Xiaodong Jia[10]* and Chengliang Yin[11]**

[1] School of Mathematical Sciences, Harbin Normal University, Harbin, China, [2] Department of Emergency, The First Medical Center of Chinese PLA General Hospital, Beijing, China, [3] Department of Oncology, The Second Medical Center of Chinese PLA General Hospital, Beijing, China, [4] National Engineering Laboratory for Medical Big Data Application Technology, Chinese PLA General Hospital, Beijing, China, [5] Medical Big Data Research Center, Medical Innovation Research Division of Chinese PLA General Hospital, Beijing, China, [6] Department of Orthopedics, Weifang Traditional Chinese Hospital, Weifang, China, [7] Laboratory of Translational Medicine, Medical Innovation Research Division of Chinese PLA General Hospital, Beijing, China, [8] The 78th Group Army Hospital of Chinese PLA, Mudanjiang, China, [9] Clinical Biobank Center, Medical Innovation Research Division of Chinese PLA General Hospital, Beijing, China, [10] Department of Liver Disease, Fifth Medical Center of Chinese PLA General Hospital, Beijing, China, [11] Faculty of Medicine, Macau University of Science and Technology, Macau, China

**Background:** N6-methyladenosine (m$^6$A) RNA modification is vital for cancers because methylation can alter gene expression and even affect some functional modification. Our study aimed to analyze m$^6$A RNA methylation regulators and m$^6$A-related genes to understand the prognosis of early lung adenocarcinoma.

**Methods:** The relevant datasets were utilized to analyze 21 m$^6$A RNA methylation regulators and 5,486 m$^6$A-related genes in m$^6$Avar. Univariate Cox regression analysis, random survival forest analysis, Kaplan–Meier analysis, Chi-square analysis, and multivariate cox analysis were carried out on the datasets, and a risk prognostic model based on three feature genes was constructed.

**Results:** Respectively, we treated GSE31210 ($n$ = 226) as the training set, GSE50081 ($n$ = 128) and TCGA data ($n$ = 400) as the test set. By performing univariable cox regression analysis and random survival forest algorithm in the training group, 218 genes were significant and three prognosis-related genes (*ZCRB1*, *ADH1C*, and *YTHDC2*) were screened out, which could divide LUAD patients into low and high-risk group ($P < 0.0001$). The predictive efficacy of the model was confirmed in the test group GSE50081 ($P = 0.0018$) and the TCGA datasets ($P = 0.014$). Multivariable

OPEN ACCESS

**Edited by:**
Suman Ghosal,
National Institutes of Health (NIH), United States

**Reviewed by:**
Rituparno Sen,
Leipzig University, Germany
Vishal Midya,
Icahn School of Medicine at Mount Sinai, United States

***Correspondence:**
Chengliang Yin
chengliangyin@163.com
Xiaodong Jia
feixiang.5420@163.com
Weiliang Zeng
zengwl@hrbnu.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:** 20 January 2021
**Accepted:** 21 April 2021
**Published:** 11 June 2021

**Citation:**
Guo B, Zhang H, Wang J, Wu R, Zhang J, Zhang Q, Xu L, Shen M, Zhang Z, Gu F, Zeng W, Jia X and Yin C (2021) Identification of the Signature Associated With m$^6$A RNA Methylation Regulators and m$^6$A-Related Genes and Construction of the Risk Score for Prognostication in Early-Stage Lung Adenocarcinoma. Front. Genet. 12:656114. doi: 10.3389/fgene.2021.656114

cox manifested that the three-gene signature was an independent risk factor in LUAD. Furthermore, genes in the signature were also externally validated using the online database. Moreover, YTHDC2 was the important gene in the risk score model and played a vital role in readers of m$^6$A methylation.

**Conclusion:** The findings of this study suggested that associated with m$^6$A RNA methylation regulators and m$^6$A-related genes, the three-gene signature was a reliable prognostic indicator for LUAD patients, indicating a clinical application prospect to serve as a potential therapeutic target.

**Keywords: lung adenocarcinoma, m$^6$A, prognostic signature, m$^6$A-related genes, RNA methylation regulators**

## INTRODUCTION

Lung adenocarcinoma (LUAD) is a type of non-small cell cancer. In the 2018 Global Cancer Report, lung cancer ranked top 1 with the highest incidence and mortality among all cancers (Bray et al., 2018).

N6-methyladenosine (m$^6$A) RNA methylation is the most abundant epigenetic modification in eukaryotic mRNA. M$^6$A methylation regulators of each modified RNA require a writer to place, an eraser to erase, and a reader to read. Based on these proteins, m$^6$A affected RNA splicing (He et al., 2019), translation, and RNA stability (Wang et al., 2014; He et al., 2019). Evidence is now mounting that m$^6$A methylation underlies the progression of tumors and affects specific biological processes through non-coding RNA modification (Xiao et al., 2019). Moreover, the over-expression of YTHDF1 in the reader might affect the prognosis of ovarian cancer patients (Liu et al., 2020). In the writer family, high expression of METTL3 promoted the proliferation of bladder cancer (Cheng et al., 2019) and led to a poor prognosis (Han et al., 2019). Over-expression knockdown of ALKBH5 could effectively reduce cell proliferation in pancreatic cancer in erasers family (Tang et al., 2020). Meanwhile, m$^6$A has many functions in cancer (He et al., 2019; Ma et al., 2019; Ma and Ji, 2020), such as reduced m$^6$A has a relationship with phenotypes of gastric cancer (Zhang et al., 2019), KIAA1429 is associated with prognosis of liver cancer (Lan et al., 2019), and FTO could facilitate the development of breast cancer (Niu et al., 2019). However, to our knowledge, there are few studies related to m$^6$A methylation in early LUAD, and this may be a novel treatment strategy for patients with early LUAD.

In this study, GEO and TCGA data were used to explore the influence of m$^6$A methylation genes and their regulated genes on the prognosis of early lung adenocarcinoma. The signature was conducted for identifying new therapeutic biomarkers and treatment strategy development.

**Abbreviations:** m$^6$A, N6-methyladenosine; LUAD, lung adenocarcinoma; ROC, receiver operating characteristic; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; m$^1$A, N1-methyladenosine; RSFVH, random survival forest algorithm; lncRNAs, long chain non-coding RNA; OS, overall survival; GO, gene ontology; HR, hazard ratio; CI, confidence interval; KM, Kaplan–Meier.

## MATERIALS AND METHODS

### Expression Data

Data was downloaded from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) public databases. GSE31210 ($n$ = 226) was used as the training set, GSE50081 ($n$ = 128) as the validation set 1, and TCGA ($n$ = 400) data as the validation set 2. Three independent datasets were used for model construction and model verification. Each independent dataset included the clinical characteristics: survival status, survival time, age, sex, and clinical TNM stage. In GEO data, TNM clinical stage was divided into stages I and II, which were shown in **Table 1**. Besides, the GPL570 chip platform was re-annotated by the probe to get the final expression profile of the GEO data (Fan et al., 2018). Only mRNA probes were selected, and 8,597 mRNA expression profiles were obtained.

### Selection of m$^6$A RNA Methylation Regulatory Factors and m$^6$A-Related Genes

We collected 21 m$^6$A methylated genes through literature investigation (**Supplementary Table 1**) (Zhang et al., 2020). We found that among these 21 genes, 14 genes were expressed in

**TABLE 1 |** Clinical information of the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) datasets.

| Characteristic | GSE31210 | GSE50081 | TCGA |
|---|---|---|---|
| **Age (years)** | | | |
| >61 | 104 | 104 | 251 |
| ≤61 | 122 | 24 | 149 |
| **Sex** | | | |
| Female | 121 | 63 | 217 |
| Male | 105 | 65 | 183 |
| **Vital status** | | | |
| Alive | 191 | 76 | 278 |
| Dead | 35 | 52 | 122 |
| **Pathological stage** | | | |
| Stage I | 168 | 92 | 280 |
| Stage II | 58 | 36 | 120 |

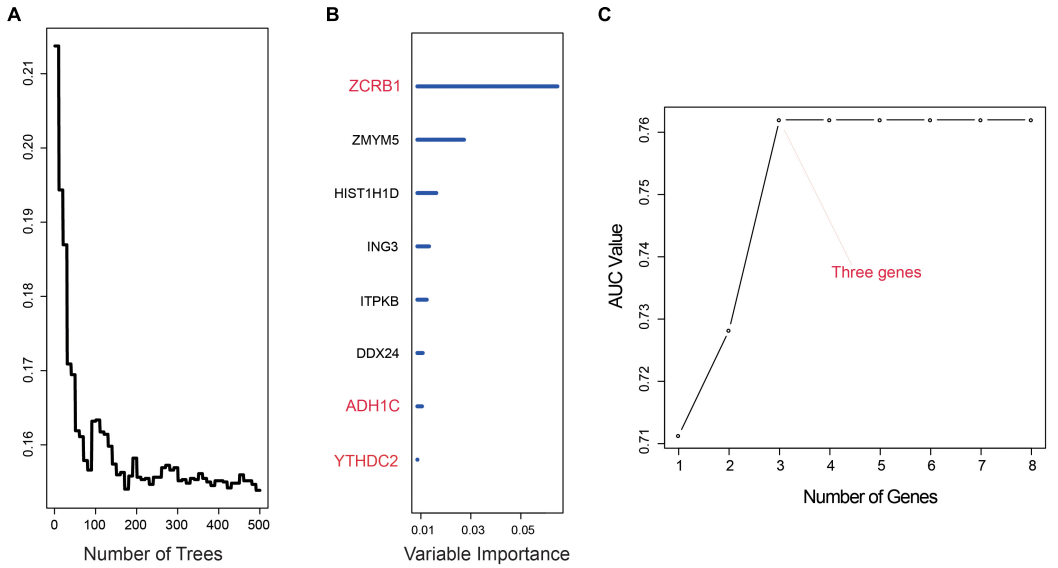



**FIGURE 1 |** Random survival forest analysis. **(A,B)** random survival forests variable hunting analysis reveals the error rate for the data as a function of trees and uses the associated score to filter N6-methyladenosine (m$^6$A) RNA methylation regulators and m$^6$A-related genes. **(C)** Receiver operating characteristic (ROC) for selected prognostic signature from all 255 signatures.

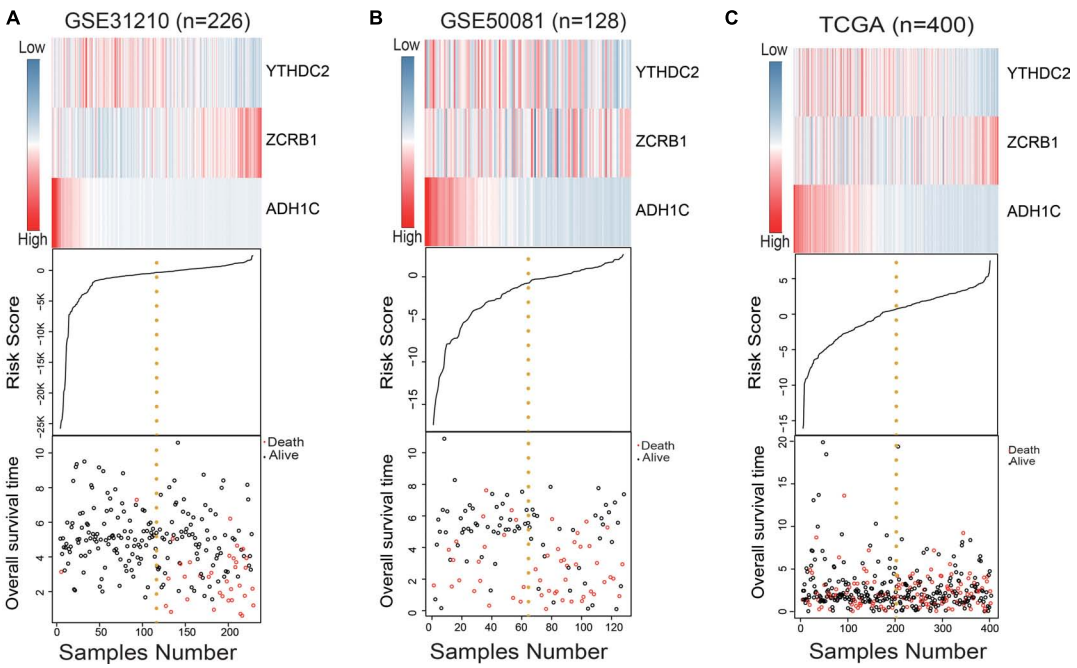



**FIGURE 2 |** Evaluation of the risk predictive model in the training set and test set. **(A–C)** The distribution of m$^6$A RNA methylation regulators and m$^6$A-related gene expression level, patients' survival status, and risk score between high- and low-risk group.

the training set (GSE31210). In LUAD, a total of 5,486 m$^6$A-regulated genes were downloaded from the m$^6$Avar database[1] (Zheng et al., 2018). Among the 5,507 genes, 2,615 genes were expressed in GSE31210.

[1]http://m6avar.renlab.org/

Discovery of the m$^6$A RNA Methylation Regulators and m$^6$A-Related Genes and Establishment of the m$^6$A Methylation Risk Score Model.

We obtained prognostic-related gene sets through survival analysis [univariate cox and Kaplan–Meier (KM)] and receiver operating characteristic (ROC) curve. In the training set

GSE31210, we used the random survival forest (RSF) (Ishwaran et al., 2008) to establish a prognostic model related to patient overall survival (OS). Methods of analyzing survival data were often parametric, nonlinear effects of variables, and modeled by expanding matrix for specialized functions. Identifying multiple variable interactions was also problematic. These difficulties could be effectively solved using RSF (Ishwaran et al., 2008). Its basic formula is:

$$\mathrm{RSF} = \sum_{i=1}^{N} \mathrm{Expi} \times \mathrm{Coefi}$$

The meanings of the parameters in this formula are: RS is the risk score, *N* is the number of genes, Exp is the expression amount of genes in the data, and Coef is the coefficient of cox analysis for the genes resulting from the random survival forest. We used gene combinations to select the largest AUC to construct a prognostic model. Based on the median of the risk scores, the data were divided into two groups: the high-risk group and the low-risk group. The KM curve was used to compare the difference between the high and low-risk groups. In the three datasets, using the median score divided two groups.

## Estimation of Outcome Signature for Patients' Prognosis and Its Relationship With Clinical Characteristics

To assess the characteristics of the patients' prognosis and its relationship with clinical features, we used chi-square analysis to judge the correlation between the model and clinical data. KM survival curve and log-rank test were used to describe the relationship between the model and OS. Furthermore,

**TABLE 2 |** Prognosis of the three genes in the signature.

| ENSEMBL ID | Symbol ID | Gene name | Coef | *P*-value | Prognostic indicator |
|---|---|---|---|---|---|
| ENSG00000047188 | YTHDC2 | YTH domain containing 2 | −2.02 | 0.00 | low |
| ENSG00000139168 | ZCRB1 | zinc finger CCHC-type and RNA binding motif containing 1 | 1.73 | 0.00 | high |
| ENSG00000248144 | ADH1C | alcohol dehydrogenase 1C (class I), gamma polypeptide | −1.96 | 0.00 | low |

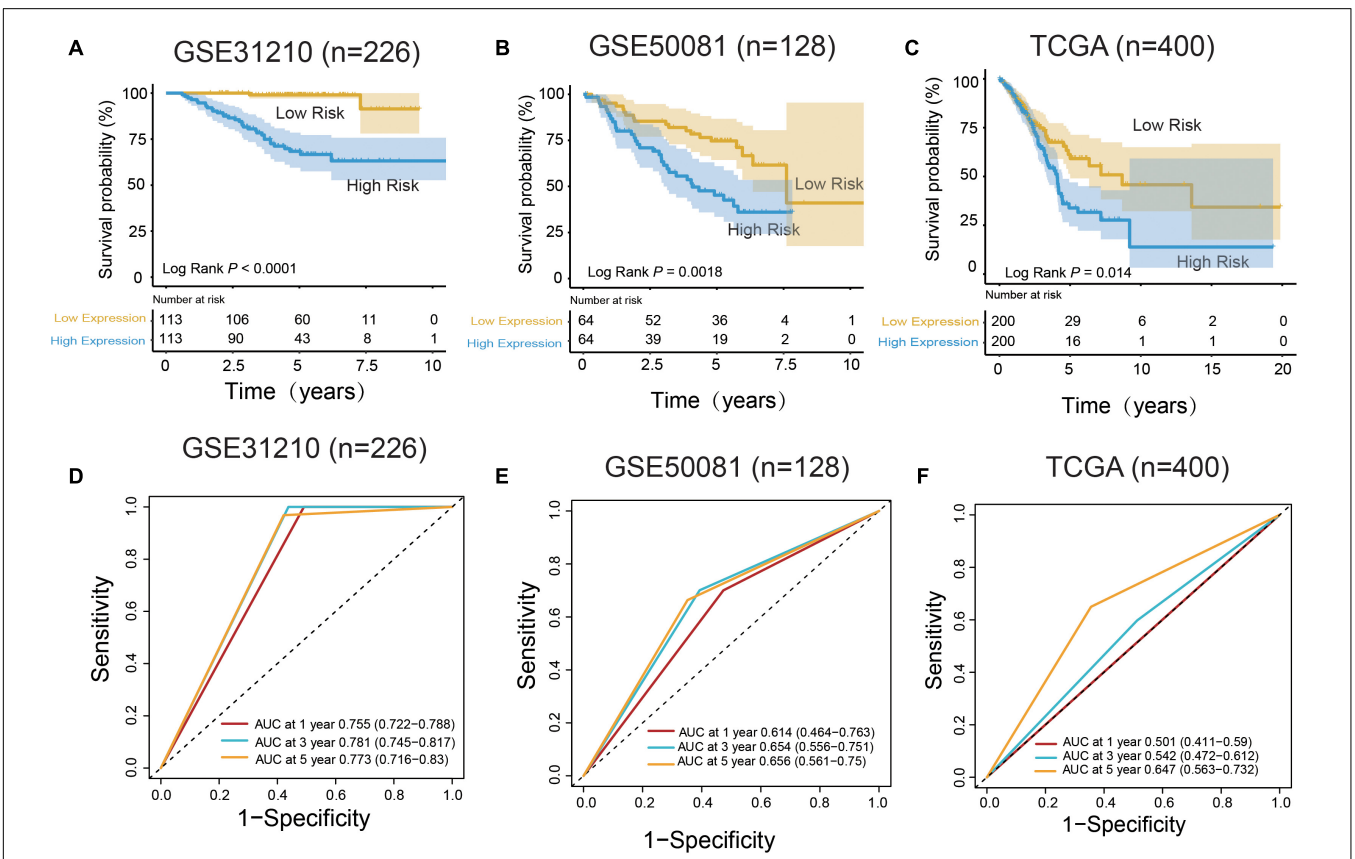



**FIGURE 3 |** m$^6$A RNA methylation regulators and m$^6$A-related genes signature predict overall survival of patients of LUAD. **(A–C)** Kaplan–Meier survival curves classify patients into high- and low-risk groups by the m$^6$A RNA methylation regulators and m$^6$A-related genes signature in the training dataset (GSE31210), and test dataset (GSE50081 and TCGA). *P* values were calculated by log-rank test. **(D–F)** m$^6$A RNA methylation regulators and m$^6$A-related genes signatures were used for predicting survival in 1, 3, and 5 years by TimeROC analysis.

multivariate Cox regression analysis was used to study whether the clinical data (age, gender, and pathological stage) were related to OS in the training set and validation set. We used univariate Cox analysis to judge whether the clinical information had prognostic value.

## External Validation of the Genes in the Gene Signature

Furthermore, four online tools were used to verify the gene expression levels (Oncomine database[2]; TIMER database[3], and GEPIA database, Gene Expression Profiling Interactive Analysis[4]) and protein levels [The Human Protein Atlas (HPA) database[5]] in the model. Meanwhile, online public databases (The cBioPortal for Cancer Genomics[6]) were used to analyze and understand the gene influence on early treatment of LUAD.

## Function Notes and Protein–Protein Interaction

The R package "clusterProfiler" was used to annotate the function and select the statistically significant pathways. The relationship between proteins was analyzed by using the online website STRING[7] (Szklarczyk et al., 2019). Cytoscape was used to visualize the network. Then the main networks were chosen by a degree of the gene in the net analysis.

## Statistical Analysis

In three independent datasets, all KM and cox analyses were performed using the R package "survival". Cox analysis was used to select prognostic genes and test models. "ROC" and "TimeROC" were available to verify the feasibility of the model. Functional annotations were made using the R package "Clusterprofiler." All of our analyses (besides online website analysis) were performed in the R language. R packages were used as follows: "pROC," "TimeROC," "survival," "clusterProfiler," and "randomForestSRC." The *P* values of the above analyses were all $<0.05$ as statistically significant.

# RESULTS

## Patient Population

All 226, 128, and 400 patients diagnosed with LUAD were collected from the GEO (GSE31210 and GSE50081) and TCGA database, respectively. A total of 2,615 m$^6$A-related genes out of the genes expressed were identified in the GSE31210 dataset. In **Table 1**, the median age of the enrolled samples was 61 years. The ratio of male vs. female was 1.15:1, with 191 live cases and 35 death cases. The longest survival was 10 years. Each sample data was only distributed in stages I–II of LUAD. The study workflow is demonstrated in **Supplementary Figure 1**.

[2]https://www.oncomine.org/resource/main.html
[3]https://cistrome.shinyapps.io/timer/
[4]http://gepia.cancer-pku.cn/index.html
[5]http://www.proteinatlas.org
[6]https://www.cbioportal.org/
[7]https://string-db.org/

**TABLE 3 |** Clinical information and signature Chi-square table.

| Variables | Status | low | high | *P* |
|---|---|---|---|---|
| **GSE31210 dataset (*n* = 226)** | | | | |
| Age | | | | 0.89 |
| | ≤61 | 62 | 60 | |
| | >61 | 51 | 53 | |
| Gender | | | | 0.18 |
| | Female | 66 | 55 | |
| | Male | 47 | 58 | |
| Pathological stage | | | | 0.00 |
| | I | 97 | 71 | |
| | II | 16 | 42 | |
| **GSE50081 dataset (*n* = 128)** | | | | |
| Age | | | | 0.82 |
| | ≤61 | 11 | 13 | |
| | >61 | 53 | 51 | |
| Gender | | | | 1.00 |
| | Female | 31 | 32 | |
| | Male | 33 | 32 | |
| Pathological stage | | | | 0.03 |
| | I | 52 | 40 | |
| | II | 12 | 24 | |
| **TCGA dataset (*n* = 400)** | | | | |
| Age | | | | 0.00 |
| | ≤61 | 60 | 89 | |
| | >61 | 140 | 111 | |
| Gender | | | | 0.69 |
| | Female | 111 | 106 | |
| | Male | 89 | 94 | |
| Pathological stage | | | | 0.10 |
| | I | 148 | 132 | |
| | II | 52 | 68 | |

## Construction of the Risk Score Model, the m$^6$A RNA Methylation Regulators, and m$^6$A-Related Genes Risk Score

After we used univariate cox analysis and ROC curve, 218 prognostic-related genes were selected, and the screening criteria were $P < 0.01$ and AUC $> 0.6$ in **Supplementary Table 2**. Furthermore, gene screening was performed by the importance scores of the random survival forest analysis. We permuted and combined the eight genes selected from the random survival forest, obtaining $2^8 - 1 = 255$ prediction models (**Figures 1A,B**). The 255 models were evaluated by AUC, and the optimal predictive ability was found in the combination of three genes, ZCRB1, ADH1C, and YTHDC2. As a prediction model, the AUC of the three-gene model was 0.762 (**Figure 1C**). The risk score of the model was RSF = (1.725151 × ZCRB1) + (−1.964326 × ADH1C) + (−2.015378 × YTHDC2). Each gene name represented its expression level in a certain sample.

We used the RSF formula to calculate the risk score of each sample and plotted the heat map of the three genes (**Figures 2A–C**), finding that in the high-risk group, ADH1C and YTHDC2

basically had low expression, while ZCRB1 obviously had high expression. This was particularly evident in the training group (GSE31210) (**Figure 2A**).

The results made it clear that ADH1C had high expression in the low-risk group as a protection factor by cox analysis (**Table 2**).

## The Validation of Performance in Predicting Overall Survival

In the training set, the median risk score divided all patients into two groups: high-risk group ($n$ = 113) and low-risk group ($n$ = 113) (**Figure 3A**). The KM survival curve and log rank test showed that our model had an excellent predictive power. In the validation set, the median risk score was also used to divide the patients into two groups in GSE50081 ($n$ = 128), and there were 64 patients in the high-risk group and 64 patients in the low-risk group (**Figure 3B**). The KM survival curve showed that there was a significant difference between the high-risk group and low-risk group (Log rank $P$ = 0.0018). Grouped by median risk score in TCGA ($n$ = 400), there were 200 patients in the high-risk group and 200 samples in the low-risk group, with log rank $P$ = 0.014 (**Figure 3C**).

In the training group (GSE31210), the 5-years survival rate was 53.10% in the low-risk group and 38.05% in the high-risk group (**Figure 3A**). Additionally, the overall survival rate was 45.58%, indicating that the risk score could differentiate the data correctly. Survival was significantly improved in the two independent validation data (GSE50081 and TCGA). Moreover, in GSE50081, the low-risk group was 56.25% and the high-risk group was 29.69% (**Figure 3B**). The overall survival rate was 42.97% and the grouping label was also evident. Meanwhile, in TCGA, we selected a sample data of TNM stage (I+II) (a total of 400 cases) (**Figure 3C**). Five-years survival rate was calculated in the high- and low-risk groups, and the rates were 14.5 and 8%, respectively. The overall 5-years survival rate was 11.25% in both low- and high-risk groups, and the survival rate in the low-risk group had markedly improved. Using time ROC in 5-years survival circumstances, we found that the label had an excellent prediction effect (**Figures 3D–F**). In GSE31210 and GSE50081, the AUC was 0.773 and 0.656, respectively, and the AUC was 0.647 in the TCGA.

## The Relationship Between the Signature and Clinical Characteristics

The association was demonstrated between the model and clinical information through the chi-square test in **Table 3**. There was a significant relationship between the pathological stage and the model ($P < 0.05$) in the GEO independent dataset rather than the TCGA dataset. Besides, there were 401 females in 754 cases, accounting for 53.18% of the total. A multivariate Cox test was utilized to determine if the signature had an independent prognostic value as a factor. The results in **Table 4** showed that the signature was a risk factor, and it was statistically significant (high- vs. low-risk, GSE31210, HR = 16.24, 95% CI 3.85–68.58, $P < 0.001$, $n$ = 226; GSE50081, HR = 2.23, 95% CI 1.24–4.02, $P$ = 0.008, $n$ = 128; TCGA, HR = 1.50, 95% CI 1.03–2.18, $P$ = 0.036, $n$ = 400). Univariate Cox also indicated that the signature was a risk factor.

## Functional Annotation and Protein–Protein Interaction

A 218 gene set, obtained from survival analysis and AUC analysis, was used for functional annotation and PPI network analysis. Among the 218 genes, including m$^6$A RNA methylation

**TABLE 4 |** Univariable and multivariable Cox regression analysis of the signature and clinical information with lung adenocarcinoma (LUAD) survival.

| Variables | | Univariable cox | | | | Multivariable cox | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | HR | 95% CI of HR | | P | HR | 95% CI of HR | | P |
| | | | right | left | | | right | left | |
| **GSE31210 ($n$ = 226)** | | | | | | | | | |
| Age | >61 vs. ≤61 | 1.43 | 0.73 | 2.78 | 0.29 | 1.49 | 0.76 | 2.92 | 0.24 |
| Sex | Male vs. female | 1.52 | 0.78 | 2.96 | 0.22 | 1.03 | 0.51 | 2.08 | 0.92 |
| Pathological stage | II vs. I | 4.23 | 2.17 | 8.24 | 0.00 | 2.73 | 1.35 | 5.50 | 0.00 |
| Signature | High risk vs. low risk | 20.48 | 4.91 | 85.43 | 0.00 | 16.24 | 3.85 | 68.58 | 0.00 |
| **GSE50081 ($n$ = 128)** | | | | | | | | | |
| Age | >61 vs. ≤61 | 2.09 | 0.89 | 4.89 | 0.09 | 2.04 | 0.86 | 4.80 | 0.10 |
| Sex | Male vs. female | 1.35 | 0.78 | 2.34 | 0.29 | 1.51 | 0.86 | 2.64 | 0.15 |
| Pathological stage | II vs I, | 2.53 | 1.45 | 4.44 | 0.00 | 2.09 | 1.17 | 3.73 | 0.01 |
| Signature | High risk vs. low risk | 2.40 | 1.36 | 4.23 | 0.00 | 2.23 | 1.24 | 4.02 | 0.01 |
| **TCGA ($n$ = 400)** | | | | | | | | | |
| Age | >61 vs. ≤61 | 1.20 | 0.83 | 1.75 | 0.33 | 1.34 | 0.91 | 1.96 | 0.14 |
| Sex | Male vs. female | 1.03 | 0.72 | 1.47 | 0.87 | 1.03 | 0.72 | 1.48 | 0.88 |
| Pathological stage | II vs. I | 2.49 | 1.73 | 3.57 | 0.00 | 2.40 | 1.66 | 3.45 | 0.00 |
| Signature | High risk vs. low risk | 1.57 | 1.09 | 2.26 | 0.01 | 1.50 | 1.03 | 2.18 | 0.04 |

*CI, confidence interval; HR, hazard ratio.*

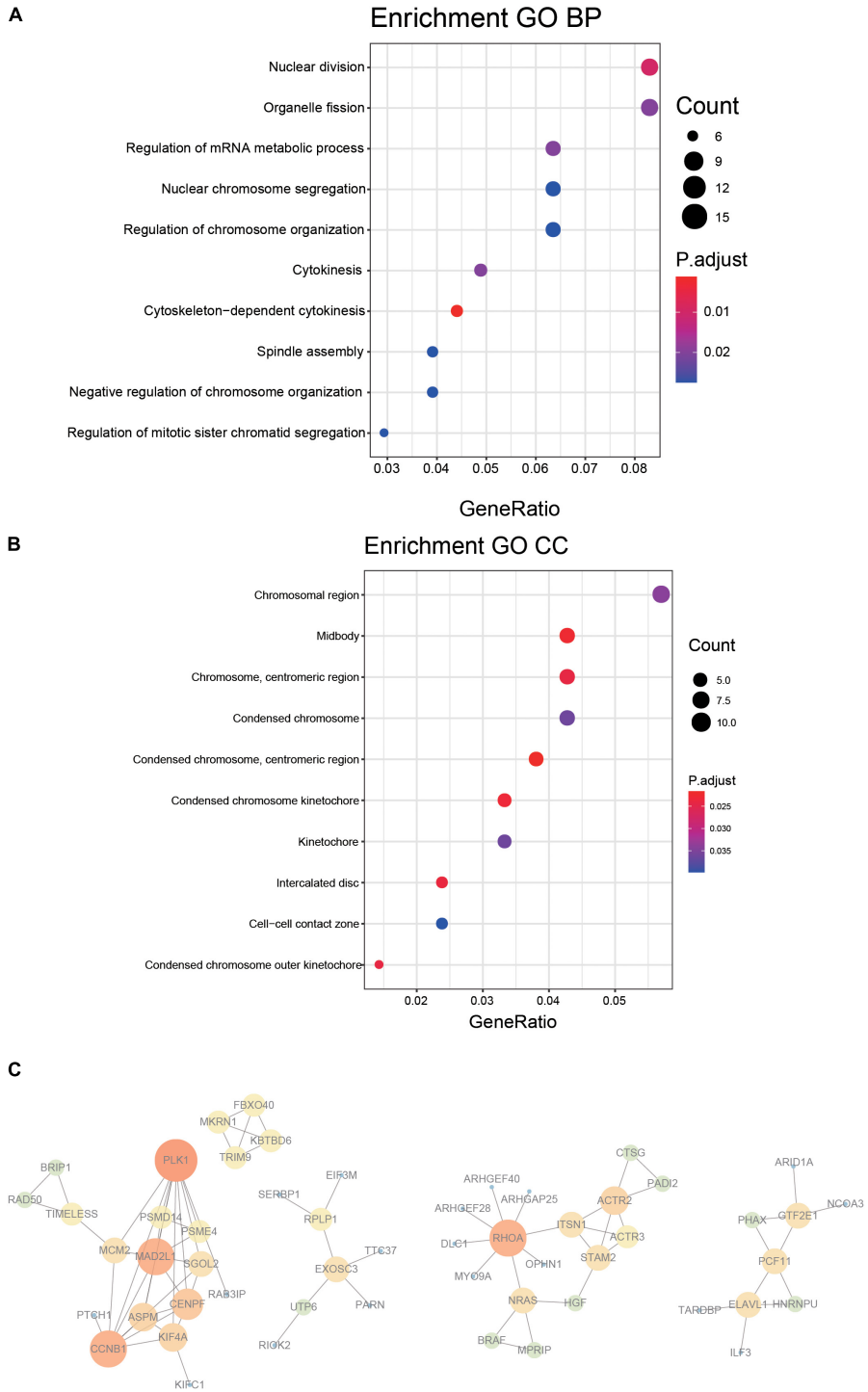



**FIGURE 4 |** Functional annotation and protein–protein interaction for the genes with significant prognosis. **(A,B)** Function prediction (BP, biological process; CC, cellular component). **(C)** Protein–protein interaction.

regulatory factors, ElAVL1, METTL14, and YTHDC2 were significantly associated with OS in LUAD. Top 10 biological processes (BPs) and cellular components (CCs) were selected by functional annotation of 218 genes, among which several results of BPs were related to division (nuclear division, organelle fission) and regulation (regulation of mRNA metabolic process and regulation of chromosome organization). The primary outcome of CCs was linked to the chromosome (**Figures 4A,B**).

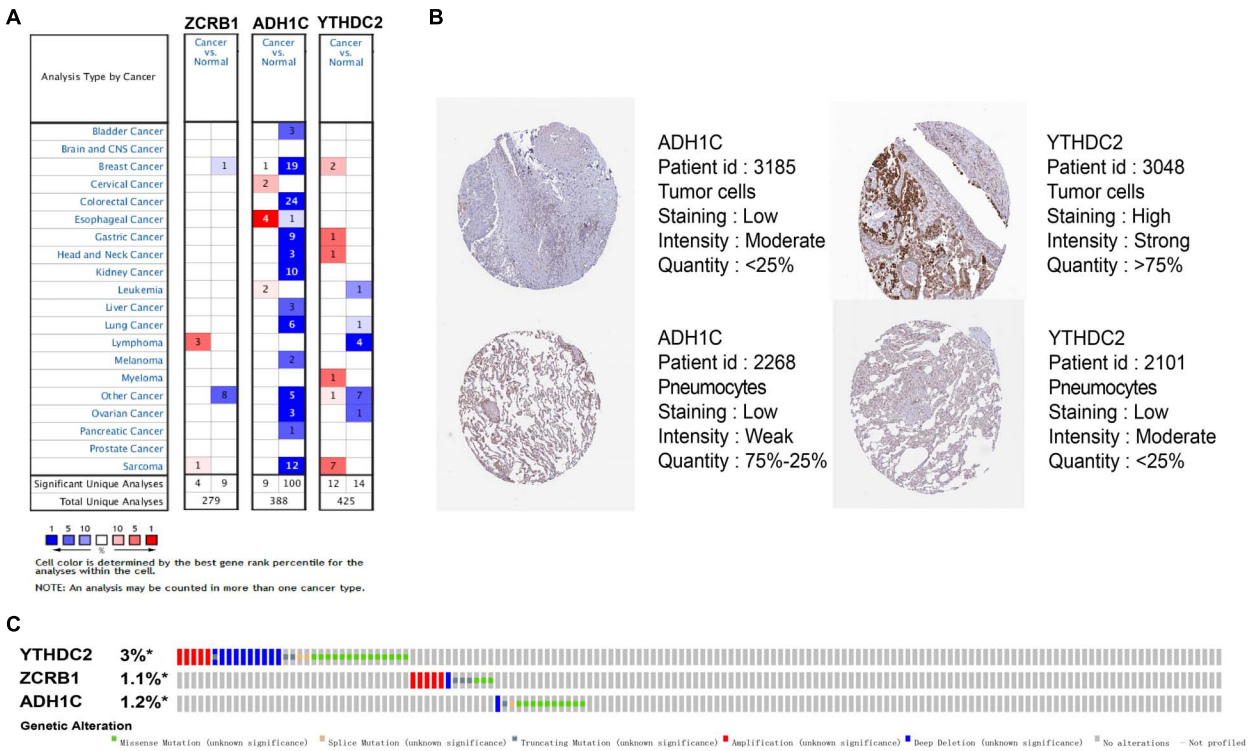



**FIGURE 5 |** Expression and genetic alterations of the three predictive genes. **(A)** The expression profiles of the three genes in the Oncomine database. **(B)** The representative protein expression of the three genes in LUAD and normal lung tissue in the Human Protein Atlas database. Data of ZCRB1 were not found in the database. **(C)** Genetic alterations of the three genes in LUAD in the cBioportal for Cancer Genomics.

PPI network was constructed by STRING, generated and visualized in Cytoscape. The combined score of the PPI criteria was >0.9. The PPI network had 88 relationships, and some genes were removed that were not part of the network (**Figure 4C**). Many key genes were observed in the network, such as PLK1, CCNB1, MAD2L1, RHOA, and ACTR2.

## External Validation Using Online Database About Genes in the Signature

The results of external validation data were consistent with our results. In LUAD, two genes YTHDC2 and ADH1C were lowly expressed in the three sets of independent data (**Figure 5A**), which was almost the same in both the TIMER database (**Figure 6**) and the GEPIA database (**Supplementary Figure 2**). Interestingly, the aberrant expression of the three genes were frequently observed in various cancers and showed some tissue-dependent patterns. For example, ZCRB1 was overexpressed in lymphoma, and ADH1C in cervical cancer and esophageal cancer, and YTHDC2 in breast cancer, gastric cancer, head and neck cancer, myeloma, and sarcoma (**Figure 5A**).

Survival analyses for each gene in the signature (ZCRB1, ADH1C, and YTHDC2) were performed in the cohorts of GSE31210, GSE50081, and TCGA datasets (**Figure 7**). ZCRB1 low-expression patient group displayed more OS than ZCRB1 high-expression patient group in GSE31210. While, ADH1C and YTHDC2 high-expression patient group displayed more OS than low-expression patient group not only in GSE31210 but also in GSE50081 and TCGA dataset.

We then reviewed the proteomic data and found YTHDC2 protein was reported significantly underexpressed in non-small cell lung cancer (Sun et al., 2020). The representative protein expression of ADH1C and YTHDC2 was explored in the human protein profiles and is shown in **Figure 5B**. Nevertheless, ZCRB1 was not found in the HPA website. YTHDC2 possessed the most frequent genetic alterations (3%) among the three genes. Meanwhile, amplification mutation, missense mutation, and deep deletion were the most common alterations among the three genes (**Figure 5C**). In summary, we further verified the abnormal expression of these three genes in LUAD, and genetic changes may help explain the aberrant expression of these genes to a certain extent.

# DISCUSSION

At the post-transcriptional level, more than 160 kinds of chemical modifications were discovered in various RNAs (Roundtree et al., 2017; Boccaletto et al., 2018). Among these modifications, more and more evidence showed that m$^6$A modification had an essential effect on some underlying

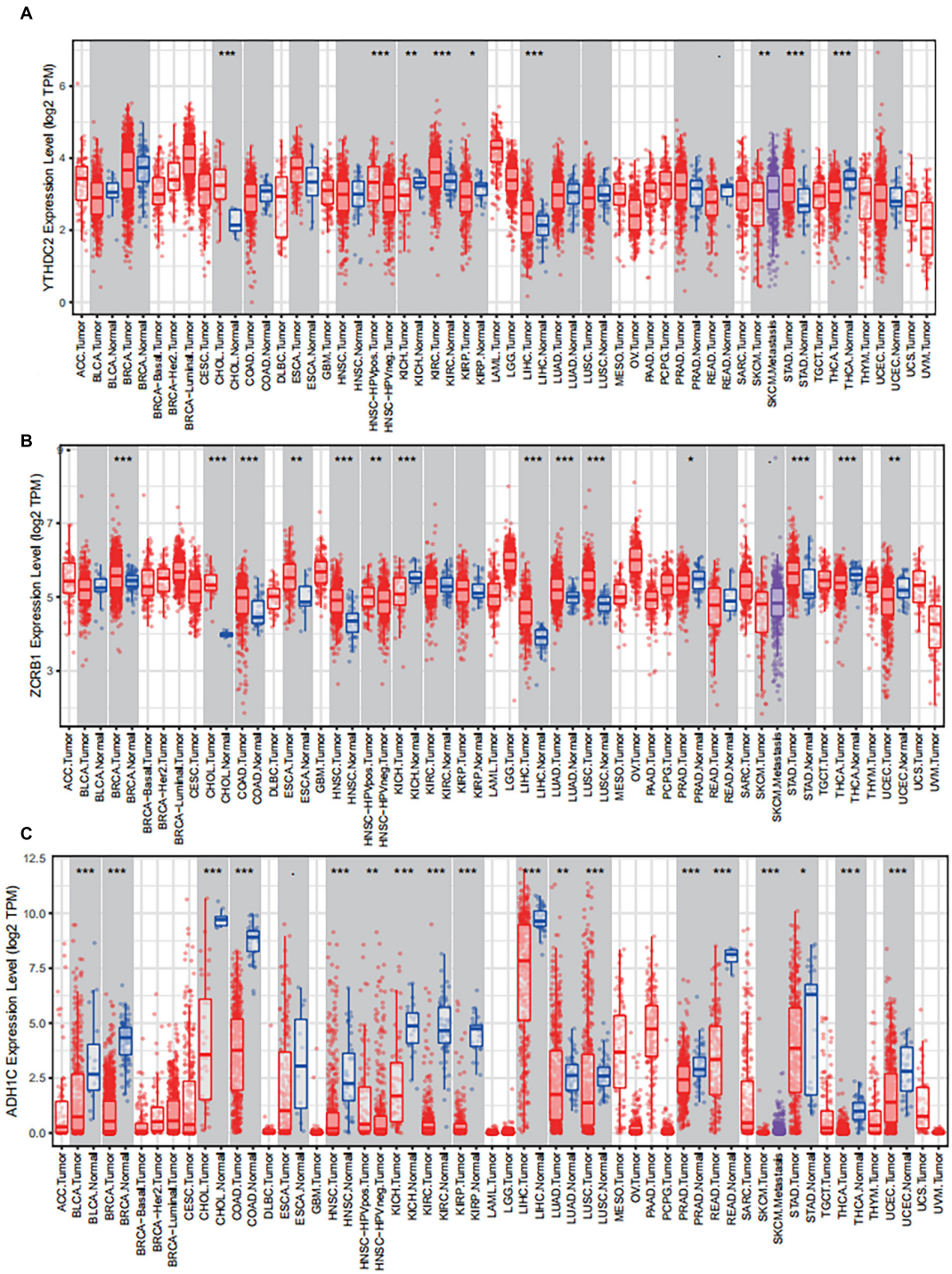



**FIGURE 6 |** The expression of the three predictive genes in cancers via Tumor IMmune Estimation Resource (TIMER, https://cistrome.shinyapps.io/timer/). **(A)** YTHDC2 expression level in tumor tissues vs normal tissues. **(B)** ZCRB1 expression level in tumor tissues vs normal tissues. **(C)** ADH1C expression level in tumor tissues vs normal tissues. ACC, adrenocortical carcinoma; BLCA, bladder carcinoma; BRCA, breast carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangio carcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUAC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

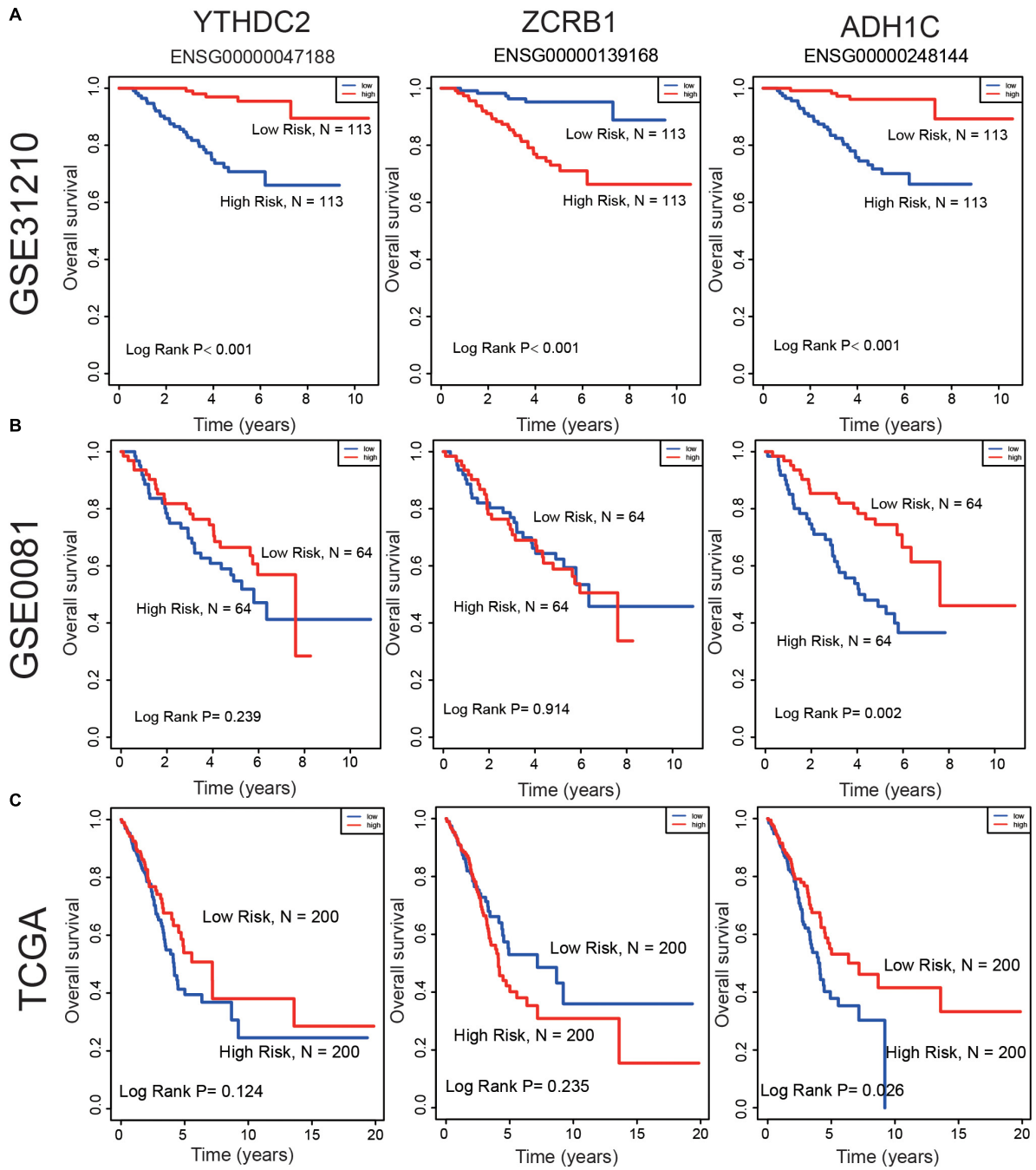



**FIGURE 7 |** Compare the low and high expression of the three predictive genes in overall survival in **(A)** GSE31210 dataset, **(B)** GSE50081 dataset, and **(C)** TCGA dataset.

diseases and prognosis of tumors. Therefore, the identification of m$^6$A RNA methylation regulators and m$^6$A-related genes in fatal LUAD may offer valuable therapeutic targets to us and clinicians. Doctors usually diagnosed LUAD as advanced, and there was a high death rate with it. Many studies illuminated that the m$^6$A process was linked to lung cancer, which made m$^6$A RNA methylation regulators and m$^6$A-related genes potential biomarkers for clinical practice. According to our research, the classification of m$^6$A-related genes in LUAD patients was in association with prognosis. We identified a signature that consisted of one m$^6$A RNA methylation regulator (YTHDC2) and two m$^6$A-related genes (ZCRB1 and ADH1C) using different statistical and machine learning methods.

Up to now, little is known about the role of YTHDC2 in tumorigenesis. Even less so in LUAD, two studies were found on the role of YTHDC2 in LUAD in recent studies. In a mouse model, low YTHDC2 expression was associated with poor prognosis in LUAD patients, and YTHDC2 improves the prognosis of LUAD patients by inhibiting the independent antioxidant function of SLC7A11 (Ma et al., 2021). In non-small cell lung cancer, a research analyzed a series of publicly available online databases and found that low YTHDC2 expression was associated with lymph node metastasis and poor prognosis (Sun et al., 2020).

ZCRB1 is a zinc finger CCHC-type and RNA binding motif containing 1. A previous study found that it was U12-type splicing playing a pivotal role by RefSeq analysis. However, the function of ZCRB1 was rarely reported in cancer, only in two studies. For example, ZCRB1's high expression can improve viral replication and transcription (Tan et al., 2012). Through genome-wide analysis of lung adenocarcinoma and healthy subjects, it was found that ZCRB1 may encode viral receptors. COVID-19 has infected plenty of people around the world, and ZCRB1 high expression may impact patients' prognosis (Cotroneo et al., 2021).

ADH1C is alcohol dehydrogenase 1C (class I). Many reports showed that drinking had an effect on some diseases. High expression of ADH1C was found to protect patients of non-small cell lung cancer (Wang et al., 2018). Using machine learning algorithms, the researchers found that ADH1C could be a prognostic marker (Shen et al., 2019).

For the reversible effect of m$^6$A on mRNA expression, we believe that m$^6$A-related genes may have different functional patterns and networks when participating in malignant tumors. Thus, m$^6$A-related genes may have different expression patterns in LUAD. In previous research, little was known about the interaction of m$^6$A-related genes. Moreover, m$^6$A RNA methylation regulators (WTAP, RBM15, KIAA1429, YTHDF1, and YTHDF2) were linked with TP53 and highly expressed in TP53 mutant LUAD (Zhuang et al., 2020). However, it is worth nothing that whether the TP53 mutant affects the expression of ZCRB1, ADH1C, and YTHDC2 is still unclear, and more evidence is needed to clarify their mechanism.

## CONCLUSION

In conclusion, our study systematically analyzed the expression, prognostic value, protein–protein interaction, and potential function of m$^6$A RNA methylation regulators and m$^6$A-related genes. We found that the expression of m$^6$A RNA methylation regulators and m$^6$A-related genes was closely related to the clinicopathologial characteristics of LUAD. A three-gene signature was identified that might effectively identify new therapeutic targets or strategies for LUAD. In summary, our study provided important clues for further studies on the role of RNA m$^6$A methylation regulators and m$^6$A-related genes in 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

CY, XJ, and WZ: study conception and design. BG, HZ, and CY: manuscript writing. BG, RW, and CY: literature review. All authors: data interpretation, discussion, final editing, and approval of the manuscript in its present form.

## FUNDING

This work was supported by Medical Big Data and AI R&D Project of the Chinese PLA General Hospital (2019MBD-025 and 2019MBD-001), National Natural Science Foundation of China (81902495), and Beijing Natural Science Foundation (7212099).

## ACKNOWLEDGMENTS

We thank all individuals who took part in this research.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.656114/full#supplementary-material

**Supplementary Figure 1 |** Flowchart of the study.

**Supplementary Figure 2 |** The three predictive genes expression levels in LUAD. Data was from the GEPIA database. T, tumor; *N*, normal tissue.

**Supplementary Table 1 |** The list of the 21 m$^6$A RNA methylation regulators from publications.

**Supplementary Table 2 |** m$^6$A methylation regulators and m$^6$A-related genes set of univariate cox regression analysis in the GSE31210 dataset ($P < 0.01$, AUC $> 0.6$, $n = 226$).

**Supplementary Table 3 |** 255 signatures in the GSE31210 dataset ($n = 226$).

## REFERENCES

Boccaletto, P., Machnicka, M. A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi: 10.1093/nar/gkx1030

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

Cheng, M., Sheng, L., Gao, Q., Xiong, Q., Zhang, H., Wu, M., et al. (2019). The m(6)A methyltransferase METTL3 promotes bladder cancer

progression via AFF4/NF-kappaB/MYC signaling network. *Oncogene* 38, 3667–3680.

Cotroneo, C. E., Mangano, N., Dragani, T. A., and Colombo, F. (2021). Lung expression of genes putatively involved in SARS-CoV-2 infection is modulated in cis by germline variants. *Eur. J. Hum. Genet.* 1–18. doi: 10.1038/s41431-021-00831-y

Fan, Z., Gao, S., Chen, Y., Xu, B., Yu, C., Yue, M., et al. (2018). Integrative analysis of competing endogenous RNA networks reveals the functional lncRNAs in heart failure. *J. Cell. Mol. Med.* 22, 4818–4829. doi: 10.1111/jcmm.13739

Han, J., Wang, J. Z., Yang, X., Yu, H., Zhou, R., Lu, H. C., et al. (2019). METTL3 promote tumor proliferation of bladder cancer by accelerating pri-miR221/222 maturation in m6A-dependent manner. *Mol. Cancer* 18:110. doi: 10.1186/s12943-019-1036-9

He, L., Li, H., Wu, A., Peng, Y., Shu, G., and Yin, G. (2019). Functions of N6-methyladenosine and its role in cancer. *Mol. Cancer* 18:176. doi: 10.1186/s12943-019-1109-9

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* 2, 841–860. doi: 10.1214/08-aoas169

Lan, T., Li, H., Zhang, D., Xu, L., Liu, H., Hao, X., et al. (2019). KIAA1429 contributes to liver cancer progression through N6-methyladenosine-dependent post-transcriptional modification of GATA3. *Mol. Cancer* 18:186. doi: 10.1186/s12943-019-1106-z

Liu, T., Wei, Q., Jin, J., Luo, Q., Liu, Y., Yang, Y., et al. (2020). The m6A reader YTHDF1 promotes ovarian cancer progression via augmenting EIF3C translation. *Nucleic Acids Res.* 48, 3816–3831. doi: 10.1093/nar/gkaa048

Ma, L., Chen, T., Zhang, X., Miao, Y., Tian, X., Yu, K., et al. (2021). The m(6)A reader YTHDC2 inhibits lung adenocarcinoma tumorigenesis by suppressing SLC7A11-dependent antioxidant function. *Redox Biol.* 38:101801. doi: 10.1016/j.redox.2020.101801

Ma, S., Chen, C., Ji, X., Liu, J., Zhou, Q., Wang, G., et al. (2019). The interplay between m6A RNA methylation and noncoding RNA in cancer. *J. Hematol. Oncol.* 12:121. doi: 10.1186/s13045-019-0805-7

Ma, Z., and Ji, J. (2020). N6-methyladenosine (m6A) RNA modification in cancer stem cells. *Stem Cells*. 1–9. doi: 10.1002/stem.3279

Niu, Y., Lin, Z., Wan, A., Chen, H., Liang, H., Sun, L., et al. (2019). RNA N6-methyladenosine demethylase FTO promotes breast tumor progression through inhibiting BNIP3. *Mol. Cancer* 18:46. doi: 10.1186/s12943-019-1004-4

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Shen, X. Y., Liu, X. P., Song, C. K., Wang, Y. J., Li, S., and Hu, W. D. (2019). Genome-wide analysis reveals alcohol dehydrogenase 1C and secreted phosphoprotein 1 for prognostic biomarkers in lung adenocarcinoma. *J. Cell Physiol.* 234, 22311–22320. doi: 10.1002/jcp.28797

Sun, S., Han, Q., Liang, M., Zhang, Q., Zhang, J., and Cao, J. (2020). Downregulation of m(6) A reader YTHDC2 promotes tumor progression and predicts poor prognosis in non-small cell lung cancer. *Thorac. Cancer* 11, 3269–3279. doi: 10.1111/1759-7714.13667

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tan, Y. W., Hong, W., and Liu, D. X. (2012). Binding of the 5′-untranslated region of coronavirus RNA to zinc finger CCHC-type and RNA-binding motif 1 enhances viral replication and transcription. *Nucleic Acids Res.* 40, 5065–5077. doi: 10.1093/nar/gks165

Tang, B., Yang, Y., Kang, M., Wang, Y., Wang, Y., Bi, Y., et al. (2020). m(6)A demethylase ALKBH5 inhibits pancreatic cancer tumorigenesis by decreasing WIF-1 RNA methylation and mediating Wnt signaling. *Mol. Cancer* 19:3. doi: 10.1186/s12943-019-1128-6

Wang, P., Zhang, L., Huang, C., Huang, P., and Zhang, J. (2018). Distinct prognostic values of alcohol dehydrogenase family members for non-small cell lung cancer. *Med. Sci. Monit.* 24, 3578–3590. doi: 10.12659/MSM.910026

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Xiao, S., Cao, S., Huang, Q., Xia, L., Deng, M., Yang, M., et al. (2019). The RNA N(6)-methyladenosine modification landscape of human fetal tissues. *Nat. Cell Biol.* 21, 651–661. doi: 10.1038/s41556-019-0315-4

Zhang, B., Wu, Q., Li, B., Wang, D., Wang, L., and Zhou, Y. L. (2020). m(6)A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. *Mol. Cancer* 19:53. doi: 10.1186/s12943-020-01170-0

Zhang, C., Zhang, M., Ge, S., Huang, W., Lin, X., Gao, J., et al. (2019). Reduced m6A modification predicts malignant phenotypes and augmented Wnt/PI3K-Akt signaling in gastric cancer. *Cancer Med.* 8, 4766–4781. doi: 10.1002/cam4.2360

Zheng, Y., Nie, P., Peng, D., He, Z., Liu, M., Xie, Y., et al. (2018). m6AVar: a database of functional variants involved in m6A modification. *Nucleic Acids Res.* 46, D139–D145. doi: 10.1093/nar/gkx895

Zhuang, Z., Chen, L., Mao, Y., Zheng, Q., Li, H., Huang, Y., et al. (2020). Diagnostic, progressive and prognostic performance of m(6)A methylation RNA regulators in lung adenocarcinoma. *Int. J. Biol. Sci.* 16, 1785–1797. doi: 10.7150/ijbs.39046

**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

METHODS
published: 28 June 2021
doi: 10.3389/fgene.2021.639872

# Detecting lncRNA–Cancer Associations by Combining miRNAs, Genes, and Prognosis With Matrix Factorization

*Huan Yan[1,2], Hua Chai[3] and Huiying Zhao[1,2]**

[1] Department of Medical Research Center, Sun Yat-sen Memorial Hospital, Guangzhou, China, [2] Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangzhou, China, [3] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

**Motivation:** Long non-coding RNAs (lncRNAs) play important roles in cancer development. Prediction of lncRNA–cancer association is necessary for efficiently discovering biomarkers and designing treatment for cancers. Currently, several methods have been developed to predict lncRNA–cancer associations. However, most of them do not consider the relationships between lncRNA with other molecules and with cancer prognosis, which has limited the accuracy of the prediction.

**Method:** Here, we constructed relationship matrices between 1,679 lncRNAs, 2,759 miRNAs, and 16,410 genes and cancer prognosis on three types of cancers (breast, lung, and colorectal cancers) to predict lncRNA–cancer associations. The matrices were iteratively reconstructed by matrix factorization to optimize low-rank size. This method is called detecting lncRNA cancer association (DRACA).

**Results:** Application of this method in the prediction of lncRNAs–breast cancer, lncRNA–lung cancer, and lncRNA–colorectal cancer associations achieved an area under curve (AUC) of 0.810, 0.796, and 0.795, respectively, by 10-fold cross-validations. The performances of DRACA in predicting associations between lncRNAs with three kinds of cancers were at least 6.6, 7.2, and 6.9% better than other methods, respectively. To our knowledge, this is the first method employing cancer prognosis in the prediction of lncRNA–cancer associations. When removing the relationships between cancer prognosis and genes, the AUCs were decreased 7.2, 0.6, and 5% for breast, lung, and colorectal cancers, respectively. Moreover, the predicted lncRNAs were found with greater numbers of somatic mutations than the lncRNAs not predicted as cancer-associated for three types of cancers. DRACA predicted many novel lncRNAs, whose expressions were found to be related to survival rates of patients. The method is available at https://github.com/Yanh35/DRACA.

**Keywords: lncRNA, cancer, prognosis, survival, mutation**

OPEN ACCESS

***Edited by:***
*Rosalba Giugno, University of Verona, Italy*

***Reviewed by:***
*Francesco Russo, Statens Serum Institut (SSI), Denmark*
*Jianzhao Gao, Nankai University, China*

****Correspondence:***
*Huiying Zhao*
*zhaohy8@mail.sysu.edu.cn*

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

***Received:*** *10 December 2020*
***Accepted:*** *15 April 2021*
***Published:*** *28 June 2021*

***Citation:***
*Yan H, Chai H and Zhao H (2021) Detecting lncRNA–Cancer Associations by Combining miRNAs, Genes, and Prognosis With Matrix Factorization. Front. Genet. 12:639872. doi: 10.3389/fgene.2021.639872*

## INTRODUCTION

The human genome consists of protein-encoding mRNA and non-coding RNAs (ncRNAs), but only a small portion of the human genome corresponds to the protein-coding genes (PCGs; Atkinson et al., 2012; Ezkurdia et al., 2014). Among ncRNA, long non-coding RNAs (lncRNAs) are transcription length over 200 nucleotides (Wilusz et al., 2009; Evans et al., 2016) that play important roles in a variety of biological processes and pathological conditions of cancers. The abnormal transcriptions of lncRNA may cause changes in the expression of target genes related to cancer pathways (Prensner and Chinnaiyan, 2011; de Lena et al., 2017). For example, lncRNA *PTENP1* is a pseudogene of the tumor suppressor *PTEN*, which inhibits the induction of autophagy in liver cancers (Chen et al., 2015). Another lncRNA *GAS5* has been shown to regulate cancer proliferation in many human cancer systems (Mazar et al., 2017). In recent years, a portion of lncRNAs has gradually been used as biomarkers of cancers. For example, in human hepatocellular carcinoma cells (HCCs), the lncRNA, uc002mbe.2, is expressed at lower levels than normal cells, but its expression can be increased 300-fold after treatment with histone deacetylase inhibitor Trichostatin A (TSA, Yang et al., 2013). The lncRNA *SChLAP1* is a tissue biomarker that can be used to identify prostate cancer patients at high risk of fatal progression, according to a study of prostate cancer patients in the United States (Mehra et al., 2016). Unfortunately, efficiently identifying lncRNAs–cancers associations is a challenge due to the complexity of relationships between them.

Detecting associations of lncRNAs and common cancers is important for early diagnosis and improving overall survival rate. Currently, breast, lung, and colorectal cancers are the most frequently diagnosed cancers. Although the overall survival rate of breast cancer has improved significantly, it is still an important cause of global death (Kalimutho et al., 2019). Therefore, it is necessary to identify lncRNAs associated with cancers for improving the early diagnosis. In recent years, a growing number of evidences demonstrate that lung cancer is one of the main causes of cancer death in men and women all around the world (Jemal et al., 2011). Simultaneously, colorectal cancer is the third most common cancer worldwide, with 1.36 million people diagnosed in 2012 (Ferlay et al., 2015). Thus, the occurrence of these three types of cancers is a serious threat to human health. Predicting potential lncRNAs associated with these cancers can provide useful information for prevention, diagnosis, and treatment.

Many lncRNAs play important roles through interacting with miRNAs. miRNA is a class of single-stranded RNAs with about 22 long chains of nucleotides, which act as either oncogene or tumor suppressor (Bartel, 2004). Accumulating evidences demonstrated that lncRNA–miRNA crosstalk has emerged as core roles in the pathogenesis and development of human cancer (Xue et al., 2017). Thus, constructing lncRNA–miRNA relationship may help to identify lncRNA–cancer associations.

By using interactions between lncRNA with other molecules, many methods have been developed to predict potential lncRNA–cancer associations (Chen et al., 2017). Liu et al. (2015) proposed a method that utilized the expression profiles of lncRNAs and PCGs in cancers to construct lncRNA–PCG bipartite network, which was then used to identify cancer-associated lncRNAs *via* random walks. It has previously used human phenotypic ontologies to annotate disease to improve the predictive power of lncRNA associated with disease (Le and Dao, 2018). Recently, based on the relationships of lncRNA or miRNA with other molecules, matrix factorization methods were used to predict lncRNA–disease associations (Fu et al., 2018) and miRNA–disease associations (Xuan et al., 2019). LION model applied the characteristics of lncRNAs, genes, and diseases to predict the relationships between lncRNAs and diseases through network diffusion (Sumathipala et al., 2019). At the same time, there are also related study based on heterogeneous clustering methods to predict the unknown relationships between lncRNAs and diseases based on the relationship network constructed by diseases, lncRNAs, microRNAs, and genes (Barracchia et al., 2018). LP-HCLUS uses multi-type hierarchical clustering methods to predict potentially lncRNA–disease relationships (Barracchia et al., 2020). However, all these methods only discriminate disease-associated lncRNAs without relating the lncRNAs with specific cancer types.

Moreover, all these methods overlooked the relationships between lncRNAs and cancer prognosis. The presence of lncRNAs in cancers can be an important factor clinically determining the prognosis of patients. Recently, an approach has been proposed to estimate the relationship between genes and the cancer prognosis by analyzing multi-omics data and clinical information from The Cancer Genome Atlas (TCGA) database (Wang et al., 2018). More recently, a method was presented to determine the gene and patient prognosis for 13 types of cancers (Chai et al., 2019), which reminds us to use the relationships between genes and the prognosis of three types of cancers in the prediction of lncRNA–cancer association.

In this study, we constructed a method, called detecting lncRNA cancer association (DRACA), to predict associations between lncRNAs and three common cancers. This method integrated the relationships between lncRNAs, cancer prognosis, miRNAs, genes, and cancers into a matrix and utilized matrix factorization to fuse multiple effective biological features in the prediction. This is the first method using cancer prognosis to detect lncRNA–cancer associations, which was indicated as a critical feature in the prediction. Further analyses indicated that the predicted cancer-associated lncRNAs contain significantly more somatic mutations than the average. In addition, several novel cancer-associated lncRNAs predicted by this study were significantly correlated with the survival rates of cancer patients and were expressed to be significantly different in cancer tissues and paracarcinomatous tissues. Thus, the predicted lncRNAs are biologically meaningful in the cancer process.

## METHODS

### Matrix Factorization

The matrices were constructed by the relationships between $N$ ($N$ = 5) kinds of features. The main

framework of the model is to optimize the equation:

$$\min_{G \geq 0} O(G, S, W) = \sum_{R_{ij} \in \mathcal{R}} W_{ij} ||R_{ij} - G_i S_{ij} G_j^T||_F^2 + \alpha ||vec(W)||_F^2 \quad (1)$$

$$s.t.\ W \geq 0,\ \sum vec(W) = 1$$

where α is used to control the complexity of $vec(w)$ (set as $1 \times 10^5$ in the study), $R_{ij}$ is a collection of relations across data sources that include $R_{LM}$, $R_{LG}$, $R_{LC}$, $R_{GP}$, $R_{MG}$, $R_{MC}$, and $R_{GC}$ (**Table 1**), $i$ and $j$ are the $i$th and $j$th features from two different data sources, respectively, $R_{ij}$ is reconstructed as $G_i S_{ij} G_j^T$ by singular vector decomposing (SVD), $W$ is calculated by Equation 2, $i$ and $j$ are two kinds of features, and $|| \cdot ||_F^2$ is the Frobenius norm.

The low-rank size of reconstructed matrix in Equation 1 was optimized according to the prediction of lncRNA–cancer relationships in the training set by giving appropriate weights ($W_{ij}$). $W_{ij}$ was calculated by Equation 2, where γ is the Lagrangian multipliers. Here, the performance of the prediction was evaluated by Area Under Curve (AUC). To avoid overfitting, 10-fold cross-validation was employed.

$$w_{ij} = \begin{cases} \frac{\gamma - H_{ij}}{2\alpha}, if\ \gamma - H_{ij} > 0\ and\ R_{ij} \in R \\ 0,\ if\ \gamma - H_{ij} \leq 0\ and\ R_{ij} \notin R \end{cases} \quad (2)$$

$$(H_{ij} = ||R_{ij} - G_i S_{ij} G_j^T||_F^2)$$

## Dataset Construction

The dataset includes five kinds of features and their relationships, which are lncRNAs, miRNAs, genes, cancers, and cancer prognosis. The relationships between these features were collected from public databases. The lncRNA–miRNA relationships ($R_{LM}$) were downloaded from starBase v2.0 (Li et al., 2014); the lncRNA–gene interactions ($R_{LG}$) were from lncReg (Zhou et al., 2015); the lncRNA–cancer associations ($R_{LG}$) were from lncRNADisease (Bao et al., 2018); the miRNA–gene relationships ($R_{MG}$) were from miRTarbase (Chou et al., 2018); the miRNA–cancer relationships ($R_{MC}$) were from MNDR v2.0 (Cui et al., 2018); the gene–cancer ($R_{GC}$) relationships were from DisGeNet (Pinero et al., 2017).

**TABLE 1 |** The matrix size and the number of associations in the dataset.

| Relationships | Matrices | Size | Associations |
|---|---|---|---|
| lncRNA–miRNA | $R_{LM}$ | 1,679 × 2,759 | 10,120 |
| lncRNA–gene | $R_{LG}$ | 1,679 × 16,410 | 511 |
| lncRNA–cancer | $R_{LC}$ | 1,679 × 3 | 542 |
| miRNA–gene | $R_{MG}$ | 2,759 × 16,410 | 380,639 |
| miRNA–cancer | $R_{MC}$ | 2,759 × 3 | 3,343 |
| Gene–cancer | $R_{GC}$ | 16,410 × 3 | 9,015 |
| Gene–prognosis | $R_{GP}$ | 16,410 × 3 | 1,169 |

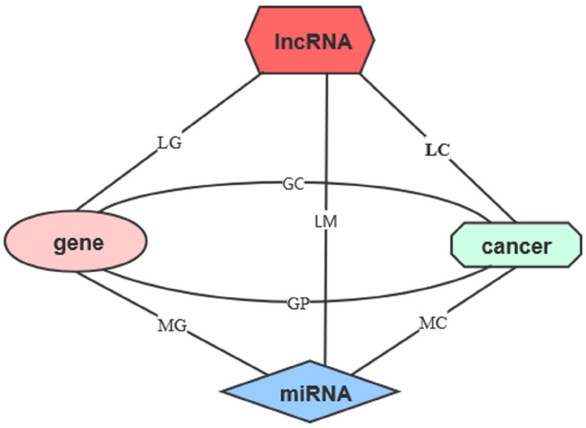



**FIGURE 1 |** The network of five features. The five features include lncRNAs, miRNAs, genes, cancers, and cancer prognosis. The line represents the relationship matrices.

Additionally, we calculated the gene–prognosis relationships ($R_{GP}$) by integrating multi-omics data from TCGA as described in a previous study (Chai et al., 2019). Briefly, we downloaded multi-omics data including RNA expression data, DNA methylation data, and copy number variation data of 614 breast cancer patients, 733 lung cancer patients, and 255 colorectal cancer patients from TCGA dataset[1]; then, we employed Autoencoder to rebuild composite features that were subsequently used by Cox proportional hazard model to estimate the prognosis risk of patients. Finally, XGboost was used to classify the prognosis of patients into high and low risks by scoring relationships between genes and the prognosis. The scores of genes were ranged from 0 to 1. The genes with scores higher than 0.5 were defined as highly correlated. The relationships between the genes and the prognosis of three kinds of cancers were included in the matrix factorization model. In summary, this study constructed a dataset including 1,679 lncRNAs, 2,759 miRNAs, 16,410 genes, and 16,410 genes–prognosis relationships and three kinds of cancers (breast, lung, and colorectal).

The relationships between these data are provided in **Table 1**. By using these relationships, we constructed lncRNA–cancer network as shown in **Figure 1**. The lncRNA–cancer relationships in lncRNADisease were used as golden standards to determine the lncRNA–cancer associations. As shown in **Table 1**, **542** lncRNA–cancer associations in the database were considered as the positive dataset, and 4,495 lncRNA–cancer with no relationships were included as the negative dataset. Briefly, 185, 179, and 178 lncRNAs associated with breast cancer, lung cancer, or colorectal cancer were collected as the positive dataset, whereas 1,494, 1,500, and 1,501 lncRNAs not associated with breast cancer, lung cancer, or colorectal cancer were collected as the negative dataset.

[1]https://www.cancer.gov/tcga

## Statistical Measurements in Evaluating the Methods

The 10-fold cross-validation was used to evaluate the performance of DRACA. We randomly divided positive and negative genes into 10-fold and used nine-fold as training and one-fold for testing. This process was repeated for 10 times. The prediction AUC was calculated for the testing fold. The average AUC was used as 10-fold cross-validation result of the model. In this study, we used AUC, maximum Matthews correlation coefficient (MCC), accuracy (ACC), precision, sensitivity, and specificity to evaluate the performance of DRACA. Calculations of these measurements were shown in Equations 3–7.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$precision = \frac{TP}{TP+FP} \quad (5)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (6)$$

$$specificity = \frac{TN}{FP+TN} \quad (7)$$

# RESULTS

## The Influences of the Low-Rank Size (k)

The low-rank size ($k$) of decomposed matrix in Equation 1 was optimized according to the performance of prediction. The performance was evaluated by AUC. In this study, $k_1$ was the low-rank size of $R_{[lncRNA]}$ that was the relationship between lncRNA with other features and was kept as 1,679; $k_4$ and $k_5$ were the low-rank sizes of $R_{[cancer]}$ and $R_{[cancer\ prognosis]}$ that were the relationships between cancers with other features and were kept as 3. $k_2$ and $k_3$ were the low-rank sizes of $R_{[miRNA]}$ and $R_{[gene]}$ that were relationships between miRNA and gene with other molecules and cancers, respectively. $k_2$ and $k_3$ were optimized.

The $k_2$ was optimized from 10 to 2,759 by a step of 100 and keeping $k_3$ as 50 to reduce the computational cost. As a result, when $k_2$ = 1,610, the highest AUC of 0.787 was achieved. Then, $k_3$ was trained by keeping $k_2$ = 1,610. The best AUC of 0.789 was provided when $k_3$ = 1,810. Then, we examined the performance of the model in predicting the lncRNA associations with breast cancer, lung cancer, and colorectal cancer, respectively. AUC values of 0.806, 0.801, and 0.778 were achieved, respectively, for three types of cancers.

We expected that the model gave a better performance when it was trained for a specific cancer. Here, this model was trained for prediction of associations between lncRNA and breast cancer, lncRNA and lung cancer, and lncRNA and colorectal cancer, respectively. In the training procedure, $k_2$ and $k_3$ were optimized, and 10-fold cross-validation was applied to avoid over training. For breast cancer, when $k_2$ = 2,210 and $k_3$ = 2,510, the highest AUC of 0.810 was obtained, which was slightly higher than the AUC of 0.806 obtained by the model trained for predicting all associations between the cancers and lncRNA. For lung cancer, when $k_2$ = 1,110 and $k_3$ = 3,110, the AUC was 0.796 that was a marginal decrease compared with 0.801 obtained by the model trained for prediction of all associations between the cancers and lncRNA. For colorectal cancer, $k_2$ = 1,610 and $k_3$ = 710 provided the highest AUC of 0.795 that was higher than the AUC of 0.778 reached by predicting all associations between the cancers and lncRNA. The results are shown in **Table 2**. We further used this method in liver hepatocellular carcinoma. Result indicated that the 10-fold cross-validation AUC achieved 0.749 and MCC achieved 0.313 (**Table 2**).

## Measuring the Contribution of the Features

To measure the contribution of each feature in the prediction, we individually removed the relationships between features and examined their influence on AUC areas. For prediction of breast cancer-associated lncRNAs, when the relationship between genes and cancer prognosis ($R_{GP}$) was removed, the AUC of DRACA was reduced from 0.810 to 0.738 (7.20%). In removing the relationship $R_{GP}$ in the prediction of lung cancer, the AUC was reduced from 0.796 to 0.790 (0.60%). In the prediction of lncRNA–colorectal cancer association, the removal of $R_{GP}$ dramatically reduced the AUC values from 0.795 to 0.745 (5.00%). We also examined the contributions of the relationships, $R_{LM}$, $R_{LG}$, and $R_{MG}$, in the prediction of the associations of lncRNA with three types of cancers, respectively. The results are shown in **Table 3**. As shown in **Table 3**, the lncRNA–miRNA ($R_{LM}$) was the most important feature in the prediction. Meanwhile, we found that removing the gene–cancer relationships or miRNA–cancer relationships can also reduce the prediction.

**TABLE 2 |** The performance of DRACA in the prediction of associations between lncRNA and three types of cancers.

| Cancer | AUC (AUC[a]) | MCC | ACC | Precision | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Breast cancer | 0.810 (0.806) | 0.336 | 0.658 | 0.232 | 0.910 | 0.625 |
| Lung cancer | 0.796 (0.801) | 0.404 | 0.764 | 0.294 | 0.858 | 0.764 |
| Colorectal cancer | 0.795 (0.778) | 0.371 | 0.714 | 0.254 | 0.888 | 0.694 |
| Liver hepatocellular carcinoma | 0.749 | 0.313 | 0.676 | 0.236 | 0.841 | 0.656 |

[a] *The AUC values of the DRACA model that was trained to predict the association between lncRNA and three cancers.*

**TABLE 3 |** The AUCs and MCCs for DRACA predictions after removing the associations between features.

| | Breast cancer | | Lung cancer | | Colorectal cancer | |
|---|---|---|---|---|---|---|
| | **AUC** | **MCC** | **AUC** | **MCC** | **AUC** | **MCC** |
| All | 0.81 | 0.336 | 0.796 | 0.404 | 0.795 | 0.371 |
| $-R_{LM}$ | 0.57 | 0.048 | 0.585 | 0.056 | 0.549 | −0.010 |
| $-R_{LG}$ | 0.749 | 0.333 | 0.756 | 0.356 | 0.731 | 0.312 |
| $-R_{MG}$ | 0.668 | 0.258 | 0.685 | 0.313 | 0.569 | 0.154 |
| $-R_{GP}$ | 0.738 | 0.347 | 0.79 | 0.387 | 0.745 | 0.303 |
| $-R_{MC}$ | 0.715 | 0.338 | 0.734 | 0.339 | 0.722 | 0.294 |
| $-R_{GC}$ | 0.5 | 0 | 0.5 | 0 | 0.5 | 0 |

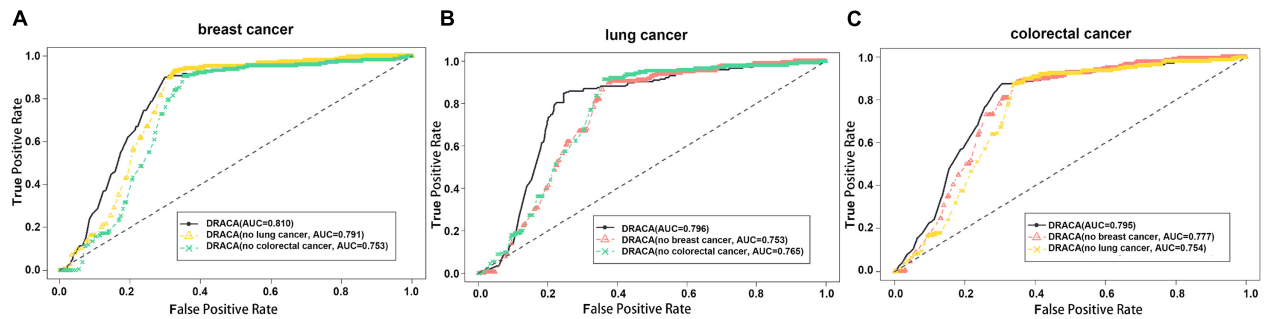



**FIGURE 2 |** The influences of cancer types in detecting lncRNA cancer association (DRACA) prediction. The black lines represent the ROC curves of DRACA without removing any features in predicting associations between lncRNA and cancers. The red and green lines denote the ROC curves of DRACA removing cancer information. **(A)** The ROC curves to show the influences of removing lung cancer information (yellow curve) or removing colorectal cancer information (green curve) in predicting breast cancer associated lncRNAs; **(B)** The ROC curves to show the influences of removing breast cancer (red curve) or removing colorectal cancer information (green curve) in predicting lung cancer associated lncRNAs; **(C)** The ROC curves to show the influences of removing breast cancer information (red curve) or lung cancer (yellow curve) in predicting colorectal cancer associated lncRNAs.

When all the miRNA-related features (lncRNA–miRNA, miRNA–gene, and miRNA–cancer features) were removed from the prediction or all the gene-related features (gene–cancer, gene–prognosis, gene–cancer, and miRNA–gene features) were removed from the prediction, the AUC values of DRACA are close to random. More details are included in **Supplementary Table 1**.

## The Impact of Other Cancers on the Prediction

This study constructed DRACA by including the information of three types of cancers that may have influences on the prediction. These influences were tested through excluding cancer information individually. As shown in **Figure 2**, in the prediction of lncRNA–breast cancer associations, removing the lung cancer and removing the colorectal cancer individually resulted in the AUCs of 0.791 and 0.753, respectively, which are lower than the AUC value 0.810 obtained by using all the features. **Figure 2** also describes the impacts of breast cancer and colorectal cancer in the prediction of lung cancer-associated lncRNA and the impacts of breast cancer and lung cancer in the prediction of colorectal cancer-associated lncRNAs. When removing breast cancer or colorectal cancer information in predicting lung cancer-associated lncRNAs, the AUC values were decreased from 0.796 to 0.753 or from 0.796 to 0.765, respectively. The contributions of breast cancer and lung cancer in the prediction of lncRNAs associated with colorectal cancer were indicated by the reduced AUCs from 0.795 to 0.777 and to 0.754, respectively. Thus, colorectal cancer contributed more in the predictions of lncRNA–breast cancer and lncRNA–lung cancer associations than two other cancers. Moreover, removing lung cancer had reduced more AUC values in predicting lncRNA–colorectal cancer associations than in removing breast cancer.

## Comparison With Other Methods

Detecting lncRNA cancer association was compared with the Naïve Bayesian classifier to predict potential lncRNA–disease associations (NBCLDA; Yu et al., 2018) in terms of MCC on the same dataset by 10-fold cross-validation. NBCLDA is a method constructing a global tripartite network that combines lncRNA–cancer, miRNA–cancer, and miRNA–lncRNA associations, including gene–miRNA interactions, gene–lncRNA associations, and gene–disease interactions, to predict potential lncRNA–disease associations. **Table 4** uncovers that DRACA always performed better in MCCs (0.336, 0.404, and 0.371) than NBCLDA (0.265, 0.256, and 0.245).

We also compared the predictions of DRACA with the method developed by integratinglncRNA—disease network, lncRNA functional similarity network, and the disease semantic similarity network (BPLLDA, Xiao et al., 2018). This method inferred the lncRNA–disease association according to the paths connecting

**TABLE 4 |** Comparing DRACA with three methods on MCC values.

| | Breast cancer | Lung cancer | Colorectal cancer |
|---|---|---|---|
| DRACA | 0.336 | 0.404 | 0.371 |
| NBCLDA | 0.265 | 0.256 | 0.245 |
| BPLLDA | 0.330 | 0.248 | 0.393 |
| MFLDA | 0.161 | 0.141 | 0.057 |

them and their lengths in the network. BPLLDA was developed based on a database including 156 lncRNAs and their associated diseases. Among these lncRNAs, 56 were included in the DRACA database, which were used to compare these two methods. The comparison was performed by 10-fold cross-validation and measured by MCC. As shown by **Table 4**, DRACA performed significantly better than BPLLDA in the prediction of lncRNA–breast cancer associations, lncRNA–lung cancer associations, and lncRNA–colorectal cancer associations.

Furthermore, we compared DRACA with the method developed to predict the lncRNA-disease associations based on matrix factorization approaches MFLDA (Fu et al., 2018). It is different from DRACA in two respects. First, it is a method without considering the relationship between lncRNA and cancer prognosis. Second, it has been constructed by 214 lncRNAs that is much less than the number of lncRNAs in DRACA. Out of 214 lncRNAs, 98 were from the DRACA database, which were used for the comparison. The results indicated that DRACA was superior to MFLDA in predicting the relationships between lncRNAs and three types of cancers.

In summary, DRACA was compared with three recently developed methods in predicting lncRNA–cancer associations. The results indicated that DRACA performed always better than NBCLDA, BPLLDA, and MFLDA in the prediction of three types of cancers. Moreover, DRACA has been constructed by 1,679 lncRNAs that are 7 and 11 times more than lncRNAs in BPLLDA and MFLDA, respectively. Thus, DRACA can potentially discover more novel lncRNA–cancer associations.

## Testing the Predicted lncRNA–Cancer Associations

Detecting lncRNA cancer association gives each lncRNA a score to indicate its relationship with certain cancer. The higher the score, the higher the probability that the lncRNA and the cancer are related. In order to select candidate lncRNAs, we used the maximum MCC to obtain the score threshold. The MCC was calculated by Equation 3. The best MCCs of 0.336, 0.404, and 0.371 were achieved for breast cancer, lung cancer, and colorectal cancer, respectively. When DRACA achieved the best MCC, we also calculated other statistical measurements including accuracy (ACC), precision, sensitivity, and specificity, as shown in **Table 2**.

By using the thresholds given by the best MCCs for the three types of cancers (0.785, 0.965, and 0.815), 636, 521, and 616 lncRNAs were predicted as related to breast cancer, lung cancer, and colorectal cancer, respectively. From them, we checked the top 20 candidate lncRNAs (a total of 60 lncRNAs for three types of cancers) that were not collected in the lncRNADisease database. We searched these lncRNAs in PubMed to obtain the literatures regarding their relationships with cancers. For breast cancer, lung

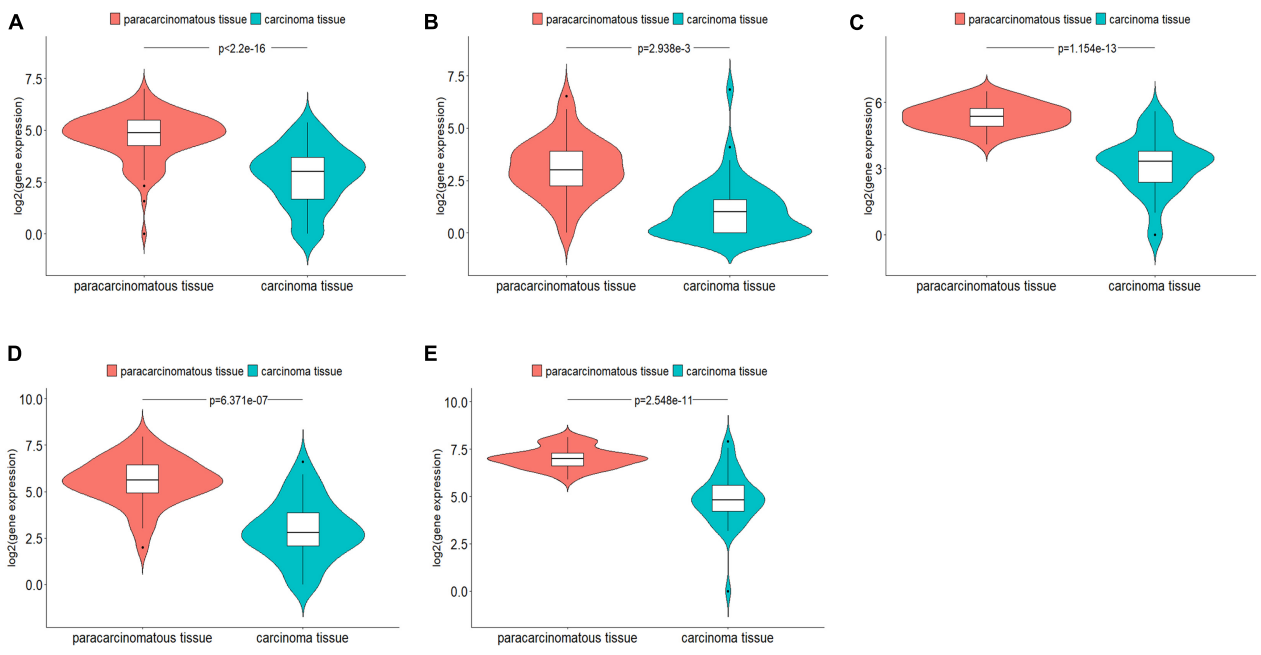



**FIGURE 3 |** Five genes that were predicted as cancer-associated by DRACA were found expressed significantly different between carcinoma tissues and paracarcinomatous tissues. **(A)** *Lnc-LAMC2-1:1* was found expressed significantly different in breast cancer tissues and paracarcinomatous tissues; **(B)** *DGKK* expressed significantly different in lung cancer tissues and paracarcinomatous tissues; **(C–E)** *EPB41L4A-AS2*, *MANCR*, and *lnc-HOXC4-3:1* expressed significantly different in colorectal cancer tissues and paracarcinomatous tissues.

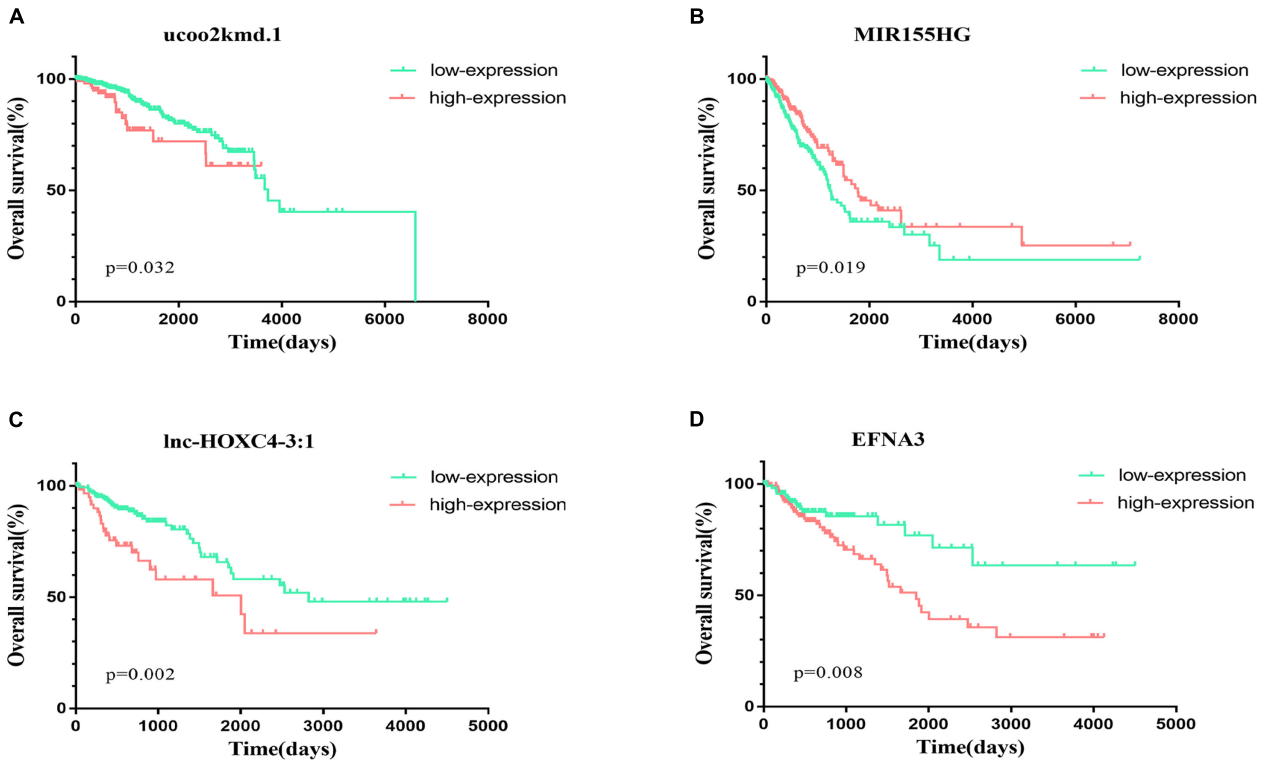



**FIGURE 4 |** The survival curves of two groups of patients who highly and lowly expressed *ucoo2kmd.1*, *MIR155HG*, *lnc-HOXC4-3:1*, and *EFNA3*, respectively. The patients were divided into two groups using the *surv_cutpoint* function of the *survminer R* package according to the gene expression levels, which were represented as "High-expression" and "Low-expression," respectively. The red lines denote the relationship between survival time and overall survival for the patients in the "High-expression" group, and the blue lines represent the relationship between the survival time and the overall survival for the patients in the "Low-expression" group. **(A)** The survival curves of two groups of the breast cancer patients who highly and lowly expressed *ucoo2kmd.1* gene, respectively; **(B)** The survival curves of two groups of the lung cancer patients who highly and lowly expressed *MIR155HG* gene, respectively; **(C)** The survival curves of two groups of the colorectal cancer patients who highly and lowly expressed *lnc-HOXC4-3:1* gene, respectively. **(D)** The survival curves of two groups of the colorectal cancer patients who highly and lowly expressed *EFNA3* gene, respectively.

cancer, and colorectal cancer, respectively, 10, 10, and 13 out 20 lncRNAs were reported as related with cancers. More details are included in **Supplementary Tables 2–4**.

For these predicted new lncRNAs, we examined if they were expressed to be significantly different in carcinoma tissues and paracarcinomatous tissues. Out of 60 predicted top cancer-associated lncRNAs, 20 were included in TCGA database, which included seven predicted as associated with breast cancer, five predicted as associated with lung cancer, and eight predicted as associated with colorectal cancer. From TCGA database, we downloaded gene expression data for 106 breast cancer patients, 52 lung cancer patients, and 38 colorectal patients. By comparing the gene expression data of these 20 lncRNAs in the carcinoma tissues and the paracarcinomatous tissues using *edgeR R* package ($FDR < 0.05$, $|\log FC| > 1$), five lncRNAs were found to be expressed significantly different, which included one lncRNA for breast cancer, one lncRNA for lung cancer, and three lncRNAs for colorectal cancer (**Figure 3**). The statistical evaluations on the differences of gene expression are shown in **Supplementary Table 5**.

We also analyzed the relationships between 20 lncRNAs and the patient survival rates. From TCGA database, we downloaded survival information for 611 breast cancer patients, 439 lung cancer patients, and 251 colorectal cancer patients. Patients were divided into the high-expression group and low-expression group by using the *surv_cutpoint* function of the *survminer R* package according to the gene expression. Then, we compared the overall survival rates of two groups. The results were shown in Kaplan–Meier plots (**Figure 4**). The differences of the survival rates were tested by the log-rank (Mantel–Cox) test. Here, the overall survival rates were the numbers of cases living for a certain period divided by the total numbers of patients in this group at the beginning. Genes were defined as significantly related with patient survival rates if the Mantel–Cox test *P*-value is lower than 0. Out of 20 genes, 5 were found to be significantly related with the patient survival rates. Briefly, patients in the low-expression and high-expression groups of *ucoo2kmd.1* were found to be significantly different in survival rates according to Mantel–Cox test (*P*-value = 0.032) as shown in **Figure 4A**. Similarly, the expression of *MIR155HG* (**Figure 4B**) was found to be significantly (*P*-value = 0.019) associated with the overall

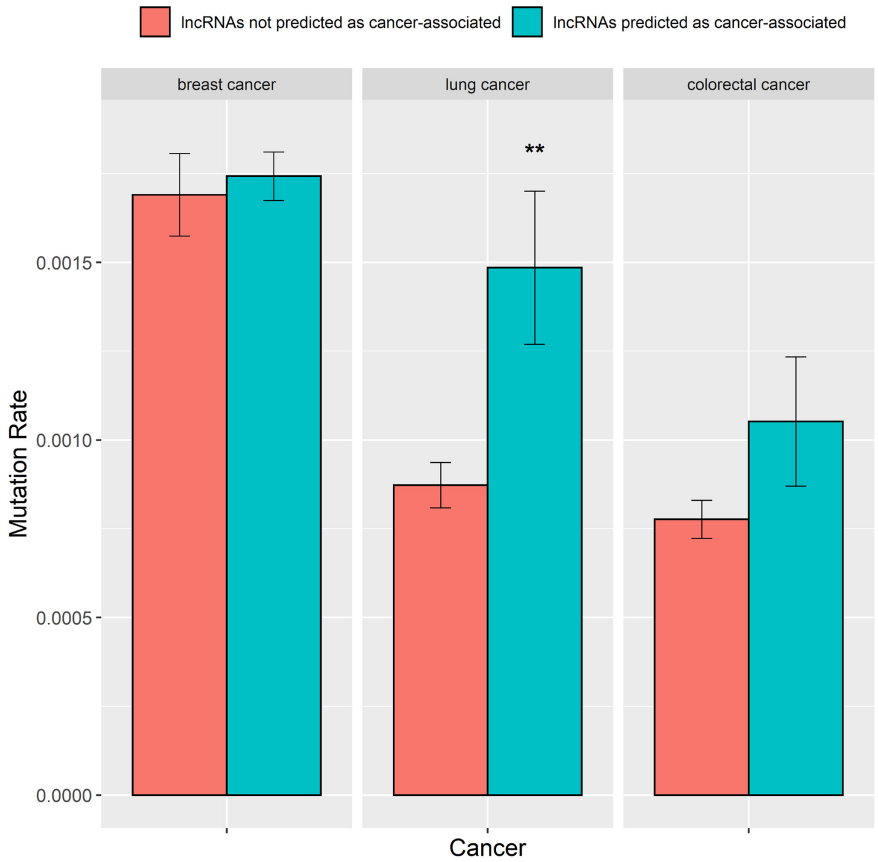



**FIGURE 5 |** The mutation rates in the lncRNAs predicted as cancer-associated by DRACA are higher than in the lncRNAs not predicted as cancer-associated. "*" denotes *t*-test *P*-value < 5.0–2E; "**" represents *t*-test *P*-value < 1.0–2E.

survival of lung cancer. At the same time, the expressions of *lnc-HOXC4-3:1* (**Figure 4C**), *EFNA3* (**Figure 4D**), and *LINC00520* (**Supplementary Figure 6**) were identified to be significantly related with the overall survival of colorectal cancer patients with *P*-values of 0.002, 0.008, and 0.021, respectively. Among these genes, *lnc-HOXC4-3:1* and *EFNA3* were also found to be expressed significantly different in carcinoma tissues and paracarcinomatous tissues as shown in **Figure 3C**.

## The Numbers of Somatic Mutations in lncRNAs Predicted as Cancer-Associated by Detecting lncRNA Cancer Association

A greater number of mutations in lncRNAs raise their probability for causing cancers (Beroukhim et al., 2010; Huarte, 2015). Hence, we explored whether the predictions of the DRACA model are correlated with the number of mutations in lncRNAs. We collected somatic mutation data from the international cancer genome consortium (ICGC) database, which contained somatic mutations of 651 lncRNAs for breast cancer, 568 lncRNAs for lung cancer, and 526 lncRNAs for colorectal cancer. Then, we examined the difference between the number of mutations in the lncRNAs that were predicted as cancer-associated and in the lncRNAs that were not predicated as cancer-associated by DRACA. The lncRNAs were defined as cancer-associated if their scores were higher than the threshold giving the best MCC. For three types of cancers, the numbers of mutations in the lncRNAs that are predicted as cancer-associated are higher than those in the lncRNAs that are not predicted as cancer-associated. The lncRNAs predicted as breast cancer-, lung cancer-, and colorectal cancer-associated were indicated with more somatic mutations than the lncRNAs not predicted as cancer related with *P*-values, 3.5e-1, 3.5e-3, and 7.4e-2 (**Figure 5**). Thus, the lncRNAs predicted as cancer-associated tend to occur with more somatic mutations.

## CONCLUSION

In this study, we presented a method, DRACA, that is an approach using miRNAs, genes, lncRNAs, and cancer prognosis to construct matrices in the prediction of lncRNA–cancer associations. DRACA utilizes matrix factorization technology to decompose different heterogeneous data matrices into low-rank matrices by tri-factorization and optimizing weight for matrices.

Using 10-fold cross-validation, we searched the appropriate sizes of low-rank matrices and verified the validity of the features. In a 10-fold cross-validation experiment, the method obtains AUCs of 0.810, 0.796, and 0.795 in predicting lncRNA-related breast cancer, lung cancer, and colorectal cancer. DRACA was compared with three methods, NBCLDA, BPLLDA, and MFLDA, and was indicated with significantly better performances. To illustrate the biological meaning of the prediction, we compared the predicted score with the number of somatic mutations in each lncRNA. We found that the lncRNAs predicted as cancer-associated have more somatic mutations than the lncRNAs not predicted as cancer-associated. Thus, integrating the relationships among lncRNAs, miRNAs, genes, and cancer prognosis with matrix factorization technology can accurately predict potential lncRNA–cancer associations. Moreover, among 20 novel lncRNAs predicted as cancer-associated by DRACA, nine were indicated to be expressed significantly different between the carcinoma tissues and the paracarcinomatous tissues, and five were significantly correlated with the survival rates of patients.

## DISCUSSION

lncRNAs had been viewed as "junk" in the genome. Recently, lncRNAs have attracted much attention due to the discovery that they are key regulators of cancer transformation and progression. Thus, discovering novel lncRNA–cancer association has possibilities to lead to early diagnosis and new treatment of cancers. Despite the rapid increase in the catalog of roles reported for lncRNAs, one of the greatest challenges is in the identification of cancer risk lncRNAs efficiently.

In this study, we presented an approach, DRACA, to predict lncRNAs associated with three specific cancers. DRACA is different from previously developed methods in several aspects. DRACA includes the feature of cancer prognosis, which greatly improves prediction ability but was missed by other methods. We used AUC to train the model and calculated the best MCC for each model. AUC and MCC are commonly used for evaluating the reliability of the model (Chicco and Jurman, 2020). However, MCC is easy to be fluctuated because MCC value is dependent on the prediction of score of each gene.

## 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

HZ designed and supervised the study. HY and HC conducted the analyses. HY wrote the manuscript. All authors contributed to the final revision of the manuscript.

## FUNDING

This work was supported by the National Key R&D Program of China (2018YFC0910500), GD Frontier & Key Tech Innovation Program (2019B020228001), National Natural Science Foundation of China (61772566, U1611261, 81801132, and 81971190), program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06D211), Guangdong Province Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation (2017B030314026), and Natural Science Foundation of Guangdong, China (2019A1515012207).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.639872/full#supplementary-material

## REFERENCES

Atkinson, S. R., Marguerat, S., and Bahler, J. (2012). Exploring long non-coding RNAs through sequencing. *Semin. Cell Dev. Biol.* 23, 200–205. doi: 10.1016/j.semcdb.2011.12.003

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2018). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037. doi: 10.1093/nar/gky905

Barracchia, E. P., Pio, G., D'Elia, D., and Ceci, M. (2020). Prediction of new associations between ncRNAs and diseases exploiting multi-type hierarchical clustering. *BMC Bioinformatics* 21:70. doi: 10.1186/s12859-020-3392-2

Barracchia, E. P., Pio, G., Malerba, D., and Ceci, M. (2018). "Identifying lncRNA-Disease Relationships via Heterogeneous Clustering," in *New Frontiers in Mining Complex Patterns. NFMCP 2017. Lecture Notes in Computer Science*, Vol. 10785, eds A. Appice, C. Loglisci, G. Manco, E. Masciari, and Z. Ras (Cham: Springer International Publishing), 35–48. doi: 10.1007/978-3-319-78680-3_3

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899–905. doi: 10.1038/nature08822

Chai, H., Zhou, X., Cui, Z., Rao, J., Hu, Z., Lu, Y., et al. (2019). Integrating multi-omics data with deep learning for predicting cancer prognosis. *bioRxiv* [Preprint] doi: 10.1101/807214

Chen, C. L., Tseng, Y. W., Wu, J. C., Chen, G. Y., Lin, K. C., Hwang, S. M., et al. (2015). Suppression of hepatocellular carcinoma by baculovirus-mediated expression of long non-coding RNA PTENP1 and MicroRNA regulation. *Biomaterials* 44, 71–81. doi: 10.1016/j.biomaterials.2014.12.023

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. doi: 10.1186/s12864-019-6413-7

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* 46, D371–D374. doi: 10.1093/nar/gkx1025

de Lena, P. G., Paz-Gallardo, A., Paramio, J. M., and Garcia-Escudero, R. (2017). Clusterization in head and neck squamous carcinomas based on lncRNA expression: molecular and clinical correlates. *Clin. Epigenetics* 9:36. doi: 10.1186/s13148-017-0334-6

Evans, J. R., Feng, F. Y., and Chinnaiyan, A. M. (2016). The bright side of dark matter: lncRNAs in cancer. *J. Clin. Invest.* 126, 2775–2782. doi: 10.1172/jci84421

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., et al. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* 23, 5866–5878. doi: 10.1093/hmg/ddu309

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi: 10.1002/ijc.29210

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nat. Med.* 21, 1253–1261. doi: 10.1038/nm.3981

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Kalimutho, M., Nones, K., Srihari, S., Duijf, P. H. G., Waddell, N., and Khanna, K. K. (2019). Patterns of genomic instability in breast cancer. *Trends Pharmacol. Sci.* 40, 198–211. doi: 10.1016/j.tips.2019.01.005

Le, D. H., and Dao, L. T. M. (2018). Annotating diseases using human phenotype ontology improves prediction of disease-associated long non-coding RNAs. *J. Mol. Biol.* 430, 2219–2230. doi: 10.1016/j.jmb.2018.05.006

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Liu, Y., Zhang, R., Qiu, F., Li, K., Zhou, Y., Shang, D., et al. (2015). Construction of a lncRNA-PCG bipartite network and identification of cancer-related lncRNAs: a case study in prostate cancer. *Mol. Biosyst.* 11, 384–393. doi: 10.1039/c4mb00439f

Mazar, J., Rosado, A., Shelley, J., Marchica, J., and Westmoreland, T. J. (2017). The long non-coding RNA GAS5 differentially regulates cell cycle arrest and apoptosis through activation of BRCA1 and p53 in human neuroblastoma. *Oncotarget* 8, 6589–6607. doi: 10.18632/oncotarget.14244

Mehra, R., Udager, A. M., Ahearn, T. U., Cao, X., Feng, F. Y., Loda, M., et al. (2016). Overexpression of the long non-coding rna schlap1 independently predicts lethal prostate cancer. *Eur. Urol.* 70, 549–552. doi: 10.1016/j.eururo.2015.12.003

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Prensner, J. R., and Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer Discov.* 1, 391–407. doi: 10.1158/2159-8290.Cd-11-0209

Sumathipala, M., Maiorino, E., Weiss, S. T., and Sharma, A. (2019). Network diffusion approach to predict LncRNA disease associations using multi-type biological networks: LION. *Front. Physiol.* 10:888. doi: 10.3389/fphys.2019.00888

Wang, M., Li, L., Liu, J., and Wang, J. (2018). A gene interaction network-based method to measure the common and heterogeneous mechanisms of gynecological cancer. *Mol. Med. Rep.* 18, 230–242. doi: 10.3892/mmr.2018.8961

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi: 10.1101/gad.1800909

Xiao, X., Zhu, W., Liao, B., Xu, J., Gu, C., Ji, B., et al. (2018). BPLLDA: predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. Genet.* 9:411. doi: 10.3389/fgene.2018.00411

Xuan, Z., Li, J., Yu, J., Feng, X., Zhao, B., and Wang, L. (2019). A probabilistic matrix factorization method for identifying lncRNA-disease associations. *Genes (Basel)* 10:126. doi: 10.3390/genes10020126

Xue, M., Zhuo, Y., and Shan, B. (2017). "MicroRNAs, Long Noncoding RNAs, and Their Functions in Human Disease," in *Bioinformatics in Microrna Research*, eds J. Huang, G. M. Borchert, D. Dou, J. Huan, W. Lan, M. Tan, et al. (New York, NY: Humana Press), 1–25. doi: 10.1007/978-1-4939-7046-9_1

Yang, H., Zhong, Y., Xie, H., Lai, X., Xu, M., Nie, Y., et al. (2013). Induction of the liver cancer-down-regulated long noncoding RNA uc002mbe.2 mediates trichostatin-induced apoptosis of liver cancer cells. *Biochem. Pharmacol.* 85, 1761–1769. doi: 10.1016/j.bcp.2013.04.020

Yu, J., Ping, P., Wang, L., Kuang, L., Li, X., and Wu, Z. (2018). A novel probability model for LncRNA-disease association prediction based on the naive bayesian classifier. *Genes* 9:345. doi: 10.3390/genes9070345

Zhou, Z., Shen, Y., Khan, M. R., and Li, A. (2015). LncReg: a reference resource for lncRNA-associated regulatory networks. *Database* 2015:bav083. doi: 10.1093/database/bav083

**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: 09 July 2021
doi: 10.3389/fgene.2021.642608

# Significance of Tumor Mutation Burden Combined With Immune Infiltrates in the Progression and Prognosis of Advanced Gastric Cancer

*Xiong Guo[1†], Xiaolong Liang[1†], Yujun Wang[2], Anqi Cheng[1], Han Zhang[3], Chuan Qin[1,4] and Ziwei Wang[1*]*

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

Gastric cancer (GC) is a serious malignant tumor with high mortality and poor prognosis. The prognosis and survival are much worse for advanced gastric cancer (AGC). Recently, immunotherapy has been widely promoted for AGC patients, and studies have shown that tumor mutation burden (TMB) is closely related to immunotherapy response. Here, RNA-seq data, matched clinical information, and MAF files were downloaded from the cancer genome atlas (TCGA)-STAD project in the TCGA database. The collation and visual analysis of mutation data were implemented by the "maftools" package in R. We calculated the TMB values for AGC patients and divided the patients into high- and low-TMB groups according to the median value of TMB. Then, the correlation between high or low TMB and clinicopathological parameters was calculated. Next, we examined the differences in gene expression patterns between the two groups by using the "limma" R package and identified the immune-related genes among the DEGs. Through univariate Cox regression analysis, 15 genes related to prognosis were obtained. Furthermore, the two hub genes (APOD and SLC22A17) were used to construct a risk model to evaluate the prognosis of AGC patients. ROC and survival curves and GEO data were used as a validation set to verify the reliability of this risk model. In addition, the correlation between TMB and tumor-infiltrating immune cells was examined. In conclusion, our results suggest that AGC patients with high TMB have a better prognosis. By testing the patient's TMB, we could better guide immunotherapy and understand patient response to immunotherapy.

**Keywords: advanced gastric cancer, tumor mutation burden, immune infiltration, prognosis, bioinformatics analysis**

OPEN ACCESS

***Edited by:***
*Shaoli Das,*
*National Institutes of Health (NIH), United States*

***Reviewed by:***
*Rituparno Sen,*
*Leipzig University, Germany*
*Binbin Wang,*
*National Cancer Institute, National Institutes of Health (NIH), United States*
*Provas Das,*
*Baylor University, United States*

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

*† These authors have contributed equally to this work*

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

***Received:*** *09 February 2021*
***Accepted:*** *17 June 2021*
***Published:*** *09 July 2021*

***Citation:***
*Guo X, Liang X, Wang Y, Cheng A, Zhang H, Qin C and Wang Z (2021) Significance of Tumor Mutation Burden Combined With Immune Infiltrates in the Progression and Prognosis of Advanced Gastric Cancer. Front. Genet. 12:642608. doi: 10.3389/fgene.2021.642608*

## INTRODUCTION

Gastric cancer (GC) is a common malignant tumor worldwide, with the fifth and third highest morbidity and mortality, respectively, of all cancers (Chen, 2016). This disease seriously threatens human health. The 5-year survival rate of advanced gastric cancer (AGC) is less than 25% (Ajani et al., 2017). In recent years, with the improvement of diagnosis and treatments, there has been a steady decline in the incidence and mortality rates of this cancer. However, despite the decline in incidence in most countries, clinicians are still expected to see more cases of GC in the future due to the aging population. On the other hand, because the onset of gastric cancer is insidious, it is frequently at an advanced stage at diagnosis, and resulting in a high mortality rate (Cascinu, 2020). At present, the best treatment for patients with GC is surgery, but aging patients cannot tolerate surgery, and in some cases the tumor is discovered too late for surgery to be effective. Therefore, palliative care is particularly important for these patients. In addition to radiotherapy and chemotherapy, immunotherapy has made great progress in recent years, and bringing hope to patients with AGC.

Traditionally, patients with advanced inoperable gastric cancer are treated with sequential chemotherapy, mainly platinum and fluoropyrimidine combination drugs (Song et al., 2017). However, the median survival is still less than 1 year. Recently, immune checkpoint inhibitors (ICIs), such as anti-programmed cell death-1 (PD-1) or programmed cell death ligand-1 (PD-L1) monoclonal antibodies, have improved the overall survival (OS) of various types of cancers, including AGC (Kim and Oh, 2018). To date, two anti-PD-1 inhibitors have been approved for AGC in Japan: nivolumab as third- or later-line treatment for AGC and pembrolizumab for previously treated patients with microsatellite instability-high tumors (Kawazoe et al., 2020). However, some gastric cancers may not be sensitive to immune checkpoint inhibitor monotherapies, so patients with gastric cancer may require combination therapy to improve the response to anti-PD-1 therapy. Therefore, methods to predict and improve patient response to immunotherapy or novel treatment methods are highly desired for AGC (Cascinu, 2020). A recent study suggested that predicting the response to immunotherapy on the basis of the tumor mutation burden (TMB) load may be a new opportunity (Morrison et al., 2018).

Tumor mutation burden is defined as the total number of somatic gene coding errors, base insertions, substitutions, or deletion errors detected per million bases (Yarchoan et al., 2017). Mutations in driver genes can lead to cancers. However, if a large number of somatic cell mutations occur, new antigens will be produced to activate CD8+ cytotoxic T cells, and triggering T-cell-mediated antitumor activity (Bi et al., 2020). Therefore, as the TMB increases, more new antigens are produced, and the tumors are more easily recognized by immune cells in the tumor microenvironment. TMB was used as a biomarker for anti-PD-1 treatment in colorectal cancer, and a higher TMB was associated with a better response to immunotherapy (Le et al., 2015). Recently, Tian et al. (2020) constructed a novel TMB estimation model that can be used as a prognostic biomarker for patients with non-small cell lung cancer. TMB can predict not only the response to immunotherapy but also patient survival. However, there are few studies on the relationship between TMB and immune infiltration in AGC.

In this study, we calculated the TMB of 338 AGC patients with complete clinical information, revealing the mutation characteristics of AGC patients. Then, we studied the correlation between the clinicopathological parameters and the normalized TMB value. Two TMB-related gene signatures were used to construct a risk model that could predict the survival of AGC patients. Moreover, we explored the relationship between TMB and the tumor microenvironment and provided new targets for immunotherapy for GC.

## MATERIALS AND METHODS

### Data Acquisition and Processing

The transcriptome data were obtained using the Illumina (San Diego, CA, United States) HiSeq 2000 RNA sequencing platform, and the genetic mutation data were downloaded from the cancer genome atlas (TCGA) database[1]. The transcriptome profiles are HTseq-Count files. The mutation data are in Annotated Somatic Mutation format, and the workflow type is "VarScan2 Annotation." Clinical data for the corresponding GC patients were also retrieved from the STAD project in TCGA database, which included age, tumor stage, sex, and survival information. The patient's clinical information was provided in **Supplementary Table 1**. We excluded patients with incomplete clinical information and a survival time of less than 30 days and then selected patients with AGC for analysis based on the clinical information. The "maftools" package in R software was used to visually analyze the mutation annotation format (MAF) file (Mayakonda et al., 2018). Gene chip data of gastric cancer was downloaded from the NCBI (National Center for Biotechnology Information) GEO database as the data for the validation set. The chip number is GSE84437, submitted by Yong-Min Huh and others. The study included transcriptome results and complete clinical information of 433 gastric cancers. In addition, the list of immune-related genes was obtained from the resources section of the ImmPort database[2].

### Calculation of the Tumor Mutation Burden

Tumor mutation burden was defined as the number of somatic coding insertion/deletion mutations and non-synonymous base replacements per megabase of the genome, and it was estimated by estimating the number of somatic mutations and dividing the total length of the exons. First, we used Perl scripts to extract tumor mutation data from AGC patient sequences and then used R software to calculate the TMB value according to the following formula for each patient:

$$TMB = Sn \times 1000000/n$$

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

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

where, Sn represents the absolute number of somatic mutations and n represents the number of exon bases with coverage depth $\geq 100\times$ (Jiang et al., 2019). The calculated TMB value of the patient is provided in **Supplementary Table 2**.

## Prognostic Analysis of TMB Value

After calculating the TMB value for each patient, the TMB value was combined with the patient clinical information, including survival status and survival time. Then, all patients were assigned to either the high- or low-TMB group, with the median value of TMB as the cutoff. Kaplan-Meier (K-M) survival analysis and log-rank tests were performed to evaluate the difference in the OS rate between the above two groups. Additionally, we explored the relationship between TMB and clinical features, including sex, age, tumor grade, and TNM stage. The patients were divided into two groups according to clinical characteristics, and the Wilcoxon rank-sum test was used for statistical analysis.

## Identification of TMB-Related Differentially Expressed Genes and Functional Enrichment Analysis

The gene expression data from AGC patients were standardized by the "limma" R package, and then the DEGs between the high- and low-TMB groups were identified using the Wilcoxon test. $|\ \mathrm{Log_2\_fold\ change\ (FC)}|\ > 1.0$ and false discovery rate (FDR) $< 0.05$ were used as cutoffs to identify qualified DEGs for subsequent analyses, and volcano maps and heat maps were used for visual analysis using the "pheatmap" R package. In addition, we carried out gene ontology (GO) and kyoto encyclopedia of genes and genomes (KEGG) pathway functional enrichment analyses by using the "clusterProfiler" R package and visualized the enrichment results (Yu et al., 2012).

## Construction and Verification of Risk Score Model

We took the intersection of the previously obtained immune-related gene list with the TMB-related differential genes and obtained the immune genes that were differentially expressed in the low- and high-TMB groups. Since these genes are related to immunity and TMB in AGC, they were used for further analysis. First, univariate Cox regression analysis was used to identify candidate genes associated with survival. Next, the "glmnet" package in R software was used to further filter the risk model with least absolute shrinkage and selection operator (LASSO) Cox regression analysis. Finally, multiple Cox regression analysis was used to further screen the optimal prognostic genes for the construction of risk models, and a time-dependent receiver operating characteristic (ROC) curve was used to assess the accuracy of the constructed model (Guo et al., 2020). The expression of genes and the regression coefficients obtained in the regression model were used to calculate the patients' risk scores. The calculation formula is as follows. Risk score (patients) $=$ $\Sigma$ Coefficient (gene $i$) $*$ expression value (gene $i$). Where, n, i, coefficient, and expression value represent the number of selected genes, gene number, regression coefficient value, and gene expression value, respectively.

Meanwhile, the log-rank test was used to analyze the survival data between the low- and high-TMB groups. In addition, GSE84437 data were downloaded from the GEO database as a validation set, and the risk model was used to analyze the prognosis of gastric cancer patients. The clinical information of patients in the GSE84437 database was provided in **Supplementary Table 3**. A nomogram was constructed by gene expression based on this model to predict the different annual survival rates of patients for TCGA and GEO data.

## Evaluation of Immune Cell Infiltration

CIBERSORT is a deconvolution algorithm that combines the labeled genomes of different immune cell subpopulations to calculate the proportions of 22 immune cells in tissues. The 22 types of immune cells include various myeloid cells, NK cells, 3 types of B cells, and 7 types of T cells (Bi et al., 2020). In this study, we analyzed tumor immune cell infiltration in the tumor microenvironment of AGC patients in the low- and high-TMB groups. Samples with a CIBERSORT output $p$-value $< 0.05$ were screened for further analysis.

Furthermore, the tumor immune estimation resource (TIMER) web server was used to precalculate the abundance of six tumor-infiltrating immune subsets (Kang et al., 2020). The modules in TIMER were used to explore the association of immune infiltration with gene expression and survival outcomes in the current study[3].

## Evaluation of the Value of Genes in the Model in a Pan-Cancer Panel

The cancer genome atlas pancancer data (ACC, BLCA, RCA, CESC, CCA, COAD, DLBC, GBM, HNSC, KIRC, KICH, KIRP, LGG, LAML, LIHC, LUSC, LUAD, MESO, OV, PAAD, PRAD, PCPG, READ, SKCM, SARC, TGCT, THYM, THCA, UCS, UCEC, and UVM), including RNA-Seq, stemness scores based on mRNA (RNAss) and DNA methylation (DNAss) and matched clinical information, were downloaded from the Xena browser[4]. We calculated the expression of APOD and SLC22A17 in the 33 cancers in the pancancer dataset, and through univariate Cox regression analysis, the risk values of these two genes for these 33 cancers were calculated. The Pearson correlation test method was used to calculate the correlation between gene expression and stromal scores, RNAss, and DNAss of 33 different cancer types based on the ESTIMATE algorithm. The drug responses to 262 FDA-approved drugs or drugs in clinical trials were included in the correlation analysis. The data were downloaded from the NCI-60 database, which contains data on 60 different cancer cell lines from 9 different tumors[5] (Zhang X. et al., 2020).

## Statistical Analyses

All data were processed with Perl (5.30.1) and R (version 3.6.2) software. Survival analyses were performed using the K-M method and the log-rank test. Pearson's correlation test was used for the correlation analysis between two groups. The Wilcoxon

[3]https://cistrome.shinyapps.io/timer/

[4]https://xenabrowser.net/datapages/

[5]http://bioinformatics.mdanderson.org/estimate/

rank-sum test was used for differential analyses of subgroups. All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

# RESULTS

## Somatic Mutation Analysis in Advanced Gastric Cancer

To identify somatic mutations in AGC patients in the TCGA database, we used the "maftools" package in R software to visually analyze the mutation data. Complete somatic mutation data were available for 251 AGC patients, of which 222 (88.45%) had somatic mutations. The 30 genes with the highest mutation rates in patients with AGC are displayed in the waterfall plot (**Figure 1A**) and include well-known cancer-related genes such as TTN (49%), TP53 (44%), and MUC16 (28%). Among them, missense mutations were the most common variant classification, single-nucleotide polymorphisms (SNPs) were the most common variant type, and $C > T$ mutations accounted for the vast majority of single nucleotide variations (SNVs) (**Supplementary Figures 1A–C**). The maximum number of mutations in one sample was 5137 (**Supplementary Figure 1D**), and the median number of mutations was 90 (**Supplementary Figure 1E**). In addition, we showed the number of each variant in the different samples through box plots (**Supplementary Figure 1F**). And the correlation calculations for top 20 mutated genes are shown in **Figure 1B**. Moreover, we classified these mutant genes and identified their enrichment in different pathways (**Supplementary Figure 1G**) and mutations in all samples of AGC (**Supplementary Figure 1H**). The most mutated pathways were RTK-RAS (77/85, 90.59%), WNT (66/68, 97.06%), and NOTCH (57/71, 80.28%). In addition, 55.78% of the patients had mutations in the RTK-RAS pathway (140/251), 43.82% (110/251) had mutations in the WNT pathway, and 42.63% (107/251) had mutations in the NOTCH pathway. These are the key signaling pathways in cancer progression. The mutant genes in RTK-RAS, WNT, and NOTCH pathway in patients with AGC are shown in the waterfall chart, respectively (**Supplementary Figures 1I–K**).

## Correlation Between TMB and Clinicopathological Characteristics of AGC Patients

To explore the prognostic function of TMB, we calculated and visualized the TMB value of gastric cancer samples in the TCGA database (**Figure 2A**). Then, we divided patients into low-TMB and high-TMB groups according to the median value of TMB. The TMB values for each patient were shown in **Supplementary Table 2**. The survival rate of the two groups was plotted by using K-M curves. Interestingly, we found that the survival rate of patients in the high TMB group was superior to that of patients in the low TMB group (**Figure 2B**). To further investigate the correlation between TMB and the clinical characteristics of gastric cancer patients. We downloaded the clinical information and detected the relationship between TMB and clinical features. The results showed that TMB is positively correlated with patient age. In addition, TMB was negatively correlated with sex and N stage. It means female patients with age $< 65$ have less TMB value than the other people. In addition, patients with no lymph node metastasis might have less TMB. There were no correlations between TMB and T stage, M stage, stage, or tumor grade (**Figures 2C–I**).

## Variation in the Genes Related to TMB and Functional Analysis

One of the ways in which TMB functions is to affect gene expression. To obtain the DEGs related to TMB, we divided

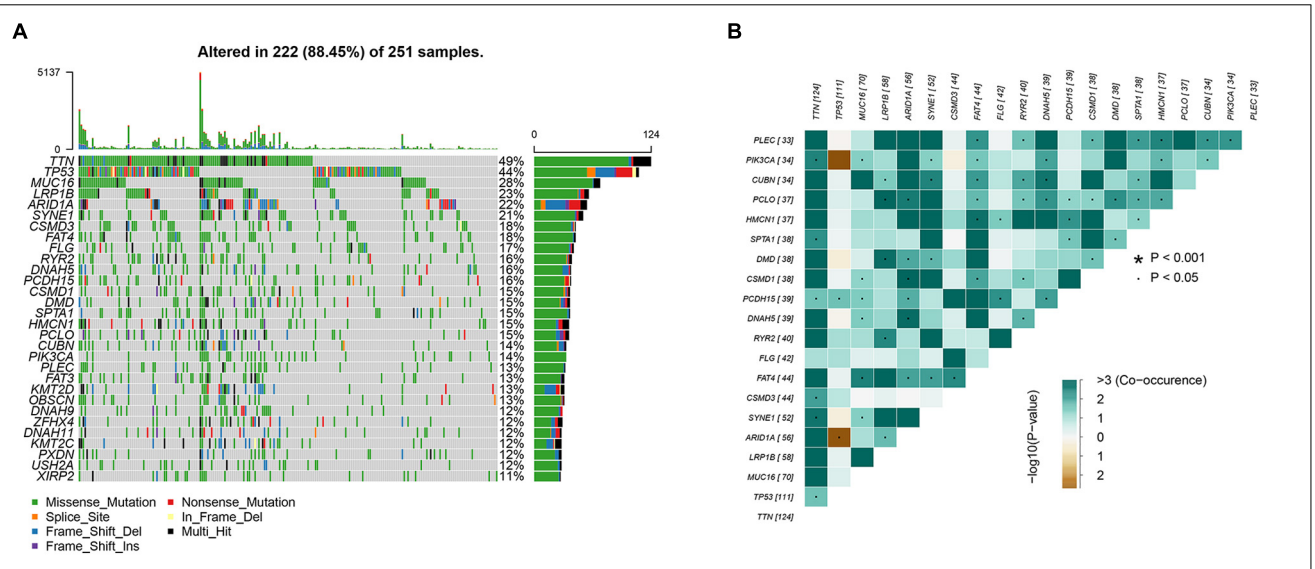



**FIGURE 1 |** Analyses of somatic mutation profiles in advanced gastric cancer. **(A)** Waterfall plot of detailed mutation information of top 30 genes in each sample, with various color annotations to distinguish different mutation types. **(B)** Correlation between the top 20 mutated genes.

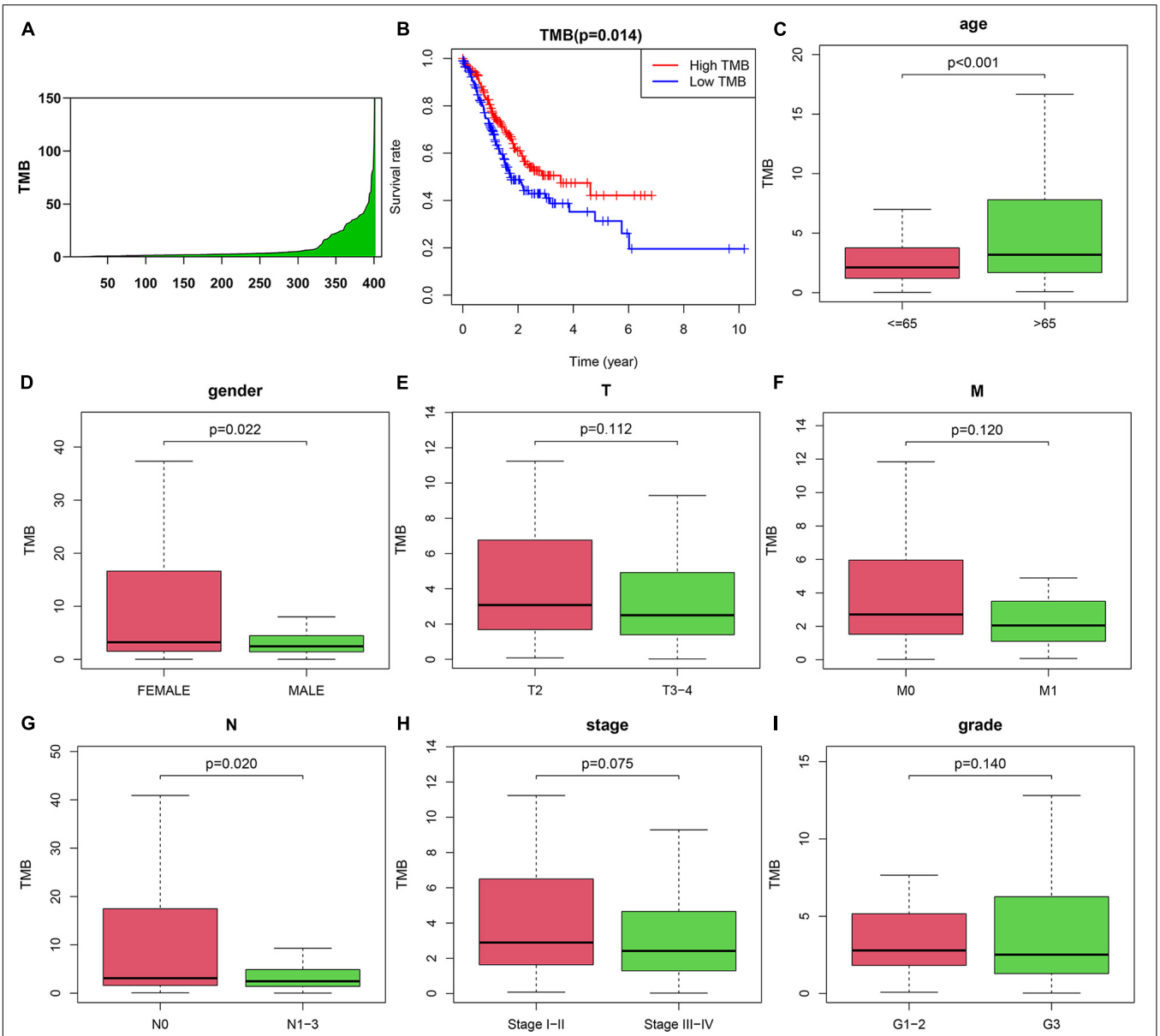



**FIGURE 2 |** Correlation between tumor mutation burden (TMB) and clinicopathological characteristics of AGC patients. **(A)** TMB value in advanced gastric cancer samples. **(B)** Survival analysis between high-TMB and low-TMB patients. **(C–I)** Correlation between TMB and **(C)** age, **(D)** gender, **(E)** T, **(F)** M, **(G)** N, **(H)** Stage, and **(I)** grade of patients.

patients into a high TMB group and a low TMB group according to the median TMB value. Then, the "limma" package in R software was used to identify genes that were differentially expressed between the two groups. We found 847 DEGs, including 796 upregulated genes and 51 downregulated genes, in the high TMB group compared with the low TMB group. The top 40 most DEGs were visualized by using a heat map (**Figure 3A**). A volcano map was plotted to exhibit the DEGs (**Figure 3B**). For GO analysis, we revealed that DEGs were mainly enriched in muscle system process, collagen-containing extracellular matrix and receptor ligand activity processes (**Figure 3C**). In addition, we conducted KEGG analysis based on DEGs. We found that DEGs mainly belonged to the neuroactive ligand-receptor interaction, cAMP signaling pathway, calcium signaling pathway, and vascular smooth muscle contraction and cell adhesion molecules categories (**Figure 3D**).

## Construction and Validation of Prognostic Model

To determine the relationship between TMB and immune infiltration in patients with AGC, we obtained immune-related DEGs by intersecting the 847 DEGs related to TMB with 1881 immune-related genes. A total of 107 immune-related

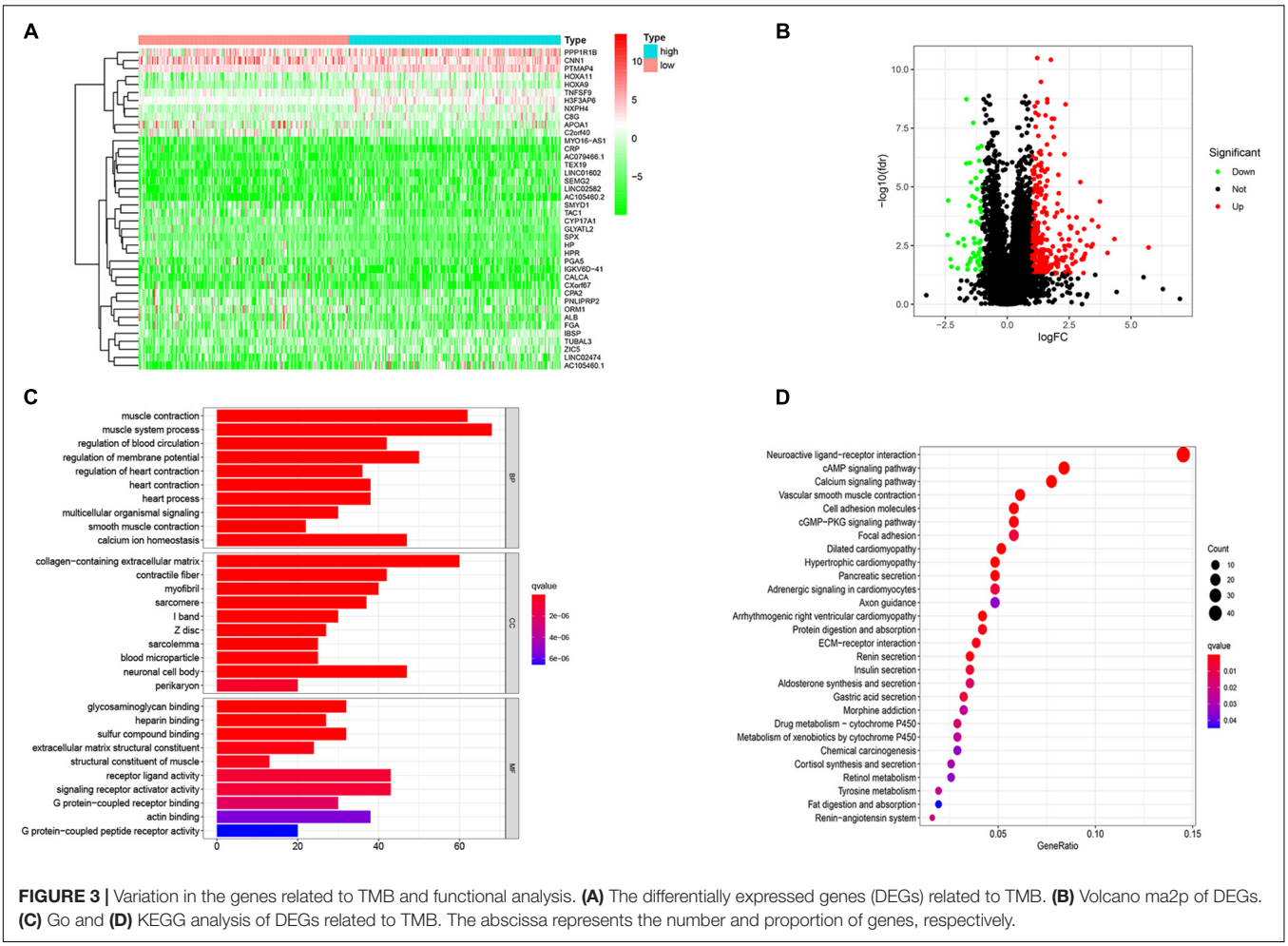



**FIGURE 3 |** Variation in the genes related to TMB and functional analysis. **(A)** The differentially expressed genes (DEGs) related to TMB. **(B)** Volcano ma2p of DEGs. **(C)** Go and **(D)** KEGG analysis of DEGs related to TMB. The abscissa represents the number and proportion of genes, respectively.

DEGs were identified for further analysis (**Figure 4A**). Then, we identified 15 genes as candidate prognosis-related genes by using univariate analysis (**Figure 4B**). The hazard ratio of prognostic genes was shown in **Table 1**. LASSO regression was subsequently performed on 15 candidate prognosis-related genes, and two genes were retained for constructing the prognostic model (**Figures 4C,D**). TCGA and GEO data were downloaded to verify the accuracy of the model. We first validated the accuracy of the model in the TCGA dataset. After ranking the patients according to the calculated risk score, patients were divided into a low-risk group and a high-risk group according to the median risk score. Low-risk group patients had better outcomes in terms of survival probability (**Figure 4E**). A ROC curve was plotted to validate the accuracy of the prognostic model (**Figure 4F**). Then, patients were ranked based on risk score (**Figure 4G**). The risk score for each patient was provided in **Supplementary Table 4**. We found that patients had longer survival times in the low-risk group, and more patients died in the high-risk group (**Figure 4H**). The expression of the two genes in each group was visualized by a heat map, and gene expression increased in parallel with the risk score (**Figure 4I**). Then, the GSE84437 data in the GEO database was used as the validation set, and we got similar results (**Supplementary Figures 2A–E**). This confirmed the reliability of our model.

## APOD and SLC22A17 Are Related to Patient Survival, TMB, and Patient Clinical Characteristics

We obtained two key genes, APOD and SLC22A17, from the prognostic model. To determine whether APOD and SLC22A17 affect the survival probability of patients, we performed K-M survival analysis to explore the survival rates of the two groups. It can be observed that higher expression of ADPO and SLC22A17 correlated with worse prognosis (**Figures 5A,B**). In addition, we found that patients in both the APOD low group and SLC22A17 low group had the better prognosis. Conversely, if the two genes both are highly expressed at the same time, the patient prognosis is even worse (**Figure 5C**). The expression of SLC22A17 and APOD in TMB-high and TMB-low group was shown in **Supplementary Figure 4**. These results indicated that APOD and SLC22A17 can be applied simultaneously for predicting patient prognosis. We further detected the relationship among the expression level of the two genes, TMB and clinical characteristics. The results showed that

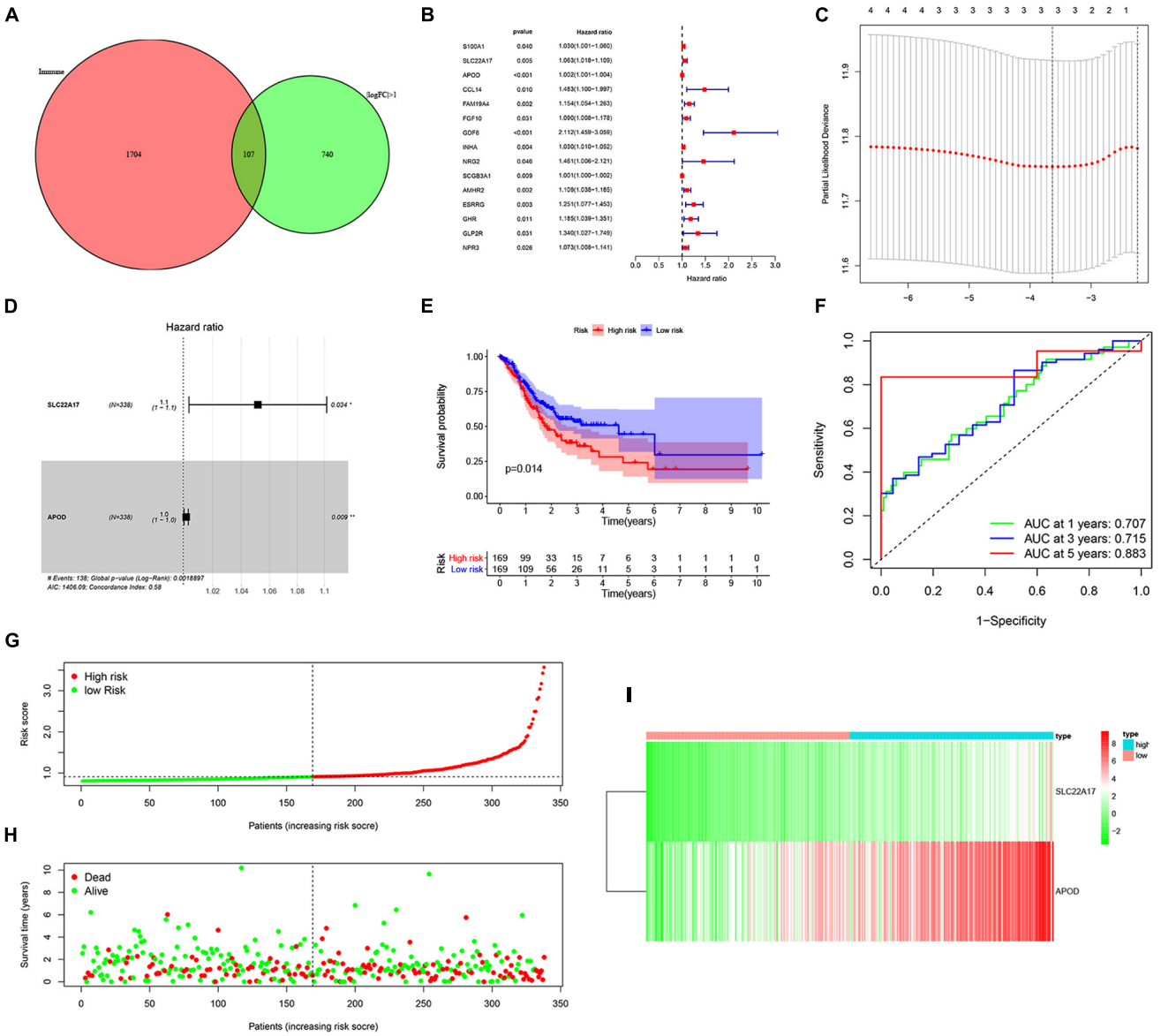



**FIGURE 4 |** Construction and validation of prognostic model. **(A)** Venn analysis of immune-related differentially expressed genes. **(B)** 15 candidate prognosis-related genes were obtained using univariate analysis. **(C,D)** Two prognosis-related genes were obtained by using LASSO regression and used for the construction of prognostic model. **(E,F)** High-risk group correlated with poor survival outcome, with $p = 0.014$. **(E)** Survival analysis of high-risk and low-risk groups. **(F)** ROC curves of 1, 2-, and 5-year survival prediction, with AUC = 0.707, 0.715, and 0.883, respectively. **(G,H)** The distribution of risk score and gene expression levels among patients in the cancer genome atlas (TCGA) data. **(I)** The expression of two prognostic genes between high-risk and low-risk patients in TCGA training set.

the expression of the two genes was lower in the high-TMB group (**Figure 5D**). The relationship between SLC22A17 and APOD gene expression and each clinical feature such as age, gender, grade, stage, and TNM-stage were shown in **Supplementary Figure 3**. We only found that the expression of SLC22A17 is related to the patient's age. In addition, a nomogram was further constructed according to the gene expression levels of APOD and SLC22A17 in the TCGA datasets. The patients' 1-, 2-, and 3-year survival could be predicted by using a nomogram (**Figure 5E**). At the same time, the calibration curves of the model also confirmed that the predicted 1-year survival rate was relatively consistent with the actual 1-year survival rate (**Figure 5F**).

## Relation of TMB and Prognostic Model Genes With Immune Cell Infiltration

Patients with higher TMB scores have been reported to manifest better response to immunotherapy. However, whether TMB is associated with immune infiltration remains unclear. In order explore the underlying association, we detected the proportions of 22 types of infiltrating immune cells in gastric cancer samples

**TABLE 1 |** Univariate COX regression analysis of TMB related prognostic genes in advanced gastric cancer.

| Gene symbol | HR | (95%CI) | *p*-Value |
|---|---|---|---|
| *S100A1* | 1.0301 | (1.0012–1.0597) | 0.0404 |
| *SLC22A17* | 1.0627 | (1.0184–1.1089) | 0.0051 |
| *APOD* | 1.0022 | (1.0009–1.0035) | 0.0006 |
| *CCL14* | 1.4825 | (1.1003–1.9974) | 0.0096 |
| *FAM19A4* | 1.1536 | (1.0536–1.2631) | 0.0019 |
| *GDF6* | 2.1123 | (1.4587–3.0587) | 7.52e-05 |
| *INHA* | 1.0304 | (1.0096–1.0516) | 0.0039 |
| *NRG2* | 1.4612 | (1.0064–2.1214) | 0.00160 |
| *SCGB3A1* | 1.0012 | (1.0003–1.0022) | 0.0088 |
| *GHR* | 1.1847 | (1.0390–1.3509) | 0.0113 |
| *GLP2R* | 1.3403 | (1.0274–1.7486) | 0.0307 |
| *NPR3* | 1.0725 | (1.0083–1.1408) | 0.0261 |
| *FGF10* | 1.0897 | (1.0080–1.1779) | 0.0306 |
| *AMHR2* | 1.1091 | (1.0381–1.1849) | 0.0021 |
| *ESRRG* | 1.2509 | (1.0767–1.4532) | 0.0034 |

*HR, hazard ratio; CI, confidence interval.*

by using the CIBERSORT algorithm. The results are shown in a bar plot map (**Figure 6A**). Then, we compared the distributions of the 22 types of infiltrating immune cells in the high-TMB and low-TMB groups. The results were visualized in a heat map (**Figure 6B**). We found that naive B cells, resting memory CD4 T cells, regulatory T cells (Tregs), activated NK cells, monocytes, resting dendritic cells and resting mast cells had higher levels of infiltration in the low-TMB group. In contrast, activated memory T cells, follicular helper T cells, resting NK cells, M0 macrophages, M1 macrophages, activated mast cells, and neutrophils were more abundant in the high-TMB group (**Figure 6C**). Next, we detected the correlations among 22 types of infiltrating immune cells and visualized them in a matrix based on the Pearson correlation coefficient (**Figure 6D**).

Furthermore, we calculated the correlation between the infiltration of each of the 22 types of immune cells and the expression of APOD and SLC22A17 (**Figure 7A**). Based on the correlation matrix, we found that APOD ($R = -0.28$, $p = 9.4E\text{-}06$) and SLC22A17 ($R = -0.22$, $p = 0.00072$) were negatively associated with T cell CD4 memory activation (**Figures 7B,C**). The TIMER, containing the abundance of six tumor-infiltrating immune subsets, was further utilized to detect the correlation between copy number variation and the infiltration level of immune cells. We found that the infiltration level was broadly decreased in patients with APOD and SLC22A17 copy number variation compared with the diploid/normal group (**Figures 7D,E**). To determine whether the infiltration levels of these six immune cells affect patient survival rate, we performed survival analysis to explore the association of immune infiltration with gene expression and survival outcomes. We observed that patients with low levels of macrophage infiltration had better survival outcomes (**Figure 7F**).

## Evaluation of the Value of TMB-Related Prognostic Model Genes Across Cancers

APOD and SLC22A17 are dysregulated and can be used for prognosis in gastric cancer patients. However, whether these two genes exert functions in other cancers is not known. To detect the value of the two genes in other cancers, we downloaded TCGA pancancer data. Then, we analyzed the expression levels of APOD and SLC22A17 in 33 types of cancers. We observed that APOD was dysregulated in 17 types of cancers and that SLC22A17 was dysregulated in 16 types of cancers, with significant *p*-values (**Figures 8A,B**). Univariate Cox regression analysis was subsequently used to identify the prognostic value in the 33 cancers (**Figure 8C**). The ESTIMATE algorithm was used to detect the correlation between gene expression and stromal scores, RNAss, and DNAss in 33 different cancer types. Not surprisingly, we found that APOD and SLC22A17 have a wide range of stromal scores in association with 33 different cancer types. In addition, in terms of the correlation between the two genes and cancer stemness, APOD and SLC22A17 had various degrees of association with the RNAss and DNAss in 33 types of cancers (**Figure 8D**). Interestingly, we observed that the APOD and SLC22A17 genes were negatively correlated with RNAss and DNAss in almost all of the cancer types. In contrast, SLC22A17 and APOD were positively associated with RNAss in patients with ACC, GBM, LGG, PCPG, and DLBC. In addition, SLC22A17 is strongly positively associated with DNAss in GBM, HNSC, THYM, USC, and UVM patients. APOD was strongly positively related to DNAss in CHOL, DLBC, KIRC, READ, SKCM, THCA, THYM, UCEC, and UVM patients.

Pearson correlation was subsequently performed to detect the correlation coefficient between the two genes and RNAss, DNAss, StromalScore, ImmuneScore, and ESTIMATEScore in patients with STAD. The SLC22A17 and APOD genes were negatively associated with RNAss and DNAss, which is consistent with the results of univariate Cox regression analysis. However, SLC22A17 and APOD had positive relationships with the StromalScore, ImmuneScore and ESTIMATEScore (**Figure 8E**). For the correlation between SLC22A17, APOD, and tumor drug resistance, we next determined the effect of SLC22A17 and APOD on drug sensitivity. Drugs approved by the FDA or drugs in clinical trials were selected for the correlation analysis. Interestingly, APOD exerts a greater role in drug sensitivity analysis. We found that APOD is positively related to sensitivity to vemurafenib, PD-98059, dabrafenib, hypothemycin, selumetinib, bafetinib, denileukin diftitox (Ontak), cobimetinib, and okadaic acid. By contrast, APOD is negatively associated with sensitivity to pyrazoloacridine, pralatrexate, batracylin, docetaxel, and floxuridine. However, SLC22A17 only had a negative relationship with the sensitivity to palbociclib and sunitinib (**Figure 8F**).

# DISCUSSION

Gastric cancer is a malignant tumor with a high recurrence rate and ranks as the third leading cause of cancer-related death worldwide (Al-Mahrouqi et al., 2011). In recent years, enormous

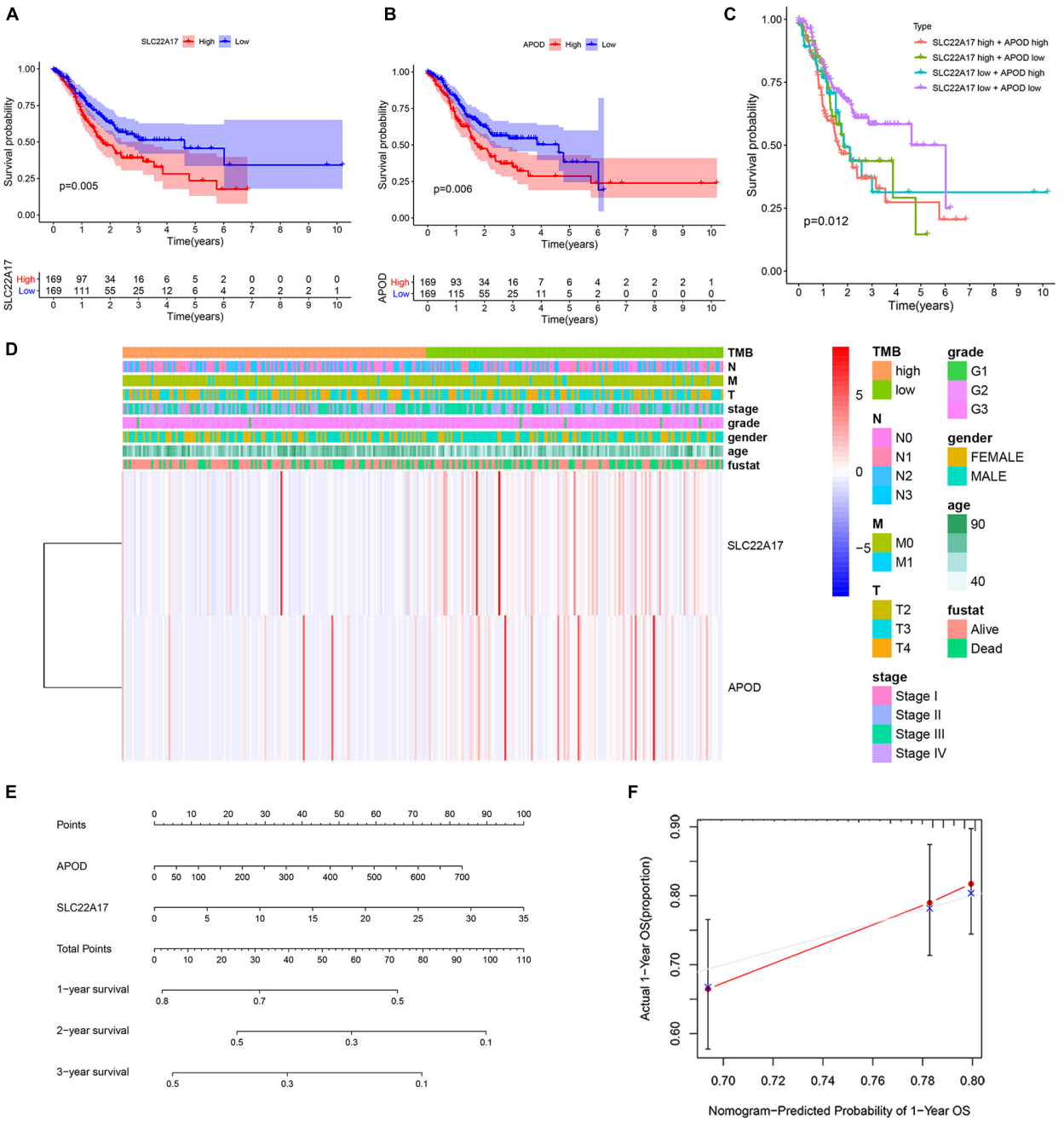



**FIGURE 5 |** Two genes in prognostic model were associated with patients' survival and clinical characteristics. **(A,B)** Survival analysis of **(A)** SLC22A17 and **(B)** APOD genes in patients with AGC. **(C)** Survival analysis of AGC patients with different expressions group of SLC22A17 and APOD. **(D)** The expression of SLC22A17 and APOD are associated with patients' TMB and clinical characteristics. **(E)** The patients' 1-, 2-, and 3-year survival were predicted by using a nomogram. **(F)** Calibration curves for the survival probability at 1 year.

progress has been made in the diagnosis and treatment of gastric cancer. However, the mortality of GC, and especially of AGC, remains high. Therefore, it is of great significance to explore the molecular subtypes of AGC and find effective targeted therapy strategies for specific subtypes.

Gene mutation is closely associated with the initiation and development of cancer (Ikediobi et al., 2006). For example, it has been reported that mutation in BRCA2 is closely related to patient survival, chemotherapy response, and genome instability (Yang et al., 2011). APC mutations are common in colorectal cancers (Nishisho et al., 1991). In addition, mutation of APC is related to the stage of colorectal cancer (Robles et al., 2016). Mutations in cancer-related genes also affect treatment strategies (Hu H. et al., 2018). TMB is a vital biological indicator

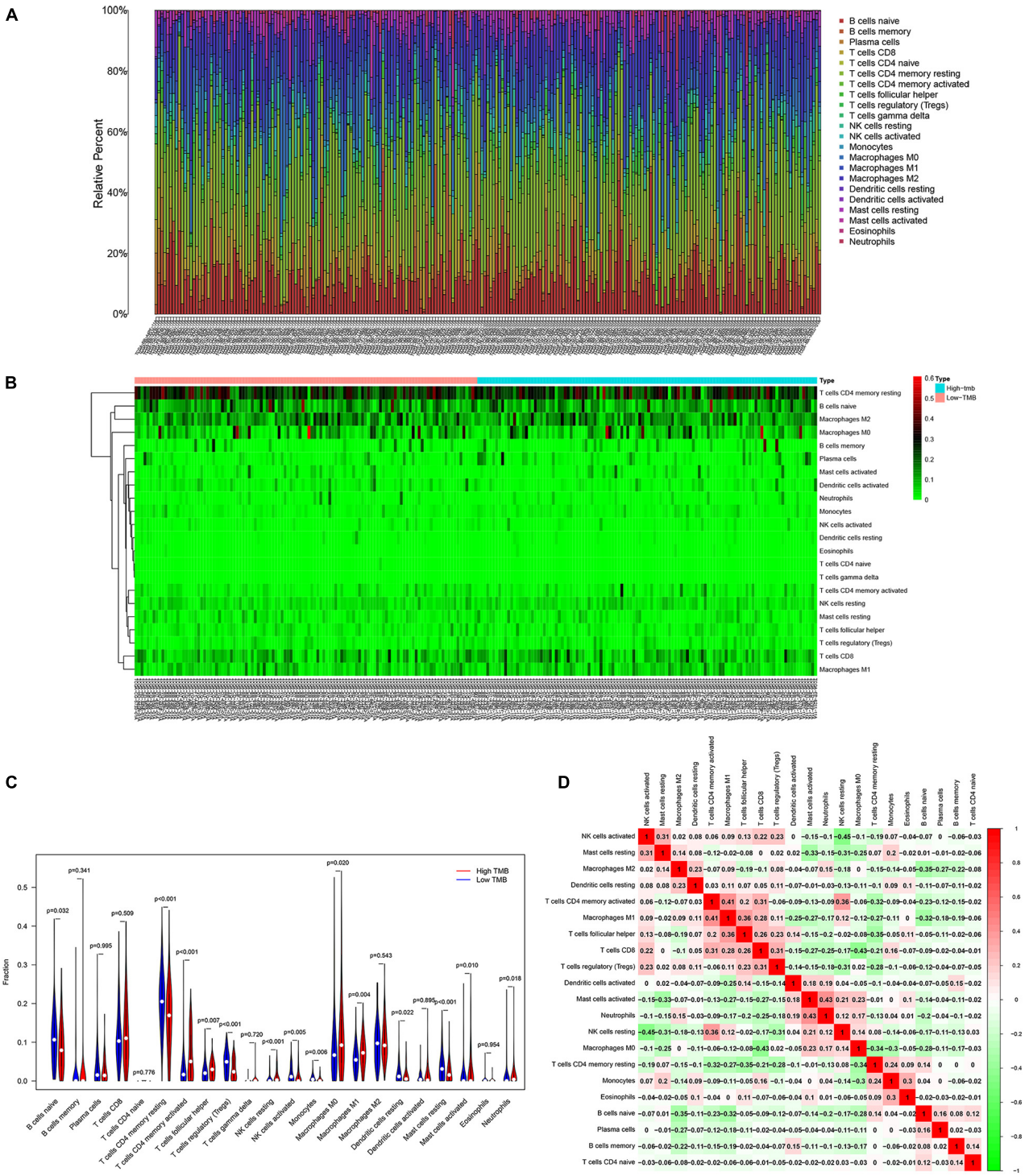



**FIGURE 6 |** Patients with various degree of TMB have different features of immune cell infiltration. **(A)** Proportion of immune infiltrating cells in gastric cancer samples. **(B)** Heat map and **(C)** bar graph of immune infiltrating cells between high-TMB and low-TMB patients. **(D)** Correlation analysis of 22 kinds of immune cells.

reflecting the degree of tumor mutation. TMB varies widely among cancer patients. Alexandrov LB reported that TMB could affect the immunotherapy effect of cancer (Alexandrov et al., 2013). Recently, TMB was identified as an immunotherapy biomarker (Chan et al., 2019). With regard to how TMB affects immunotherapy outcomes, Chen DS reported that there are more proteins produced by high-TMB patients, and these proteins can be recognized by the immune system. Immune cells are more

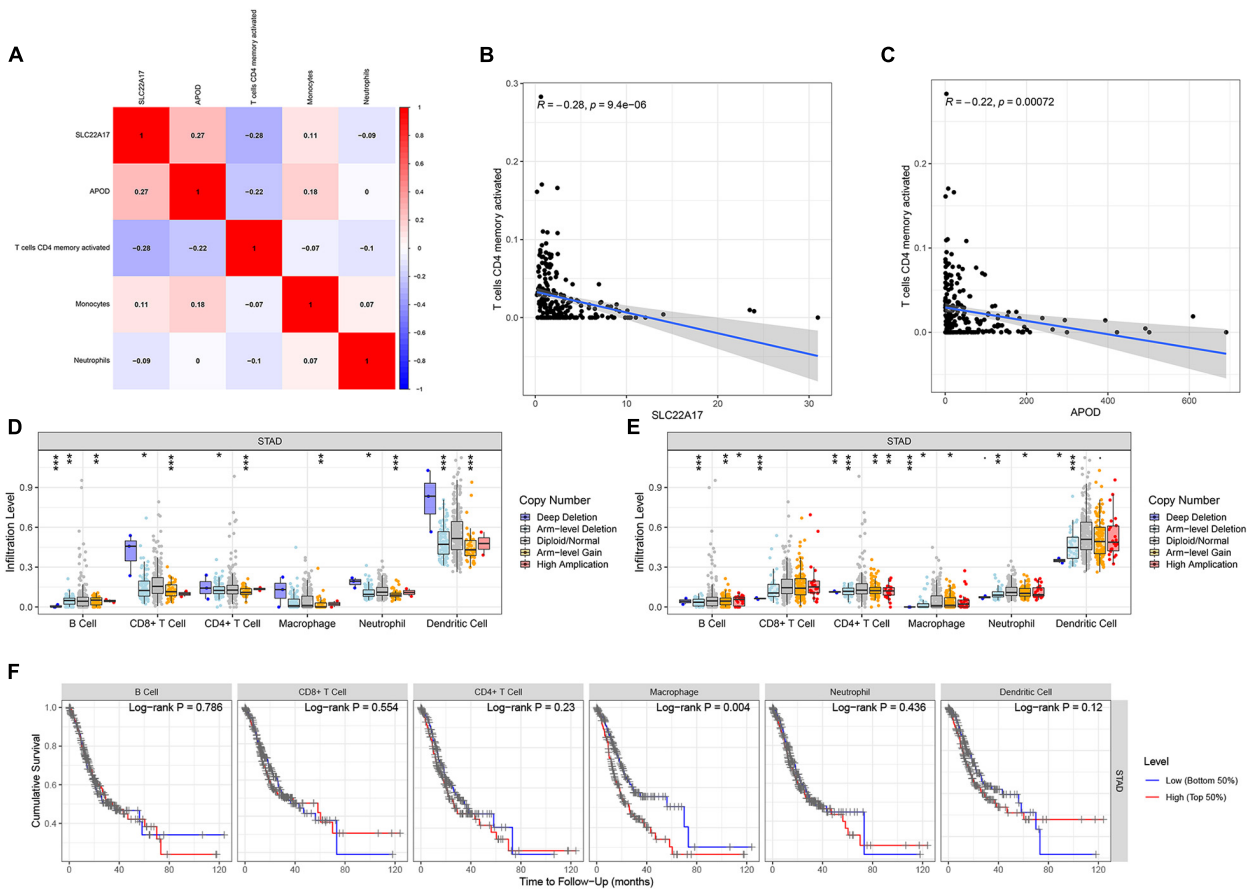



**FIGURE 7 |** Relation of TMB and prognostic model genes with immune cell infiltration. **(A)** Correlation between prognostic related immune infiltrating cells and prognostic model constructed by SLC22A17 and APOD. **(B,C)** T cells CD4 memory activated was negative associated with the expression of **(B)** SLC22A17 and **(C)** APOD. **(D,E)** Immune infiltration level among gastric cancer patients with diverse degree of copy number variation. **(F)** Survival probability of patients with low and high immune infiltration level of six immune cells. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

easily able to identify and eliminate those tumor cells with high TMB (Chen and Mellman, 2017; Chan et al., 2019). Further research on the association of TMB and immunity will be helpful to identify the critical biomarkers and pathways of AGC.

To explore the association of TMB with AGC, we analyzed somatic mutations in AGC patient samples. A total of 222 (88.45%) patients were identified to have somatic mutations. We ranked the top 30 most common mutations in these patients. The TTN, TP53, and MUC16 genes had the highest mutation frequencies. TTN mutation has been reported to be correlated with prognosis in lung cancer and gastric cancer (Cheng et al., 2019; Yang et al., 2020). MUC16 has also been reported to be associated with prognosis and immunotherapy efficiency in gastric cancer (Yang et al., 2020). TP53 mutation is common and affects treatment strategies in various cancers (Jiao et al., 2018; Kaur et al., 2018; Barbosa et al., 2019; Ahn et al., 2020). The mutant genes are enriched in key pathways involved in cancer progression. The WNT, NOTCH, and RTK-RAS signaling pathways are often dysregulated and can be employed as therapeutic targets in diverse cancers (Nusse and Clevers, 2017; Imperial et al., 2019; Krishna et al., 2019). According to the degree of TMB, we divided patients into a high-TMB group and a low-TMB group. Patients in the high-TMB group had better survival outcomes, which is consistent with the results in other cancers (Devarakonda et al., 2018). Patients aged over 65 have higher TMB. We attributed this to the weak ability of patients aged over 65 to eliminate mutations. The DEGs related to TMB were identified according to the degree of TMB. The results showed that these genes were mainly enriched in neuroactive ligand-receptor interactions, the cAMP signaling pathway and the calcium signaling pathway.

Differentially expressed genes related to TMB were intersected with 1881 immune-related genes. Then, we constructed a prognostic model with two prognostic genes, SLC22A17 and APOD. Based on the prognostic model, TCGA and GEO datasets were used to test the efficiency of the model. As expected, patients in the two low-risk cohorts had better survival outcomes. These results indicated that the prognostic model of differentially expressed TMB-related genes combined with immune-related genes functions well in gastric cancer. In addition, a nomogram was employed to predict the survival rate in gastric cancer. Then, we determined the prognostic function of SLC22A17 and APOD. The relationship between the expression levels of the two genes and patient clinical characteristics was visualized using a heat

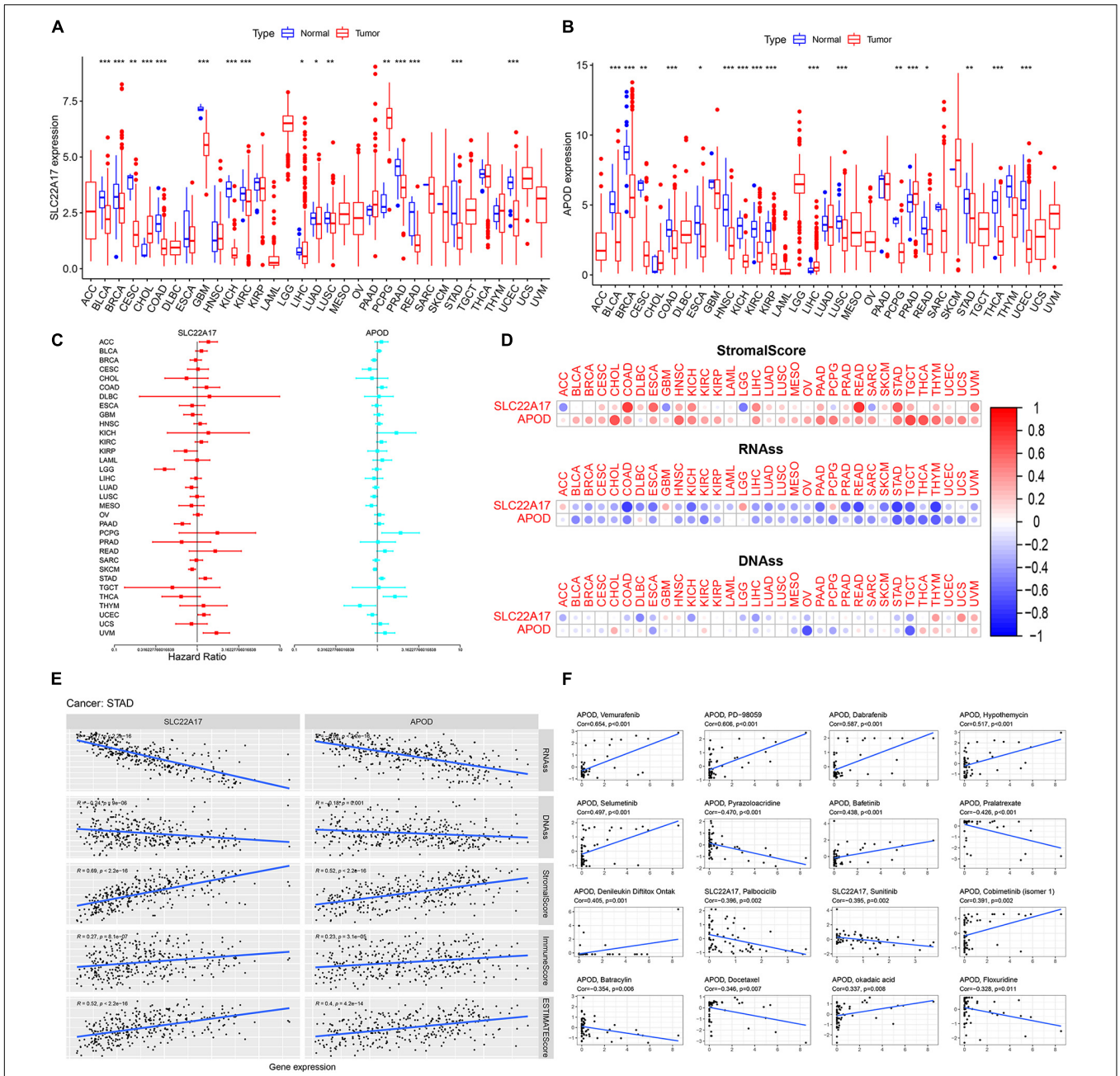



**FIGURE 8 |** SLC22A17 and APOD are dysregulated in multi-types cancer cells and related to cancer stemness and drug resistance. **(A,B)** Expression of **(A)** SLC22A17 and **(B)** APOD in multi-types cancer cells. **(C)** The prognostic value of SLC22A17 and APOD in the 33 cancers was identified by using univariate cox regression analysis. **(D,E)** SLC22A17 and APOD are associated with cancer stemness in various cancer types, including gastric cancer. **(F)** The correlation between SLC22A17, APOD, and tumor drug resistance. The abscissa and ordinate represent drug sensitivity score and gene expression, respectively. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

map. These two genes can be considered prognostic biomarkers in gastric cancer. APOD was reported to be the prognostic factor of gastric. Patients with high expression of APOD might have a shorter OS time. Two authors have also reported that SLC22A17 could be a prognosis biomarker of gastric cancer. Specifically, SLC22A17 was identified as a prognosis gene which may affect immune cell infiltration and iron metabolism in gastric (Hu C. et al., 2018; Wang et al., 2020; Wei et al., 2020). Although these two genes have been reported to be involved in gastric cancer, the specific mechanism of their regulation of gastric cancer is still unclear, which needs further research. In addition, whether these two genes possess prognosis function across different types of cancers remains unclear. Hence, we detected the expression of SLC22A17 and APOD in 33 types of cancers and determined

the association of the two genes with cancer stemness-related indicators (Zhang X. et al., 2020). SLC22A17 and APOD were found to be dysregulated in diverse cancers. In almost all cancers, SLC22A17 and APOD have positive relationships with the StromalScore, ImmuneScore and ESTIMATEScore. In contrast, the SLC22A17 and APOD genes were negatively associated with RNAss and DNAss in most cancers. Regarding drug resistance, we observed that APOD exerted a greater role in drug sensitivity. APOD has a strong positive relationship with resistance to many drugs. All these results indicated that these two genes have the same expression pattern and exhibit a similar correlation with StromalScore, RNAss, and DNAss in nearly all cancers. However, the predictive performance of these genes for other specific cancers requires more research.

Tumor mutation burden affects the degree of immune infiltration and efficacy of immune therapy in several cancers (Wu et al., 2019; Kang et al., 2020; Zhang L. et al., 2020). To explore the underlying association in gastric cancer, we analyzed the distribution of 22 infiltrating immune cells in tumor samples. The results showed that the proportions of infiltrating immune cells varied between the high-TMB group and the low-TMB group. Some kinds of infiltrating immune cells increased in tumor samples with high TMB. However, numerous infiltrating immune cells were decreased in tumor samples with low TMB. More research is needed to determine whether the infiltration of each type of immune cell is caused by TMB. To further clarify the association of TMB and immune infiltration in AGC, we analyzed the immune infiltration level in samples with diverse TMBs and found that the infiltration level was broadly decreased in patients with higher copy number variation compared with the diploid/normal group, which is consistent with other studies (Hu H. et al., 2018; Chan et al., 2019). Interestingly, we observed that patients with low infiltration had better survival outcomes. We speculate that this may be related to the poor prognosis of patients with AGC; the stage of patients diagnosed with AGC and the available therapeutic strategies may also account for this difference. More experiments are needed to clarify the association between TMB and immune infiltration.

## CONCLUSION

Our results indicate that immune-related genes generated from TMB-related differential expression analysis are involved in the progression of AGC. A prognostic model constructed with SLC22A17 and APOD might have vital roles across multiple types of cancers. Detection of TMB combined with immune infiltrating cells in AGC patients could be an effective method in guiding cancer therapy strategies, especially immunotherapy.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

## AUTHOR CONTRIBUTIONS

XG, YW, and XL designed the study. XG, XL, CQ, and HZ collected and analyzed the data. XL, XG, and AC wrote and revised the manuscript. ZW was responsible for supervising the study. All authors read and gave final approval of the manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China (81974385).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.642608/full#supplementary-material

## REFERENCES

Ahn, I., Tian, X., and Wiestner, A. (2020). TP53Ibrutinib for chronic lymphocytic leukemia with alterations. *N. Engl. J. Med.* 383, 498–500. doi: 10.1056/NEJMc2005943

Ajani, J., Lee, J., Sano, T., Janjigian, Y., Fan, D., and Song, S. (2017). Gastric adenocarcinoma. *Nat. Rev. Dis. Prim.* 3:17036. doi: 10.1038/nrdp.2017.36

Alexandrov, L., Nik-Zainal, S., Wedge, D., Aparicio, S., Behjati, S., Biankin, A., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

Al-Mahrouqi, H., Parkin, L., and Sharples, K. (2011). Incidence of stomach cancer in oman and the other gulf cooperation council countries. *Oman Med. J.* 26, 258–262. doi: 10.5001/omj.2011.62

Barbosa, K., Li, S., Adams, P., and Deshpande, A. (2019). The role of TP53 in acute myeloid leukemia: challenges and opportunities. *Genes Chromos. Cancer* 58, 875–888. doi: 10.1002/gcc.22796

Bi, F., Chen, Y., and Yang, Q. (2020). Significance of tumor mutation burden combined with immune infiltrates in the progression and prognosis of ovarian cancer. *Cancer Cell Int.* 20:373. doi: 10.1186/s12935-020-01472-9

Cascinu, S. (2020). Lenvatinib and pembrolizumab in advanced gastric cancer. *Lancet Oncol.* 21, 1004–1005. doi: 10.1016/s1470-2045(20)30336-3

Chan, T. A., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S. A., Stenzinger, A., et al. (2019). Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* 30, 44–56. doi: 10.1093/annonc/mdy495

Chen, L.-L. (2016). The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* 17, 205–211. doi: 10.1038/nrm.2015.32

Chen, D., and Mellman, I. (2017). Elements of cancer immunity and the cancer-immune set point. *Nature* 541, 321–330. doi: 10.1038/nature21349

Cheng, X., Yin, H., Fu, J., Chen, C., An, J., Guan, J., et al. (2019). Aggregate analysis based on TCGA: TTN missense mutation correlates with favorable prognosis in lung squamous cell carcinoma. *J. Cancer Res. Clin. Oncol.* 145, 1027–1035. doi: 10.1007/s00432-019-02861-y

Devarakonda, S., Rotolo, F., Tsao, M., Lanc, I., Brambilla, E., Masood, A., et al. (2018). Tumor mutation burden as a biomarker in resected non-small-cell lung cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 36, 2995–3006. doi: 10.1200/jco.2018.78.1963

Guo, X., Wang, Y., Zhang, H., Qin, C., Cheng, A., Liu, J., et al. (2020). Identification of the prognostic value of immune-related genes in esophageal cancer. *Front. Genet.* 11:989. doi: 10.3389/fgene.2020.00989

Hu, C., Zhou, Y., Liu, C., and Kang, Y. (2018). A novel scoring system for gastric cancer risk assessment based on the expression of three CLIP4 DNA methylation-associated genes. *Int. J. Oncol.* 53, 633–643. doi: 10.3892/ijo.2018.4433

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.e1618. doi: 10.1016/j.cell.2018.09.038

Ikediobi, O., Davies, H., Bignell, G., Edkins, S., Stevens, C., O'Meara, S., et al. (2006). Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Mol. Cancer Therap.* 5, 2606–2612. doi: 10.1158/1535-7163.mct-06-0433

Imperial, R., Toor, O., Hussain, A., Subramanian, J., and Masood, A. (2019). Comprehensive pancancer genomic analysis reveals (RTK)-RAS-RAF-MEK as a key dysregulated pathway in cancer: Its clinical implications. *Semin. Cancer Biol.* 54, 14–28. doi: 10.1016/j.semcancer.2017.11.016

Jiang, T., Shi, J., Dong, Z., Hou, L., Zhao, C., Li, X., et al. (2019). Genomic landscape and its correlations with tumor mutational burden, PD-L1 expression, and immune cells infiltration in Chinese lung squamous cell carcinoma. *J. Hematol. Oncol.* 12:75. doi: 10.1186/s13045-019-0762-1

Jiao, X., Qin, B., You, P., Cai, J., and Zang, Y. (2018). The prognostic value of TP53 and its correlation with EGFR mutation in advanced non-small cell lung cancer, an analysis based on cBioPortal data base. *Lung cancer (Amster. Nether.)* 123, 70–75. doi: 10.1016/j.lungcan.2018.07.003

Kang, K., Xie, F., Mao, J., Bai, Y., and Wang, X. (2020). Significance of tumor mutation burden in immune infiltration and prognosis in cutaneous melanoma. *Front. Oncol.* 10:573141. doi: 10.3389/fonc.2020.573141

Kaur, R., Vasudeva, K., Kumar, R., and Munshi, A. (2018). Role of p53 gene in breast cancer: focus on mutation spectrum and therapeutic strategies. *Curr. Pharmaceut. Design* 24, 3566–3575. doi: 10.2174/1381612824666180926095709

Kawazoe, A., Shitara, K., Boku, N., Yoshikawa, T., and Terashima, M. (2020). Current status of immunotherapy for advanced gastric cancer. *Jpn. J. Clin. Oncol.* 51, 20–27. doi: 10.1093/jjco/hyaa202

Kim, H. J., and Oh, S. C. (2018). Novel systemic therapies for advanced gastric cancer. *J. Gastric. Cancer* 18, 1–19. doi: 10.5230/jgc.2018.18.e3

Krishna, B., Jana, S., Singhal, J., Horne, D., Awasthi, S., Salgia, R., et al. (2019). Notch signaling in breast cancer: From pathway analysis to therapy. *Cancer Lett.* 461, 123–131. doi: 10.1016/j.canlet.2019.07.012

Le, D., Uram, J., Wang, H., Bartlett, B., Kemberling, H., Eyring, A., et al. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* 372, 2509–2520. doi: 10.1056/NEJMoa1500596

Mayakonda, A., Lin, D., Assenov, Y., Plass, C., and Koeffler, H. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28, 1747–1756. doi: 10.1101/gr.239244.118

Morrison, C., Pabla, S., Conroy, J., Nesline, M., Glenn, S., Dressman, D., et al. (2018). Predicting response to checkpoint inhibitors in melanoma beyond PD-L1 and mutational burden. *J. Immunother. Cancer* 6:32. doi: 10.1186/s40425-018-0344-8

Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., et al. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science (New York N. Y.)* 253, 665–669. doi: 10.1126/science.1651563

Nusse, R., and Clevers, H. (2017). Wnt/β-catenin signaling, disease, and emerging therapeutic modalities. *Cell* 169, 985–999. doi: 10.1016/j.cell.2017.05.016

Robles, A., Traverso, G., Zhang, M., Roberts, N., Khan, M., Joseph, C., et al. (2016). Whole-exome sequencing analyses of inflammatory bowel disease-associated colorectal cancers. *Gastroenterology* 150, 931–943. doi: 10.1053/j.gastro.2015.12.036

Song, Z., Wu, Y., Yang, J., Yang, D., and Fang, X. (2017). Progress in the treatment of advanced gastric cancer. *Tumour. Biol.* 39:1010428317714626. doi: 10.1177/1010428317714626

Tian, Y., Xu, J., Chu, Q., Duan, J., Zhang, J., Bai, H., et al. (2020). A novel tumor mutational burden estimation model as a predictive and prognostic biomarker in NSCLC patients. *BMC Med.* 18:232. doi: 10.1186/s12916-020-01694-8

Wang, M., Li, Z., Peng, Y., Fang, J., Fang, T., Wu, J., et al. (2020). Identification of immune cells and mRNA associated with prognosis of gastric cancer. *BMC Cancer* 20:206. doi: 10.1186/s12885-020-6702-1

Wei, J., Gao, X., Qin, Y., Liu, T., and Kang, Y. (2020). An iron metabolism-related SLC22A17 for the prognostic value of gastric cancer. *Onco. Targets Ther.* 13, 12763–12775. doi: 10.2147/ott.S287811

Wu, Y., Xu, J., Du, C., Wu, Y., Xia, D., Lv, W., et al. (2019). The predictive value of tumor mutation burden on efficacy of immune checkpoint inhibitors in cancers: a systematic review and meta-analysis. *Front. Oncol.* 9:1161. doi: 10.3389/fonc.2019.01161

Yang, D., Khan, S., Sun, Y., Hess, K., Shmulevich, I., Sood, A., et al. (2011). Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA* 306, 1557–1565. doi: 10.1001/jama.2011.1456

Yang, Y., Zhang, J., Chen, Y., Xu, R., Zhao, Q., and Guo, W. (2020). MUC4, MUC16, and TTN genes mutation correlated with prognosis, and predicted tumor mutation burden and immunotherapy efficacy in gastric cancer and pan-cancer. *Clin. Transl. Med.* 10:e155. doi: 10.1002/ctm2.155

Yarchoan, M., Hopkins, A., and Jaffee, E. (2017). Tumor mutational burden and response rate to PD-1 inhibition. *N. Engl. J. Med.* 377, 2500–2501. doi: 10.1056/NEJMc1713444

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

Zhang, L., Li, B., Peng, Y., Wu, F., Li, Q., Lin, Z., et al. (2020). The prognostic value of TMB and the relationship between TMB and immune infiltration in head and neck squamous cell carcinoma: a gene expression-based study. *Oral oncology* 110:104943. doi: 10.1016/j.oraloncology.2020.104943

Zhang, X., Klamer, B., Li, J., Fernandez, S., and Li, L. (2020). A pan-cancer study of class-3 semaphorins as therapeutic targets in cancer. *BMC Med. Genom.* 13:45. doi: 10.1186/s12920-020-0682-5

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 July 2021
doi: 10.3389/fgene.2021.707299

# Esophageal Cancer Associated Immune Genes as Biomarkers for Predicting Outcome in Upper Gastrointestinal Tumors

*Chuanhui Zhu[1,2†], Qianqian Xia[3†], Bin Gu[4], Mengjing Cui[3], Xing Zhang[3], Wenjing Yan[3], Dan Meng[3], Siyuan Shen[3], Shuqian Xie[3], Xueliang Li[1*‡], Hua Jin[5*‡] and Shizhi Wang[3*‡]*

[1] Department of Gastroenterology, The First Affiliated Hospital, Nanjing Medical University, Nanjing, China, [2] Department of Gastroenterology, Nanjing BenQ Medical Center, The Affiliated BenQ Hospital, Nanjing Medical University, Nanjing, China, [3] Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing, China, [4] Department of Neurosurgery, Zhongda Hospital, School of Medicine, Southeast University, Nanjing, China, [5] Clinical Laboratory, Affiliated Tumor Hospital of Nantong University (Nantong Tumor Hospital), Nantong, China

OPEN ACCESS

**Edited by:**
Rosalba Giugno,
University of Verona, Italy

**Reviewed by:**
Cheng Liang,
Shandong Normal University, China
Haoyun Lei,
Carnegie Mellon University,
United States

***Correspondence:**
Xueliang Li
ligakur@aliyun.com
Hua Jin
ntmgjh@163.com
Shizhi Wang
shizhiwang2009@seu.edu.cn

[†] These authors have contributed equally to this work

[‡] These authors have supervised this work

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

**Received:** 09 May 2021
**Accepted:** 28 June 2021
**Published:** 19 July 2021

**Citation:**
Zhu C, Xia Q, Gu B, Cui M, Zhang X, Yan W, Meng D, Shen S, Xie S, Li X, Jin H and Wang S (2021) Esophageal Cancer Associated Immune Genes as Biomarkers for Predicting Outcome in Upper Gastrointestinal Tumors. Front. Genet. 12:707299. doi: 10.3389/fgene.2021.707299

Esophageal cancer (EC) is the seventh most common tumor in the world, ranking the sixth leading cause of cancer death, with a 5-year survival rate of 15-25%. Therefore, reliable prognostic biomarkers are needed to effectively predict the prognosis of EC. In this study, the gene profile information of the EC cohort served as a training set, which was derived from TCGA and Immport databases. GO and KEGG enrichment analysis was performed on the differential genes in normal and tumor groups of EC. The immune genes in differentially expressed genes (DEGs) were further obtained for univariate and multivariate Cox and Lasso regression analysis, and 6 independent immune genes (*S100A3*, *STC2*, *HSPA6*, *CCL25*, *GPER1*, and *OSM*) associated with prognosis were obtained to establish an immune risk score signature (IRSS). The signature was validated using head and neck cancers (HNSC) and gastric cancer (GC)in upper gastrointestinal malignancies as validation sets. The Kaplan-Meier results showed that the prognosis of the high-risk group was significantly favorable than that of the low-risk group in both the training set ($P < 0.001$; HR = 3.68, 95% CI = 2.14−6.35) and the validation set ($P = 0.010$; HR = 1.43, 95% CI = 1.09−1.88). A nomogram combining multiple clinical information and IRSS was more effective than a single independent prognostic factor in predicting outcome. This study explored the potential link between immunity and EC, and established and validated prognostic biomarkers that can effectively predict the prognosis of EC, HNSC and GC based on six immune genes.

**Keywords: esophageal cancer, prognostic biomarker, head and neck cancers, gastric cancer, the upper gastrointestinal tumors**

## INTRODUCTION

Esophageal cancer (EC) is the 7th most common tumor in the world (Global Burden of Disease Cancer Collaboration et al., 2018), ranking the 6th leading cause of cancer death, which seriously threatens human health (Bray et al., 2018). According to the data, it is estimated that 456,000 new cases of EC were reported worldwide in 2012, half of which were in China (Zhu et al., 2016).

EC mainly includes two histological subtypes, esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EA), accounting for more than 95% of esophageal malignancies, of which ESCC is more common (Enzinger and Mayer, 2003). Smoking, alcohol consumption, chronic gastroesophageal reflux disease, obesity are critical risk factors for the occurrence of the disease (Huang and Yu, 2018). Unfortunately, most patients are already at an advanced stage at diagnosis, therefore the curative ratio is low and the prognosis is poor (Ferlay et al., 2015). In recent years, despite the application of new diagnostic and therapeutic techniques that have improved the survival rate of EC patients (Vendrely et al., 2018), the 5-year overall survival (OS) rate is still unsatisfactory, fluctuating between 15 and 25% (Short et al., 2017). Therefore, it is urgent to find robust biomarkers to predict the prognosis of EC patients and provide potential therapeutic targets.

Inflammation has been well known to be a complex biological response in which the human immune system attempts to eliminate the stimulus of inflammation and initiate repair and regeneration (Wallach et al., 2014; Karin and Clevers, 2016). Inflammatory response plays a pivotal role in tumorigenesis, development and metastasis (Taniguchi and Karin, 2018). For instance, the expression of immune-related genes such as interleukin (IL)-6 members, including IL-11, IL-27, IL-31, leukemia inhibitory factor, and oncostatin M (*OSM*), affect tumor cell proliferation, survival, inflammation, and metabolism (Taniguchi and Karin, 2014). The occurrence of EC is closely correlated to inflammation. It is well known that EA is inflammation-related cancer (O'Sullivan et al., 2014). Chronic inflammation has also been proved to be a crucial factor in the development of ESCC. On the one hand, oxidative and genotoxic stresses caused by smoking, drinking and carcinogens trigger inflammation, on the other hand, oral microbiota disorders, human papillomavirus (HPV) infection, and improper diet can also cause inflammation. EC cells can inhibit the body's anti-tumor immunity through inflammation-related mechanisms such as immune checkpoints, secretory factors and negatively regulated immune cells (Diakowska and Krzystek-Korpacka, 2020).

Since immune inflammation is a vital process in triggering tumorigenesis, identifying whether immunity affects the prognosis of patients remains an active area of research. Several studies have reported that tumor prognosis-related models have been established to predict patient survival (Huang et al., 2019; Shen et al., 2019; Qu et al., 2020). However, there are few studies on the establishment of prognostic models for EC, let alone immune-related ones. In the present study, we used the Cancer Genome Atlas (TCGA) database to explore the correlation between immune mechanisms and the occurrence of EC and established a novel risk score signature based on immune genes to effectively predict the outcome of EC patients as well as provide a potential clinical combination therapy. Taken together, our findings highlight the functional role of immune-related signatures and reveal potential prognostic biomarkers for ECs to predict the prognosis of upper gastrointestinal tumors.

# MATERIALS AND METHODS

## Data Collection and Processing

The datasets of esophageal cancer (TCGA-ESCA) and head and neck cancer (TCGA-HNSC), including their gene expression profiles, clinic information and survival information, were downloaded from the UCSC database[1]. EC samples with prognostic information were collected as a training set, consisting of 162 tumor samples and 11 normal samples. And a total of 500 patients with HNSC containing prognostic information were collected as a validation set. Patients with an OS of fewer than 60 days were removed because their cause of death may not be attributable to tumors.

From the Gene List module of the Immunology Database and Analysis Portal (ImmPort) database[2], we downloaded complete gene names directly, totaling 2483 immune-related genes (**Supplementary Table 1**).

## Differential Expression Analysis

Based on the expression of genes in EC, we first performed a differential expression analysis to identify genes differentially expressed in normal and tumor groups. Briefly, differentially expressed genes (DEGs) were obtained using the "limma" software package in R. Among them, $\text{Log}_2|\text{FC}| > 1$ and false discovery rate (FDR) $< 0.25$ were the criteria. "ggplot2," "Cairo," and "ggrepel" packages in the R were used to plot volcanoes to visualize the DEGs.

## Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Analysis

To identify potential biological processes and enrichment pathways of DEGs, GO, and KEGG was performed using the cluster Profilter R package. KEGG is a type of gene annotation, a database that integrates genomic, chemical and systematic functional information. Go database mainly describes gene characteristics in different dimensions and levels, involving cell composition, biological process and molecular function. The adjusted *P*-value less than 0.05 was considered statistically significant.

## Establishment of Immune Risk Scoring Signature (IRSS) for Prognosis

A total of 1734 immune genes were expressed in EC and intersected with DEGs to obtain differentially expressed immune genes (DEIGs). Subsequently, DEIGs were used in univariate Cox regression analysis to identify significant prognosis-related immune genes, followed by Least absolute shrinkage and selection operator (LASSO) regression analysis to obtain independent prognostic genes. LASSO regression can improve the accuracy and interpretability of the model and also exclude the problem of collinearity between independent variables (Alhamzawi and Ali, 2018). Multivariate Cox regression analysis

[1]https://tcga.xenahubs.net

[2]https://immport.niaid.nih.gov/

was conducted to obtain regression coefficients for independent prognostic factors. Finally, an immune risk score signature (IRSS) was established based on the multivariate Cox regression coefficient beta value, and the formula is as follows: an immune risk score signature (IRSS) = EXP*gene*1* β1 + EXP*gene*2 *β2 + EXP*gene*3*β3 + . . . + EXP*gene*n*βn, where EXP means expression level and β represents the regression coefficient from the multivariate Cox (Zeng et al., 2017).

By calculating the risk score for each sample of TCGA-ESCA, patients were divided into low- and high-risk groups using the median as the cut-off value. Furthermore, visualization of the Kaplan-Meier (KM) curve was utilized to compare OS between the two groups by the log-rank test. The area under the receiver operating characteristic (ROC) curve (AUC) was adopted for analyzing the prognostic predictive value of IRSS in patients with EC. The ROC curves are all referred to as the receiver operating characteristic curves, with sensitivity as the ordinate and 1-specificity as the abscissa (DeLong et al., 1988). The AUC is the probability value, which ranges from 0.5 to -1, used to evaluate the accuracy of the model prediction, and a larger area means higher accuracy. In the present study, the larger its value, the higher the degree to which the predicted overall survival agreed with the actual overall survival.

## Immune Risk Score Signature Combined With Clinicopathological Information

We screened for prognostic predictive factors, including clinical characteristics and established IRSS. Specifically, the univariate Cox proportional hazard model was employed to analyze the correlation between IRSS and OS, and the multivariate Cox regression analysis was used to evaluate whether the established IRSS could serve as an independent prognostic predictor. Further, to comprehensively assess patient survival, we constructed a nomogram integrating distinct clinicopathological information, including age, sex, disease type, stage, smoking, alcohol, BMI and IRSS, using the "rms" package. Additionally, the concordance index (C-index) was used to evaluate the predictive accuracy of the nomogram. Similarly, the decision curve analysis (DCA) of 2, 3, and 5 years was calculated to evaluate whether the synthetic nomogram established by us is suitable for clinical application. The x-axis represents the percentage of the threshold probability, and the y-axis represents the net income.

## Validation of IRSS

To assess the general applicability of the signature, Considering the anatomical and histological similarities, we selected the TCGA-HNSC ($n = 500$) to further validate the established model. The risk scores of each patient in the HNSC cohort were calculated and ranked using the formula of the IRSS established in TCGA-ESCA. HNSC samples that had been sorted by scores were divided into high- and low-risk groups according to the cut-off values obtained in the TCGA-ESCA cohort. KM curves were used for comparison of the survival differences between the two groups, and ROC curves were used to assess the accuracy of the signature prediction.

Similarly, the nomogram was used to comprehensively assess the survival probability of patients with HNSC, incorporating clinical information including age, gender, stage, smoking, alcohol, lymph nodes, and IRSS. Calibration curves (2-, 3-, and 5-year) were drawn to assess whether the predictive effect of the nomogram was accurate, and its 45° line represented the best predictive effect. In addition, the C-index was used to compare the accuracy of traditional TNM-stage, IRSS, and nomogram prediction. DCA was performed to evaluate the clinical value of the comprehensive nomogram for HNSC.

Further, to evaluate the prognostic value of the IRSS in gastric cancer (GC), which is an upper gastrointestinal tumor, we utilized the KM plotter online analysis website to validate the model[3]. This website contains multiple GEO databases of GC involving GSE62245, GSE14210, GSE15459, GSE22377, GSE29272, and GSE51105. We combined these databases to provide a prognostic assessment of overall survival based on genes in the IRSS in 631 patients with GC, respectively (Szasz et al., 2016).

## Statistical Analysis

Simple mathematical analysis and processing were completed by Excel software. Multivariate Cox regression analysis was performed by SPSS 20.0, with a probability of stepwise entry of 0.05 and removal of 0.1. Further data analysis and visualization are mainly accomplished by R (v3.6.1). Survival ROC curves were drawn by the "survival ROC" package in R. "Survival" packages were used to plot KM curves, C-index, as well as clinical univariate and multivariate regression analyses in R. Besides, visualization of DEGs, was accomplished by volcanoes drawn by the "ggplot2," "Cairo," and "ggrepel" packages. The *P*-value less than 0.05 was considered a statistically significant criterion.

# RESULTS

## Differential Analysis

A total of 156 EC patients with prognostic and gene expression data and survival longer than 60 days were included in the training set, as well as 11 matched normal samples. To investigate a biomarker that can effectively predict the prognosis of EC, we established a risk score model based on immune genes to evaluate the outcomes of patients with EC. Specifically, we performed a differential analysis between normal and tumor groups and obtained genes significantly associated with EC. And a total of 1479 DEGs were identified, as shown in **Supplementary Table 2**, and visualized with volcano maps (**Figure 1A**).

## GO and KEGG Analysis

To explore the potential association between gene expression and immunity in normal and tumor groups in the TCGA-ESCA cohort, we performed GO and KEGG enrichment pathway analysis. The DEGs in normal and tumor groups were enriched in a variety of processes, most of which were in immune-related pathways. Specifically, **Figure 2** shows the cytokine-cytokine receptor interaction and IL-17 signaling

[3]http://kmplot.com/analysis/

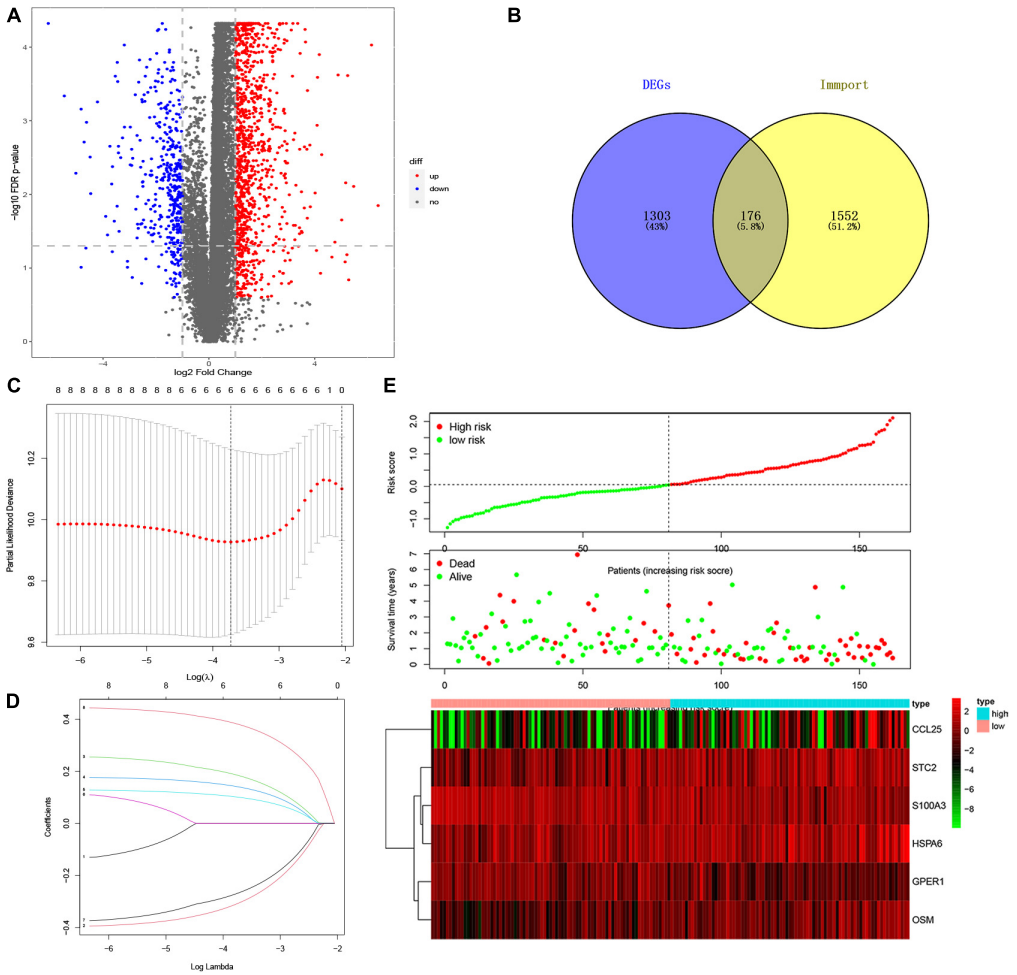



**FIGURE 1 |** Establishment of IRSS Signature. **(A)** Volcano plot of 1479 differentially expressed genes. **(B)** Venn diagram of the intersection of DEGs and immune genes. **(C)** Ten-time cross-validation for tuning parameter selection in the LASSO model. **(D)** LASSO coefficient profiles. **(E)** The risk score, survival status, and heat map of six immune genes in patients with EC.

pathway in KEGG enrichment analysis and the molecular functional modules involved in chemokine activity, cytokine activity, chemokine and receptor binding in GO enrichment analysis. Therefore, these findings revealed that the occurrence of EC was related to the expression level of immune genes. Detailed enrichment analysis results are presented in the **Supplementary Tables 3**, **4**.

## Construction and Prognostic Value of IRSS

To explore whether immune genes could be used as effective biomarkers to indicate the prognosis of EC, we selected immune-related genes from the DEGs for further analysis. The Venn diagram (**Figure 1B** and **Supplementary Table 5**) showed that 176 DEIGs were screened from the overlap of immune genes and DEGs. Subsequent univariate Cox regression analysis yielded 8 immune genes significantly associated with prognosis (**Supplementary Table 6**), followed by LASSO regression analysis. Combining the results of **Figures 1C,D**, it was considered that the model fit the best when the penalty coefficient was 6, and the corresponding six immune genes were selected into the model, which was *S100A3*, *STC2*, *HSPA6*, *CCL25*, *GPER1*, and *OSM* (**Figure 1E** and **Supplementary Table 7**). moreover, multivariate Cox regression analysis was performed on the six immune genes, which were still able to enter the equation as a prognostic predictor (**Supplementary Table 8**). Moreover, the corresponding regression coefficients were obtained, β*1*-β*6*, which were −0.400, 0.246, 0.177, 0.127, −0.349, and 0.442, respectively. According to the formula mentioned above, combined with the beta value of multivariate Cox regression, the IRSS was finally established:

$$\begin{aligned}\text{IRSS} = {} & \text{EXP}\,\mathit{S100A3}^{*} - 0.400 + \text{EXP}\,\mathit{STC2}^{*}\,0.246 \\ & + \text{EXP}\,\mathit{HSPA6}^{*}\,0.177 + \text{EXP}\,\mathit{CCL25}^{*}\,0.127 \\ & + \text{EXP}\,\mathit{GPER1}^{*} - 0.349 + \text{EXP}\,\mathit{OSM}^{*}\,0.442\end{aligned}$$

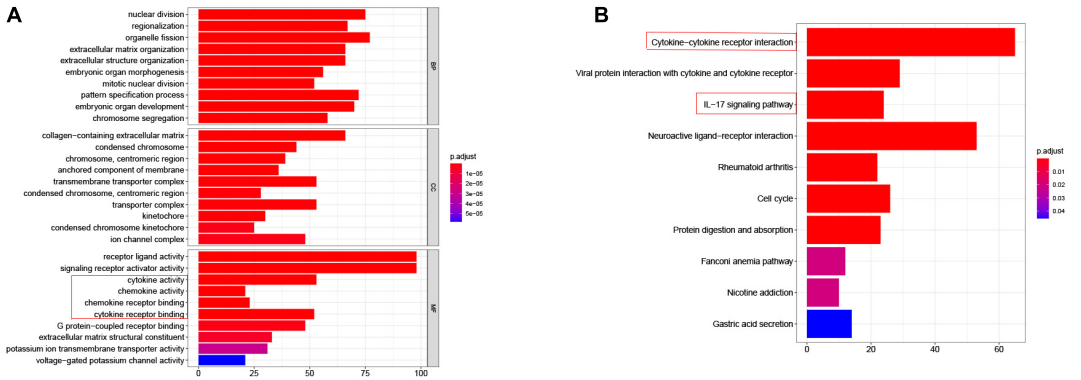



**FIGURE 2 |** Differential genes expressed in normal and tumor groups of esophageal cancer are involved in immune-related pathways. **(A)** GO, Gene Ontology, immune related pathways marked by red boxes; **(B)** KEGG, Kyoto Encyclopedia of Genes and Genomes, immune related pathways marked by red boxes.

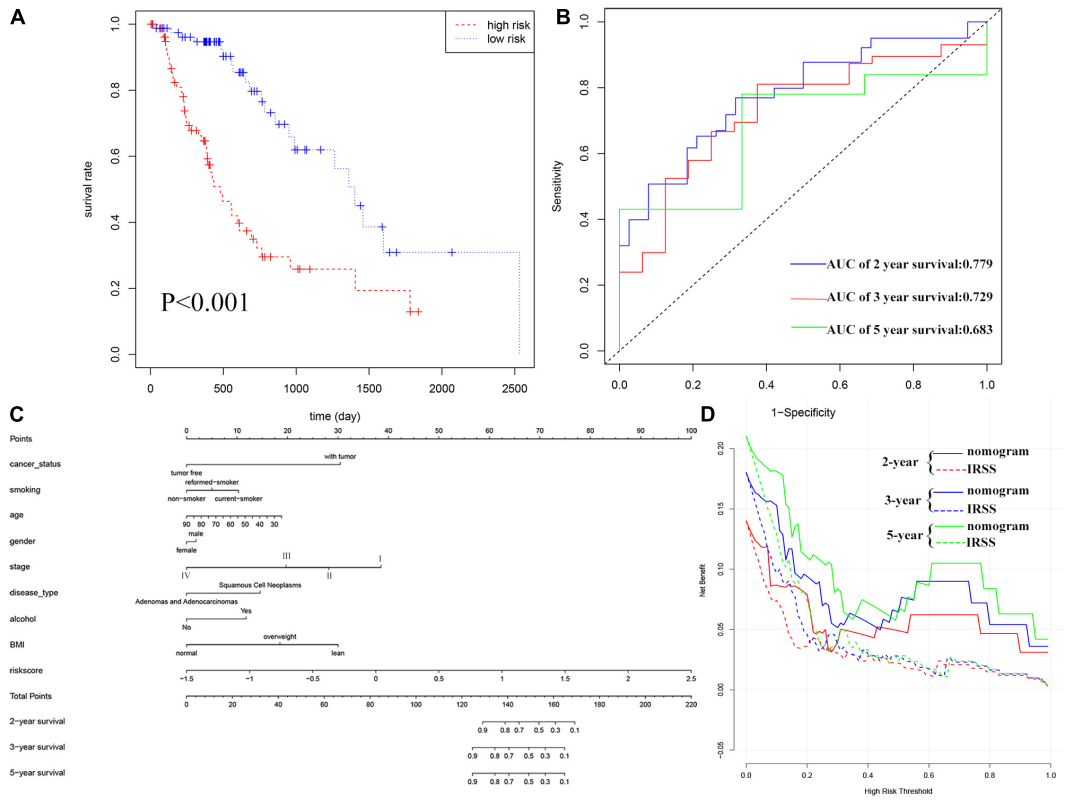



**FIGURE 3 |** Evaluation of IRSS signatures and establishment and evaluation of nomograms. **(A)** Kaplan-Meier curves show that OS was significantly different between the high- and low-risk groups in TCGA-ESCA. **(B)** The signature is shown by the time-dependent ROC curve for predicting 2, 3, and 5-year survival. **(C)** 2-, 3-, and 5-year nomogram for predicting OS of EC. There are nine components in this nomogram: cancer status, smoking, age, gender, stage, histological type, alcohol, BMI, and IRSS. **(D)** Decision curve analysis for the evaluation of the net benefits of IRSS and nomogram at 2, 3, and 5 years.

Furthermore, according to the above formula, the risk score of each EC patient was directly calculated. And then, the samples were divided into high- and low-risk groups, which were grouped according to the median and interquartile range [M(IRQ) = 0.040 (−0.321, 0.588)]. The results of the KM curve showed that the prognosis of the high-risk group was worse than that of the low-risk groups (**Figure 3A**, log-rank $P < 0.001$; HR = 3.68, 95% CI = 2.14−6.35). ROC curves were employed to assess the accuracy of established models for predicting OS in patients with EC. As shown in **Figure 3B**, the AUC values of 2, 3, and 5 years were 0.779, 0.729, and 0.683, respectively, indicating the robustness and accuracy of the model in predicting patient prognosis.

## The Value of IRSS in Clinical Characteristics

To determine whether the established IRSS has prognostic significance, we further performed univariate and multivariate Cox regression analysis. Univariate Cox regression analysis showed that risk score, cancer status and stage were prognostic predictors in TGGA-ECSA, but not smoking, alcohol, age, sex, disease type, BMI, and radiation therapy. More importantly, the risk score was also observed to be the only independent predictor in multivariate Cox regression analysis (**Table 1**). The above results show that our established IRSS could serve as a robust and novel biomarker for predicting prognosis.

Nomograms, which simplify statistical prediction models to single numerical estimates of event probabilities tailored to individual patient profiles, are widely used for prognostic assessment of tumors (Iasonos et al., 2008). A variety of clinical features have prognostic value in clinical practice. Therefore, in order to accurately evaluate the prognosis of patients, we established a nomogram containing multiple clinicopathological characteristics as well as IRSS. As shown in **Figure 3C**, scores for each variable could be calculated and combined to comprehensively predict the prognosis of patients with EC.

The C-index of the established nomogram, risk signature, and TNM-stage was 0.881, 0.721, and 0.693 (**Table 2**), respectively. In summary, the predictive ability of our IRSS was stronger than that of the traditional TNM-stage, however, the predictive accuracy of the nomogram integrating multiple clinical information was the most robust. Consistent with this result, the DCA figure (**Figure 3D**) also proved that the nomogram combined with various clinical features has better clinical application value.

**TABLE 2 |** The C-index values of the nomogram, TNM-stage, and IRSS.

| Cohorts | Variables | C-index (95%CI) |
|---|---|---|
| EC | TNM-stage | 0.693 (0.657,0.731) |
| | IRSS | 0.721(0.688,0.754) |
| | nomogram | 0.881 (0.822,0.940) |
| HNSC | TNM-stage | 0.512 (0.493,0.531) |
| | IRSS | 0.558 (0.534,0.582) |
| | nomogram | 0.781 (0.759,0.803) |

## Validation of Other Cancer Species

It is well known that HNSC and EC belong to malignant epithelial tumors of the upper gastrointestinal tract, which are characterized by early dissemination and poor prognosis (Sproll et al., 2018). To verify the general applicability of the IRSS, the data of the TCGA-HNSC cohort with similar tissue and

**TABLE 1 |** Univariate/multivariate Cox regression analysis of clinicopathological features of EC associated with OS.

| Variables | | Patient *N* (156) | Univariate analysis | | Multivariate analysis | |
|---|---|---|---|---|---|---|
| | | | HR[a] [95% CI] | *P* | HR [95% CI[b]] | *P* |
| Age | <65 | 93 | 1 | | | |
| | ≥65 | 63 | 0.896 [0.536,1.497] | 0.675 | | |
| BMI | Normal | 8 | 1 | | | |
| | Lean | 57 | 2.322 [0.894,6.035] | 0.084 | | |
| | Overweight | 82 | 1.355 [0.734,2.503] | 0.332 | | |
| Stage | Stage i | 15 | 1 | | 1 | |
| | Stage ii | 67 | 2.859 [0.661,12.360] | 0.16 | | |
| | Stage iii | 47 | 7.483 [1.699,32.966] | 0.008 | | |
| | Stage iv | 8 | 22.130 [4.493,109.001] | <0.0001* | 1.286 [0.438,3.778] | 0.647 |
| Cancer status | Tumor free | 60 | 1 | | 1 | |
| | With tumor | 35 | 3.673 [1.394,9.676] | 0.008* | 2.164 [1.182,3.964] | 0.516 |
| Histological type | EA | 79 | 1 | | | |
| | ESCC | 77 | 0.812 [0.482,1.366] | 0.433 | | |
| Gender | Male | 23 | 1 | | | |
| | Female | 133 | 2.236 [0.892,5.603] | 0.381 | | |
| Smoking | Non-smoker | 45 | 1 | | | |
| | Current-smoker | 32 | 1.677 [0.736,3.819] | 0.218 | | |
| | Reformed-smoker | 61 | 1.615 [0.785,3.319] | 0.193 | | |
| Alcohol | No | 44 | 1 | 0.181 | | |
| | Yes | 110 | 0.703 [0.419,1.178] | | | |
| Radiation therapy | No | 61 | 1 | | | |
| | Yes | 21 | 1.489 | 0.439 | | |
| IRSS | | 156 | 2.149 [1.666,2.772] | <0.0001* | 2.319[1.615,3.330] | 0.012* |

[a] *HR, hazard ratio.*
[b] *CI, confidence interval.*
**P < 0.05.*

anatomical characteristics was downloaded as the validation cohort. First, according to the IRSS formula obtained in the TCGA-ESCA cohort, the risk score was calculated for each patient in the HNSC cohort. Further, HNSC samples were divided into high and low-risk groups according to the median IRSS of the EC cohort.

The results of KM analysis (**Figure 4A**) could confirm that the low-risk group was associated with a better prognosis, while the high-risk group predicted a worse prognosis ($P$ = 0.010; HR = 1.43, 95% CI = 1.09−1.88), which was consistent with the results of EC. The AUC of survival ROC curve shows that the model had good consistency in predicting OS and actual OS (**Figure 4B**, 0.535, 0.561, and 0.613 at 2, 3, and 5 years, respectively). Clinical characteristics and IRSS were used to establish a predictive nomogram for predicting the prognostic survival probability of HNSC patients at 2, 3, and 5 years (**Figure 4C**). The calibration curve results confirm that there is good consistency between the actual survival probability and the predicted probability (**Figure 4D**). The results of the C-index and DCA showed that IRSS had a better prognostic predictive ability for HNSC than traditional TNM-stage, but the comprehensive nomogram was the best (**Figures 4E,F** and **Table 2**).

To investigate the prognostic predictive value of the model in GC with upper gastrointestinal tumors, the IRSS model was further validated in gastric cancer. Combining multiple GEO databases of gastric cancer, KM-plot results indicated that 6 immune genes in IRSS were highly associated with the prognosis of GC, and each independent gene could likewise serve as a biomarker for predicting the outcome of GC (**Supplementary Figure 1**).

Taken together, the established IRSS had good applicability and could not only predict the prognosis of EC but also serve as a prognostic predictive biomarker for the upper gastrointestinal tumors.

## DISCUSSION

Esophageal cancer remains one of the most lethal malignancies in the world with a poor prognosis (Ferlay et al., 2015). Over the past decades, the incidence of EC has increased markedly in many countries (Simard et al., 2012), ranking fourth among cancer deaths in China (Chen et al., 2016). Owing to the lack of early-onset symptoms, EC is usually diagnosed at an advanced stage. A variety of studies have found that the carcinogenic process of EC is closely correlated with the immune-inflammatory response (Lin et al., 2016). A major mechanism of inflammation-induced esophageal carcinogenesis is through structural activation of inflammatory signaling pathways (Abdel-Latif et al., 2009). EC cells are rich in tumor antigens, including tumor-associated antigens and neoantigens, and can initiate dendritic cell-mediated cytotoxic T lymphocytes early in tumorigenesis (Huang and Fu, 2019). Environmental exposure can trigger chronic esophageal inflammation, further promoting the activation of pro-inflammatory signaling pathways for survival and proliferation (Lin et al., 2016). The induction of these pathways leads to the activation of downstream gene transcription and enzyme activity, which play a key role in tumor growth and survival. Tumor immunotherapy is a promising new method for the treatment of EC, and different studies on EC immunotherapy have been carried out in recent years (Kelly, 2019). However, EC immunotherapy always results in mixed outcomes, partly because of the lack of reliable markers to predict treatment response (Huang and Fu, 2019). In the current study, we aim to establish immune-related biomarkers to effectively predict the outcome of EC.

To explore the relationship between EC and immune mechanisms, we selected the TCGA-ECSA database as a training set for analysis. To find the DEGs between the normal group and tumor group of EC to obtain gene annotation information, the differential analysis was carried out first. We then performed GO and KEGG enrichment analysis on the DEGs and the results showed that immune and tumor-related signaling pathways were significantly enriched. This is consistent with previous findings that immune inflammation induction is an important mechanism of esophageal carcinogenesis (Abdel-Latif et al., 2009). Therefore, we will further explore the potential role of immunological biomarkers in tumor prognosis.

Next, we select the immune genes among the DEGs and obtain 6 independent immune genes related to prognosis according to the Cox proportional hazard model and lasso regression analysis. These six immune genes were integrated to construct an IRRS that can effectively predict prognosis. Among these genes, *S100A3* belongs to the S100 family and is considered to be associated with a good prognosis of ovarian cancer (Bai et al., 2018), which is similar to our results. However, in gastric cancer, the high expression of *S100A3* is closely in relation to the poor survival of patients (Wang et al., 2019). *STC2* (stanniocalcin 2), whose expression in ESCA was higher than that in corresponding normal tissues, was significantly associated with lymph node metastasis, lymphatic invasion and distant metastasis (Kita et al., 2011; Kashyap et al., 2012), and has been reported as a prognostic glycolysis-related gene in HNSCC (Ferreira do Carmo et al., 2020; Liu and Yin, 2020). *HSPA6*, a heat shock protein, was considered to be associated with the recurrence of human hepatocellular carcinoma in the study of Yang et al. (2015). Zhang et al. (2016) have reported that *CCL25* (C-C chemokine receptor ligand 25) may promote the migration and invasion of cancer cells by affecting several Epithelial-mesenchymal transition (EMT) markers and providing the chemotactic ability for hepatocytes and breast cancer cells through the CCL25/CCR9 signaling pathway. *GPER1* (G-protein-coupled estrogen receptor 1) is recognized as a key regulator of immune-mediated events in breast, pancreatic, prostate and hepatocellular carcinomas, as well as melanoma (Notas et al., 2020). *OSM* has been reported to have diagnostic, prognostic, and therapeutic capabilities in a variety of diseases (Verstockt et al., 2019). For example, Tawara et al. (2018) argue that early therapeutic inhibition of *OSM* in breast cancer patients is thought to prevent breast cancer metastasis.

In this study, the results of KM analysis showed that IRSS was an effective biomarker for predicting the prognosis of EC. Significant differences in OS were observed between the high- and the low-risk group, implying that the high-risk group was associated with adverse outcomes. Furthermore,

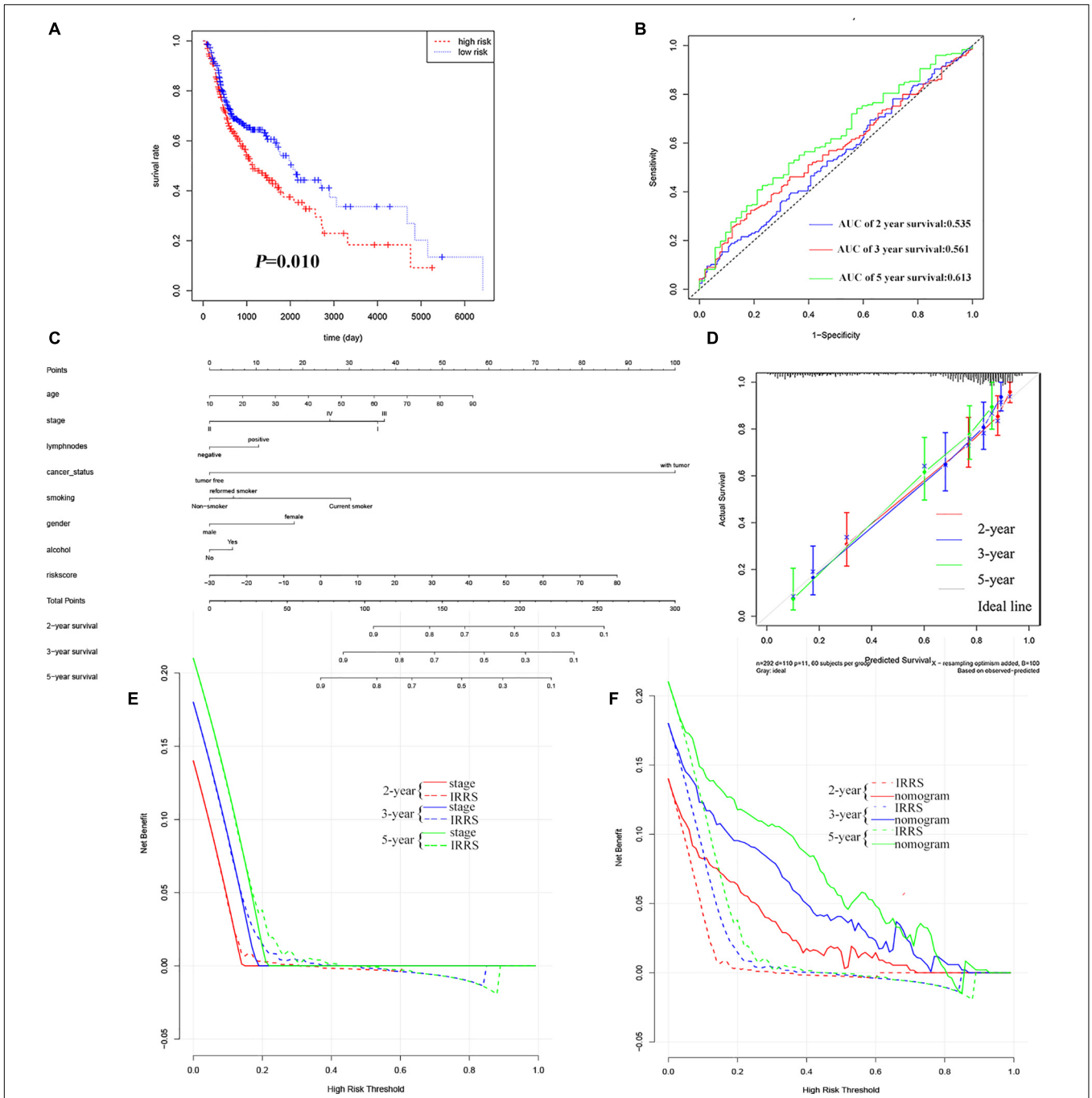



**FIGURE 4 |** Validation of IRSS signature with TCGA-HNSC. **(A)** Kaplan-Meier curves show that OS in the low-risk was significantly higher than in the high-risk group. **(B)** Time-dependent ROC curve analysis of the IRSS at 2, 3, and 5 years. **(C)** 2-, 3-, and 5-year nomogram for predicting OS of HNSC. **(D)** The Calibration curve of the nomogram for predicting OS rate at 2, 3, and 5 years. **(E,F)** Decision curve analysis for the evaluation of the net benefits of TNM-stage, IRSS and nomogram at 2, 3, and 5 years.

the survival ROC results showed that the predictive effect of our model on prognosis was in good agreement with the actual results. Additionally, survival analysis with multiple clinicopathological information, age, sex, tissue type, stage, smoking, alcohol consumption, BMI and radiation therapy as covariates, demonstrated that the established model remained a robust independent prognostic predictor. In order to evaluate the prognosis comprehensively, we combined a variety of clinical information and established a nomogram to score the survival probability of each patient. The results of the DCA and C-index showed that the prediction accuracy of IRSS was higher than that of the traditional TNM-stage, however, the nomogram

integrating multiple clinical information could predict the prognosis of EC patients more accurately.

Head and neck cancers, mainly including two histological subtypes of head and neck adenocarcinoma (HNA) and head and neck squamous cell carcinoma (HNSCC) (Andreasen et al., 2019). HNSCC is not only close to EC in histological classification and anatomical location but also has many similar carcinogenic factors. Chronic inflammation and microbial dysbiosis, including HPV infection (de Villiers et al., 2004), Porphyromonas gingivalis infection, and their synergistic effects with alcohol and tobacco (Olsen and Yilmaz, 2019), are closely associated with the occurrence of oral and digestive cancers, including (larynx, throat, lip, mouth, and salivary glands) and ESCA. Additionally, overexpression of the Dek oncogene in SCC (squamous cell carcinoma)-derived human keratinocytes can promote the development of ESCA and HNSC *in vivo* (Matrka et al., 2018). Considering the similarity of histological type, anatomical location and pathogenic factors, we utilized TCGA-HNSC as a validation cohort to evaluate the prognostic predictive value of the established model for these two tumors. Interestingly, the IRSS we constructed can not only be used as a prognostic biomarker for EC but also be used to predict the outcome of HNSC, which shows that the signature has wide robustness and applicability. Moreover, this may provide a new idea for the treatment of EC.

Currently, potential biomarkers for predicting prognosis have been widely used in EC and other cancers (Li et al., 2019; Lu et al., 2020). For instance, Qu et al. (2020). comprehensively analyzed the tumor microenvironment of cutaneous melanoma by using ESTIMATE and identified genes associated with the tumor microenvironment as biomarkers and their correlation with the immune system (Pan et al., 2019). As we prepared this paper, a study on the immune risk model of EC has been established and published (Guo et al., 2020). However, compared with this literature, our differential analysis screening criteria are more stringent. The number of prognosis-related immune genes obtained was different due to different screening criteria, but the overlapping two genes, *OSM* and *HSPA6*, confirmed the reliability of our established model. In addition, the clinicopathological factors considered in our nomogram, including smoking, alcohol consumption, disease type, BMI, and tumor status, enable a comprehensive assessment of the prognostic survival probability of patients with esophageal cancer. Besides, the dataset TCGA-HNSC was used as a validation set to confirm the applicability, robustness, and prognostic value of the model in upper gastrointestinal malignancies. Therefore, compared with the former, our study has further research progress and clinical significance.

In this study, the potential relationship between immunity and EC was explored. Based on six immune genes, a novel and robust biomarker for predicting the prognosis of EC and HNSC was established and validated. The signature proved to be an independent prognostic biomarker, which may provide a potential therapeutic target for the clinical treatment of upper gastrointestinal cancers such as EC, GC and HNSC, as well as ideas for the study of their correlation.

## 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/s.

## AUTHOR CONTRIBUTIONS

CZ and QX conceived the study, performed the data analysis, and wrote the manuscript. BG downloaded the gene expression data of esophageal cancer. XL, HJ, and SW critically revised the manuscript for research content and administrative support. All authors read and approved the final manuscript.

## FUNDING

This study was supported by the National Natural Science Foundation of China (81872684), the Fundamental Research Funds for the Central Universities, Southeast University "Zhongying Young Scholars" Project, the Six Talent Peaks Project in Jiangsu Province (wsw-201), the "SIX ONE" Talent Research Project for the High-level Health Personnel of Jiangsu Province (LGY2018037), the Fifth Scientific Research Project of Nantong ("226 Project"), the Research Project from the Nantong Commission of Health (MB2020018), and the Nanjing Science and Medical Development Foundation (YKK17251).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.707299/full#supplementary-material

## REFERENCES

Abdel-Latif, M. M., Duggan, S., Reynolds, J. V., and Kelleher, D. (2009). Inflammation and esophageal carcinogenesis. *Curr. Opin. Pharmacol.* 9, 396–404. doi: 10.1016/j.coph.2009.06.010

Alhamzawi, R., and Ali, H. T. M. (2018). The Bayesian adaptive lasso regression. *Math. Biosci.* 303, 75–82. doi: 10.1016/j.mbs.2018.06.004

Andreasen, S., Kiss, K., Mikkelsen, L. H., Channir, H. I., Plaschke, C. C., Melchior, L. C., et al. (2019). An update on head and neck cancer: new entities and their histopathology, molecular background, treatment, and outcome. *APMIS* 127, 240–264. doi: 10.1111/apm.12901

Bai, Y., Li, L. D., Li, J., and Lu, X. (2018). Prognostic values of S100 family members in ovarian cancer patients. *BMC Cancer* 18:1256. doi: 10.1186/s12885-018-5170-5173

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

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

de Villiers, E. M., Gunst, K., Stein, H., and Scherubl, H. (2004). Esophageal squamous cell cancer in patients with head and neck cancer: prevalence of human papillomavirus DNA sequences. *Int. J. Cancer* 109, 253–258. doi: 10.1002/ijc.11685

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845.

Diakowska, D., and Krzystek-Korpacka, M. (2020). Local and systemic interleukin-32 in esophageal, gastric, and colorectal cancers: clinical and diagnostic significance. *Diagnostics (Basel)* 10:785. doi: 10.3390/diagnostics10100785

Enzinger, P. C., and Mayer, R. J. (2003). Esophageal cancer. *N. Engl. J. Med.* 349, 2241–2252. doi: 10.1056/NEJMra035010

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi: 10.1002/ijc.29210

Ferreira do Carmo, A., Dourado, M. R., Ervolino, de Oliveira, C., Bastos, D. C., et al. (2020). Stanniocalcin 2 contributes to aggressiveness and is a prognostic marker for oral squamous cell carcinoma. *Exp. Cell Res.* 393:112092. doi: 10.1016/j.yexcr.2020.112092

Global Burden of Disease Cancer Collaboration, Fitzmaurice, C., Akinyemiju, T. F., Al, Lami FH, Alam, T., Alizadeh-Navaei, R., et al. (2018). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA Oncol.* 4, 1553–1568. doi: 10.1001/jamaoncol.2018.2706

Guo, X., Wang, Y., Zhang, H., Qin, C., Cheng, A., Liu, J., et al. (2020). Identification of the prognostic value of immune-related genes in esophageal Cancer. *Front. Genet.* 11:989. doi: 10.3389/fgene.2020.00989

Huang, F. L., and Yu, S. J. (2018). Esophageal cancer: risk factors, genetic association, and treatment. *Asian J. Surg.* 41, 210–215. doi: 10.1016/j.asjsur.2016.10.005

Huang, H., Liu, Q., Zhu, L., Zhang, Y., Lu, X., Wu, Y., et al. (2019). Prognostic value of preoperative systemic immune-inflammation index in patients with cervical Cancer. *Sci. Rep.* 9:3284. doi: 10.1038/s41598-019-39150-39150

Huang, T. X., and Fu, L. (2019). The immune landscape of esophageal cancer. *Cancer Commun. (Lond)* 39:79. doi: 10.1186/s40880-019-0427-z

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

Karin, M., and Clevers, H. (2016). Reparative inflammation takes charge of tissue regeneration. *Nature* 529, 307–315. doi: 10.1038/nature17039

Kashyap, M. K., Pawar, H. A., Keerthikumar, S., Sharma, J., Goel, R., Mahmood, R., et al. (2012). Evaluation of protein expression pattern of stanniocalcin 2, insulin-like growth factor-binding protein 7, inhibin beta a and four and a half LIM domains 1 in esophageal squamous cell carcinoma. *Cancer Biomark* 12, 1–9. doi: 10.3233/CBM-120289

Kelly, R. J. (2019). The emerging role of immunotherapy for esophageal cancer. *Curr. Opin. Gastroenterol.* 35, 337–343. doi: 10.1097/MOG.0000000000000542

Kita, Y., Mimori, K., Iwatsuki, M., Yokobori, T., Ieta, K., Tanaka, F., et al. (2011). STC2: a predictive marker for lymph node metastasis in esophageal squamous-cell carcinoma. *Ann. Surg. Oncol.* 18, 261–272. doi: 10.1245/s10434-010-1271-1271

Li, D., Zhang, L., Liu, Y., Sun, H., Onwuka, J. U., Zhao, Z., et al. (2019). Specific DNA methylation markers in the diagnosis and prognosis of esophageal cancer. *Aging (Albany NY)* 11, 11640–11658. doi: 10.18632/aging.102569

Lin, E. W., Karakasheva, T. A., Hicks, P. D., Bass, A. J., and Rustgi, A. K. (2016). The tumor microenvironment in esophageal cancer. *Oncogene* 35, 5337–5349. doi: 10.1038/onc.2016.34

Liu, Y., and Yin, S. (2020). A novel prognostic index based on the analysis of glycolysis-related genes in head and neck squamous cell carcinomas. *J. Oncol.* 2020:7353874. doi: 10.1155/2020/7353874

Lu, Z., Yan, W., Liang, J., Yu, M., Liu, J., Hao, J., et al. (2020). Nomogram based on systemic immune-inflammation index to predict survival of tongue cancer patients who underwent cervical dissection. *Front. Oncol.* 10:341. doi: 10.3389/fonc.2020.00341

Matrka, M. C., Cimperman, K. A., Haas, S. R., Guasch, G., Ehrman, L. A., Waclaw, R. R., et al. (2018). Dek overexpression in murine epithelia increases overt esophageal squamous cell carcinoma incidence. *PLoS Genet.* 14:e1007227. doi: 10.1371/journal.pgen.1007227

Notas, G., Kampa, M., and Castanas, E. (2020). G Protein-Coupled estrogen receptor in immune cells and its role in immune-related diseases. *Front. Endocrinol. (Lausanne)* 11:579420. doi: 10.3389/fendo.2020.579420

Olsen, I., and Yilmaz, O. (2019). Possible role of *Porphyromonas gingivalis* in orodigestive cancers. *J. Oral. Microbiol.* 11:1563410. doi: 10.1080/20002297.2018.1563410

O'Sullivan, K. E., Phelan, J. J., O'Hanlon, C., Lysaght, J., O'Sullivan, J. N., and Reynolds, J. V. (2014). The role of inflammation in cancer of the esophagus. *Expert Rev. Gastroenterol. Hepatol.* 8, 749–760. doi: 10.1586/17474124.2014.913478

Pan, X. B., Lu, Y., Huang, J. L., Long, Y., and Yao, D. S. (2019). Prognostic genes in the tumor microenvironment in cervical squamous cell carcinoma. *Aging (Albany NY)* 11, 10154–10166. doi: 10.18632/aging.102429

Qu, Y., Zhang, S., Zhang, Y., Feng, X., and Wang, F. (2020). Identification of immune-related genes with prognostic significance in the microenvironment of cutaneous melanoma. *Virchows Arch.* 478, 943–959. doi: 10.1007/s00428-020-02948-2949

Shen, S., Wang, G., Zhang, R., Zhao, Y., Yu, H., Wei, Y., et al. (2019). Development and validation of an immune gene-set based Prognostic signature in ovarian cancer. *EBioMedicine* 40, 318–326. doi: 10.1016/j.ebiom.2018.12.054

Short, M. W., Burgers, K. G., and Fry, V. T. (2017). Esophageal Cancer. *Am. Fam. Phys.* 95, 22–28.

Simard, E. P., Ward, E. M., Siegel, R., and Jemal, A. (2012). Cancers with increasing incidence trends in the United States: 1999 through 2008. *CA Cancer J. Clin.* 62, 118–128. doi: 10.3322/caac.20141

Sproll, C., Fluegen, G., and Stoecklein, N. H. (2018). Minimal residual disease in head and neck cancer and esophageal Cancer. *Adv. Exp. Med. Biol.* 1100, 55–82. doi: 10.1007/978-3-319-97746-1_4

Szasz, A. M., Lanczky, A., Nagy, A., Forster, S., Hark, K., Green, J. E., et al. (2016). Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* 7, 49322–49333. doi: 10.18632/oncotarget.10337

Taniguchi, K., and Karin, M. (2014). IL-6 and related cytokines as the critical lynchpins between inflammation and cancer. *Semin. Immunol.* 26, 54–74. doi: 10.1016/j.smim.2014.01.001

Taniguchi, K., and Karin, M. (2018). NF-kappaB, inflammation, immunity and cancer: coming of age. *Nat. Rev. Immunol.* 18, 309–324. doi: 10.1038/nri.2017.142

Tawara, K., Bolin, C., Koncinsky, J., Kadaba, S., Covert, H., Sutherland, C., et al. (2018). OSM potentiates preintravasation events, increases CTC counts, and promotes breast cancer metastasis to the lung. *Breast Cancer Res.* 20:53. doi: 10.1186/s13058-018-0971-975

Vendrely, V., Launay, V., Najah, H., Smith, D., Collet, D., and Gronnier, C. (2018). Prognostic factors in esophageal cancer treated with curative intent. *Dig. Liver Dis.* 50, 991–996. doi: 10.1016/j.dld.2018.08.002

Verstockt, S., Verstockt, B., and Vermeire, S. (2019). Oncostatin M as a new diagnostic, prognostic and therapeutic target in inflammatory bowel disease (IBD). *Expert Opin. Ther. Targets* 23, 943–954. doi: 10.1080/14728222.2019.1677608

Wallach, D., Kang, T. B., and Kovalenko, A. (2014). Concepts of tissue injury and cell death in inflammation: a historical perspective. *Nat. Rev. Immunol.* 14, 51–59. doi: 10.1038/nri3561

Wang, C., Luo, J., Rong, J., He, S., Zhang, L., and Zheng, F. (2019). Distinct prognostic roles of S100 mRNA expression in gastric cancer. *Pathol. Res. Pract.* 215, 127–136. doi: 10.1016/j.prp.2018.10.034

Yang, Z., Zhuang, L., Szatmary, P., Wen, L., Sun, H., Lu, Y., et al. (2015). Upregulation of heat shock proteins (HSPA12A, HSP90B1, HSPA4, HSPA5 and HSPA6) in tumour tissues is associated with poor outcomes from HBV-related early-stage hepatocellular carcinoma. *Int. J. Med. Sci.* 12, 256–263. doi: 10.7150/ijms.10735

Zeng, J. H., Liang, L., He, R. Q., Tang, R. X., Cai, X. Y., Chen, J. Q., et al. (2017). Comprehensive investigation of a novel differentially expressed lncRNA expression profile signature to assess the survival of patients with colorectal adenocarcinoma. *Oncotarget* 8, 16811–16828. doi: 10.18632/oncotarget.15161

Zhang, Z., Sun, T., Chen, Y., Gong, S., Sun, X., Zou, F., et al. (2016). CCL25/CCR9 signal promotes migration and invasion in hepatocellular and breast cancer cell lines. *DNA Cell Biol.* 35, 348–357. doi: 10.1089/dna.2015.3104

Zhu, H., Jin, H., Pi, J., Bai, H., Yang, F., Wu, C., et al. (2016). Apigenin induced apoptosis in esophageal carcinoma cells by destruction membrane structures. *Scanning* 38, 322–328. doi: 10.1002/sca.21273

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 July 2021
doi: 10.3389/fgene.2021.687979

# Identification of Potential Prognostic Biomarkers Associated With Cancerometastasis in Skin Cutaneous Melanoma

*Yang Li[1†], Shanshan Lyu[2†], Zhe Gao[1], Weifeng Zha[1], Ping Wang[1], Yunyun Shan[1], Jianzhong He[3]* and Suyang Huang[1]**

[1] Dermatology, The Third People's Hospital of Hangzhou, Hangzhou, China, [2] Department of Pathology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China, [3] Department of Pathology, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China

OPEN ACCESS

**Edited by:**
Suman Ghosal,
National Institutes of Health (NIH), United States

**Reviewed by:**
Jin Li,
Harbin Medical University, China
Shaoli Das,
National Institutes of Health (NIH), United States
Sourish Ghosh,
National Institutes of Health (NIH), United States

***Correspondence:**
Jianzhong He
hejizh2010@163.com
Suyang Huang
hsy716@163.com

[†] These authors have contributed equally to this work

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

**Received:** 30 March 2021
**Accepted:** 18 June 2021
**Published:** 21 July 2021

**Citation:**
Li Y, Lyu S, Gao Z, Zha W, Wang P, Shan Y, He J and Huang S (2021) Identification of Potential Prognostic Biomarkers Associated With Cancerometastasis in Skin Cutaneous Melanoma. Front. Genet. 12:687979. doi: 10.3389/fgene.2021.687979

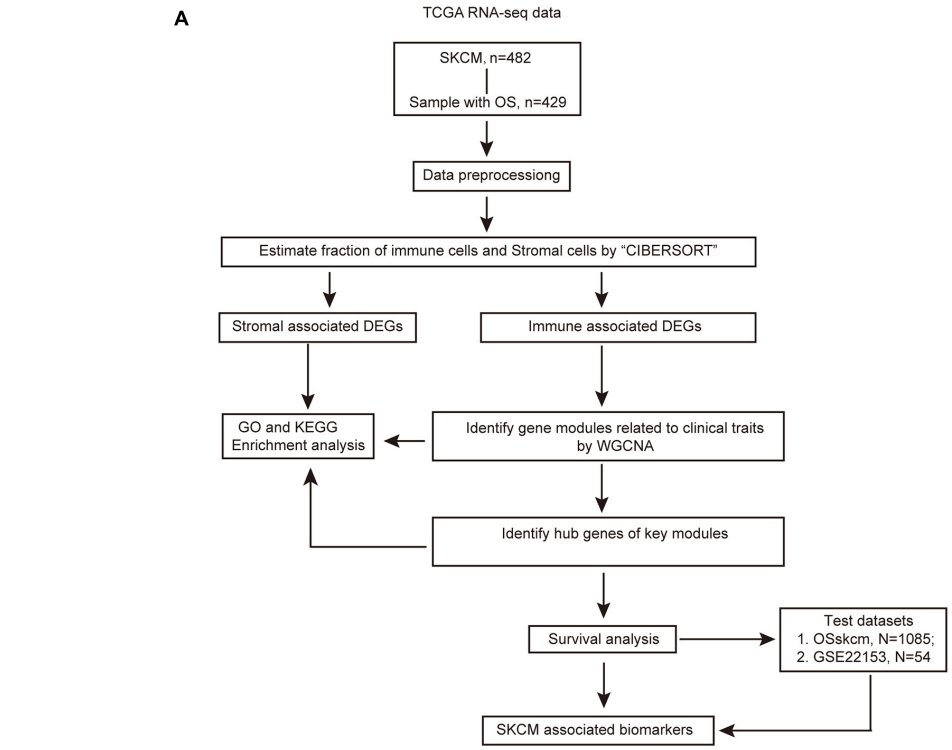
Skin cutaneous melanoma (SKCM) is a highly aggressive tumor. The mortality and drug resistance among it are high. Thus, exploring predictive biomarkers for prognosis has become a priority. We aimed to find immune cell-based biomarkers for survival prediction. Here 321 genes were differentially expressed in immune-related groups after ESTIMATE analysis and differential analysis. Two hundred nineteen of them were associated with the metastasis of SKCM *via* weighted gene co-expression network analysis. Twenty-six genes in this module were hub genes. Twelve of the 26 genes were related to overall survival in SKCM patients. After a multivariable Cox regression analysis, we obtained six of these genes (PLA2G2D, IKZF3, MS4A1, ZC3H12D, FCRL3, and P2RY10) that were independent prognostic signatures, and a survival model of them performed excellent predictive efficacy. The results revealed several essential genes that may act as significant prognostic factors of SKCM, which could deepen our understanding of the metastatic mechanisms and improve cancer treatment.

**Keywords: skin cutaneous melanoma, immune microenvironment, WGCNA, metastasis, prognostic biomarkers**

## INTRODUCTION

Skin cutaneous melanoma (SKCM) is a high-mortality-rate malignant tumor caused by abnormal melanocyte proliferation in neural crest cells (Bray et al., 2018; Siegel et al., 2020). According to the GLOBOCAN database (gco.iarc.fr), there were more than 200,000 new cases of SKCM over the world, and a quarter of them died in 2018 (Bray et al., 2018). The leading cause of death from this cancer is the metastasis of multiple organs (Zhu et al., 2016). The mortality rate of SKCM patients was significantly higher than that of other malignant tumors (Ekwueme et al., 2011). Therefore, SKCM seriously threatens public health and has become one of the evilest tumors worldwide (Gershenwald et al., 2017). The risk factors of SKCM included atypical mole or dysplastic nevus patterns and increased mole count (Chen et al., 2013). The treatment of the tumor microenvironment (TME) as a new treatment strategy has attracted public attention (Yang et al., 2018). It is composed of numerous cell types and is involved in the occurrence and invasion of tumors (Hanahan and Weinberg, 2000). With the development of tumor cytology and molecular biology, a deeper understanding of TME is essential to reveal improved immunotherapy

(Li et al., 2017; Qian et al., 2018). An algorithm called ESTIMATE could estimate the abundance of immune cells according to the gene expression level of tissues (Yoshihara et al., 2013; Li et al., 2016). Research shows that targeting stromal cells and connective tissue cells can be a new way to overcome drug resistance effectively (Hemminki et al., 2020).



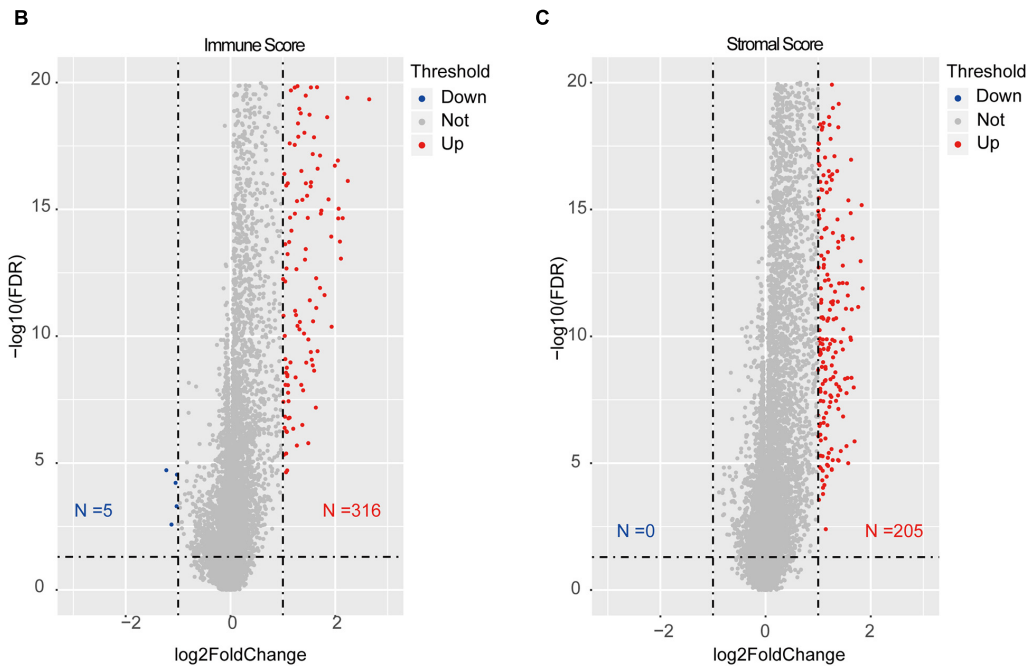




**FIGURE 1 |** Overview of the integration analysis. **(A)** Workflow of the analysis. **(B)** Volcano plot showing the differentially expressed genes (DEGs) between high- and low-immune-score samples. **(C)** Volcano plot showing the DEGs in high and low stromal score groups. The red color indicates the up-regulated genes, while blue represents the down-regulated ones. The horizontal dotted line represents a false discovery rate equal to 0.05, and the vertical dotted line represents a fold change equal to 2 or 0.5.

Weighted gene co-expression network analysis (WGCNA) is a computational method often used to explore the relationship between genes and clinical characteristics (Langfelder and Horvath, 2008; Yuan et al., 2020). The significant dominance of WGCNA is to combine genes into co-expression modules and build the relationship between clinical traits and genes (Luo et al., 2019). WGCNA could analyze a mass of genes and identify expression modules related to clinical features and critical genes for further verification (Luo et al., 2019; Radulescu et al., 2020).

This study obtained several modules and hub genes with significant differences in tumor microenvironment based on WGCNA and identified potential biomarkers that can predict SKCM prognosis (**Figure 1A**).

# MATERIALS AND METHODS

## Data Sources

Any ethical issue did not involve this study because it used public data which has already been published. We extracted the expression matrix of 473 SKCM patients and their clinical information from TGGA. Only 429 SKCM patients with complete overall survival information were selected. The clinical information of these patients (including gender, weight, pathologic stage, and so on) are shown in **Table 1** and **Supplementary Table 1**. The gene expression profiles were quantified by fragments per kilobase of transcript per million mapped reads and normalized through log2-based transformation. Besides that, the immune and stromal scores of each sample were calculated by the ESTIMATE analysis. The high-immune-score group represented the high proportion of immune cells in the tumor microenvironment, and the low-immune-score group represented conversely. The stromal score plays the same role but represents the stromal cell. An independent test dataset that contains 54 SKCM patients was downloaded from the Gene Expression Omnibus database.

**TABLE 1 |** Clinicopathological characteristics of 429 skin cutaneous melanoma patients in The Cancer Genome Atlas dataset.

| Clinical and pathological indices | Case no. | OS (%) | *P*-value[a] |
|---|---|---|---|
| Specimens | 429 | | |
| Mean age | 58 | | |
| Age (years) | | | <0.001 |
| ≤58 | 216 | 52.8 | |
| >58 | 213 | 54.5 | |
| Gender | | | 0.135 |
| Male | 266 | 49.2 | |
| Female | 163 | 60.7 | |
| pTNM stage | | | <0.001 |
| I | 114 | 47.4 | |
| II | 127 | 55.9 | |
| III | 166 | 56.0 | |
| IV | 22 | 54.5 | |
| Sample type | | | <0.001 |
| Metastatic | 349 | 48.9 | |
| Primary | 80 | 73.8 | |

[a]*Log-rank test using the Kaplan–Meier method. P < 0.05 was considered significant.*
*OS, overall survival.*

## Differential Analysis

The patients were classified into high- or low-immune-score groups and stromal score groups based on the median score of the ESTIMATE analysis. Then, differential analyses were used to filter the differentially expressed genes (DEGs) between the high and low groups. Finally, the raw *P*-value was corrected by false discovery rate (FDR). The differential analysis was performed through the "limma" R package, and the threshold was FDR $< 0.05$ and $|\log2\ FC| \geq 1$ (**Supplementary Table 2**).

## Constructed WGCNA Network and Identified Modules

We performed WGCNA analysis on the immune-related DEGs by the "WGCNA" R package. First, the *pickSoftThreshold* function was used to select the soft threshold (power) to construct the non-scale network. In this study, the power was set at 10. Second, modules were detected by the hierarchical clustering function "blockwiseModules." Then, the modules were associated with clinical characteristics by calculating gene significance (GS) and module membership (MM). Although a correlation between traits and modules has been found and the most relevant modules can be selected for analysis, the modules themselves still contain a large number of genes, so it is necessary to further search for the most important genes. All modules can be correlated with genes, and all continuous traits can also be correlated with gene expression levels. If genes significantly associated with traits are also significantly associated with a particular module, then those genes are likely to be crucial. Finally, the crucial genes in the candidate modules were filtered for further analysis. The cutoff for screening important genes was GS $> 0.25$ and MM $> 0.8$ (Liang et al., 2020).

## Enrichment Analysis

All the DEGs and candidate genes were subjected to an enrichment analysis using the "clusterProfile" R package (Yu et al., 2012). The functional background datasets contained the Gene Ontology (GO) terms (Dennis et al., 2003) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al., 2017). Functions with a FDR $< 0.05$ were selected for further discussion.

## Validation of Candidate Genes

GEPIA[1] was used to validate the immune-related DEGs. The web server collected the expression data of 9,736 tumor patients and 8,587 normal samples from The Cancer Genome Atlas (TCGA) and the GTEx projects. For the transcriptional level validation in SKCM, we set the criteria of significant results to $|\log2\ FC| \geq 0.585$ and $P < 0.05$. We used the TIMER web server to verify whether the crucial genes are associated with the immune cell infiltrate levels.

[1]http://gepia.cancerpku.cn/

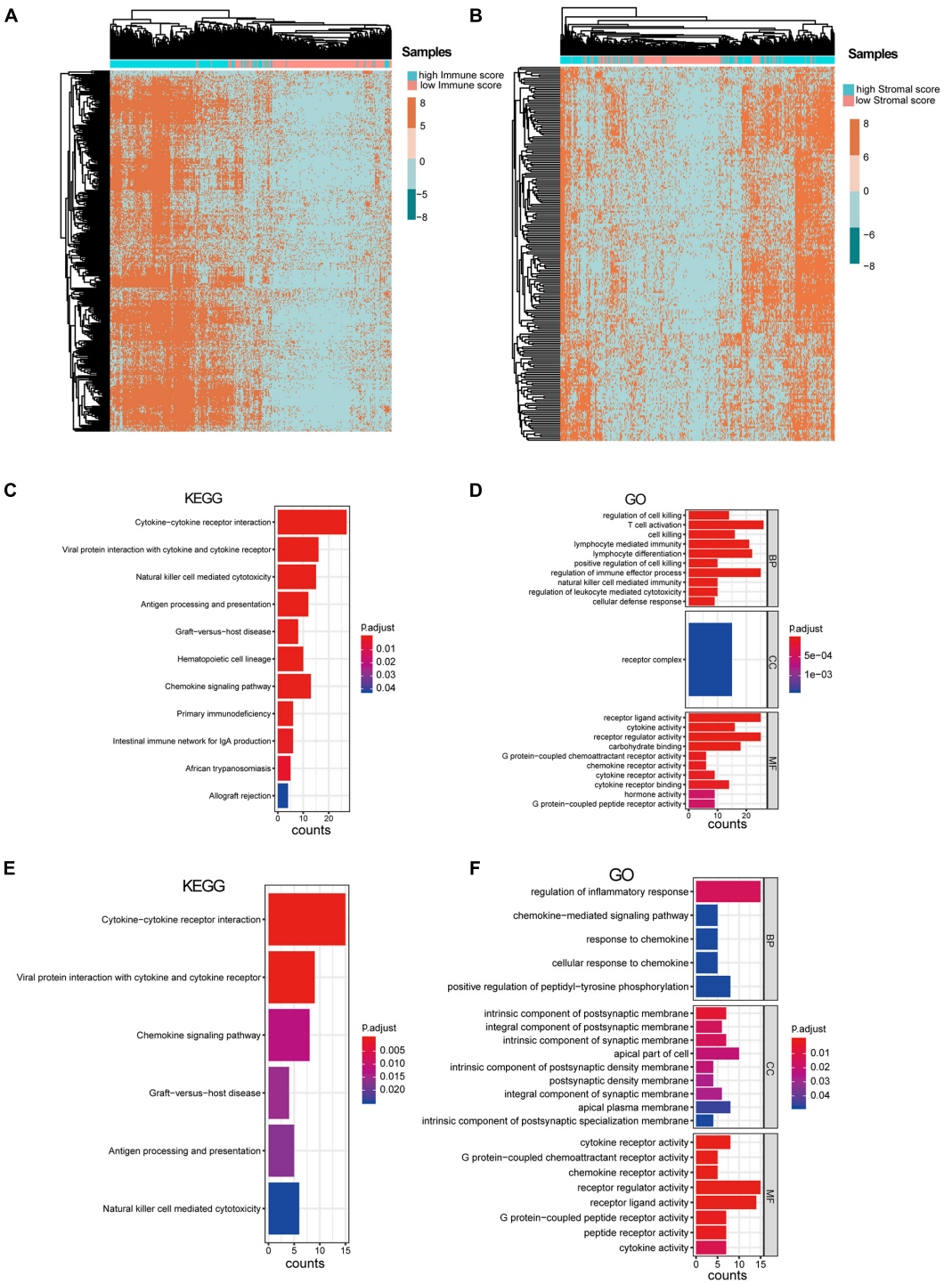



**FIGURE 2 |** Analysis of immune-associated differentially expressed genes (DEGs) and stromal-associated DEGs. **(A)** Heat map of immune-associated DEGs in skin cutaneous melanoma (SKCM) samples. **(B)** Heat map of stromal-associated DEG genes in SKCM samples. It shows that only immune-associated DEGs could nicely separate the low- vs. high-immune-score samples. **(C,D)** Enrichment analysis of the immune-associated DEGs. **(E,F)** Enrichment analysis of the stromal-associated DEGs.

## Survival Analysis

Survival analysis was used to filter vital prognostic biomarkers through the "survival" R package. The signature was filtered as independent of other clinical features through multivariable Cox regression analysis. Then, the independent clinical genes were used to combine a new survival signature by the Cox regression

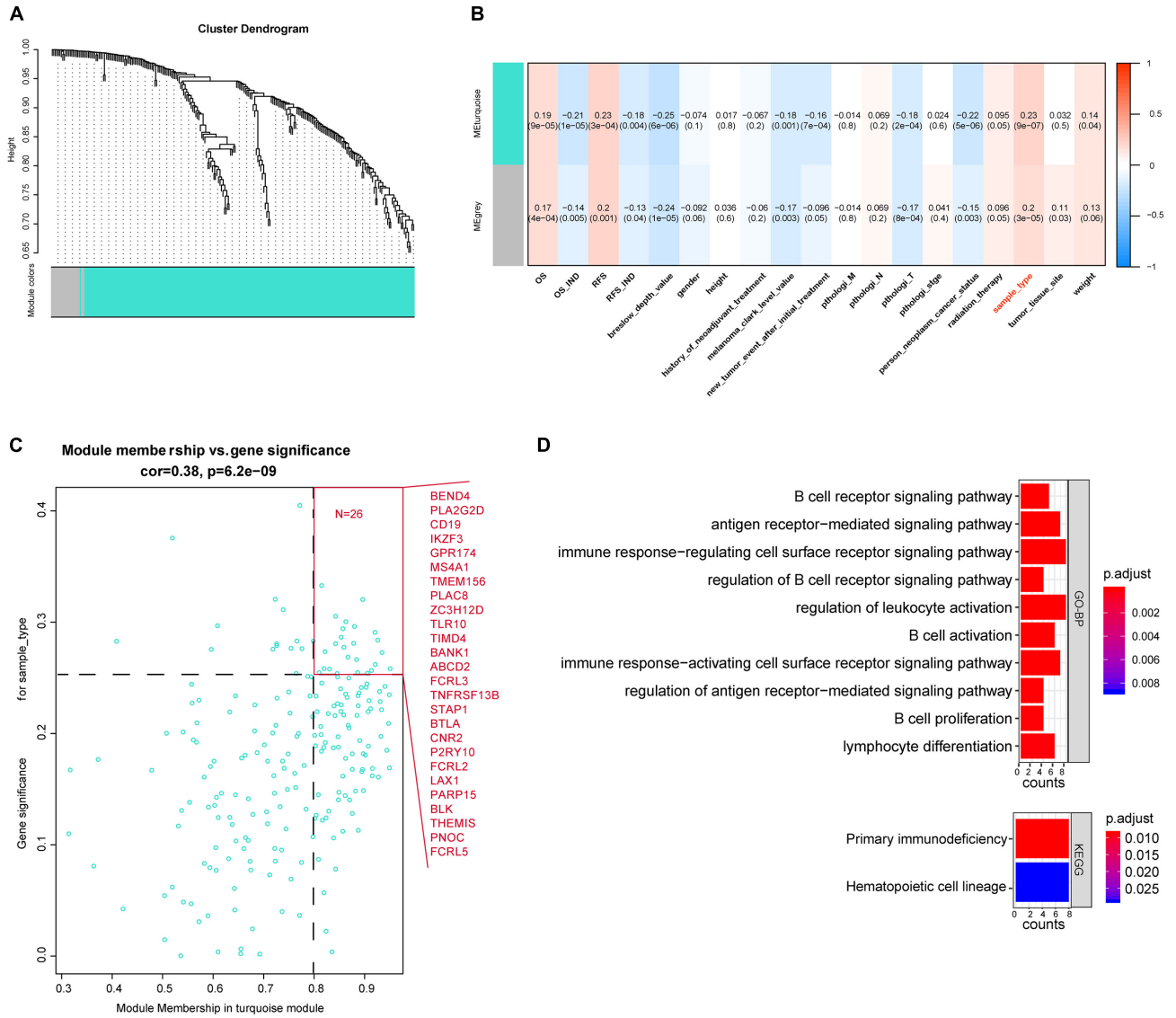



**FIGURE 3 |** Weighted gene co-expression network analysis (WGCNA) of immune-associated DEGs. **(A)** Cluster plot of genes based on the topological analysis of WGCNA. It shows that these immune-associated DEGs had two expression patterns. **(B)** Relationships between the module and clinical traits. Each cell described the relationship coefficients and *P*-values. **(C)** The gene significance and module membership of the blue module associated with metastasis. Hub genes are shown in red front. **(D)** Kyoto Encyclopedia of Genes and Genomes and Gene Ontology analysis of metastasis-associated hub genes.

model. The risk score was calculated by the expression of selected crucial genes, and correlation was estimated by Cox regression coefficients through the following formula:

$$\text{Risk score} = (\text{exp gene1} * \text{coef gene1}) + (\text{exp gene2} * \text{coef gene2}) + \ldots + (\text{exp geneN} * \text{coef geneN})$$

Then, we performed an area under the receiver operating characteristics (ROC) curve index to explore the prognostic efficiency of this signature using the "pROC" R package. The OSskcm Tool, which combined the survival information of more than 1,000 SKCM patients, was used to test the prognostic ability of the candidate genes (Zhang et al., 2020). An independent test dataset that contains 54 SKCM patients was used to verify the prognostic efficacy of the survival model (GSE22153).

# RESULTS

## Identification of Immune- and Stromal-Associated DEGs

After excluding the patients with no survival information, 429 qualified patients of the TCGA SKCM dataset were selected. Corresponding clinical traits that include overall survival

information were also downloaded. Based on the ESTIMATE analysis results, we divided the SKCM patients into high- and low-immune-score or high- and low-stromal-score groups. Then, we identified DEGs between these high- and low-score groups. According to the immune scores, 321 genes were differentially expressed, including 316 up-regulated genes and five down-regulated genes (**Figure 1B**). Similarly, there were 205 DEGs based on stromal scores; interestingly, all of them are up-regulated (**Figure 1C**). We found no intersection between these immune-related DEGs and the stromal-related DEGs. It suggested that the two types of DEGs performed different functions in SKCM. Then, the heat maps hinted at the gene expression patterns of DEGs (**Figures 2A,B**), and it was found that the immune-related DEGs had better classification efficiency. Then, the enrichment results showed that the immune-related DEGs were mainly enriched in the chemokine signaling pathway, cytokine–cytokine receptor interaction, primary immunodeficiency (KEGG) (**Figure 2C**), lymphocyte-mediated immunity, T cell activation, and regulation of immune effector processes (GO terms) (**Figure 2D**). It showed that the results of the ESTIMATE analysis are credible. The stromal DEGs were mainly enriched in cytokine–cytokine receptor interactions, antigen processing and presentation, natural killer cell-mediated cytotoxicity (KEGG) (**Figure 2E**), regulation of inflammatory response, and cellular response to chemokine (GO terms) (**Figure 2F**). These results indicate that the DEGs we screened are closely related to the immune response in SKCM patients, which may be used as new biomarkers for SKCM. Because the immune-related DEGs had a better classification efficiency by the cluster analysis, we use the 321 immune-related DEGs for further analysis.

## Identification of Gene Co-expression Modules That Associated With Clinical Traits

After differential analyses, we selected the 321 immune-related DEGs to build the gene co-expression network by WGCNA. The cutoff of soft power was set at 10 because it could make the scale-free topology model fit $R^2$ reach 0.85, and the mean connectivity is less than 20. This indicates that we have built a scale-free network (**Supplementary Figures 1A–D**). Then, we set the minimum module size at 30 to filter the co-expression modules. Finally, turquoise and gray co-expression modules were built (**Figure 3A**). The heat map described the topological overlap matrix (TOM) of input genes and showed the relationship between the two modules (**Supplementary Figure 1E**). The results showed that the 321 immune-related DEGs were expressed in two patterns.

## Identification of Crucial Modules

We calculated the relationship between the two modules and clinical traits (**Supplementary Figures 1E,F**) and then selected the essential genes. The results showed that the turquoise module, which contains 219 genes, was significantly associated with sample type (**Figure 3B**). Sample type stands for the primary tumor or the metastatic one. Based on the cutoff (GS > 0.25 and MM > 0.8), we identified 26 crucial genes out of the 219 turquoise module genes (**Figure 3C**). The enrichment result of the 26 genes showed that they were enriched in the primary immunodeficiency pathway and immune cell-associated signaling pathways. It suggested that these genes may play a crucial role in the metastasis of SKCM (**Figure 3D**).

## Validation of the Crucial Candidate Genes

We used the GEPIA (see text footnote 1) database to screen the 26 candidate DEGs that were not only immune-related DEGs but also differentially expressed between SKCM patients and normal samples. This screening procedure can help us obtain the biomarkers with more potential for clinical application. Finally, we obtained 12 crucial genes that are differentially expressed in cancer patients compared with normal samples and correlated with the tumor immune microenvironment (**Figure 4** and **Table 2**).

**TABLE 2 |** Basic information of the 12 crucial genes.

| Gene symbol | Full title | Module membership in turquoise module | Gene significance | *P*-value of differential analysis | |
|---|---|---|---|---|---|
| | | | | High immune score vs. Low immune score | *SKCM* vs. *normal* |
| PLA2G2D | Phospholipase A2 group IID | 0.896203345 | 0.320391963 | 2.74E–51 | 9.16E–64 |
| CD19 | CD19 molecule | 0.842601708 | 0.305343554 | 5.57E–35 | 5.81E–40 |
| IKZF3 | IKAROS family zinc finger 3 | 0.862856409 | 0.300451054 | 6.19E–44 | 9.03E–52 |
| MS4A1 | Membrane Spanning 4-Domains A1 | 0.857009326 | 0.294092449 | 3.77E–36 | 3.12E–32 |
| TMEM156 | Transmembrane Protein 156 | 0.920464075 | 0.290862446 | 5.28E–55 | 7.99E–40 |
| PLAC8 | Placenta associated 8 | 0.853632449 | 0.286478015 | 3.81E–45 | 2.47E–43 |
| ZC3H12D | Zinc finger CCCH-type containing 12D | 0.926001735 | 0.283865057 | 2.74E–58 | 0.0151 |
| FCRL3 | Fc receptor like 3 | 0.934510176 | 0.272729855 | 2.08E–51 | 5.46E–46 |
| TNFRSF13B | TNF receptor superfamily member 13B | 0.867106521 | 0.268605275 | 1.70E–43 | 7.14E–32 |
| P2RY10 | P2Y receptor family member 10 | 0.950493884 | 0.258242908 | 5.55E–55 | 2.55E–36 |
| PNOC | Prepronociceptin | 0.845668597 | 0.253861151 | 2.65E–38 | 2.06E–48 |
| FCRL5 | Fc receptor like 5 | 0.846566936 | 0.250761762 | 1.12E–37 | 3.66E–47 |

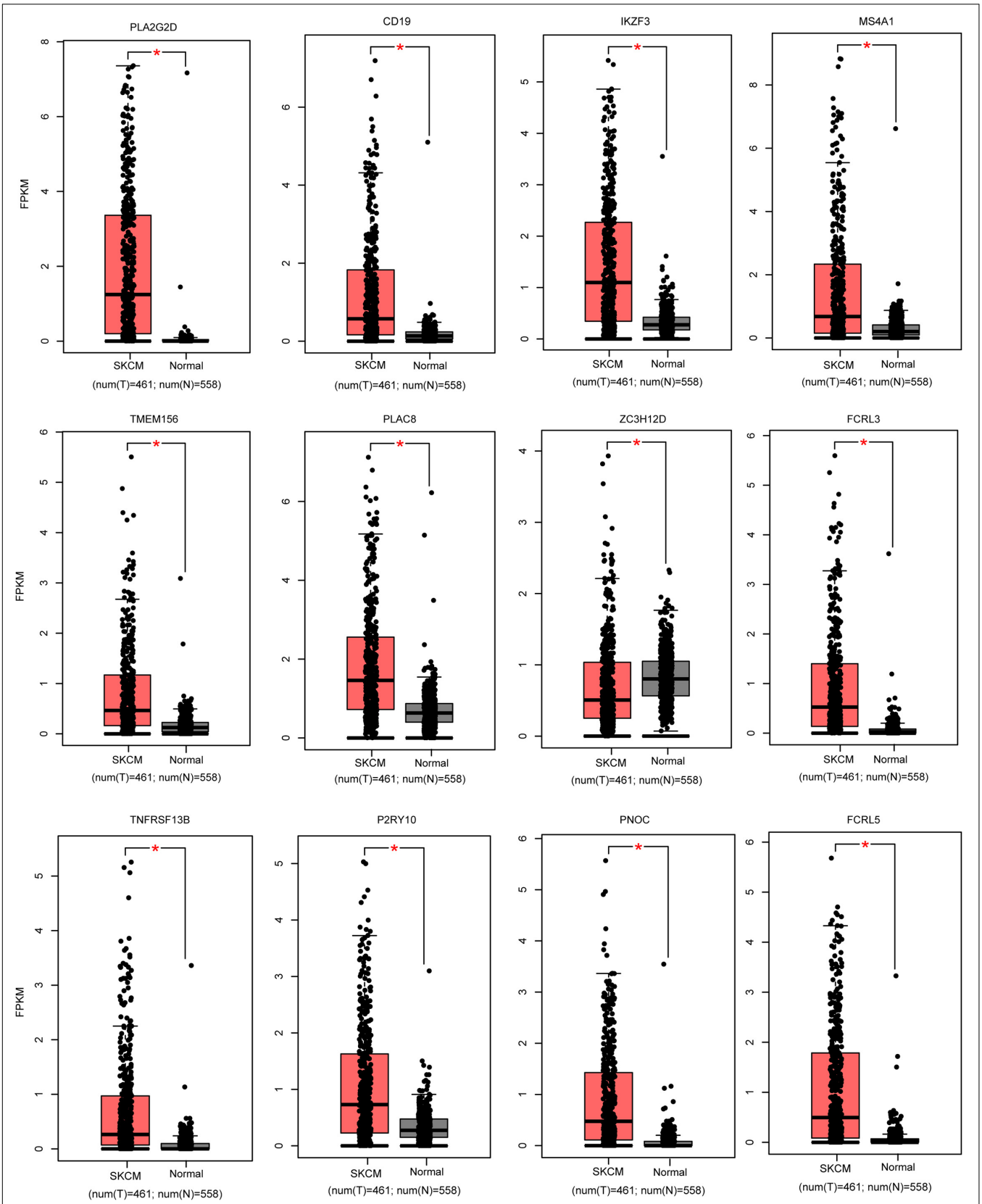



**FIGURE 4 |** Validation of the expression pattern of hub genes in melanoma. The red color represents the melanoma samples, while black indicates normal samples. *$P < 0.05$.

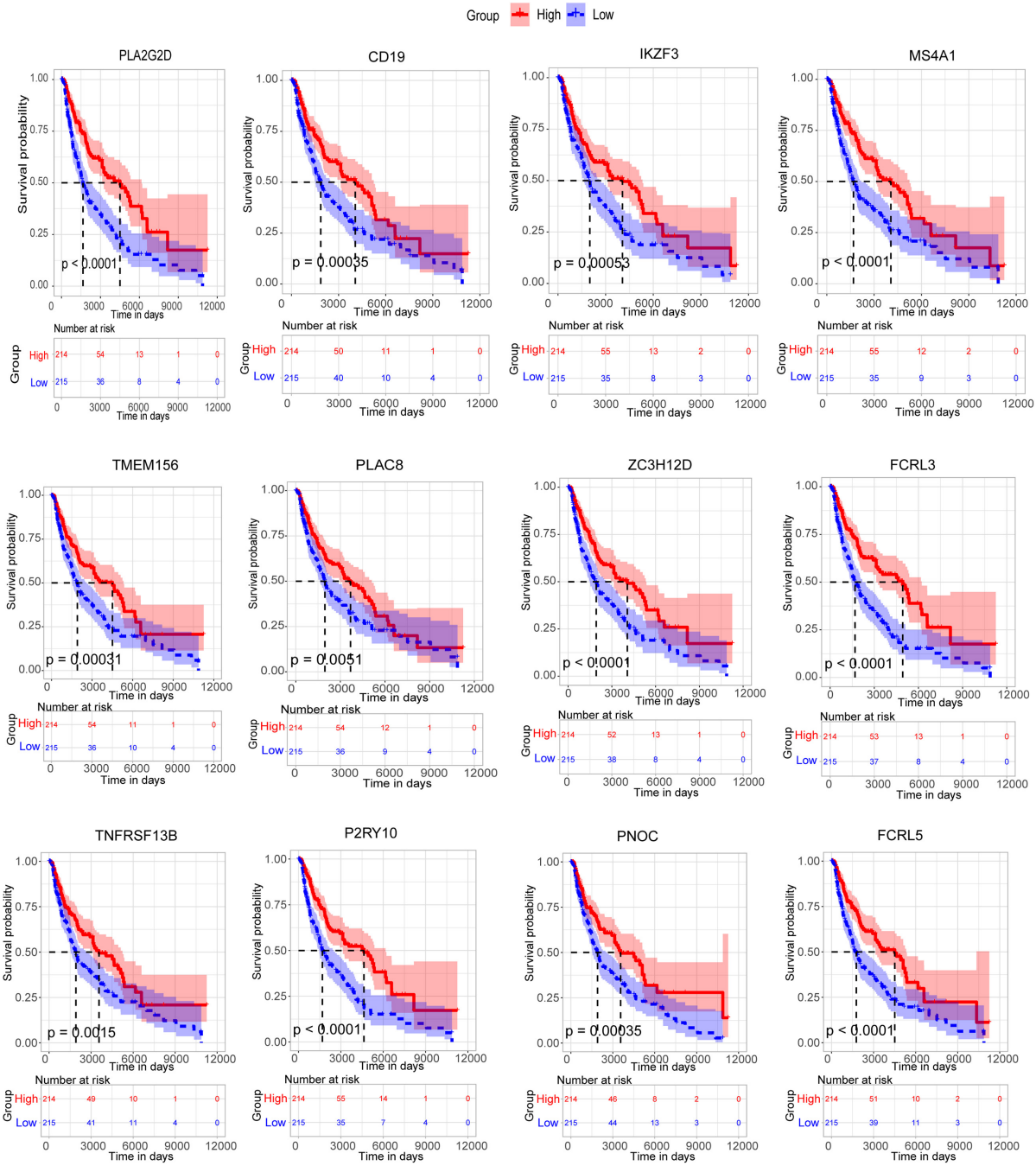



**FIGURE 5 |** Survival analysis of 12 candidate metastasis-associated key genes.

## The Crucial Genes are Potential Prognostic Biomarkers

Then, all the 12 essential genes (PLA2G2D, IKZF3, FCRL3, FCRL5, PNOC, PLAC8, P2RY10, TMEM156, ZC3H12D, MS4A1, CD19, and TNFRSF13B) were tested by survival analysis. We divided the patients into high- or low-expression groups based on the median expression level of the genes and performed a survival test. It found that all of them have a good prognostic efficacy in SKCM (survival $P < 0.05$). Interestingly, all the 12 genes are protective factors (**Figure 5**). A test dataset including 1,085 SKCM patients in the OSskcm Tool also testified the prognostic ability of these candidate genes (**Supplementary Figure 2A**). Then, we performed a multivariable Cox regression analysis and found that six of these genes (PLA2G2D, IKZF3, MS4A1, ZC3H12D, FCRL3, and P2RY10) were independent prognostic signatures (**Figure 6A**). Next, we combined these genes into

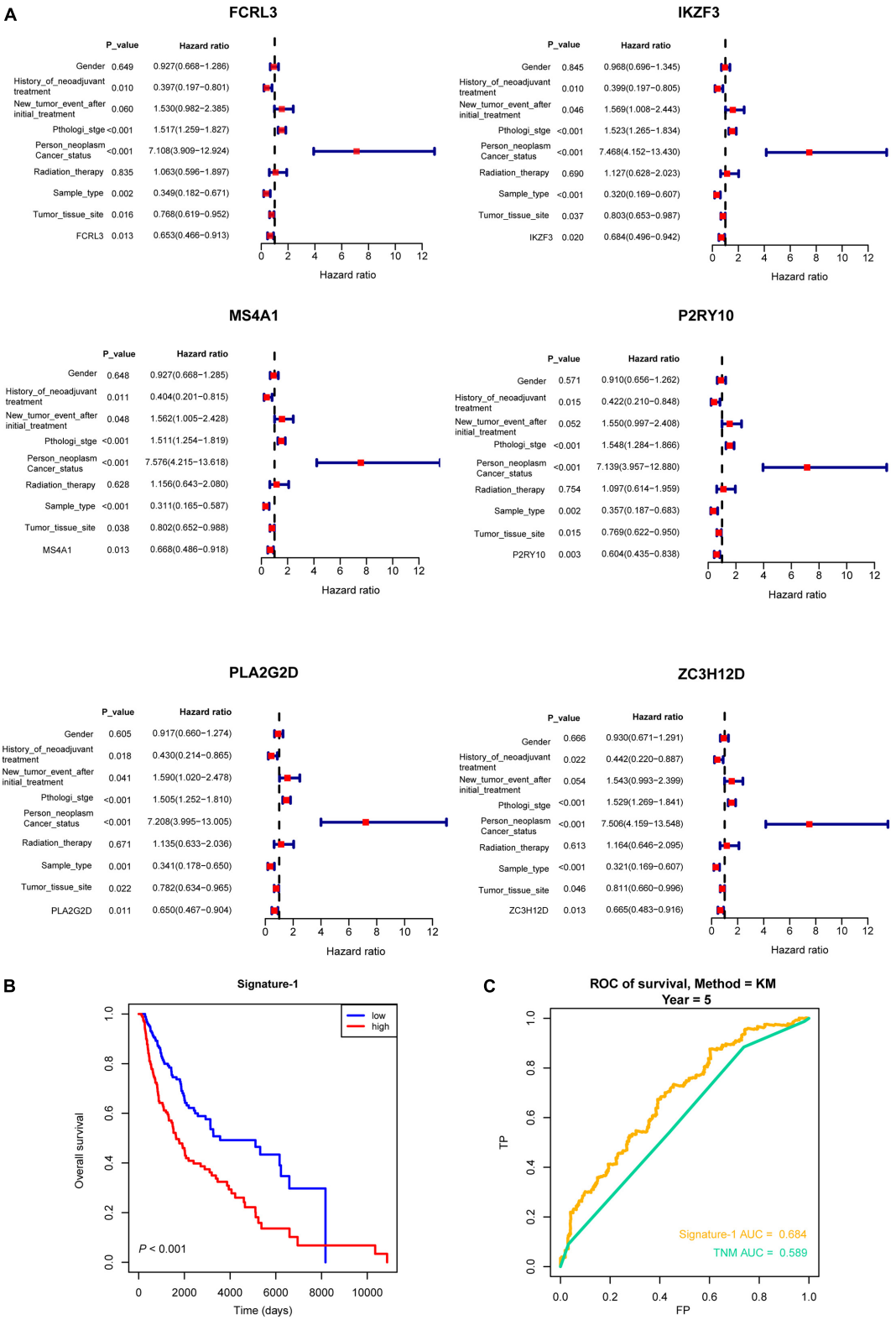



**FIGURE 6 |** Multivariable Cox regression analysis of crucial genes. **(A)** Forest plot showing the six hub genes (PLA2G2D, IKZF3, MS4A1, ZC3H12D, FCRL3, and P2RY10) that were independent prognostic factors in skin cutaneous melanoma. **(B)** The survival model (Signature-1) constructed by the six genes and the Kaplan–Meier curve which showed that it was survival-associated. **(C)** Receiver operating characteristic analysis showing that Signature-1 performed a better prognostic efficacy than the TNM stage.

a Cox proportional hazard model to construct a survival signature (Signature-1). We explored the survival efficiency of Signature-1 ($p < 0.05$, **Figure 6B**). Then, an ROC analysis was used to compare the prognostic value between Signature-1 and TNM stage, and we found that Signature-1 had a better prognostic efficacy than the TNM stage (**Figure 6C**). Next, the test dataset that contained 54 SKCM patients was used to verify the efficacy of the signature, and the result showed that Signature-1 kept its prognostic value in the test dataset (**Supplementary Figure 2B**). All the results hinted that the six genes had an excellent prognostic efficacy in SKCM, and we developed a survival model associated with tumor microenvironment and metastasis which may be applied to the clinic.

## DISCUSSION

Thousands of people worldwide suffer melanoma every year, and the number of SKCM is growing faster than any other type of malignancy. The numbers of research demonstrate the role of the immune cells on tumor cells, and the immune components in melanoma tissue can be used to evaluate the therapeutic and prognostic efficacy in melanoma (Ladanyi, 2015). Patients with primary tumors usually have higher than a 5-year survival rate (Balch et al., 2009). Bioinformatics analysis is widely used in the discovery of various biomarkers (Chen et al., 2020). Thus, obtaining predictive biomarkers for prognosis has become a priority.

WGCNA is an algorithm used to find crucial modules from a gene expression (Luo et al., 2019). Candidate therapeutic biomarkers are identified based on the relationship between the modules and the phenotype. Here we constructed the co-expression modules *via* WGCNA using the DEGs in high-immune-score SKCM patients compared with low-immune-score SKCM patients. Then, we obtained 12 crucial genes associated with the metastasis of SKCM, and six of them were independent prognostic biomarkers. The survival model of the six genes had a good predictive efficacy. We also used the TIMER web to verify the association between the six genes and immune cells (**Supplementary Figure 3**); all of them are associated with immune cell infiltrate levels. In addition, FCRL3 can promote IL-10 expression in B cells through the SHP-1 and p38 MAPK signaling pathways and is highly expressed on CD4 + CD26- T cells (Wysocka et al., 2014; Cui et al., 2020). IKZF3 is a predictor for survival in multiple myeloma stage III patients (Awwad et al., 2018). MS4A1 is associated with apoptosis of B-cell lymphoma Ramos cells (Kawabata et al., 2013). P2RY10 has been reported to be a tumor microenvironment-associated gene and a potential diagnostic biomarker of metastatic melanoma (Wang et al., 2018, 2020). PLA2G2D has been reported to moderate inflammation and could be a potential biomarker for treating inflammatory disorders (Miki et al., 2013). ZC3H12D is associated with inflammation (Huang et al., 2018). In SKCM, we first found that these crucial genes are involved in metastasis and perform similar functions in our WGCNA network. At the same time, they have a good prognostic efficacy. All of these genes have potential clinical applications as key prognostic biomarkers.

All in all, our findings may improve our fundamental knowledge of the molecular mechanisms of SKCM, and these prognostic biomarkers may improve the treatment of this cancer.

## CONCLUSION

Firstly, we filtered the immune-associated DEGs by the ESTIMATE analysis and got a metastasis-associated module through WGCNA. We then obtained overlapping DEGs in SKCM patients compared with normal samples and in the immune microenvironment, and 12 genes were screened. Next, we used survival analysis to obtain crucial prognostic biomarkers, and six genes with independent prognostic efficacy were filtered. The results may be helpful for future studies concerning SKCM to find potential prognostic targets.

## 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/s.

## AUTHOR CONTRIBUTIONS

YL, SL, and SH conceived and devised the study. YL and SL performed the bioinformatic and statistical analysis. ZG, WZ, PW, and YS found related data and analysis tools. YL, JH, and SH supervised the research and wrote the manuscript. All authors contributed to the article and approved the submitted version.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.687979/full#supplementary-material

**Supplementary Figure 1 |** Weighted gene co-expression network analysis (WGCNA) in the study. **(A,B)** Topology of the co-expression network. **(C,D)** Scale-free topology based on the cutoff of power (power = 10). **(E)** Visualization of the WGCNA network. Heat map showing the TOM among all modules. **(F)** The clinical trait information of 429 skin cutaneous melanoma patients.

**Supplementary Figure 2 |** Test datasets were used to present the prognostic efficacy of six crucial genes. **(A)** Survival analysis of six crucial genes in a dataset of 1,085 SKCM patients. **(B)** A test dataset used to show the prognostic efficacy of Signature-1.

**Supplementary Figure 3 |** Correlation of six hub genes with immune infiltration in melanoma.

## REFERENCES


Awwad, M. H. S., Kriegsmann, K., Plaumann, J., Benn, M., Hillengass, J., Raab, M. S., et al. (2018). The prognostic and predictive value of IKZF1 and IKZF3 expression in T-cells in patients with multiple myeloma. *Oncoimmunology* 7:e1486356. doi: 10.1080/2162402X.2018.1486356

Balch, C. M., Gershenwald, J. E., Soong, S. J., Thompson, J. F., Atkins, M. B., Byrd, D. R., et al. (2009). Final version of 2009 AJCC melanoma staging and classification. *J. Clin. Oncol.* 27, 6199–6206. doi: 10.1200/JCO.2009.23.4799

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

Chen, S. T., Geller, A. C., and Tsao, H. (2013). Update on the epidemiology of melanoma. *Curr. Dermatol. Rep.* 2, 24–34. doi: 10.1007/s13671-012-0035-5

Chen, Y., Liao, L. D., Wu, Z. Y., Yang, Q., Guo, J. C., He, J. Z., et al. (2020). Identification of key genes by integrating DNA methylation and next-generation transcriptome sequencing for esophageal squamous cell carcinoma. *Aging* 12, 1332–1365. doi: 10.18632/aging.102686

Cui, X., Liu, C. M., and Liu, Q. B. (2020). FCRL3 promotes IL-10 expression in B cells through the SHP-1 and p38 MAPK signaling pathways. *Cell Biol. Int.* 44, 1811–1819. doi: 10.1002/cbin.11373

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

Ekwueme, D. U., Guy, G. P. Jr., Li, C., Rim, S. H., Parelkar, P., and Chen, S. C. (2011). The health burden and economic costs of cutaneous melanoma mortality by race/ethnicity-United States, 2000 to 2006. *J. Am. Acad. Dermatol.* 65(5 Suppl. 1), S133–S143. doi: 10.1016/j.jaad.2011.04.036

Gershenwald, J. E., Scolyer, R. A., Hess, K. R., Sondak, V. K., Long, G. V., Ross, M. I., et al. (2017). Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J. Clin.* 67, 472–492. doi: 10.3322/caac.21409

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/s0092-8674(00)81683-9

Hemminki, K., Huang, W., Sundquist, J., Sundquist, K., and Ji, J. (2020). Autoimmune diseases and hematological malignancies: exploring the underlying mechanisms from epidemiological evidence. *Semin. Cancer Biol.* 64, 114–121. doi: 10.1016/j.semcancer.2019.06.005

Huang, W. Q., Yi, K. H., Li, Z., Wang, H., Li, M. L., Cai, L. L., et al. (2018). DNA Methylation Profiling Reveals the Change of Inflammation-Associated ZC3H12D in Leukoaraiosis. *Front. Aging Neurosci.* 10:143. doi: 10.3389/fnagi.2018.00143

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kawabata, K. C., Ehata, S., Komuro, A., Takeuchi, K., and Miyazono, K. (2013). TGF-beta-induced apoptosis of B-cell lymphoma Ramos cells through reduction of MS4A1/CD20. *Oncogene* 32, 2096–2106. doi: 10.1038/onc.2012.219

Ladanyi, A. (2015). Prognostic and predictive significance of immune cells infiltrating cutaneous melanoma. *Pigment Cell Melanoma Res.* 28, 490–500. doi: 10.1111/pcmr.12371

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9:559. doi: 10.1186/1471-2105-9-559

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17:174. doi: 10.1186/s13059-016-1028-7

Li, G., Qin, Z., Chen, Z., Xie, L., Wang, R., and Zhao, H. (2017). Tumor microenvironment in treatment of glioma. *Open Med.* 12, 247–251. doi: 10.1515/med-2017-0035

Liang, W., Sun, F., Zhao, Y., Shan, L., and Lou, H. (2020). Identification of susceptibility modules and genes for cardiovascular disease in diabetic patients using WGCNA analysis. *J. Diabetes Res.* 2020:4178639. doi: 10.1155/2020/4178639

Luo, Z., Wang, W., Li, F., Songyang, Z., Feng, X., Xin, C., et al. (2019). Pan-cancer analysis identifies telomerase-associated signatures and cancer subtypes. *Mol. Cancer* 18:106. doi: 10.1186/s12943-019-1035-x

Miki, Y., Yamamoto, K., Taketomi, Y., Sato, H., Shimo, K., Kobayashi, T., et al. (2013). Lymphoid tissue phospholipase A2 group IID resolves contact hypersensitivity by driving antiinflammatory lipid mediators. *J. Exp. Med.* 210, 1217–1234. doi: 10.1084/jem.20121887

Qian, J., Wang, C., Wang, B., Yang, J., Wang, Y., Luo, F., et al. (2018). The IFN-gamma/PD-L1 axis between T cells and tumor microenvironment: hints for glioma anti-PD-1/PD-L1 therapy. *J. Neuroinflam.* 15:290. doi: 10.1186/s12974-018-1330-2

Radulescu, E., Jaffe, A. E., Straub, R. E., Chen, Q., Shin, J. H., Hyde, T. M., et al. (2020). Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol. Psychiatry* 25, 791–804. doi: 10.1038/s41380-018-0304-1

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

Wang, L. X., Li, Y., and Chen, G. Z. (2018). Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. *PLoS One* 13:e0190447. doi: 10.1371/journal.pone.0190447

Wang, S., Zheng, X., Chen, X., Shi, X., and Chen, S. (2020). Prognostic and predictive value of immune/stromal-related gene biomarkers in renal cell carcinoma. *Oncol. Lett.* 20, 308–316. doi: 10.3892/ol.2020.11574

Wysocka, M., Kossenkov, A. V., Benoit, B. M., Troxel, A. B., Singer, E., Schaffer, A., et al. (2014). CD164 and FCRL3 are highly expressed on CD4+CD26- T cells in Sezary syndrome patients. *J. Invest. Dermatol.* 134, 229–236. doi: 10.1038/jid.2013.279

Yang, L., Song, X., Gong, T., Jiang, K., Hou, Y., Chen, T., et al. (2018). Development a hyaluronic acid ion-pairing liposomal nanoparticle for enhancing anti-glioma efficacy by modulating glioma microenvironment. *Drug Deliv.* 25, 388–397. doi: 10.1080/10717544.2018.1431979

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

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

Yuan, Y., Chen, J., Wang, J., Xu, M., Zhang, Y., Sun, P., et al. (2020). Identification hub genes in colorectal cancer by integrating weighted gene co-expression network analysis and clinical validation in vivo and vitro. *Front. Oncol.* 10:638. doi: 10.3389/fonc.2020.00638

Zhang, L., Wang, Q., Wang, L., Xie, L., An, Y., Zhang, G., et al. (2020). OSskcm: an online survival analysis webserver for skin cutaneous melanoma based on 1085 transcriptomic profiles. *Cancer Cell Int.* 20:176. doi: 10.1186/s12935-020-01262-3

Zhu, Z., Liu, W., and Gotlieb, V. (2016). The rapidly evolving therapies for advanced melanoma–Towards immunotherapy, molecular targeted therapy, and beyond. *Crit. Rev. Oncol. Hematol.* 99, 91–99. doi: 10.1016/j.critrevonc.2015.12.002


**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.

frontiers in Genetics

REVIEW
published: 27 July 2021
doi: 10.3389/fgene.2021.671057

# Detection of Cell Types Contributing to Cancer From Circulating, Cell-Free Methylated DNA

*Megan E. Barefoot[1], Netanel Loyfer[2], Amber J. Kiliti[1,3], A. Patrick McDeed IV[4], Tommy Kaplan[2] and Anton Wellstein[1]**

[1] Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, United States, [2] School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel, [3] Department of Biochemistry and Molecular and Cellular Biology, Georgetown University, Washington, DC, United States, [4] Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, DC, United States

Detection of cellular changes in tissue biopsies has been the basis for cancer diagnostics. However, tissue biopsies are invasive and limited by inaccuracies due to sampling locations, restricted sampling frequency, and poor representation of tissue heterogeneity. Liquid biopsies are emerging as a complementary approach to traditional tissue biopsies to detect dynamic changes in specific cell populations. Cell-free DNA (cfDNA) fragments released into the circulation from dying cells can be traced back to the tissues and cell types they originated from using DNA methylation, an epigenetic regulatory mechanism that is highly cell-type specific. Decoding changes in the cellular origins of cfDNA over time can reveal altered host tissue homeostasis due to local cancer invasion and metastatic spread to distant organs as well as treatment responses. In addition to host-derived cfDNA, changes in cancer cells can be detected from cell-free, circulating tumor DNA (ctDNA) by monitoring DNA mutations carried by cancer cells. Here, we will discuss computational approaches to identify and validate robust biomarkers of changed tissue homeostasis using cell-free, methylated DNA in the circulation. We highlight studies performing genome-wide profiling of cfDNA methylation and those that combine genetic and epigenetic markers to further identify cell-type specific signatures. Finally, we discuss opportunities and current limitations of these approaches for implementation in clinical oncology.

**Keywords: Cell-free DNA (cfDNA), cellular damage, circulating tumor DNA (ctDNA), deconvolution, liquid biopsy, tissue-of-origin, tumor microenvironment**

OPEN ACCESS

***Edited by:***
*Suman Ghosal, National Institutes of Health (NIH), United States*

***Reviewed by:***
*Xiaobin Zheng, Carnegie Institution for Science (CIS), United States*
*Sergey Aganezov, Johns Hopkins University, United States*

****Correspondence:***
*Anton Wellstein*
*wellstea@georgetown.edu*

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

***Received:*** *22 February 2021*
***Accepted:*** *17 May 2021*
***Published:*** *27 July 2021*

***Citation:***
*Barefoot ME, Loyfer N, Kiliti AJ, McDeed AP IV, Kaplan T and Wellstein A (2021) Detection of Cell Types Contributing to Cancer From Circulating, Cell-Free Methylated DNA. Front. Genet. 12:671057. doi: 10.3389/fgene.2021.671057*

## LIQUID BIOPSIES AND CELL-FREE DNA (CFDNA) IN ONCOLOGY

Liquid biopsies are emerging as a minimally invasive approach to complement and potentially advance the traditional standards of care in oncology (Bronkhorst et al., 2019). Tissue biopsies are taken as part of routine clinical care for most solid cancers and used to identify the molecular determinants of disease that can inform both diagnosis and prognosis. However, tissue biopsies are invasive and limited by inaccuracies due to sampling locations, restricted sampling frequency, and

poor representation of local tumor heterogeneity as well as dispersed cancerous lesions. To address these limitations, liquid biopsy technologies are rapidly advancing to provide analysis of tumors using circulating biomarkers in fluids such as the blood. One of the main advantages of liquid biopsies is its capacity for serial sampling by simple blood draws. The increased sampling frequency is helpful to monitor clonal evolution of tumor subpopulations as well as to assess evolutionary dynamics influencing treatment response and resistance as well as disease recurrence (Corcoran and Chabner, 2018). Also, liquid biopsies are capable of capturing systemic changes to provide an organism-wide picture of disease progression including the local primary tumor as well as distant metastatic sites and treatment responses across different sites. Finally, liquid biopsies are uniquely able to capture tumor heterogeneity over time, and thus complement traditional tissue biopsies that can only sample locally and at accessible sites (**Figure 1**).

Similar to tissue biopsies, the major purpose of liquid biopsies in oncology is to identify circulating analytes that provide molecular information about the cancer. In this context, there are a multitude of molecules that may be isolated from biological fluids and targeted for analysis. Until recently, the main focus has been on circulating molecules that can be directly tied back to the primary tumor, including circulating tumor cells (CTCs), cell-free tumor DNA (ctDNA), tumor-educated platelets (TEPs), and tumor secreted vesicles (exosomes, oncosomes, apoptotic bodies) (Best et al., 2015; Rapisuwon et al., 2016). However, as comprehensive approaches gain traction, there has been an expansion to include molecules reflective of dynamic changes to the host, tumor microenvironment and distant metastatic sites as well. Both tumor cells and normal host-derived cells release cell-free DNA (cfDNA) into the circulation as a result of physiological processes. cfDNA is thought to originate from the genomes of dying cells, including cells within tumors, and is reflective of cell turnover rates at steady state as well as altered homeostasis throughout the body with disease (Kustanovich et al., 2019; Heitzer et al., 2020; Rostami et al., 2020). Thus, circulating tumor DNA (ctDNA) is a subset of cfDNA that has different biological characteristics (**Table 1**). There is still much to be learned about the biology of cfDNA release, distribution, and elimination mechanisms leading to differential stability and circulation half-life in healthy compared to diseased states (Jiang and Lo, 2016; Heitzer and Speicher, 2018; Sanchez et al., 2018; Serpas et al., 2018; Han et al., 2020; Barefoot et al., 2021). The focus of this review will be on methylated cell-free DNA and its utility and applications in cancer diagnosis and management.

## INCREASED SIGNAL ABUNDANCE FROM LEVERAGING EPIGENETIC CHANGES IN BOTH TUMOR AND NON-TUMOR CELLS

There are still many challenges to overcome before liquid biopsies may be routinely implemented in the clinic. Signal abundance (fraction of target cfDNA relative to total cfDNA), sequencing depth, and breadth of genomic regions assayed by sequencing are factors that must be considered to detect signals in the circulation of cancer patients relevant to inform care (**Figure 2B**; Im et al., 2020). Strategies aimed at increasing any of these factors will improve the odds that informative signals can be detected. Signal abundance is largely a byproduct of the biology of the disease in question and therefore little can be done to modify this variable (Heitzer et al., 2019). For instance, ctDNA is highly correlated with tumor burden, with larger amounts of ctDNA found in the circulation of individuals at advanced stages of tumor progression. For this reason, mutation analysis of ctDNA is limited in its capacity to detect cancer-related signals, especially with low-volume tumors at early stage and relapse (Im et al., 2020). However, signal abundance can be increased by leveraging signals from all cfDNA molecules rather than the smaller subset of fragments containing specific tumor-related mutations (**Figures 2A,C**). This can be accomplished by targeting tumor-specific epigenetic changes that occur early on during carcinogenesis and thus are found at higher abundance in early stage cancers than tumor-related mutations (Snyder et al., 2016; Ulz et al., 2016; Wong et al., 2016; Leygo et al., 2017; Jiang et al., 2018, 2019; Cristiano et al., 2019; Gai and Sun, 2019; Ivanov et al., 2019; Panagopoulou et al., 2019; Sun et al., 2019; Van der pol and Mouliere, 2019; Sadeh et al., 2021). Further, combining tumor-cell derived signals with those from the surrounding host microenvironment can increase signal abundance (Hoadley et al., 2018; Liu et al., 2018; Haigis et al., 2019; Lam et al., 2019). Tumor DNA identified by genetic or epigenetic markers circulates admixed with non-tumor DNA. The same DNA sequence is found in all non-tumor cells and simple sequence analysis cannot be used to distinguish its cell-type origin. However, covalent and non-covalent epigenetic marks are pivotal to cell-type identity and can be used to distinguish tumor as well as non-tumor DNA, expanding the reach of molecules targeted to reflect disease-pertinent changes (Barefoot et al., 2021).

## TOWARD "THIRD-GENERATION" LIQUID BIOPSIES: FROM TARGETED TO COMPREHENSIVE APPROACHES

Despite its highly fragmented nature, advances in sequencing technologies have made comprehensive profiling of low integrity cfDNA possible. At a fixed target abundance and coverage, detection probabilities can be increased by broader sequencing, increasing the number of potential markers assayed (Im et al., 2020). Genomic analysis of ctDNA has decreased sensitivity relative to epigenetic approaches because of lower abundance at any one given marker (Leygo et al., 2017). Increasing the number of potential mutations assayed with whole-genome Next-Generation-Sequencing (NGS) applications has been shown to increase sensitivity, but there can still be a lack of sufficient markers when limited to tumor-specific mutations alone (Im et al., 2020). Comprehensive epigenetic profiling of tumor and non-tumor cfDNA has led to advances in detection of brain cancers, including gliomas, which "hide" behind the

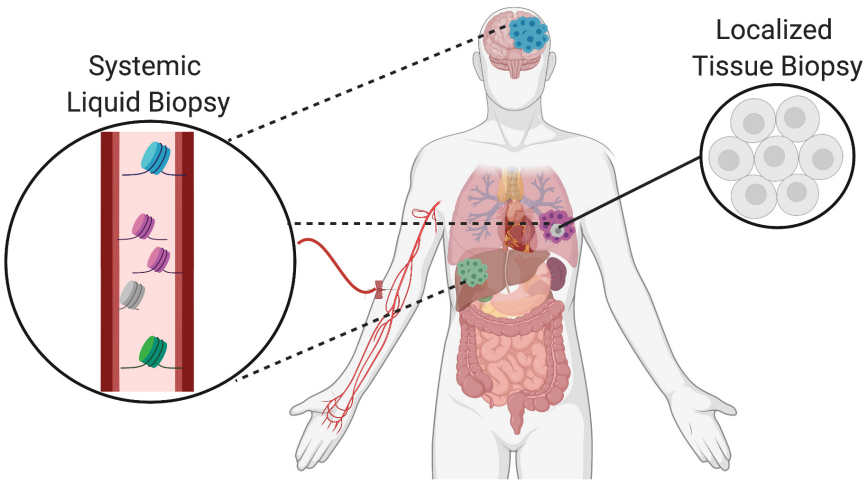



**FIGURE 1 |** Complementary role of tissue and liquid biopsies in oncology. Localized solid tissue biopsies are invasive and provide a snapshot of limited representational heterogeneity based on the small piece of tissue that is excised. In comparison, liquid biopsies are minimally invasive and allow for serial sampling to provide systemic information about the primary tumor as well as distant metastatic sites indicated in different colors. Thus, liquid biopsies complement tissue biopsies and increase representation of heterogeneity supporting the tracking of clonal evolution over time.

**TABLE 1 |** Analytes in solid vs. liquid biopsies.

| **A. Solid tissue biopsy** | **Liquid biopsy** |
|---|---|
| Invasive | Minimally invasive |
| Localized | Systemic |
| Limited sampling frequency | Serial sampling |
| Limited representation of heterogeneity | Representation of heterogeneity |
| **B. Cell-free DNA (cf-DNA)** | **Circuating tumor DNA (ct-DNA)** |
| Higher abundance | Lower abundance-subset of cfDNA (often < 1%) |
| Tumor-derived and signals derived from the surrounding microenvironment (normal cell-types) | Tumor-derived (tumor-specific genetic mutations and epigenetic abnormalities) |
| Majority hematopoietic origin | Host tissue somatic mutations are major confounder |
| Relevant to physiology and pathology | Relevant to pathology |

*(A) Comparison of solid and liquid biopsy samples. (B) Comparison of cell-free DNA (cfDNA) and circulating tumor DNA (ctDNA), two circulating analytes found in liquid biopsies.*

blood-brain barrier and have restricted access to release ctDNA into the circulation (Li et al., 2020; Nassiri et al., 2020). In these cases, integration across multiple markers allows for unparalleled sensitivity that cannot be achieved from low numbers of select targeted loci, despite high specificity and deep sequencing. Detection methods trending toward broader sequencing have been termed "third-generation" liquid biopsies and are emerging to allow for more comprehensive assessment of a multitude of signals.

Comprehensive sequencing approaches have unleashed the potential of liquid biopsies to achieve optimal sensitivity; however, there is still a need to improve the specificity and biological relevance of these assays. With the transition from targeted to comprehensive approaches come decreasing signal-to-noise ratios and new challenges to separate true biological signals from background sources of error (Ko et al., 2018). Physiological flux due to clonal hematopoiesis, inflammation, exercise, and other biological factors may dilute out relevant signals and calls for an increased understanding of the mechanisms of cell-free DNA release into the circulation and the distinct processing (Kustanovich et al., 2019; Barefoot et al., 2021). The predominating hematopoietic origins of cfDNA in healthy individuals makes it essential to identify markers separating cell-types of interest from peripheral immune cells (Barefoot et al., 2021). Machine-learning algorithms and data-science-driven approaches are being developed in tandem to reduce dimensionality and make sense of the data available to identify applicable information that may better inform clinical courses of action (Ko et al., 2018). As these approaches become increasingly complex, prior knowledge about the relevance of

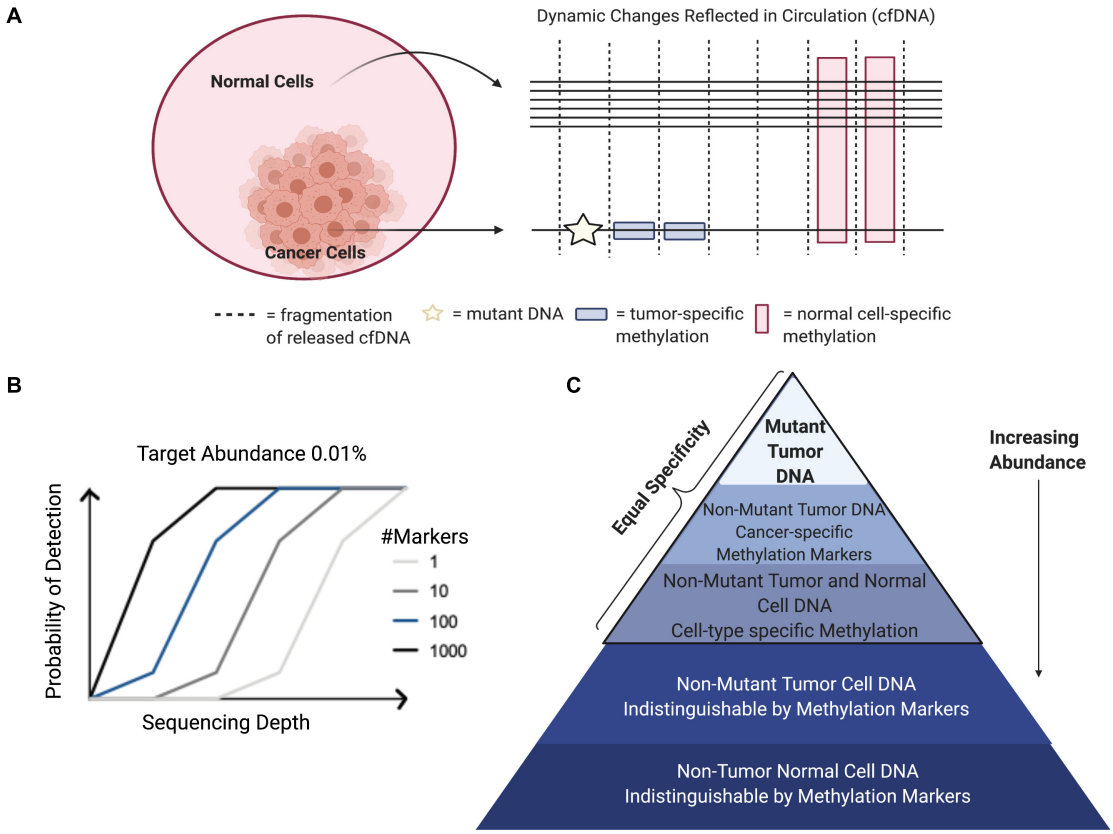



**FIGURE 2 |** Factors contributing to probability of signal detection. **(A)** Tumors release mutant genomic and epigenomic cfDNA into the circulation. Normal cell types from the surrounding microenvironment and other somatic cells also release cfDNA into the circulation that can be identified through cell-type specific epigenetic markers. Combining tumor cell- and normal host cell-derived signals can increase target abundance in the circulation to increase sensitivity, while maintaining specificity. **(B)** Target abundance, sequencing depth, and breadth of genomic regions assayed are factors that determine signal detection probability in the circulation of cancer patients. **(C)** Relative abundance of cfDNA populations and associated specificity.

the features selected will be imperative to maintain biological interpretability.

## BIOLOGICAL RELEVANCE OF CELL-FREE DNA METHYLATION PATTERNS

DNA methylation functions as an epigenetic regulatory mechanism and involves covalent addition of a methyl-group to the 5-carbon of cytosine (5mc). DNA methylation occurs most commonly in the context of CpG dinucleotides (Greenberg and Bourc'his, 2019). One of the main benefits to harnessing cell-free methylated DNA for liquid biopsy applications in cancer is the potential to exploit prior knowledge about the biological relevance of these marks. DNA methylation is an intrinsic mark of cell identity and pathologic alterations of DNA methylation are hallmarks of cancer (Greenberg and Bourc'his, 2019). The DNA methylation landscape changes in a highly regulated manner throughout development. Before embryo implantation, there is a global erasure of DNA methylation that is reset in multiple stages leading to the creation of cell-type specific methylation patterns, paralleling ongoing cell differentiation and organogenesis (Dor and Cedar, 2018). Once established, this pattern of DNA methylation is highly stable and conserved across DNA replication, making DNA methylation the predominant mechanism for inherited cellular memory during cell growth (Daniūnaitė et al., 2019). DNA methylation patterns may be selected as features that are relatively hyper- or hypo-methylated in specific cell types or in the context of specific cancers. Therefore, while there has been extensive characterization of DNA methylation changes that occur with disease and physiological aging, these changes occur only at specific locations throughout the epigenome allowing methylation states at regions critical to cell-type identity to remain constant over time (Michalak et al., 2019). This stability allows methylated cfDNA to serve as a robust biomarker in the face of patient heterogeneity, capable of being generalized across diverse patient populations (Dor and Cedar, 2018). There are many areas where liquid biopsies can be applied in clinical oncology. These include, but are not limited to, efforts aimed at early detection, assessment of prognosis, detection of minimal residual disease, metastasis, targeted-therapy selection, and treatment response monitoring (Sina et al., 2019; Luo et al., 2021). Both cancer and cell-type

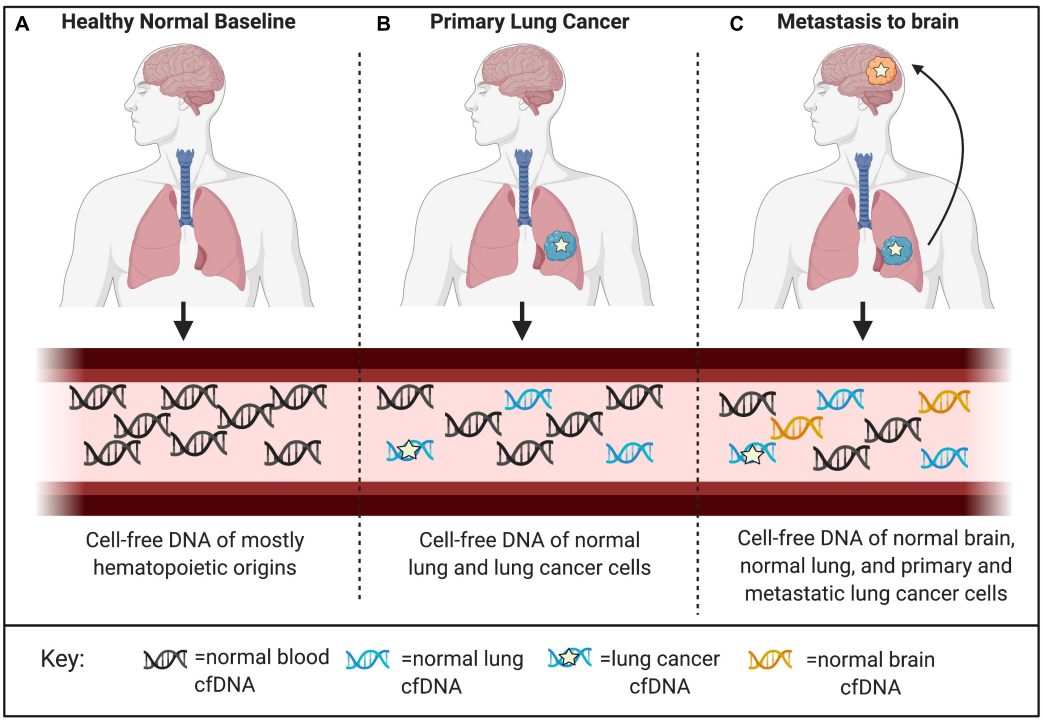



**FIGURE 3 |** Applications for detection and localization of metastasis and Cancer of Unknown Primary (CUP). **(A)** CfDNA in healthy individuals is mostly of hematopoietic origin. **(B)** The composition of cell-free DNA changes with disease. In this example, primary lung cancer results in increased levels of ctDNA identified by tumor-specific genomic and epigenomic markers, as well as increased levels of cfDNA from the surrounding lung microenvironment identified by normal cell-specific epigenetic markers. **(C)** Genomic mutations occur independently in primary and metastatic tumor sites. Liquid biopsies are capable of capturing this heterogeneity; however, mutations alone cannot localize these clonal populations to their tissue origins at the primary tumor site and distant metastatic site. As a complementary approach, normal tissue- and cell-type epigenetic markers can be used for detection and localization of metastasis and Cancers of Unknown Primary (CUP).

specific cell-free DNA methylation markers have been employed in each of these applications; however, there are important distinctions based on using disease-specific or normal cell-type specific markers that are worth noting. Specifically, cell-type specific DNA methylation markers have unique applications to localize cancers of unknown primary (CUP) as well as to detect metastases (**Figure 3**; Gai et al., 2018; Moss et al., 2018). In addition, systemic therapy-related adverse event monitoring remains one of the most promising applications.

## CELL-FREE DNA METHYLATION TECHNOLOGIES

There are many techniques that can be used to study DNA methylation as well as different strategies that can be applied to classify and quantify methylation status (Olkhov-Mitsel and Bapat, 2012; Kurdyukov and Bullock, 2016; Galardi et al., 2020; Zhao et al., 2020). These methodologies must be able to distinguish between methylated and unmethylated cytosines. This review mainly focuses on 5mc as it is the most commonly characterized epigenetic mark in cancer. However, other DNA modifications, including 5-hydroxymethylcytosine (5hmc), are thought to be more dynamic, reflecting active demethylation events, and may be complementary to characterize as well (Song et al., 2017). The different DNA methylation detection technologies and platforms are categorized in **Figure 4**. The main methods are restriction enzyme digestion, affinity enrichment, bisulfite-conversion, and enzymatic modification approaches. To date, several of these approaches have been successfully implemented to study genome-wide cfDNA methylation, highlighted in **Table 2**. Restriction enzyme-based methods cleave DNA at enzyme specific CpG sites. However, the highly fragmented nature of cfDNA and limited frequency of CpG-containing recognition sites make this approach challenging for comprehensive profiling of cfDNA (Huang and Wang, 2019). cfMeDIP-seq is an affinity-based approach that enriches for methylated DNA using 5mc-specific antibodies (Shen et al., 2018). As such, it is capable of characterizing overall methylation levels across a region, but not at single CpG sites. In addition, the majority of cell-type specific methylation markers in the human body are hypomethylated as a result of methylation resetting that takes place throughout tissue differentiation and development. These methods that specifically enrich for hypermethylated DNA may have limited detection potential at these regions of interest.

Bisulfite conversion chemically modifies DNA so that unmethylated cytosines (C) are deaminated to uracil (U) to be later replaced by thymine (T) via PCR, while unmethylated

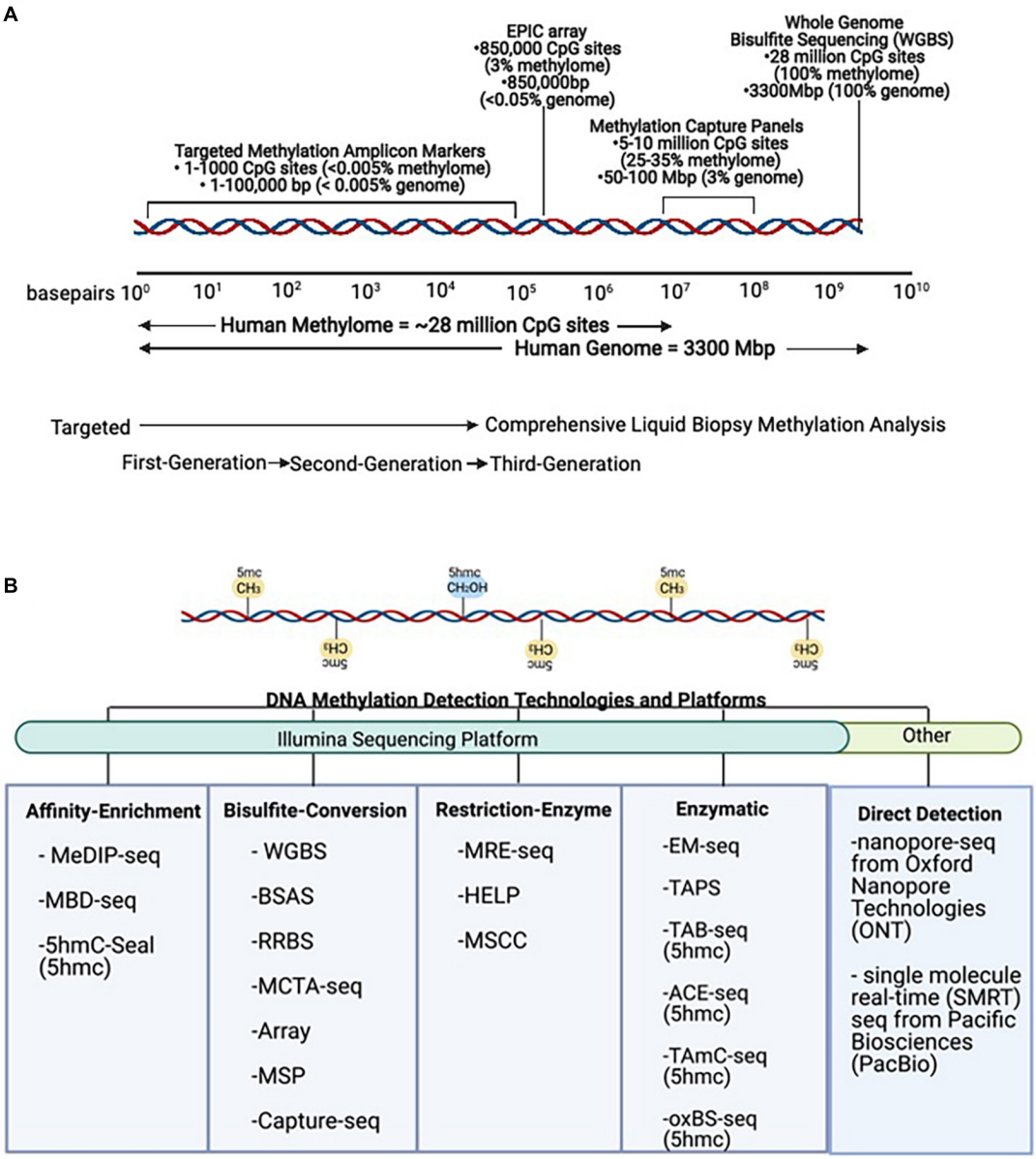



**FIGURE 4 |** DNA methylation technologies and platforms for signal detection. **(A)** Scale representation of DNA methylation technologies from targeted First- and Second-Generation toward comprehensive Third-Generation applications in liquid biopsies. **(B)** Methods for detection of DNA methylation. The same DNA sequence is found in all non-tumor cells, and simple sequence analysis cannot be used to distinguish its cell-type identity. However, these methods can be used to detect DNA methylation (5mc) and DNA hydroxymethylation (5hmc) levels in tumor and non-tumor cells. MeDIP-seq, Methylated DNA immunoprecipitation sequencing; MBD, methyl-CpG-binding domain sequencing; WGBS, Whole Genome Bisulfite Sequencing; BSAS, Bisulfite Amplicon Sequencing; RRBS, Reduced Representation Bisulfite Sequencing; MCTA-seq, methylated CpG tandem amplification and sequencing; MSP, Methylation Specific PCR; MRE-seq, methylation-sensitive restriction enzyme sequencing; HELP, Hpall-tiny fragment enrichment by ligation-mediated PCR; MSCC, Methyl-sensitive Cut Counting; EM-seq, Enzymatic Methyl-Sequencing; TAPS, TET-assisted pyridine borane sequencing; TAB-seq, TET-assisted bisulfite sequencing; ACE-seq, APOBEC-coupled epigenetic sequencing; hmc-CATCH, chemical-assistant C-to-T conversion of 5hmC sequencing; oxBS-seq, oxidative bisulfite sequencing.

cytosines are protected and remain cytosine (C) (Olova et al., 2018). The majority of comprehensive cfDNA methylation profiling has been done using bisulfite conversion methods, including Whole Genome Bisulfite Sequencing (WGBS), Reduced Representation Bisulfite Sequencing (RRBS), Methylated CpG Tandem Amplification and Sequencing (MCTA-seq), and Methylation Arrays. WGBS and RRBS are capable of detecting DNA methylation at single-base resolution. More importantly, these methods are capable of detecting read-specific DNA methylation patterns (Scott et al., 2020).

WGBS is the most comprehensive approach, but it can be costly to sequence the whole genome to an informative depth. However, sequencing costs are decreasing, making this approach more attractive. RRBS has been optimized in a few instances for accommodating highly fragmented cfDNA molecules (Guo et al., 2017; De Koker et al., 2019). Despite these modifications, the use of restriction enzymes in this sequencing approach give rise to the same limitations as restriction-enzyme based methods. MCTA-seq uses primers to preferentially amplify methylated CpG islands (CpG tandem regions) and, while being more targeted, this approach is also biased toward hypermethylated regions (Liu X. et al., 2019). Methylation hybridization arrays allow for single-base resolution but do not allow for pattern analysis of multiple CpG sites from the same molecule and have reduced genome-wide coverage of CpG sites compared to NGS approaches (Moss et al., 2018).

While bisulfite conversion has long been considered the gold standard of methylation detection, there are major limitations that recent advances in enzymatic approaches show promise in overcoming (Schutsky et al., 2018; Liu Y. et al., 2019). For instance, sodium bisulfite is a harsh chemical treatment that causes unwanted DNA degradation and fragmentation, resulting in uneven genome coverage. Enzymatic Methyl-seq (EM-seq) uses the enzyme APOBEC to deaminate unmethylated cytosines and protects methylated cytosines from conversion by utilizing TET2 as an oxidative enhancer (Vaisvila et al., 2019). This results in the same base conversions as bisulfite sequencing, but this method has been shown to cause less DNA damage and as a result is more sensitive, requiring smaller amounts of input DNA. This method is used in a recent publication to profile cytosine methylation and nucleosome occupancy at the same time, a feat made possible from retention of the original cfDNA structure without fragmentation or degradation (Erger et al., 2020).

The nuances of these different methodologies to detect DNA methylation make choosing the right method and accounting for its limitations essential toward accurate interpretation of results. Methylation detection technologies are rapidly evolving, leading to expanded potential applications. For instance, one such advancement involves the direct detection of methylation without treatment of DNA, possible with nanopore-sequencing from Oxford Nanopore Technologies (ONT) and single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) (Flusberg et al., 2010; Liu Q. et al., 2019; Ewing et al., 2020; Yuen et al., 2020; Tse et al., 2021). Although direct detection of methylation is not currently possible with cfDNA inputs, these advances point toward new possibilities in the future.

## TISSUE-OF-ORIGIN (TOO) DECONVOLUTION ANALYSIS: USING HEALTHY CELL-TYPE SIGNALS TO INFORM ABOUT DISEASE

Tissue-of-origin (TOO) analysis takes each individual cell-free DNA molecule in the circulation and routes it back to its tissue and cellular origins as a non-invasive monitoring tool for tissue damages (**Figure 5**). At steady state, cfDNA is released into the circulation reflective of cellular turnover happening throughout the human body, resulting in a complex mixture of fragments (Barefoot et al., 2021). On average, the plasma from healthy individuals has 1,500 genome equivalents or roughly 10 ng/mL cfDNA concentration (Moss et al., 2018). With cancer, cfDNA levels are thought to increase in parallel with disease progression as a result of increased proliferation and death rates of tumor cells (Kustanovich et al., 2019). However, relying on concentration of cfDNA alone to diagnose disease is too simplistic of an approach, as concentration is not an absolute indicator of disease and changes can result from a plethora of factors, including exercise, inflammation, and induction of cellular senescence. Detection of changing cell-type proportions from alterations in cfDNA composition is a more reliable approach. Shifting cfDNA makeup has been used for monitoring altered death rates of cells in different tissues, applicable to a broad spectrum of physiological and pathological conditions as well as therapeutic interventions. These include non-invasive prenatal testing, solid organ transplant, cancer, neurodegenerative and autoimmune pathologies, among many others (Sun et al., 2015; Zemmour et al., 2018; Chatterton et al., 2019; Cheng A. P. et al., 2019). To demonstrate feasibility, DNA methylation patterns specific to a variety of epithelial, endothelial, nervous, stromal, muscle, fat, and immune cell-types have been discovered and successfully applied using TOO analysis of cfDNA (Lehmann-Werman et al., 2016, 2018; Moss et al., 2018). In addition to tumor-derived DNA, changes to the host microenvironment can contribute to altered cell-type proportions of cfDNA in the circulation through cancer-related changes to normal tissue architecture. Although normal cell-specific DNA methylation markers are used, relevance to disease is inferred through abnormal detection in the circulation as a result of aberrant cell death and tissue damages (Heitzer et al., 2020). Thus, the changing proportion of normal cell types found in the circulation can be used to inform about disease states (Houseman et al., 2012; Teschendorff et al., 2017; Zheng et al., 2018; Huang et al., 2019; Barefoot et al., 2021). Recent studies demonstrating the feasibility of using cell-type specific cfDNA methylation marks for TOO analysis in cancer are described below (**Table 2**). This methodology is useful to detect damage to specific cell types in tissues and has many applications to inform diagnostics in the clinic as well as to reveal complexities of cancer pathophysiology at the cellular level.

## COMPUTATIONAL METHODS FOR CELL-MIXTURE DECONVOLUTION IN LIQUID BIOPSY

Advances in liquid biopsy technology have led to the generation of massive amounts of methylation sequencing data that can be difficult to analyze due to the extensive number of possible features in the human methylome. Computational methods, including many machine learning techniques, have been developed to better handle such data by isolating specific signals and discriminative features, thereby reducing the dimensions of the data so that it is easier to interpret (Ko et al., 2018). In

**TABLE 2 |** Feasibility of tissue-of-origin analysis in oncology using cell-free DNA methylation markers.

| Disease | Methylation data type | Marker type | Deconvolution method | Publication |
|---|---|---|---|---|
| HCC, NIPT, Transplant | WGBS | Tissue-specific | QP | Sun et al., 2015 |
| PDAC, CRC, Diabetes, Transplant, MS, TBI, IBD | BSAS | Tissue-specific | Read-specific binary classification | Lehmann-Werman et al., 2016, 2018 |
| Transplant | WGBS | Tissue-specific | QP | Cheng et al., 2017 |
| CRC, LCP | RRBS, WGBS | Both | Multi-class prediction, RF, feature extraction "haplotype blocks" | Guo et al., 2017 |
| MI, sepsis | BSAS | Tissue-specific | Read-specific binary classification | Zemmour et al., 2018 |
| CRC, BRCA, PDAC, CUP, Transplant, Sepsis | 450K array | Tissue-specific | NNLS regression | Moss et al., 2018 |
| Transplant, infection | WGBS | Tissue-specific | QP | Cheng A. P. et al., 2019 |
| Neurotrauma + neurodegenerative disease | tNGBS (multiplex 35 amplicons) | Tissue-specific | Read-specific binary classification (k-mer analysis) | Chatterton et al., 2019 |
| HCT, GVHD, transplant | WGBS | Tissue-specific | QP | Cheng et al., 2020 |
| HCC, cirrhosis, cholelithiasis, acute pancreatitis | MCTA-seq | Tissue-specific | PSO | Liu Y. et al., 2019 |
| BRCA | BSAS | Tissue-specific | Read-specific binary classification | Moss et al., 2020 |
| mCRPC | Cpature-seq/WGBS | Both | PCA | Wu et al., 2020 |
| 12 cancer types | Cpature-seq/WGBS | Both | Ensemble logistic regression | Liu et al., 2020 |
| ALS, pregnancy | WGBS | Tissue-specific | Bayesian EM algorithm (CelFiE) likelihood-based | Caggiano et al., 2020 |
| Transplant, AKI | cfNOME-seq | Tissue-specific | LSM (QP) | Erger et al., 2020 |
| COVID-19 | WGBS | Tissue-specific | NNLS regression | Cheng et al., 2021 |
| HCC, CRC, LCP | WGBS | Cancer-specific | Read-specific, likelihood-based | Kang et al., 2017; Li et al., 2018 |
| LCP, HCC. PDAC, GBM, CRC, BRCA | hMe-Seal (5hmc) | Cancer-specific | RF, Mclust | Song et al., 2017 |
| PDAC, AML, BRCA, CRC, RCC, PLC | MeDIP-seq | Cancer-specific | Limma, binomial GLM | Shen et al., 2018 |
| Pediatric MB | WGBS/CMS-IP-seq | Cancer-specific | Multivariate Cox regression linear model | Li et al., 2020 |
| Glioma, intracranial tumors | MeDIP-seq | Cancer-specific | Binomial RF | Nassiri et al., 2020 |

*HCC, Hepatocellular Cancer; NIPT, Non-Invasive Prenatal Testing; PDAC, Pancreatic Cancer; CRC, Colorectal Cancer; MS, Multiple Sclerosis; TBI, Traumatic Brain Injury; IBD, Inflammatory Bowel Disease; LCP, Lung Cancer Primary; MI, Myocardial Infarction; BRCA, Breast Cancer; CUP, Cancer Unknown Primary; GBM, Glioblastoma Multiforme; AML, Acute Myeloid Leukemia; RCC, Renal Cell Carcinoma; HBC, Hepatobiliary Cancer; NSCLC, Non-Small Cell Lung Cancer; HCT, Hematopoietic Cell Transplant; GVHD, Graft-vs.-Host Disease; AKI, Acute Kidney Injury; ALS, Amyotrophic Lateral Sclerosis; MB, Medulloblastoma; WGBS, Whole Genome Bisulfite Sequencing; BSAS, Bisulfite Amplicon Sequencing; RRBS, Reduced Representation Bisulfite Sequencing; ddPCR, Droplet Digital PCR; tNGBS, targeted Next Generation Bisulfite Sequencing; MeDIP-seq, Methylated DNA immunoprecipitation Sequencing; CMS-IP-seq, Cytosine 5-methyenesulphonate-immunoprecipitation sequencing; MCTA-seq, Methylated CpG Tandems Amplification Sequencing; cfNOME-seq, cell-free Nucleosome Occupancy and Methylation Sequencing; RF, random forest; GLM, generalized linear model; NNLS, Non-Negative Least Squares; LSM, Linear Least Squares Minimization; QP, Quadratic Programming; PSO, Particle Swarm Optimization; EM, Expectation-Maximization. Some of the materials are based on Barefoot et al. (2021).*

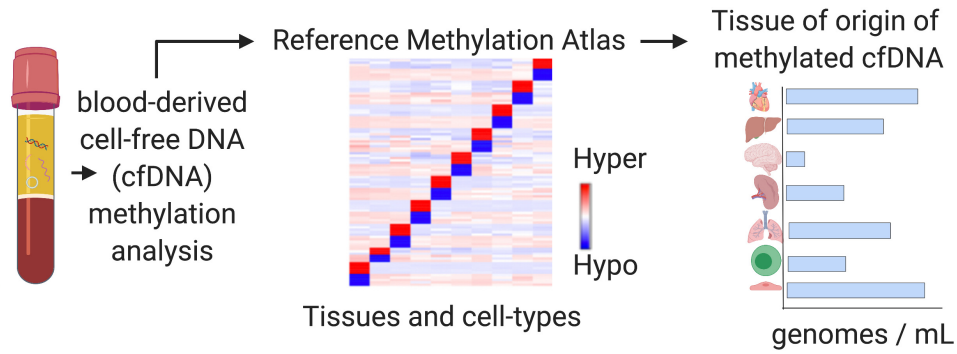



**FIGURE 5 |** Tissue-of-origin deconvolution analysis. CfDNA is a mixture of fragments released from healthy and diseased cells in different tissue types throughout the human body into the circulation. DNA methylation is highly cell-type specific and can be used to identify the cellular origins of cfDNA at specific markers. Tissue-of-origin (TOO) analysis traces cfDNA molecules back to the tissues and cell types they originated from and use changing tissue proportions to reveal altered tissue homeostasis in diseased states or during therapy.

this dimension reduction approach, the data is projected into lower-dimensional spaces, ultimately with the aim to improve prediction accuracy through increasing the signal-to-noise ratio. As previously described, the total makeup of cfDNA can be modeled as a complex mixture with TOO deconvolution analysis aiming to trace each individual cfDNA molecule back to its cellular origins as a non-invasive measure of tissue damage.

There are many computational methods that have been successfully applied to facilitate TOO deconvolution of cfDNA (**Table 2**). These include reference-based supervised learning models, which utilize labeled training and test datasets for classification tasks (Teschendorff et al., 2017). Commonly used methods include linear or logistic regression and random forests. In addition, one study uses Particle Swarm Optimization as a supervised global optimization method. While global optimization methods may be supervised, semi-supervised, or unsupervised, in this case the method is applied as a supervised learning model (Liu Y. et al., 2019). There are also several unsupervised learning models, including clustering and density estimation methods, in which the goal is to learn the inherent structure and relations of unlabeled data (Houseman et al., 2016). One advantage of unsupervised, reference-free algorithms is the ability to estimate contributions from unknown cell types, or cell types for which reference methylation data is not available. However, the biological meaning of the features selected in these models is often lost or difficult to interpret, making it more challenging to explain the relevance of results. Recently, deep learning has also been applied as a powerful modeling technique for deconvolution of DNA methylation data as these methods perform simultaneous feature extraction and classification (Levy et al., 2020; Menden et al., 2020). As a high-level overview, the computational methods for cell-mixture deconvolution can be generalized as adhering to the following format that is consistent across liquid biopsy applications. Initially, features are selected or extracted that can characterize variation among cell-type contributors in the circulation. Then, statistical models are built to estimate the mixing proportions of each cell type based on the reduced number of discriminative DNA methylation features selected. Typically, these models are trained using reference data where the mixing proportions are already known and then tested on datasets where the mixing proportions are unknown for evaluation (Feng et al., 2019). As a final step, predictive models can be developed after deconvolution, using the inferred cell-mixture proportions as predictors to estimate disease phenotypes.

Despite demonstrated success applying these computational models to cfDNA methylation deconvolution, these algorithms were originally designed to be learned from very large training datasets (Ko et al., 2018). In order to maintain predictive capabilities, modifications are necessary to optimize these models for working with smaller and often more diverse datasets, typical to the field of liquid biopsy. With this in mind, there are important biological properties of cell-free DNA methylation that can be leveraged toward this goal. First, in comparison to DNA in tissues that is artificially sheared for introduction to standard library preparation methods, the fragmentation patterns of cfDNA are biologically derived (Lo et al., 2021). The majority of cfDNA fragments are ∼167 bp, representing the length of DNA wrapped around a nucleosome and reflective of degradation by nucleases as a by-product of cell death. This fragmented nature of cfDNA lends itself to methods developed for characterizing cfDNA at the level of single molecules as opposed to population-level averages at single CpG sites (Li et al., 2018). Read-specific analysis allows for each read-pair to be modeled as an independent sample reflective of each individual cfDNA molecule. This allows the depth of sequencing to be utilized toward increasing sample size (Scott et al., 2020). In addition, the density of neighboring CpG sites varies across the human genome with highly dense organization defined as CpG islands. Methylation status at adjacent CpG sites is co-regulated in CpG islands due to the expanse of methylating and demethylating enzymes acting in the area (Marzese and Hoon, 2015). This co-dependency can be leveraged to increase specificity through modeling the methylation features selected with pattern analysis. DNA methylation detection technologies and computational approaches that take advantage of pattern analysis of individual cfDNA molecules have demonstrated to

be more robust and to increase the sensitivity and specificity of cell-type proportion estimations (Lehmann-Werman et al., 2016; Guo et al., 2017; Li et al., 2018; Zemmour et al., 2018; Chatterton et al., 2019). Overall, the biological relevance of selected DNA methylation markers and derived tissue proportions to disease can be utilized to inform analysis and model optimization.

## COMBINING GENETIC AND EPIGENETIC MARKERS TO NON-INVASIVELY MONITOR TREATMENT RESPONSE AND THERAPY-RELATED ADVERSE EVENTS

There is a need to identify predictive biomarkers for real-time monitoring of therapy-related adverse events relative to therapeutic efficacy. Combining changes to mutant ctDNA with altered proportions of cell-type specific cfDNA can reflect intervention-based changes (**Figure 6**; Erger et al., 2020; Guo et al., 2020; Wu et al., 2020). Therapy regimens for many cancers involve surgery, chemotherapy, radiotherapy, targeted therapy, and immunotherapy (Hofman et al., 2019). Each of these interventions can have a different systemic effect, and the ability to distinguish different cell types participating and potentially contributing to toxicities with cfDNA in serially drawn blood samples could significantly impact therapeutic decision making. Although imaging modalities can be used as an indirect way to gauge therapeutic efficacy, these results are often unreliable and difficult to interpret. Imaging results can be clouded by depictions of pseudoprogression, making them ineffective or crude instruments to monitor for concurrent changes necessary to guide therapy decisions (Maia et al., 2020). In contrast, the half-life of cfDNA is between 15 min and 2 h (Khier and Lohan, 2018). The rapid clearance allows for serial analysis of disease evolution over time, especially under selective pressures from ongoing therapy (O'Leary et al., 2018; Oellerich et al., 2019; Nabet et al., 2020; Peter et al., 2020). This technology allows for serial sampling to include a baseline comparison from which therapy-related relative changes may be assessed, taking into account patient specific co-morbidities at an individualized level.

Combining genetic and epigenetic analyses of cell-free DNA has many unique advantages when applied to precision therapeutics in cancer (Cheng T. H. et al., 2019; Zhang et al., 2019). Liquid biopsies have been shown to accurately characterize

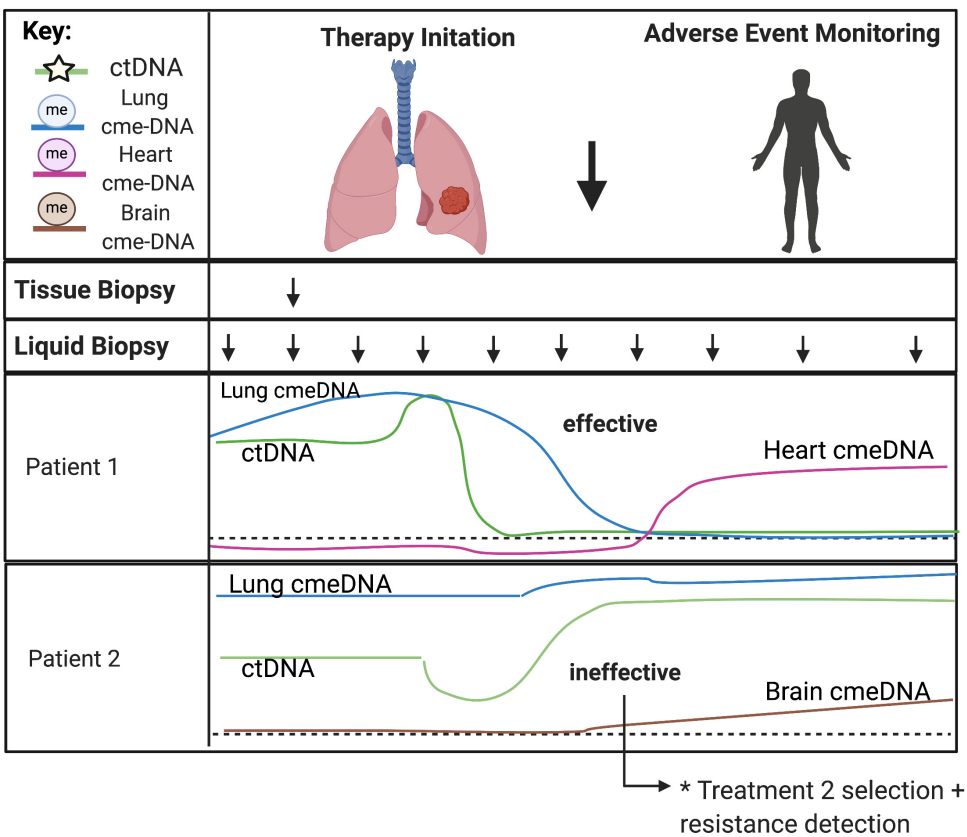



**FIGURE 6 |** Predicting treatment response and therapy-related toxicities from combined genetic and epigenetic analyses of cfDNA. The minimally invasive nature of liquid biopsies allows for serial sampling to monitor changes over time, especially under selective pressures from ongoing therapy. CtDNA can be used to track clonal heterogeneity over time to assess treatment response and detect treatment-resistant clones. Normal cell-specific cfDNA methylation patterns can be used in combination with ctDNA to assess the impact of treatment to the surrounding tumor microenvironment and to monitor for therapy-related toxicities in somatic cell-types. Acronyms: ctDNA, (circulating tumor DNA); cme-DNA, (circulating methylated cell-free DNA).

tumor genotypes and allow for molecular subtype classification to provide a comprehensive view of intratumor heterogeneity (Cohen et al., 2018; Christensen et al., 2019; Heitzer et al., 2019). High sampling frequency allows for modeling of evolutionary dynamics of tumor progression. Also, molecular changes identified after initiation of therapy can provide insight into therapy response as well as track tumor subclones that may lead to emergence of therapy resistance (Ahronian and Corcoran, 2017; Zhou et al., 2020). The systemic view provided by serial liquid biopsies is ideal to monitor widespread changes that may better inform clinical decision making in the face of uncertainty. For example, in the case of surgical removal of the tumor or therapeutic success, liquid biopsies can be used to monitor for minimal residual disease and recurrence. While ctDNA can be used to track molecular changes in the circulation, there is a benefit to monitoring the cancer-related changes to the host microenvironment in tandem requiring a combined genetic and epigenetic analysis. Cell-specific cfDNA methylation patterns of normal cells can be used in combination with ctDNA to assess the impact of treatment also on the surrounding tumor microenvironment. This is particularly useful to surveil for metastatic disease in distant tissue types from the primary tumor as well as to monitor for therapy-related toxicities in somatic cell types (Zhang et al., 2019). Further, liquid biopsies can help delineate factors that underlie clinical outcomes, providing a basis for recommending different treatments based on anticipated benefit to the patient. Liquid biopsies can identify predictive biomarkers to guide selection of treatment, recognize off-target effects, and develop individualized treatment plans for patients (Hofman et al., 2019). These applications provide a more complete picture of therapeutic response as well as tissue-specific cellular toxicity to better inform clinical care and management throughout the treatment process.

## FUTURE DIRECTIONS AND CONCLUSION

Liquid biopsies are rapidly emerging as an alternative and complementary approach to traditional solid tissue biopsies and have high utility for many applications in clinical oncology. Technology advances have made genome-wide profiling of circulating analytes possible and allow for transition from targeted to comprehensive approaches. With transition to "third-generation" liquid biopsies, DNA methylation patterns can be used to leverage signals from both tumor and non-tumor cells to increase signal abundance and discern biological relevance (Im et al., 2020). Despite great potential, comprehensive applications of liquid biopsy in oncology are still in their infancy. Additional large-scale, stratified, and randomized longitudinal studies are needed to begin to understand the complex interactions and biological significance of the comprehensive data identified from NGS technologies. Future work aimed at elucidating the biology of cell-free DNA release is needed to begin to control for co-morbidities and other confounding variables. Efforts aimed at assessing the effect of therapy regimens (chemo, radiation, immunotherapy, etc.) on tumor and non-tumor signals will become essential to determine what signals can be derived from the circulation. Tissue-of-origin analysis can be used to localize signals and generation of cell-type specific reference methylomes can improve specificity of features selected for application of TOO analysis in cancer (Moss et al., 2018; Barefoot et al., 2021). In addition, combining genetic and epigenetic markers may improve targeted-therapy selection and treatment response monitoring. These approaches are potentially synergistic, and future integration of signals across multiple genetic and epigenetic omics levels could fine-tune these applications for optimal use in precision oncology.

## AUTHOR CONTRIBUTIONS

MB and AW wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

Supported in part by funding from the National Institutes of Health United States: T32 CA009686 (MB), F30 CA250307 (MB), R01 CA231291 (AW), and P30 CA51008 (AW).

## ACKNOWLEDGMENTS

Valuable discussions and suggestions were contributed by Anne Deslattes Mays (Science and Technology Consulting LLC and Georgetown University), Habtom Ressom, Michael R. Lindberg, and Marcel O. Schmidt (all at Georgetown University). All figures in this manuscript were created with BioRender.com.

## REFERENCES

Ahronian, L. G., and Corcoran, R. B. (2017). Strategies for monitoring and combating resistance to combination kinase inhibitors for cancer therapy. *Genome Med.* 9:37. doi: 10.1186/s13073-017-0431-3

Barefoot, M. E., Lindberg, M. R., and Wellstein, A. (2021). Decoding the tissue of origin of cellular damage from cell-free dna in liquid biopsies. *Syst. Med.* 2, 365–378. doi: 10.1016/b978-0-12-801238-3.11669-1

Best, M., Sol, N., Kooi, I., Tannous, J., Westerman, B., Rustenburg, F., et al. (2015). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* 28, 666–676. doi: 10.1016/j.ccell.2015.09.018

Bronkhorst, A. J., Ungerer, V., and Holdenrieder, S. (2019). The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol. Detect. Quantif.* 17:100087. doi: 10.1016/j.bdq.2019.100087

Caggiano, C., Celona, B., Garton, F., Mefford, J., Black, B., Lomen-Hoerth, C., et al. (2020). Estimating the rate of cell type degeneration from epigenetic sequencing of cell-free DNA. *bioRxiv* [Preprint]. doi: 10.1101/2020.01.15.907022

Chatterton, Z., Mendelev, N., Chen, S., Raj, T., Walker, R., Carr, W., et al. (2019). Brain-derived circulating cell-free DNA defines the brain region and cell specific

origins associated with neuronal atrophy. *bioRxiv* [Preprint]. doi: 10.1101/538827

Cheng, A. P., Burnham, P., Lee, J. R., Cheng, M. P., Suthanthiran, M., Dadhania, D., et al. (2019). A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection. *Proc. Natl. Acad. Sci. U.S.A.* 116, 18738–18744. doi: 10.1073/pnas.1906320116

Cheng, A. P., Cheng, M. P., Gu, W., Sesing Lenz, J., Hsu, E., Schurr, E., et al. (2021). Cell-free DNA tissues of origin by methylation profiling reveals significant cell, tissue, and organ-specific injury related to COVID-19 severity. *Medicine* 2, 411.e5–422.e5. doi: 10.1016/j.medj.2021.01.001

Cheng, A. P., Cheng, M. P., Lenz, J. S., Chen, K., Burnham, P., Timblin, K. M., et al. (2020). Cell-free DNA Tissues-of-origin profiling to predict graft versus host disease and detect infection after hematopoietic cell transplantation. *bioRxiv* [Preprint]. doi: 10.1101/2020.04.25.061580

Cheng, T. H., Jiang, P., Tam, J. C., Sun, X., Lee, W., Yu, S. C., et al. (2017). Genomewide bisulfite sequencing reveals the origin and time-dependent fragmentation of urinary cfdna. *Clin. Biochem.* 50, 496–501. doi: 10.1016/j.clinbiochem.2017.02.017

Cheng, T. H., Jiang, P., Teoh, J. Y., Heung, M. M., Tam, J. C., Sun, X., et al. (2019). Noninvasive detection of bladder cancer by shallow-depth Genome-Wide Bisulfite sequencing of urinary Cell-Free DNA For methylation and copy number profiling. *Clin. Chem.* 65, 927–936. doi: 10.1373/clinchem.2018.301341

Christensen, E., Birkenkamp-Demtröder, K., Sethi, H., Shchegrova, S., Salari, R., Nordentoft, I., et al. (2019). Early detection of metastatic relapse and monitoring of therapeutic efficacy by ultra-deep sequencing of plasma cell-free DNA in patients with urothelial bladder carcinoma. *J. Clin. Oncol.* 37, 1547–1557. doi: 10.1200/jco.18.02052

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359, 926–930. doi: 10.1126/science.aar3247

Corcoran, R. B., and Chabner, B. A. (2018). Application of cell-free DNA analysis to cancer treatment. *N. Engl. J. Med.* 379, 1754–1765. doi: 10.1056/nejmra1706174

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., et al. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 570, 385–389. doi: 10.1038/s41586-019-1272-6

Daniūnaitė, K., Jarmalaite, S., and Kriukiene, E. (2019). Epigenomic technologies for deciphering circulating tumor DNA. *Curr. Opin. Biotechnol.* 55, 23–29. doi: 10.1016/j.copbio.2018.07.002

De Koker, A., Van Paemel, R., De Wilde, B., De Preter, K., and Callewaert, N. (2019). A versatile method for circulating cell-free DNA methylome profiling by reduced representation bisulfite sequencing. *bioRxiv* [Preprint]. doi: 10.1101/663195

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *Lancet* 392, 777–786. doi: 10.1016/s0140-6736(18)31268-6

Erger, F., Nörling, D., Borchert, D., Leenen, E., Habbig, S., Wiesener, M. S., et al. (2020). CfNOMe — a single assay for comprehensive Epigenetic analyses Of cell-free DNA. *Genome Med.* 12:54. doi: 10.1186/s13073-020-00750-5

Ewing, A. D., Smits, N., Sanchez-Luque, F. J., Faivre, J., Brennan, P. M., Richardson, S. R., et al. (2020). Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell* 80, 915.e5–928.e5. doi: 10.1016/j.molcel.2020.10.024

Feng, H., Jin, P., and Wu, H. (2019). Disease prediction by cell-free DNA methylation. *Brief Bioinform.* 20, 585–597. doi: 10.1093/bib/bby029

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465. doi: 10.1038/nmeth.1459

Gai, W., Ji, L., Lam, W. K., Sun, K., Jiang, P., Chan, A. W., et al. (2018). Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without Liver Metastases. *Clin. Chem.* 64, 1239–1249. doi: 10.1373/clinchem.2018.290304

Gai, W., and Sun, K. (2019). Epigenetic Biomarkers in Cell-Free DNA and applications in liquid biopsy. *Genes* 10:32. doi: 10.3390/genes10010032

Galardi, F., De Luca, F., Romagnoli, D., Biagioni, C., Moretti, E., Biganzoli, L., et al. (2020). Cell-Free DNA-Methylation-Based methods and applications in oncology. *Biomolecules* 10:1677. doi: 10.3390/biom10121677

Greenberg, M. V. C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* 20, 590–607. doi: 10.1038/s41580-019-0159-6

Guo, D., Yang, L., Yang, J., and Shi, K. (2020). Plasma cell-free DNA methylation combined with tumor mutation detection in prognostic prediction of patients with non-small cell lung cancer (Nsclc). *Medicine* 99:e20431. doi: 10.1097/md.0000000000020431

Guo, S., Diep, D., Plongthongkum, N., Fung, H. L., Zhang, K., and Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of- origin mapping from plasma DNA. *Nat. Genet.* 49, 635–642. doi: 10.1038/ng.3805

Haigis, K. M., Cichowski, K., and Elledge, S. J. (2019). Tissue-specificity in cancer: the rule, not the exception. *Science* 363, 1150–1151. doi: 10.1126/science.aaw3472

Han, D. S., Ni, M., Chan, R. W., Chan, V. W., Lui, K. O., Chiu, R. W., et al. (2020). The biology of cell-free DNA fragmentation and the roles of dnase1, dnase1l3, and dffb. *Am. J. Hum. Genet.* 106, 202–214. doi: 10.1016/j.ajhg.2020.01.008

Heitzer, E., Auinger, L., and Speicher, M. R. (2020). Cell-free DNA and apoptosis: how dead cells inform about the living. *Trends Mol. Med.* 26, 519–528. doi: 10.1016/j.molmed.2020.01.012

Heitzer, E., Haque, I. S., Roberts, C. E. S., and Speicher, M. R. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.* 20, 71–88. doi: 10.1038/s41576-018-0071-5

Heitzer, E., and Speicher, M. R. (2018). One size does not fit all: size-based plasma DNA diagnostics. *Sci. Transl. Med.* 10:eaav3873. doi: 10.1126/scitranslmed.aav3873

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291.e6–304.e6. doi: 10.1016/j.cell.2018.03.022

Hofman, P., Heeke, S., Alix-Panabières, C., and Pantel, K. (2019). Liquid biopsy in the era of immuno-oncology: is it ready for prime-time use for cancer patients? *Ann. Oncol.* 30, 1448–1459. doi: 10.1093/annonc/mdz196

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., et al. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13:86. doi: 10.1186/1471-2105-13-86

Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., and Marsit, C. J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* 17:259. doi: 10.1186/s12859-016-1140-4

Huang, C. C., Du, M., and Wang, L. (2019). Bioinformatics analysis for circulating cell-free DNA in cancer. *Cancers* 11:805. doi: 10.3390/cancers11060805

Huang, J., and Wang, L. (2019). Cell-free DNA methylation profiling analysis-technologies and bioinformatics. *Cancers* 11:1741. doi: 10.3390/cancers11111741

Im, Y. R., Tsui, D. W. Y., Diaz, L. A., and Wan, J. C. M. (2020). Next-generation liquid biopsies: embracing data science in oncology. *Trends Cancer* 7, 283–292. doi: 10.1016/j.trecan.2020.11.001

Ivanov, M., Chernenko, P., Breder, V., Laktionov, K., Rozhavskaya, E., Musienko, S., et al. (2019). Utility of cfdna fragmentation patterns in designing the liquid biopsy profiling panels to improve their sensitivity. *Front. Genet.* 10:194. doi: 10.3389/fgene.2019.00194

Jiang, P., Chan, K. C. A., and Lo, Y. M. D. (2019). Liver-derived cell-free nucleic acids in plasma: biology and applications in liquid biopsies. *J. Hepatol.* 71, 409–421. doi: 10.1016/j.jhep.2019.04.003

Jiang, P., and Lo, Y. M. D. (2016). The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. *Trends Genet.* 32, 360–371. doi: 10.1016/j.tig.2016.03.009

Jiang, P., Sun, K., Tong, Y. K., Cheng, S. H., Cheng, T. H., Heung, M. M., et al. (2018). Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10925–E10933. doi: 10.1073/pnas.1814616115

Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., et al. (2017). CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol.* 18:53. doi: 10.1186/s13059-017-1191-5

Khier, S., and Lohan, L. (2018). Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature. *Future Sci. OA* 4:FSO295.

Ko, J., Baldassano, S. N., Loh, P. L., Kording, K., Litt, B., and Issadore, D. (2018). Machine learning to detect signatures of disease in liquid biopsies - a user's guide. *Lab Chip* 18, 395–405. doi: 10.1039/c7lc00955k

Kurdyukov, S., and Bullock, M. (2016). DNA methylation analysis: choosing the right method. *Biology* 5:3. doi: 10.3390/biology5010003

Kustanovich, A., Schwartz, R., Peretz, T., and Grinshpun, A. (2019). Life and death of circulating cell-free DNA. *Cancer Biol. Ther.* 20, 1057–1067. doi: 10.1080/15384047.2019.1598759

Lam, W. K., Jiang, P., Chan, K. C., Peng, W., Shang, H., Heung, M. M., et al. (2019). Methylation analysis of plasma DNA informs etiologies of Epstein-barr virus-associated diseases. *Nat. Commun.* 10:3256. doi: 10.1038/s41467-019-11226-5

Lehmann-Werman, R., Magenheim, J., Moss, J., Neiman, D., Abraham, O., Piyanzin, S., et al. (2018). Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. *JCI Insight* 3:e120687. doi: 10.1172/jci.insight.120687

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., et al. (2016). Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1826–E1834. doi: 10.1073/pnas.1519286113

Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A., and Christensen, B. C. (2020). MethylNet: an automated and modular deep learning approach for dna methylation analysis. *BMC Bioinformatics* 21:108. doi: 10.1186/s12859-020-3443-8

Leygo, C., Williams, M., Jin, H. C., Chan, M. W., Chu, W. K., Grusch, M., et al. (2017). DNA methylation as a Noninvasive Epigenetic biomarker for the detection of cancer. *Dis. Markers* 2017, 1–13. doi: 10.1155/2017/3726595

Li, J., Zhao, S., Lee, M., Yin, Y., Li, J., Zhou, Y., et al. (2020). Reliable tumor detection by whole-genome methylation sequencing of cell-free DNA in cerebrospinal fluid of pediatric medulloblastoma. *Sci. Adv.* 6:eabb5427. doi: 10.1126/sciadv.abb5427

Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., et al. (2018). CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads Using cell-free DNA methylation sequencing data. *Nucleic Acids Res.* 46:e89. doi: 10.1093/nar/gky423

Liu, L., Toung, J., Jassowicz, A., Vijayaraghavan, R., Kang, H., Zhang, R., et al. (2018). Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification. *Ann. Oncol.* 29, 1445–1453. doi: 10.1093/annonc/mdy119

Liu, M., Oxnard, G., Klein, E., Swanton, C., Seiden, M., Liu, M. C., et al. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* 31, 745–759. doi: 10.1016/j.annonc.2020.02.011

Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C., and Wang, K. (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford nanopore sequencing data. *Nat. Commun.* 10:20. doi: 10.1038/s41467-019-10168-2

Liu, X., Ren, J., Luo, N., Guo, H., Zheng, Y., Li, J., et al. (2019). Comprehensive DNA methylation analysis of tissue of origin of plasma cell-free DNA by Methylated CpG Tandem amplification and sequencing (MCTA-Seq). *Clin. Epigenet.* 11:93. doi: 10.1186/s13148-019-0689-y

Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., et al. (2019). Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* 37, 424–429. doi: 10.1038/s41587-019-0041-2

Lo, Y. M. D., Han, D. S. C., Jiang, P., and Chiu, R. W. K. (2021). Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* 372:eaaw3616. doi: 10.1126/science.aaw3616

Luo, H., Wei, W., Ye, Z., Zheng, J., and Xu, R.-H. (2021). Liquid biopsy of methylation biomarkers in cell-free DNA. *Trends Mol. Med.* 27, 482–500. doi: 10.1016/j.molmed.2020.12.011

Maia, M. C., Salgia, M., and Pal, S. K. (2020). Harnessing cell-free DNA: plasma circulating tumour DNA for liquid biopsy in genitourinary cancers. *Nat. Rev. Urol.* 17, 271–291. doi: 10.1038/s41585-020-0297-9

Marzese, D. M., and Hoon, D. S. (2015). Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer. *Expert Rev. Mol. Diagn.* 15, 647–664. doi: 10.1586/14737159.2015.1027194

Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D. S., Kloiber, K., et al. (2020). Deep learning–based cell composition analysis from tissue expression profiles. *Sci. Adv.* 6:eaba2619. doi: 10.1126/sciadv.aba2619

Michalak, E. M., Burr, M. L., Bannister, A. J., and Dawson, M. A. (2019). The roles of dna, rna and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* 20, 573–589. doi: 10.1038/s41580-019-0143-1

Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* 9:5068. doi: 10.1038/s41467-018-07466-6

Moss, J., Zick, A., Grinshpun, A., Carmon, E., Maoz, M., Ochana, B., et al. (2020). Circulating breast-derived DNA allows universal detection and monitoring of localized breast cancer. *Ann. Oncol.* 31, 395–403. doi: 10.1016/j.annonc.2019.11.014

Nabet, B. Y., Esfahani, M. S., Moding, E. J., Hamilton, E. G., Chabon, J. J., Rizvi, H., et al. (2020). Noninvasive early identification of therapeutic benefit from immune checkpoint inhibition. *Cell* 183, 363.e13–376.e13. doi: 10.1016/j.cell.2020.09.001

Nassiri, F., Chakravarthy, A., Feng, S., Shen, S. Y., Nejad, R., Zuccato, J. A., et al. (2020). Detection and discrimination of intracranial tumors using plasma cell-free dna methylomes. *Nat. Med.* 26, 1044–1047. doi: 10.1038/s41591-020-0932-2

Oellerich, M., Schütz, E., Beck, J., and Walson, P. D. (2019). Circulating cell-free DNA—diagnostic and prognostic applications in personalized cancer therapy. *Ther. Drug Monit.* 41, 115–120. doi: 10.1097/ftd.0000000000000566

O'Leary, B., Hrebien, S., Morden, J. P., Beaney, M., Fribbens, C., Huang, X., et al. (2018). Early circulating tumor DNA dynamics and clonal selection with palbociclib and fulvestrant for breast cancer. *Nat. Commun.* 9:896. doi: 10.1038/s41467-018-03215-x

Olkhov-Mitsel, E., and Bapat, B. (2012). Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. *Cancer Med.* 1, 237–260. doi: 10.1002/cam4.22

Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R., et al. (2018). Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* 19:33. doi: 10.1186/s13059-018-1408-2

Panagopoulou, M., Karaglani, M., Balgkouranidou, I., Biziota, E., Koukaki, T., Karamitrousis, E., et al. (2019). Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. *Oncogene* 38, 3387–3401. doi: 10.1038/s41388-018-0660-y

Peter, M. R., Bilenky, M., Isserlin, R., Bader, G. D., Shen, S. Y., De Carvalho, D. D., et al. (2020). Dynamics of the cell-free DNA methylome of metastatic prostate cancer during androgen-targeting treatment. *Epigenomics* 12, 1317–1332. doi: 10.2217/epi-2020-0173

Rapisuwon, S., Vietsch, E. E., and Wellstein, A. (2016). Circulating biomarkers to monitor cancer progression and treatment. *Comput. Struct. Biotechnol. J.* 14, 211–222. doi: 10.1016/j.csbj.2016.05.004

Rostami, A., Lambie, M., Yu, C. W., Stambolic, V., Waldron, J. N., and Bratman, S. V. (2020). Senescence, necrosis, and apoptosis govern circulating cell-free DNA release kinetics. *Cell Rep.* 31:107830. doi: 10.1016/j.celrep.2020.107830

Sadeh, R., Sharkia, I., Fialkoff, G., Rahat, A., Gutin, J., Chappleboim, A., et al. (2021). Chip-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. *Nat. Biotechnol.* 39, 586–598. doi: 10.1038/s41587-020-00775-6

Sanchez, C., Snyder, M. W., Tanos, R., Shendure, J., and Thierry, A. R. (2018). New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. *NPJ Genom. Med.* 3:31.

Schutsky, E. K., DeNizio, J. E., Hu, P., Liu, M. Y., Nabel, C. S., Fabyanic, E. B., et al. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a dna deaminase. *Nat. Biotechnol.* 36, 1083–1090. doi: 10.1038/nbt.4204

Scott, C. A., Duryea, J. D., MacKay, H., Baker, M. S., Laritsky, E., Gunasekara, C. J., et al. (2020). Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol.* 21:156. doi: 10.1186/s13059-020-02065-5

Serpas, L., Chan, R. W., Jiang, P., Ni, M., Sun, K., Rashidfarrokhi, A., et al. (2018). Dnase1l3 deletion causes aberrations in length and end-motif frequencies in

plasma DNA. *Proc. Natl. Acad. Sci. U.S.A.* 116, 641–649. doi: 10.1073/pnas.1815031116

Shen, S. Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M. H., Chadwick, D., et al. (2018). Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563, 579–583. doi: 10.1038/s41586-018-0703-0

Sina, A. A. I., Carrascosa, L. G., and Trau, M. (2019). DNA methylation-based point-of-care cancer detection: challenges and possibilities. *Trends Mol. Med.* 25, 955–966. doi: 10.1016/j.molmed.2019.05.014

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., and Shendure, J. (2016). Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. *Cell* 164, 57–68. doi: 10.1016/j.cell.2015.11.050

Song, C., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., et al. (2017). 5-Hydroxymethylcytosine signatures in cell-free dna provide information about tumor types and stages. *Cell Res.* 27, 1231–1242. doi: 10.1038/cr.2017.106

Sun, K., Jiang, P., Chan, K. C., Wong, J., Cheng, Y. K., Liang, R. H., et al. (2015). Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5503–E5512. doi: 10.1073/pnas.1508736112

Sun, K., Jiang, P., Cheng, S. H., Cheng, T. H., Wong, J., Wong, V. W., et al. (2019). Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.* 29, 418–427. doi: 10.1101/gr.242719.11

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* 18:105. doi: 10.1186/s12859-017-1511-5

Tse, O. Y., Jiang, P., Cheng, S. H., Peng, W., Shang, H., Wong, J., et al. (2021). Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2019768118. doi: 10.1073/pnas.2019768118

Ulz, P., Thallinger, G. G., Auer, M., Graf, R., Kashofer, K., Jahn, S. W., et al. (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.* 48, 1273–1278. doi: 10.1038/ng.3648

Vaisvila, R., Ponnaluri, V. K., Sun, Z., Langhorst, B. W., Saleh, L., Guan, S., et al. (2019). EM-seq: detection of DNA methylation at single Base resolution from picograms of DNA. *bioRxiv* [Preprint]. doi: 10.1101/2019.12.20.884692

Van der pol, Y., and Mouliere, F. (2019). Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA. *Cancer Cell* 36, 350–368. doi: 10.1016/j.ccell.2019.09.003

Wong, F. C., Sun, K., Jiang, P., Cheng, Y. K., Chan, K. A., Leung, T. Y., et al. (2016). Cell-free DNA in maternal plasma and serum: a comparison of quantity, quality and tissue origin using genomic and epigenomic approaches. *Clin. Biochem.* 49, 1379–1386. doi: 10.1016/j.clinbiochem.2016.09.009

Wu, A., Cremaschi, P., Wetterskog, D., Conteduca, V., Franceschini, G. M., Kleftogiannis, D., et al. (2020). Genome-wide plasma DNA methylation features of metastatic prostate cancer. *J. Clin. Investig.* 130, 1991–2000. doi: 10.1172/jci130887

Yuen, Z. W.-S., Srivastava, A., Daniel, R., McNevin, D., Jack, C., and Eyras, E. (2020). Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Bioinformatics* 12:3438.

Zemmour, H., Planer, D., Magenheim, J., Moss, J., Neiman, D., Gilon, D., et al. (2018). Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. *Nat. Commun.* 9:1443. doi: 10.1038/s41467-018-03961-y

Zhang, L., Liang, Y., Li, S., Zeng, F., Meng, Y., Chen, Z., et al. (2019). The interplay of circulating tumor DNA and chromatin modification, therapeutic resistance, and metastasis. *Mol. Cancer* 18:36. doi: 10.1186/s12943-019-0989-z

Zhao, L.-Y., Song, J., Liu, Y., Song, C.-X., and Yi, C. (2020). Mapping the epigenetic modifications of DNA and RNA. *Protein Cell* 11, 792–808. doi: 10.1007/s13238-020-00733-7

Zheng, S. C., Webster, A. P., Dong, D., Feber, A., Graham, D. G., Sullivan, R., et al. (2018). A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics* 10, 925–940. doi: 10.2217/epi-2018-0037

Zhou, Q., Perakis, S. O., Ulz, P., Mohan, S., Riedl, J. M., Talakic, E., et al. (2020). Cell-free DNA analysis reveals Polr1d-mediated resistance to bevacizumab in colorectal cancer. *Genome Med.* 12:20. doi: 10.1186/s13073-020-0719-6

**Conflict of Interest:** Georgetown University filed a patent related to some of the approaches described in this manuscript. MB and AW are named as inventors on this application and declare that as a potential conflict of interest.

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

REVIEW
published: 02 August 2021
doi: 10.3389/fgene.2021.687813

# Mechanism-Centric Approaches for Biomarker Detection and Precision Therapeutics in Cancer

*Christina Y. Yu[1] and Antonina Mitrofanova[1,2]**

[1] Department of Biomedical and Health Informatics, School of Health Professions, Rutgers, The State University of New Jersey, Newark, NJ, United States, [2] Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

Biomarker discovery is at the heart of personalized treatment planning and cancer precision therapeutics, encompassing disease classification and prognosis, prediction of treatment response, and therapeutic targeting. However, many biomarkers represent passenger rather than driver alterations, limiting their utilization as functional units for therapeutic targeting. We suggest that identification of driver biomarkers through mechanism-centric approaches, which take into account upstream and downstream regulatory mechanisms, is fundamental to the discovery of functionally meaningful markers. Here, we examine computational approaches that identify mechanism-centric biomarkers elucidated from gene co-expression networks, regulatory networks (e.g., transcriptional regulation), protein–protein interaction (PPI) networks, and molecular pathways. We discuss their objectives, advantages over gene-centric approaches, and known limitations. Future directions highlight the importance of input and model interpretability, method and data integration, and the role of recently introduced technological advantages, such as single-cell sequencing, which are central for effective biomarker discovery and time-cautious precision therapeutics.

**Keywords: biomarkers, treatment response, precision medicine, predictive models, mechanism-centric approaches**

OPEN ACCESS

***Edited by:***
*Rosalba Giugno,*
*University of Verona, Italy*

***Reviewed by:***
*Vittorio Fortino,*
*University of Eastern Finland, Finland*
*Sailu Yellaboina,*
*CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, India*

***Correspondence:**
*Antonina Mitrofanova*
*amitrofa@shp.rutgers.edu*

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

***Received:*** *30 March 2021*
***Accepted:*** *28 June 2021*
***Published:*** *02 August 2021*

***Citation:***
*Yu CY and Mitrofanova A (2021) Mechanism-Centric Approaches for Biomarker Detection and Precision Therapeutics in Cancer. Front. Genet. 12:687813. doi: 10.3389/fgene.2021.687813*

## INTRODUCTION

In the past two decades, the advancement of high-throughput technologies has led to the discovery of genomic, transcriptomic, and epigenomic modalities involved in cancer initiation, progression, and treatment response. Multiple groups have started to effectively utilize molecular data produced by high-throughput oncology experiments to identify biomarkers of progression and therapeutic response in cancer patients (Sorlie et al., 2001; Zhang et al., 2001; van't Veer et al., 2002; Zhan et al., 2002, 2006; Sotiriou et al., 2003; Ayers et al., 2004; Allen et al., 2006; Jain et al., 2009; Lim et al., 2009; Petty et al., 2009; Zhao et al., 2009; Carro et al., 2010; Lefebvre et al., 2010; Shaughnessy et al., 2011; Bae et al., 2013; Aytes et al., 2014, 2018; Mitrofanova et al., 2015; Robinson et al., 2015; Wang et al., 2016; Giulietti et al., 2017; Heng et al., 2017; Hoadley et al., 2018; Abida et al., 2019; Epsi et al., 2019; Arriaga et al., 2020; Panja et al., 2020; Rahem et al., 2020). Yet, our understanding of the mechanisms involving these modalities, their upstream regulation, and effective therapeutic targeting remains incomplete.

A biomarker is an objective measure (e.g., classically a genomic/transcriptomic/epigenomic alteration, gene, protein, metabolite, or their groups), typically used to predict the incidence of disease, its progression, or treatment outcome (Strimbu and Tavel, 2010; McDermott et al., 2013). In the context of oncology, biomarkers are classically used for cancer risk assessment and screening, tumor staging, disease recurrence, selection of initial therapy, alternative therapy choices, and monitoring for therapeutic toxicities (Ludwig and Weinstein, 2005). While employed in clinical use, the existing biomarkers are still sparse and suffer from issues of reproducibility and heterogeneity, alongside a lack of understanding of their underlying regulatory mechanisms (Ludwig and Weinstein, 2005; Boutros, 2015).

One of the reasons for such a knowledge gap is the fact that the majority of biomarkers are identified from *gene-centric* approaches (we will refer to gene/protein/metabolite etc.,-centric approaches as gene-centric approaches for simplicity), where either a specific gene is investigated (based on previous biological assumptions) or a gene(s) is selected based on differential behavior without connection to the upstream and downstream molecular mechanisms. Gene-centric findings are often limited in mechanistic interpretability and connectivity to other molecular processes, positioning such biomarkers as passengers, rather than drivers, of the biological process and thus are often dataset specific (Michiels et al., 2005; Chng et al., 2016).

In classical gene-centric approaches, genes (without their connections to one another or underlying mechanisms) are utilized as inputs into white- and black-box statistical and machine learning models, which have been successfully applied to identify gene-centric markers in breast cancer (van't Veer et al., 2002; Wang et al., 2005; Zhang et al., 2013), lung cancer (Beer et al., 2002), multiple myeloma (Shaughnessy et al., 2007; Kuiper et al., 2012), colon cancer (Zhang et al., 2001; Yan et al., 2012), and prostate cancer (Garzotto et al., 2005; Erho et al., 2013), among many others. It is important to note that in white-box models (e.g., linear regression and decision trees) the relationship between input variables (i.e., genes) and output variables (i.e., disease outcomes) is understandable/explainable as they often identify linear or monotonic relationships (Zhang et al., 2001; Garzotto et al., 2005; Rosenfeld et al., 2008; Huo et al., 2017; Panja et al., 2018). On the other hand, black-box models (e.g., neural networks, gradient boosting, or ensemble models such as random forest) are able to capture non-linear/non-monotonic relationships, yet often suffer from model interpretability and subsequent limited clinical adoption (Wang et al., 2009; Ayer et al., 2010; Zhang et al., 2013). Even though both white- and black-box learning are excellent tools for predictive modeling, they mostly capture associative relationships when applied as gene-centric approaches and often miss the complexity of mechanisms inherent in biological systems, especially in the context of cancer.

Several groups have addressed this problem by developing biomarker discovery methods based on *mechanism-centric* approaches, which are not focused on single genes and take into account complex mechanisms implicated in cancer initiation, progression, and treatment response. In this review, we will discuss the mechanism-centric approaches based on construction and mining of co-expression networks (Freeman, 1977; Zhang and Horvath, 2005; Zhang and Huang, 2014; Han et al., 2016), regulatory networks (Basso et al., 2005; Lefebvre et al., 2010; Alvarez et al., 2016; Dhingra et al., 2017), protein–protein interaction (PPI) networks (Chuang et al., 2007), and molecular pathways (Epsi et al., 2019; Rahem et al., 2020; **Figure 1**). Through an in-depth understanding of upstream and downstream molecular mechanisms, such techniques open a door for the discovery of functionally interpretable molecular drivers (rather than passengers) and potential targets for precision therapeutics.

# MECHANISM-CENTRIC COMPUTATIONAL APPROACHES FOR BIOMARKER DISCOVERY

## Gene Co-expression Network Analysis

Gene co-expression networks define groups of genes that show similar/related expression patterns across an entire dataset. Highly associated genes are clustered together into modules, with the underlying rationale that co-expressed genes are likely to be co-regulated. We depict two methods, weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) and local maximal Quasi-Clique Merger (lmQCM) (Zhang and Huang, 2014), for network construction and module detection. Identified modules are defined as tightly connected groups of genes (potentially protein/gene complexes), which are then associated with clinical features to determine functionally relevant molecular structures. We also describe methods to mine such co-expression networks that include condition-specific network mining (Han et al., 2016), eigengene association (Alter et al., 2000; Zhang and Horvath, 2005), and network connectivity/hub analysis (Freeman, 1977).

### Network Construction: WGCNA and lmQCM

In general, co-expression network construction is based on a similarity matrix that describes the measure of association between a gene to all other genes (the simplest of similarity measures being correlation) (**Figure 2A**). An undirected network is constructed from the similarity matrix and is comprised of nodes denoting genes and edges denoting the associations (e.g., correlation) between genes.

One of the most well-known methods for gene co-expression network reconstruction is WGCNA, which was one of the earliest methods that proposed using weighted networks (**Figure 2B**; Zhang and Horvath, 2005). The advantage of weighted, compared to unweighted, network construction is the ability to assign meaningful weights to relationships/edges, which eliminates a need for threshold assignment and prevents information loss. WGCNA calculates correlation between pairs of genes and transforms the correlation measure into a topological overlap measure in order to minimize effects of noise and spurious associations. The resulting matrix is subjected to hierarchical clustering to determine groups of co-expressed genes, also

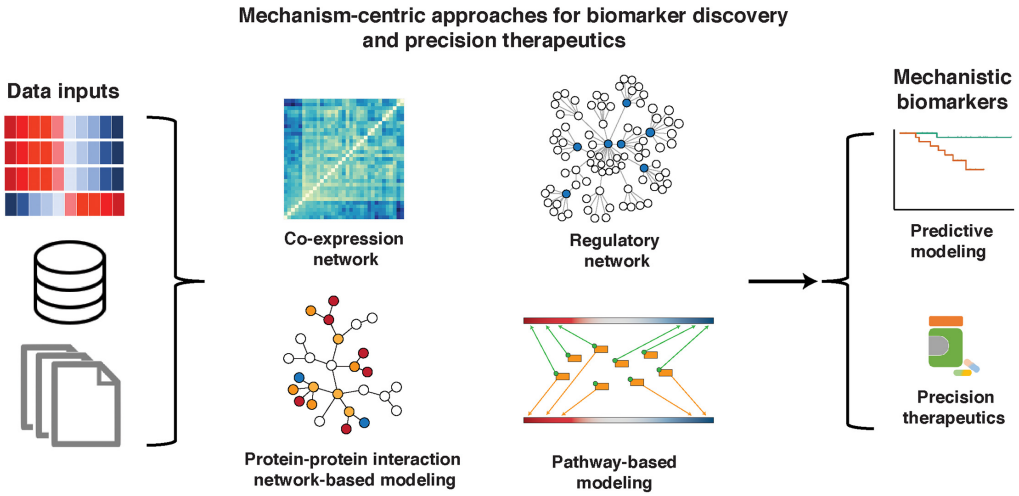



**FIGURE 1 |** Mechanism-centric approaches in biomarker discovery and precision therapeutics. A variety of data, including single- and multi-omic sources, knowledge bases, and phenotype/clinical information, can be used as inputs to mechanism-centric approaches to identify functional biomarkers of disease and therapeutic response. We describe mechanism-centric methods that are based on co-expression networks, regulatory networks, PPI networks, and molecular pathways.

referred to as gene modules. An R package for WGCNA is freely available (Langfelder and Horvath, 2008).

Because WGCNA module identification is based on hierarchical clustering, genes cannot be assigned to multiple modules, exposing WGCNA's limitation since many genes participate in multiple biological processes and often perform multiple functions. An alternative weighted co-expression method which allows genes to have multiple co-memberships in different modules is lmQCM (**Figure 2C**; Zhang and Huang, 2014). The lmQCM algorithm identifies densely connected subnetworks (i.e., quasi-cliques) using a greedy search algorithm which allows module overlaps (Ou and Zhang, 2007). In addition to allowing genes to be assigned to multiple modules, lmQCM can also identify smaller modules, which can highlight more specific and interpretable biological connections as compared to much larger modules of WGCNA that frequently contain over a thousand genes (Zhang and Huang, 2014; Yu et al., 2019). This algorithm is freely available as an R package[1] and a web-tool (Huang et al., 2021).

## Network Mining: Centered Concordance Index, Eigengenes, and Hubs

Co-expression networks can be mined to determine the functional significance of their modules or identify functionally relevant genes. Here, we discuss two techniques for module mining [Centered Concordance Index (CCI) (Han et al., 2016) and eigengenes (Alter et al., 2000; Horvath and Dong, 2008)] and two techniques to identify hub genes [intramodular connectivity (Zhang and Horvath, 2005) and betweenness centrality (Freeman, 1977)].

Centered Concordance Index has been developed to identify modules specific to each condition/phenotype. In particular, the CCI evaluates the concordance of gene expression profiles within a module based on singular value decomposition and is used to identify modules that are highly co-expressed in one condition over another (Han et al., 2016). Han et al. (2016) and Yu et al. (2019), respectively, identified several gene modules specific to lung adenocarcinoma and multiple myeloma precursors compared to non-cancer controls. The CCI is useful in identifying modules specific to phenotype conditions but has yet to be used to associate modules with continuous outcomes.

The eigengene approach transforms modules into weighted vectors, which mathematically correspond to their contribution to the first principal component in principal component analysis (Alter et al., 2000; Horvath and Dong, 2008). Eigengenes are then able to be associated with clinical features (including continuous outcomes) using correlation/association measures. For instance, Liu et al. (2015a) used the eigengene approach to identify two modules significantly associated with poor outcome in ER + breast cancer patients treated with tamoxifen. Liu et al. (2015b) and Zhang J. et al. (2020) associated module eigengenes derived from breast cancer patient data with clinical features such as survival status, tumor metastasis, and chemotherapy response. Han et al. (2019) identified module eigengenes strongly associated with patient survival in neuroblastoma.

The translational applicability of modules can be hampered by their relatively large size and might benefit from identification of hub genes within modules. Several measures have been developed to identify hubs, including intramodular connectivity and betweenness centrality. In particular, intramodular connectivity for gene $i$ is defined as the sum of edge weights between gene $i$ and the other genes in the module (Zhang and Horvath, 2005). Genes with the highest connectivity are considered hub genes and have been shown to play key roles in maintaining essential cellular functions (Jeong et al., 2001) and significantly associated with patient survival in breast cancer (Liu et al., 2015a;

[1]https://cran.r-project.org/package=lmQCM

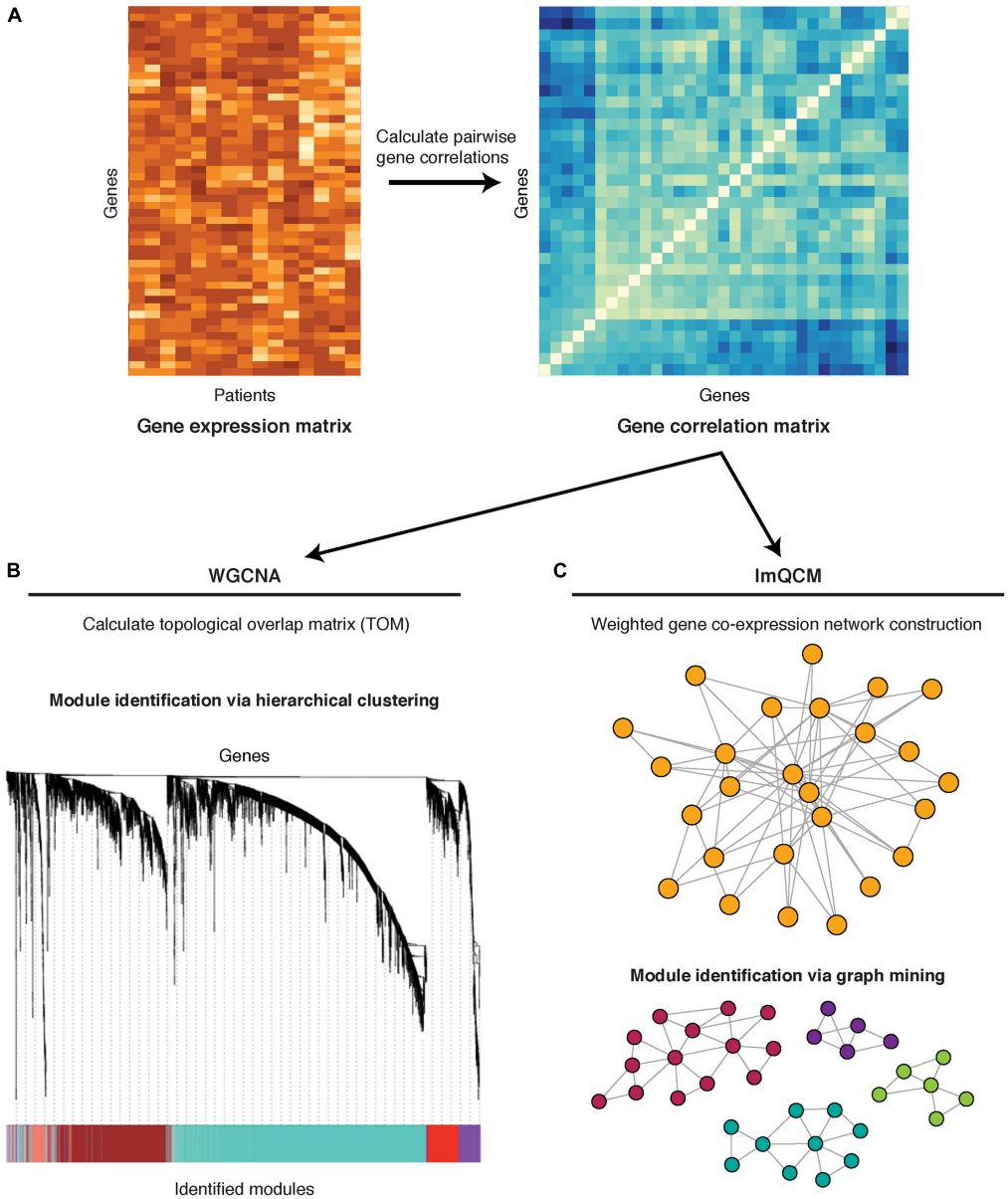



**FIGURE 2 |** Co-expression network methods: WGNCA and lmQCM. **(A)** Pairwise gene correlations are calculated from gene expression (microarray or RNA-seq) data. **(B)** The co-expression matrix is transformed into a topological overlap matrix and subjected to hierarchical clustering for module identification. A cluster dendrogram is shown, with different gene modules identified by the color bar on the bottom. **(C)** The co-expression matrix is used to construct a network, with genes as nodes and the correlation co-efficient between any two genes as the edge weight. Module identification is achieved through a greedy search for highly correlated subnetworks.

Tang et al., 2018; Jia et al., 2020; Tian et al., 2020; Zhang J. et al., 2020), glioblastoma (Horvath et al., 2006; Yang et al., 2018; Tang et al., 2019), hepatocellular carcinoma (Hu et al., 2020; Song et al., 2020), and pancreatic ductal adenocarcinoma (Giulietti et al., 2016), among others. Some of these findings have been experimentally validated, such as the ASPM hub gene in glioblastoma (Horvath et al., 2006) and FAM171A1, NDFIP1, SKP1, and REEP5 hub genes in breast cancer (Tian et al., 2020).

An alternative measure to identify hub genes is betweenness centrality, which is a network topology metric used to identify central nodes in a graph based on a shortest paths algorithm (Freeman, 1977). The betweenness centrality of gene *i* is a measure of the number of shortest paths connecting any two genes which pass through *i*. Genes with the highest betweenness scores are considered hubs and are believed to play an important role in information transfer within the network. For instance, Wang et al. analyzed modules with the betweenness centrality measure to identify eight hub genes that were significantly associated with overall survival in breast cancer patients (Wang C. C. N. et al., 2019).

## Regulatory Network Analysis

In recent years, molecular regulatory networks have received much attention from the scientific community due to their ability to capture complexity of molecular interactions present in cancer context-specific tissues (Butte and Kohane, 2000; Butte et al., 2000; Friedman et al., 2000; Basso et al., 2005; Margolin et al., 2006a,b; Werhli and Husmeier, 2007; Huynh-Thu et al., 2010; Lefebvre et al., 2010; Aytes et al., 2014). Regulatory networks define regulatory relationships between regulators (e.g., transcriptional regulators, splicing regulators, post-translational regulators, etc.), and their potential targets (e.g., genes, proteins, etc.). Such regulatory relationships provide key information about upstream and downstream regulations to infer cellular mechanisms for creating potential causal models of disease and outperform co-expression networks in their interpretability and functionally relevant determinants. Several methods have tackled reconstruction of regulatory networks using mutual information (Butte and Kohane, 2000; Basso et al., 2005; Margolin et al., 2006a), Bayesian networks (Friedman et al., 2000; Werhli and Husmeier, 2007), and regression trees (Huynh-Thu et al., 2010), to name a few. Readers are encouraged to consult the following reviews for a comprehensive overview of the different computational underpinnings employed in regulatory network analysis (Markowetz and Spang, 2007; Karlebach and Shamir, 2008; Hecker et al., 2009; Lee and Tzou, 2009; Emmert-Streib et al., 2014). Here, we focus on transcriptional [Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNe) (Margolin et al., 2006a)] and multi-omic [RegNetDriver (Dhingra et al., 2017)] regulatory networks and their mining [i.e., Master Regulator Inference Algorithm (MARINa) (Lefebvre et al., 2010), Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) (Alvarez et al., 2016), etc.] in the context of cancer biomarker studies.

### Transcriptional Regulatory Networks

The role of transcriptional regulation has been widely studied in cancer, including discovery of MYC (Gabay et al., 2014), Sox2 (Boumahdi et al., 2014), and the FOXO family (Jiramongkol and Lam, 2020) as important players in cancer initiation and progression. Transcriptional regulatory networks depict interactions between transcription factors (TFs)/co-factors (co-TFs) and their transcriptional targets, allowing the study of differential behavior in transcriptional machinery that govern oncogenic process.

#### *Network construction: ARACNe*

One of the most known and widely experimentally validated methods for transcriptional network reconstruction is ARACNe (Margolin et al., 2006a,b). This information-theoretic algorithm utilizes tissue-specific gene expression profiles to estimate pairwise mutual information between expression levels of TFs/co-TFs and expression levels of their potential (activated or repressed) targets. The advantage of using mutual information to measure such relationships lies in its ability to measure not only linear (which would be captured for example by the Pearson correlation) or monotonic (which would be captured for example by Spearman correlation) relationships, but also non-linear associations. Another novelty in transcriptional network reconstruction is introduced by the data processing inequality, which eliminates any "indirect" regulatory relationship through the principle that mutual information on the indirect path cannot exceed mutual information on any part of the direct path. Data processing inequality results in a regulatory network that includes primarily direct TF/co-TF-target interactions. ARACNe has been widely applied to several normal physiological and pathological conditions, including B-cell interactome (Basso et al., 2005), breast cancer (Lim et al., 2009; Remo et al., 2015; Walsh et al., 2017), prostate cancer (Aytes et al., 2014), colorectal cancer (Bae et al., 2013; Cordero et al., 2014; Sanz-Pamplona et al., 2014; Eskandari et al., 2018), glioma (Carro et al., 2010), T-cell acute lymphoblastic leukemia (Palomero et al., 2006), and multiple myeloma (Agnelli et al., 2011), among others. Software for ARACNe is freely available for download.[2]

#### *Network mining: MARINa and VIPER*

The ARACNe network can be effectively interrogated (i.e., mined) using MARINa (Lefebvre et al., 2010) and VIPER (Alvarez et al., 2016), two algorithms that identify TFs/co-TFs as driver biomarkers associated with specific phenotypes (e.g., cancer initiation, cancer progression, metastasis, treatment response, etc.). Specifically, MARINa (Lim et al., 2009; Lefebvre et al., 2010) requires a differentially expressed signature, defined as a ranked list of genes between any two phenotypes of interest. Then, the activated and repressed targets for each TF/co-TF (as inferred by ARACNe) are assessed for their enrichment in the over- and under-expressed parts of this signature (Lefebvre et al., 2010; **Figure 3**). Such enrichment is referred to as TF/co-TF transcriptional activity, and if it is statistically significant, the TF/co-TF is referred to as a Master Regulator (MR). As a result of this analysis, a TF/co-TF is considered an "activated" MR if its activated targets are significantly enriched in the over-expressed part of the signature and/or its repressed targets are significantly enriched in the under-expressed part of the signature. Conversely, a "repressed" MR exhibits the opposite behavior. It is important to note that TF/co-TF transcriptional activity is not defined based on the differential expression of TFs/co-TFs themselves but instead on the differential expression of their transcriptional targets. This allows the identification of TFs/co-TFs that are not necessarily differentially expressed but are modified on the post-translational level and would otherwise be missed by traditional association methods.

Master Regulator Inference Algorithm has successfully identified MRs in various cancers, including prostate cancer (Aytes et al., 2014, 2018; Mitrofanova et al., 2015; Talos et al., 2017), breast cancer (Lim et al., 2009; Fletcher et al., 2013; Remo et al., 2015), pancreatic cancer (Sartor et al., 2014), ovarian cancer (Zhang et al., 2015), glioma (Carro et al., 2010; Sonabend et al., 2014), T cell acute lymphoblastic leukemia (Della Gatta et al., 2012), and diffuse large B cell lymphoma (Ying et al., 2013; Bisikirska et al., 2016). These biomarkers also serve as valuable therapeutic targets and their silencing could potentially

[2]http://califano.c2b2.columbia.edu/aracne

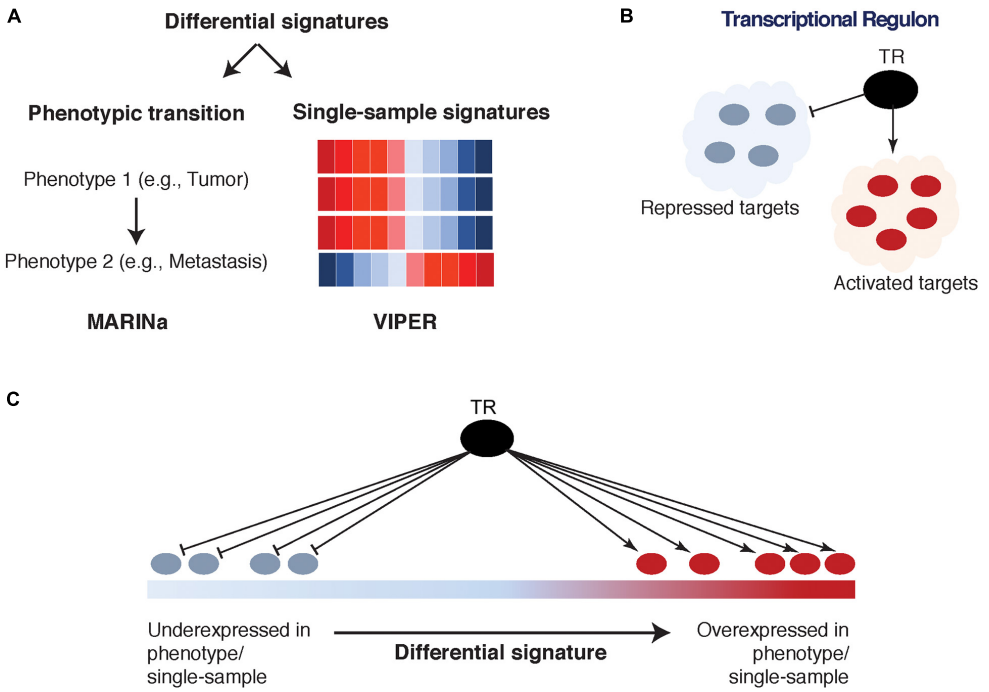



**FIGURE 3 |** Interrogation of transcriptional regulatory networks: Master Regulator Inference Algorithm (MARINa) and Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER). **(A)** A differential signature is defined between two phenotypes of interest (left) as input to MARINa; or on a single-sample level (right) as input to VIPER. **(B)** The transcriptional regulon is identified from Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNe) tissue-specific transcriptional regulatory network, which includes a transcriptional regulator (TR) and its activated and repressed targets. **(C)** The activated and repressed targets of the regulon are mapped onto the corresponding signature and used to determine the TR's transcriptional activity.

have a significant effect on inhibition of malignant phenotype. To this extent, Mitrofanova et al. developed a computational algorithm to predict drug combinations that inhibit activity levels of FOXM1 and CENPF (MRs in malignant prostate cancer) and demonstrated that their therapeutic inhibition significantly improved cancer course (Mitrofanova et al., 2015). MARINa is freely available for download.[3]

At the same time, VIPER estimates TF/co-TF transcriptional activity on an individual sample-based level, as opposed to a two-phenotype signature-based level required by MARINa (Alvarez et al., 2016; **Figure 3**). In fact, while MARINa requires carefully selected multiple samples of the same phenotype to construct a differential expression signature, VIPER is able to utilize single-sample analysis by scaling the overall patient cohort (to its average expression for each gene). Furthermore, several advantages of VIPER include estimation of TF/co-TF activity through a so-called mode of regulation (taking into account whether targets are activated, repressed, or their direction cannot be determined), inference of regulator-target interaction confidence, and accounting for target overlap between different regulators (Alvarez et al., 2016). VIPER was shown to accurately infer aberrant oncoprotein activity induced by somatic mutations, across multiple cancer types (Alvarez et al., 2016). An R package is freely available.[4]

[3]http://califano.c2b2.columbia.edu/marina
[4]http://doi.org/10.18129/B9.bioc.viper

## Multi-Omic Regulatory Network

Multi-omic data integration is another avenue to improve interpretability and discovery of functionally relevant biomarkers. Integration of different data modalities can increase the confidence of the overall findings since gene regulation is a complex process affected by multiple factors, such as gene mutations, structural variants, epigenomics, and more.

### *Network construction: RegNetDriver, step I*

RegNetDriver is an algorithm for multi-omic tissue-specific regulatory network construction and analysis (Dhingra et al., 2017; **Figure 4**). The regulatory network reconstructed by RegNetDriver represents a two-layered relationship: (i) connecting TFs to promoter/enhancer regions; and (ii) further connecting promoter/enhancer regions to their corresponding target genes. To reconstruct relationships between TFs and promoters/enhancers of potential targets, Dhingra et al. utilize tissue-specific (i.e., prostate epithelium) DNase I hypersensitive sites to define accessible regulatory DNA regions and integrate this information with promoter/enhancer annotations from ENCODE (Encode Project Consortium, 2012) and GENCODE (Harrow et al., 2012). TFs are then connected to promoters/enhancers based on the enrichment of their binding motifs. Promoters/enhancers are further connected to their target genes through significant correlation of promoter/enhancer region activity signals (estimated using bisulfite sequencing and ChIP-seq data) with target gene

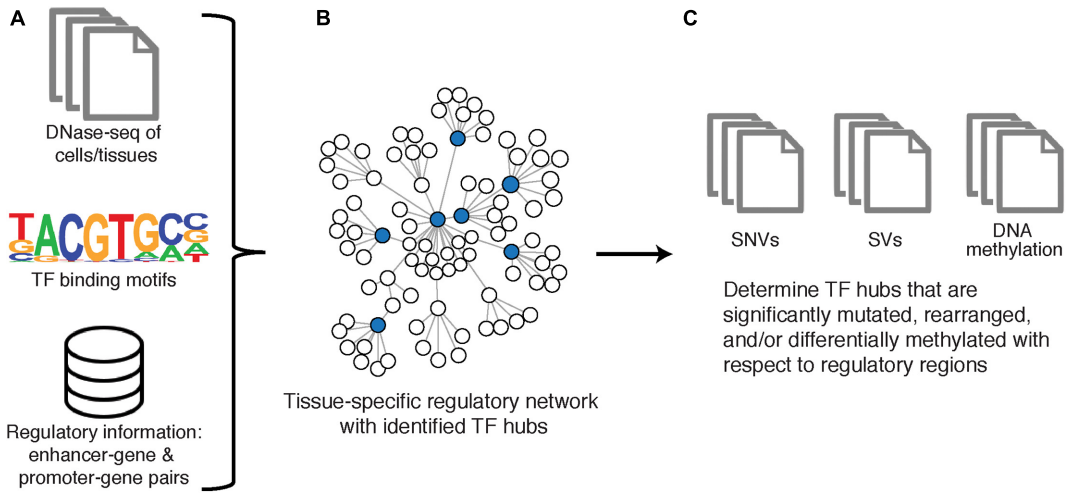



**FIGURE 4 |** RegNetDriver. **(A)** DNase-seq of DNase I hypermutation sites from a specific tissue type, information to identify TFs from binding motifs, and information of known regulatory gene pairs as used as input to reconstruct **(B)** a tissue-specific regulatory network. TF hubs are determined from nodes with the top 25% out-degree centrality. **(C)** Significantly perturbed TF hubs are identified using SNV, SV, and DNA methylation data.

expression profiles (estimated using RNA-seq data). Note that this is a directed two-layered network that estimates relationships between TFs and their transcriptional targets through their corresponding promoter/enhancer associations.

*Network mining: RegNetDriver, step II*

This network is then utilized to identify TF hubs with genomic and epigenomic alterations that can potentially cause large perturbations in this tissue-specific network. Specifically, TFs are first mined on degree centrality, such that the top 25% of TFs with the greatest number of outgoing edges are defined as hubs. Next, to identify TF hubs significantly affected on genomic and epigenomic levels in prostate cancer, they are evaluated for the presence of prostate-cancer specific genomic alterations (single nucleotide variants and structural variants) and DNA methylation changes in their coding and non-coding regulatory regions. In Dhingra et al., RegNetDriver nominated three TFs as regulatory drivers in prostate cancer, with functional validation conducted on *ERF* (Dhingra et al., 2017). RegNetDriver is freely available for download.[5]

## Protein–Protein Interaction Network-Based Analysis

Another important avenue in mechanism-centric biomarker discovery is PPIs. Such interactions elucidate putative protein complexes, which are known to perform critical functions within the cell and include for example the pre-initiation complex for RNA transcription (Greber and Nogales, 2019), the spliceosome for pre-mRNA splicing (Chen et al., 2007), and the ribosome for translation of mRNA to protein (Wilson and Doudna Cate, 2012), among others. Cancer cells in particular have been shown to deregulate protein complexes for their sustained proliferation, survival, and metastasis (Robichaud et al., 2019). In recent years, numerous public databases have cataloged networks of known and predicted PPIs, such as STRING (Szklarczyk et al., 2019), IntAct (Orchard et al., 2014), CellCircuits (Mak et al., 2007), and PINA (Cowley et al., 2012) [more comprehensive lists are described by Huang et al. (2018) and Miryala et al. (2018)]. Here, we describe the method from Chuang et al. (2007), which effectively combines PPI networks with gene expression data and evaluates these hybrid subnetworks as mechanism-centric biomarkers of breast cancer metastasis (**Figure 5**).

### Network Construction: Chuang et al., Step I

Chuang et al. introduce a hybrid approach to combine a PPI network with tissue-specific gene expression profiles across patient samples. The PPI network is comprised of nodes representing proteins and edges representing a characterized PPI, utilizing subnetworks from CellCircuits. Tissue-specific gene expression data are then overlaid onto all PPI subnetworks. For each subnetwork, its activity in each sample/patient is defined as a combination of z-scores for the subnetwork genes. This defines patient-specific vectors of subnetwork activities, which are then mined for phenotype associations.

### Network Mining: Chuang et al., Step II

Activities of subnetworks are evaluated for their association with specific phenotypes (e.g., metastatic and non-metastatic), where associations can be calculated by mutual information, t-score, or Wilcoxon score and is referred to as the subnetwork discriminative potential/score. Next, the method selects subnetworks with a locally maximal discriminative score and performs significance testing to ensure subnetworks are non-random and robust. In classification performance on a test cohort, the authors found that the subnetwork markers identified using this PPI network-based approach showed higher AUC in classifying metastatic versus non-metastatic samples compared to single-gene markers, random subnetworks, and gene sets

[5]https://khuranalab.med.cornell.edu/RegNetDriver.html

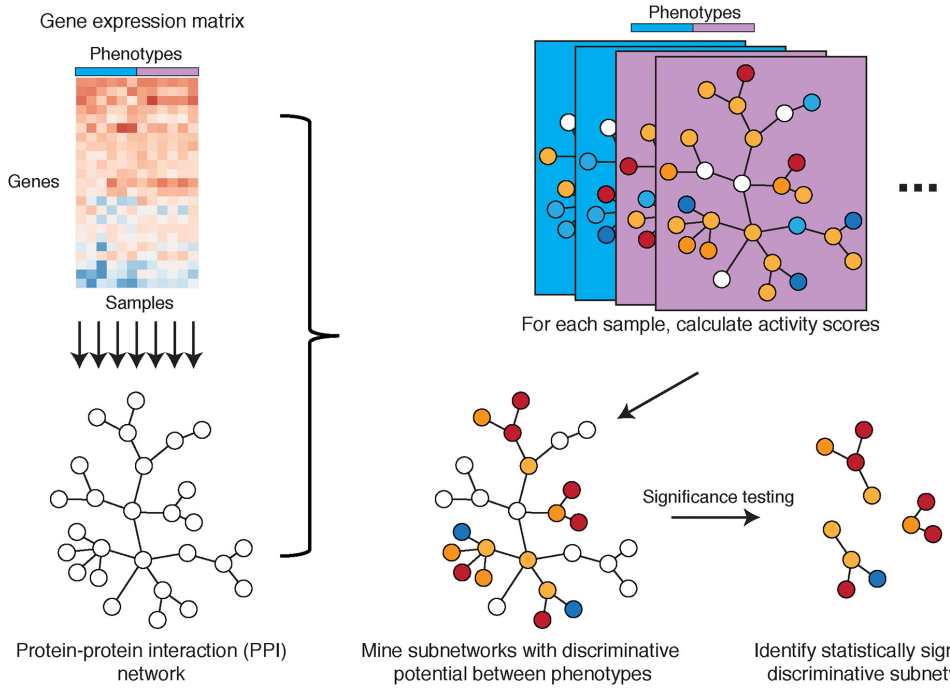



**FIGURE 5 |** Illustration of the PPI network-based approach by Chuang et al. Gene expression microarray data with phenotype information is overlaid onto a PPI network that is constructed from existing knowledge. Subnetwork activities are calculated per sample based on z-transformed gene expression values, with subnetworks defined by the PPI network. Discriminative potential for each subnetwork is determined by mutual information (or alternatively, t-score or Wilcoxon score) that measures the association between sample activities and phenotypes. Subnetworks with discriminative potential between phenotypes are identified by a greedy search for locally maximal discriminative potential scores. Discriminative subnetworks are further assessed in significance testing to identify statistically significant discriminative subnetworks.

from other annotation databases such as GO and MSigDB. Importantly, the method by Chuang et al. showed better biomarker reproducibility (i.e., higher overlap between markers) between two different breast cancer studies, outperforming gene-centric methods (Chuang et al., 2007).

## Pathway-Based Analysis: pathCHEMO and pathER

Recently, pathway-based biomarker algorithms, such as pathCHEMO (Epsi et al., 2019) and pathER (Rahem et al., 2020), have demonstrated that discovery approaches that encompass information from biological pathways significantly outperform gene-centric methods which do not take into account pathway membership.

Pathways represent a group of biochemical entities (e.g., genes, proteins, etc.), connected by interactions, relations, and reactions (including physical interactions, complex formation, transcriptional regulation, etc.), that lead to a certain product or changes in a cell. Molecular pathways have long been known to play a crucial role in cancer initiation, progression, dissemination, and therapeutic response. Some notable examples are: the role of RAS and PI3K pathways in prostate and breast cancers and their therapeutic responses (Yue et al., 2002; Haagenson and Wu, 2010), the Wnt signaling pathway in colorectal and other cancers (Zhan et al., 2017), the Hippo pathway in melanoma (Zhang X. et al., 2020), and the MYC pathway in prostate cancer progression and treatment response (Arriaga et al., 2020).

Both pathCHEMO and pathER assume that interrogation of molecular pathways, such as those present in Biocarta (Nishimura, 2001), KEGG (Kanehisa et al., 2021), and Reactome (Jassal et al., 2020), can reveal functional, biologically meaningful biomarkers that govern carcinogenesis and therapeutic response. pathCHEMO was specifically developed to compare poor versus good therapeutic response (as categorical outcomes) in cancer. In general, it evaluates differential behavior of biological pathways on both transcriptomic (RNA expression) and epigenomic (DNA methylation) levels between any two phenotypes of interest (Epsi et al., 2019). First, an RNA expression treatment response signature is defined as a list of genes ranked by their differential expression between poor and good treatment response. Then, genes in each pathway are evaluated for their enrichment in either over-expressed, under-expressed, or differentially expressed (which includes both over- and under-expressed) part of this signature. Enrichment in the over- and under-expressed parts separately allows identification of pathways where the majority of genes exhibit a similar behavior (i.e., are either over- or under-expressed), while enrichment in the differentially expressed part of the signature allows identification of pathways where some genes are over-expressed and some are under-expressed (which depicts a complex interplay of activation

and repression relationships inside a molecular pathway). This enrichment is referred to as the RNA expression-based activity level of a molecular pathway. DNA methylation-based activity for each pathway is estimated in the same manner using a DNA methylation treatment response signature. Pathways that are enriched in the RNA expression treatment response signature and the DNA methylation treatment response signature are then integrated to select those that are significantly affected on both expression and methylation levels (**Figure 6**). Activity levels of the candidate pathways are further evaluated as biomarkers of therapeutic response in independent patient cohorts. Epsi et al. showed that pathCHEMO could successfully identify molecular pathways as biomarkers of response to commonly used chemotherapy in lung adenocarcinoma, lung squamous carcinoma, and colorectal adenocarcinoma (Epsi et al., 2019). Yet, a large number of genes that participate in these pathways could potentially preclude their adoption to clinic. To overcome this limitation, "read-out" genes for each pathway were identified for which expression levels (i) correlate with pathway activity and (ii) are associated with therapeutic response. Such read-out genes were shown to produce the same predictive accuracy as the pathways themselves and constitute feasible biomarkers for clinical use (Epsi et al., 2019). pathCHEMO is freely available at http://license.rutgers.edu/technologies/2019-121_pathchemo.

As opposed to pathCHEMO, pathER applies a pathway-based approach on a single-patient level, which allows the association of pathway activity across a patient cohort to a wide range of therapeutic responses (Rahem et al., 2020). Specifically, this approach utilizes a multivariable regression Cox proportional hazards model to associate pathway activity levels with time-to-therapeutic failure, thus capturing poor, good, and medium therapeutic responses. Rahem et al. successfully applied this approach to identify both pathways and their read-out genes for tamoxifen resistance in ER-positive breast cancer (Rahem et al., 2020). pathCHEMO and pathER were compared to other approaches, including black-box machine learning techniques (such as random forest and support vector machines) and differential gene expression alone, and were shown to outperform these approaches in identifying more accurate biomarkers of therapeutic response (Epsi et al., 2019; Rahem et al., 2020).

## CHALLENGES AND LIMITATIONS OF MECHANISM-CENTRIC APPROACHES

Mechanism-centric approaches provide a powerful solution for informed biomarker discovery, yet common challenges that these methods need to account for include sufficient cohort sizes, data variability and scaling, comprehension of existing knowledge bases, and tissue-specificity (**Table 1**).

As many of these methods utilize association-based analyses (i.e., correlation, mutual information, regression, etc.), a sufficient cohort size is required to be able to accurately estimate relationships between variables. One of the direct solutions to this problem includes combining analyses in multiple datasets; however, batch effects among different acquisition methods, profiling platforms, and even institutions where datasets were collected might hamper such implementation.

In addition to a sufficient cohort size, substantial variability of expression profiles is also required to be able to accurately predict associations between variables. This task is feasible, yet it requires careful consideration, meticulous initial experimental design, and in-depth investigation of the amount of final variability necessary for successful analysis. Another challenge is the need for well-defined phenotypes, as they often require a substantially large number of samples inside each phenotype group while also demanding intra-sample homogeneity, as in the eigengene approach, MARINa, PPI network-based method by Chuang et al., pathCHEMO, etc.

At the same time, methods that rely on single-patient/sample mining (e.g., VIPER, the PPI network-based method by Chuang et al., and pathER) rely on dataset scaling to define its single-sample signatures (defined by comparing each gene to the average of its expression in the dataset of interest) making interpretation of any findings from such analyses dataset-specific.

Another known challenge is tissue-specificity, commonly faced in PPI network-based and pathway-based approaches, though some tissue- and cell-specific interaction databases are now available such as TissueNet (Basha et al., 2017), the Integrated Interactions Database (Kotlyar et al., 2019), and HumanBase (Greene et al., 2015). Tissue-specificity in these methods is usually achieved by overlaying gene expression data onto the PPI networks or molecular pathways, such as in Chuang et al., pathCHEMO, and pathER.

Furthermore, limitations of mechanism-centric approaches that utilize knowledge bases (e.g., RegNetDriver, PPI network-based approach, pathCHEMO, and pathER) lie in their reliance on known biological relationships among groups of genes/proteins/other functional units contained within a database. Various annotation, pathway, and PPI databases depend on existing information and do not include functional units that have not been previously studied, thus limiting *de novo* discoveries.

## DISCUSSION

The wide availability of large-scale data produced by high-throughput technologies has created a wealth of information for biomarker discovery. A vast majority of these biomarkers have been identified using gene-centric methods, yet their interpretability and clinical utility have been limited as they do not account for the relationships among genes. Utilizing methods that consider biological underpinnings of the data (i.e., mechanism-centric methods) can vastly improve interpretable biomarker discovery, clinical applicability and targeting, and reproducibility of results.

In particular, advantages of mechanism-centric over gene-centric approaches can be illustrated through their ability to (i) identify a tightly connected, cooperative group of genes unified by the same function, as opposed to individual genes (which might not be related); (ii) provide a mechanism-level view, which enhances the understanding of the biological mechanisms

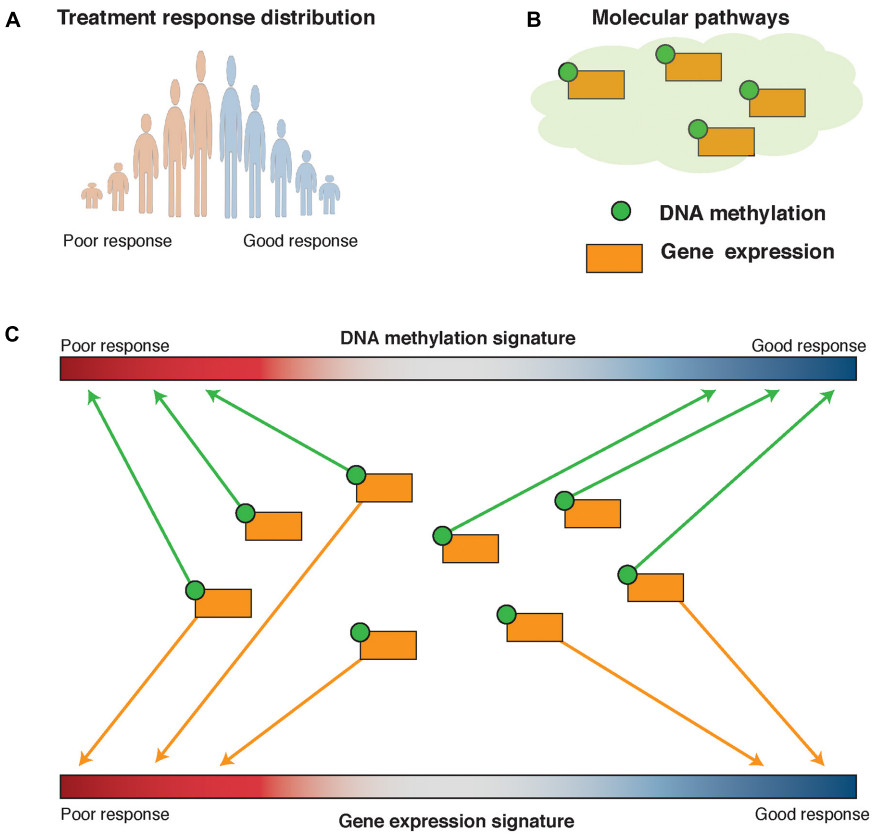



**FIGURE 6 |** Pathway-based modeling: pathCHEMO and pathER. **(A)** Therapeutic response distribution is defined based on time to therapeutic failure. Tails of this distribution are utilized in pathCHEMO and a full spectrum of therapeutic responses is utilized in pathER. **(B)** Molecular pathways are utilized as a knowledge base in pathway-based modeling. Genes in such pathways can be affected on multiple levels, such as differential expression (i.e., orange square) and DNA methylation (i.e., green satellite). **(C)** Molecular pathways are assessed for their integrated enrichment and association with therapeutic response.

implicated in a phenotype (e.g., therapeutic resistance, cancer metastasis); (iii) look at alterations in biological structures, which enhances the likelihood of identifying functionally relevant targets; (iv) identify driver as opposed to passenger markers, which allows for their effective therapeutic targeting; (v) focus on molecular structures, rather than individual genes, which decreases the chance of detecting results due to experimental noise present in biological experiments (i.e., robustness of results); and finally (vi) identify biomarkers that are more accurate and more reproducible between different cohorts.

From a computational point of view, mechanism-centric approaches can be used for interpretable feature engineering and selection (i.e., reduction), subsequently reducing the number of hypotheses to be tested. This is clearly demonstrated by gene co-expression networks, regulatory networks, PPI networks, and pathway-based methods, where cooperative groups of genes, instead of a long list of singular genes, are assessed for their association with clinical outcomes.

Mechanism-centric methods can both (i) provide interpretable inputs to white- or black-box approaches or (ii) contribute to inner model interpretability (i.e., such as in visible machine learning). First, results from mechanism-centric methods can be utilized as inputs into learning models to significantly improve predictive performance (over gene-centric inputs). One such example was demonstrated in Rahem et al., where pathway-based markers were utilized as inputs into Cox proportional hazards regression modeling and outperformed gene-centric markers for tamoxifen resistance in ER-positive breast cancer (Rahem et al., 2020). Similarly, Chuang et al. showed that markers identified by their PPI network-based method could be effectively used as inputs into a regression model and outperformed gene-centric markers in classification of metastatic breast cancer (Chuang et al., 2007). Though not in cancer, several methods have also suggested utilizing hierarchical structures (such as those inherent in Gene Ontology) as inputs for predictive models (Carvunis and Ideker, 2014; Yu et al., 2016). Second, mechanism-centric methods can potentially be incorporated into model building, such as in "visible learning," where the relationships between inputs and outputs can be interpreted (Yu et al., 2018). One such (outside of cancer) neural network method, DCell, was proposed by Ma et al., where the hierarchy of molecular relationships determined from prior knowledge (Gene Ontology and CliXO) was built into the model itself (i.e., hierarchies were utilized by nodes of the neural

**TABLE 1 |** Summary of mechanism-centric methods discussed in this review.

| Method | Data modality | Utilize knowledge base? |
|---|---|---|
| **Gene co-expression network-based** | *Identify modules of highly correlated genes*<br>+Increased interpretability at the mechanistic level<br>+Associate genes with previously uncharacterized biological functions<br>–Directionality of gene-gene interactions is unknown | |
| Centered Concordance Index (CCI) (Han et al., 2016) | *Condition-specific module identification* | |
| | Single-omic | No |
| Eigengenes (Alter et al., 2000; Zhang and Horvath, 2005) | *Identify modules associated with clinical features of interest* | |
| | Single-omic | No |
| Hubs (Freeman, 1977; Horvath and Dong, 2008) | *Hub gene identification*<br>*+Identify potential mechanism-centric target* | |
| | Single-omic | No |
| **Regulatory network-based** | *Identify regulatory relationships between a TF/co-TF and its target genes*<br>+Increased interpretability at the mechanistic level<br>+Identify potential drivers of disease<br>+Can identify non-linear relationships<br>+Tissue specific network | |
| MARINa (Lefebvre et al., 2010) | *Identify MRs from a set of samples containing two phenotypes*<br>-Need phenotype signature | |
| | Single-omic | No |
| VIPER (Alvarez et al., 2016) | *Single-sample MR identification from a cohort*<br>–Dataset scaling | |
| | Single-omic | No |
| RegNetDriver (Dhingra et al., 2017) | *Identify TF hubs that are significantly affected by single nucleotide variants, structural variants, or DNA methylation*<br>+Increase interpretability of TF hub activity through multi-omic integration<br>–Limited by information in knowledge base | |
| | Multi-omic | Yes |
| **PPI network-based** | *Use PPI subnetworks as a functional unit*<br>+Increased interpretability at the mechanistic level<br>+Connect results to the protein complex level<br>–Limited by information in knowledge base | |
| Chuang et al., 2007 | *Identify subnetworks with differential activity in metastatic breast cancer*<br>+Tissue-specificity from overlaying gene expression data<br>+Improved biomarker classification accuracy and reproducibility<br>–Dataset scaling | |
| | Multi-omic | Yes |
| **Pathway-based** | *Use molecular pathways as a functional unit*<br>+Increased interpretability at the mechanistic level<br>–Limited by information in knowledge base | |
| pathCHEMO (Epsi et al., 2019) | *Identify significantly altered pathways (at transcript and DNA methylation levels) in response to chemotherapy in lung and colorectal cancer*<br>+Improved biomarker classification accuracy and reproducibility<br>–Need phenotype signature | |
| | Multi-omic | Yes |
| pathER (Rahem et al., 2020) | *Identify pathways as markers of tamoxifen resistance in ER + breast cancer*<br>+Improved biomarker classification accuracy and reproducibility<br>–Dataset scaling | |
| | Single-omic | Yes |

*The objective of each method is detailed in italics, followed by their respective pros (+) and cons (–). Overall pros and cons for each method type are listed in a non-redundant manner. Information on data modality and if a method utilized a knowledge base is detailed as well.*

network) (Ma et al., 2018). Recently, Kuenzi et al. developed an extension of DCell, called DrugCell, which utilized chemical drug structures as a part of the neural network learning model to predict drug response in cancer cells (Kuenzi et al., 2020). This interpretable deep learning model was shown to be able to predict cell sensitivity/resistance to specific drugs, synergistic drug mechanisms, and effective drug combinations for treatment.

Further improvements in the interpretability of biological processes that inform discovery of mechanism-centric biomarkers can be made through multi-level data and method

integration. For example, several groups have combined co-expression WGCNA modules with PPI networks to uncover hubs with functional connections as biomarkers in endometrial cancer (Liu et al., 2019) and bladder cancer (Wang Y. et al., 2019). Wang et al. constructed an Active Protein-Gene network model using transcriptional regulatory and PPI networks to quantify TF activity and elucidate both upstream and downstream regulations (Wang et al., 2013). Even though this study was done in diabetes, it could be applicable to mechanism-centric biomarker discovery in cancer. Ahsen et al. embedded VIPER within a new framework (NeTFactor) to identify TFs that most likely regulate a gene-centric biomarker signature (Ahsen et al., 2019). While this method was applied to asthma and peanut allergy, it could easily be extended to cancer studies. At the same time, multi-omic integration in RegNetDriver improved the interpretability of the proposed model to explain the impact of mutations, structural variants, and DNA methylation on TF activity in prostate cancer (Dhingra et al., 2017). A recent study by Broyde et al. constructed a multi-omic lung adenocarcinoma tissue-specific oncoprotein interaction network using information obtained from ARACNe, CINDy (an algorithm identifying post-translational modulators), VIPER, and PPI predictions (Broyde et al., 2021), which depicted a complex network of interactions for KRAS and could potentially be utilized for mechanism-centric biomarker discovery. Such multi-level approaches in conjunction with mechanism-centric methods promise to uncover a deeper understanding of mechanisms involved in gene regulation and post-translational modifications in biomarker discovery.

Finally, recent technological advances, such as those seen in single-cell studies, promise to improve our understanding of intra-tumor heterogeneity, clonal evolution, and the role of microenvironment in cancer progression and therapeutic response. Single-cell gene expression offers a granular view of active pathways in a cell type-specific manner and potentially allows for the construction of cell type-specific networks. In fact, the rapid advances of single-cell sequencing technology have already allowed network analysis methods to be applied directly to data from single-cell RNA-sequencing (scRNA-seq) (Crow et al., 2016; Aibar et al., 2017; Chan et al., 2017; Fiers et al., 2018; Papili Gao et al., 2018; van Dijk et al., 2018; Lamere and Li, 2019; Jackson et al., 2020; Sekula et al., 2020; Ye et al., 2020) with integration of other data modalities for improved network inference (Aibar et al., 2017; Chan et al., 2017; Papili Gao et al., 2018; van Dijk et al., 2018; Jackson et al., 2020; Pratapa et al., 2020). Furthermore, matching single-cell and bulk patient samples could provide an invaluable resource for single-cell driven network investigations that can be compared to and related back to bulk tissues. As more single-cell data become available (e.g., RNA sequencing, targeted DNA sequencing, ATAC-seq, etc.), we foresee advances in single-cell technologies and data analysis to be central to understanding precise, clone-specific biomarkers, unveiling trajectories of tumor evolution and providing accurate ground for informed time-cautious precision therapeutics.

In summary, mechanism-centric approaches (based on gene co-expression networks, regulatory networks, PPI networks, and molecular pathways) identify biomarkers that are biologically meaningful, interpretable, reproducible, have higher translational potential, and provide greater predictive power over biomarkers identified by gene-centric methods. Thus, mechanism-centric approaches are the future of clinically relevant rational biomarker discovery, personalized treatment planning, and precision therapeutics in cancer.

## AUTHOR CONTRIBUTIONS

CY and AM conceived and wrote the manuscript. Both the authors contributed to the article and approved the submitted version.

## FUNDING

AM was supported by R01LM013236-01 and Rutgers start-up funds.

## ACKNOWLEDGMENTS

We are thankful to the Mitrofanova lab for useful discussions.

## REFERENCES

Abida, W., Cyrta, J., Heller, G., Prandi, D., Armenia, J., Coleman, I., et al. (2019). Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl. Acad. Sci. U.S.A.* 116, 11428–11436.

Agnelli, L., Forcato, M., Ferrari, F., Tuana, G., Todoerti, K., Walker, B. A., et al. (2011). The reconstruction of transcriptional networks reveals critical genes with implications for clinical outcome of multiple myeloma. *Clin. Cancer Res.* 17, 7402–7412. doi: 10.1158/1078-0432.ccr-11-0596

Ahsen, M. E., Chun, Y., Grishin, A., Grishina, G., Stolovitzky, G., Pandey, G., et al. (2019). NeTFactor, a framework for identifying transcriptional regulators of gene expression-based biomarkers. *Sci. Rep.* 9:12970.

Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463

Allen, W. L., Coyle, V. M., and Johnston, P. G. (2006). Predicting the outcome of chemotherapy for colorectal cancer. *Curr. Opin. Pharmacol.* 6, 332–336. doi: 10.1016/j.coph.2006.02.005

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10101–10106. doi: 10.1073/pnas.97.18.10101

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., et al. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* 48, 838–847. doi: 10.1038/ng.3593

Arriaga, J. M., Panja, S., Alshalalfa, M., Zhao, J., Zou, M., Giacobbe, A., et al. (2020). A MYC and RAS co-activation signature in localized prostate cancer drives bone metastasis and castration resistance. *Nat. Cancer* 1, 1082–1096. doi: 10.1038/s43018-020-00125-0

Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E. Jr., and Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited:

discrimination and calibration. *Cancer* 116, 3310–3321. doi: 10.1002/cncr.25081

Ayers, M., Symmans, W. F., Stec, J., Damokosh, A. I., Clark, E., Hess, K., et al. (2004). Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J. Clin. Oncol.* 22, 2284–2293. doi: 10.1200/jco.2004.05.166

Aytes, A., Giacobbe, A., Mitrofanova, A., Ruggero, K., Cyrta, J., Arriaga, J., et al. (2018). NSD2 is a conserved driver of metastatic prostate cancer progression. *Nat. Commun.* 9:5201.

Aytes, A., Mitrofanova, A., Lefebvre, C., Alvarez, M. J., Castillo-Martin, M., Zheng, T., et al. (2014). Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell* 25, 638–651. doi: 10.1016/j.ccr.2014.03.017

Bae, T., Rho, K., Choi, J. W., Horimoto, K., Kim, W., and Kim, S. (2013). Identification of upstream regulators for prognostic expression signature genes in colorectal cancer. *BMC Syst. Biol.* 7:86. doi: 10.1186/1752-0509-7-86

Basha, O., Barshir, R., Sharon, M., Lerman, E., Kirson, B. F., Hekselman, I., et al. (2017). The TissueNet v.2 database: a quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Res.* 45, D427–D431.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390. doi: 10.1038/ng1532

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816–824.

Bisikirska, B., Bansal, M., Shen, Y., Teruya-Feldstein, J., Chaganti, R., and Califano, A. (2016). Elucidation and pharmacological targeting of novel molecular drivers of follicular lymphoma progression. *Cancer Res.* 76, 664–674. doi: 10.1158/0008-5472.can-15-0828

Boumahdi, S., Driessens, G., Lapouge, G., Rorive, S., Nassar, D., Le Mercier, M., et al. (2014). SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature* 511, 246–250. doi: 10.1038/nature13305

Boutros, P. C. (2015). The path to routine use of genomic biomarkers in the cancer clinic. *Genome Res.* 25, 1508–1513. doi: 10.1101/gr.191114.115

Broyde, J., Simpson, D. R., Murray, D., Paull, E. O., Chu, B. W., Tagore, S., et al. (2021). Oncoprotein-specific molecular interaction maps (SigMaps) for cancer network analyses. *Nat. Biotechnol.* 39, 215–224. doi: 10.1038/s41587-020-0652-7

Butte, A. J., and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 5, 418–429.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12182–12186. doi: 10.1073/pnas.220392197

Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463, 318–325. doi: 10.1038/nature08712

Carvunis, A. R., and Ideker, T. (2014). Siri of the cell: what biology could learn from the iPhone. *Cell* 157, 534–538. doi: 10.1016/j.cell.2014.03.009

Chan, T. E., Stumpf, M. P. H., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267.e3.

Chen, Y. I., Moore, R. E., Ge, H. Y., Young, M. K., Lee, T. D., and Stevens, S. W. (2007). Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. *Nucleic Acids Res.* 35, 3928–3944. doi: 10.1093/nar/gkm347

Chng, W. J., Chung, T. H., Kumar, S., Usmani, S., Munshi, N., Avet-Loiseau, H., et al. (2016). Gene signature combinations improve prognostic stratification of multiple myeloma patients. *Leukemia* 30, 1071–1078. doi: 10.1038/leu.2015.341

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Cordero, D., Sole, X., Crous-Bou, M., Sanz-Pamplona, R., Pare-Brunet, L., Guino, E., et al. (2014). Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC Cancer* 14:708. doi: 10.1186/1471-2407-14-708

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., et al. (2012). PINA v2.0: mining interactome modules. *Nucleic Acids Res.* 40, D862–D865.

Crow, M., Paul, A., Ballouz, S., Huang, Z. J., and Gillis, J. (2016). Exploiting single-cell expression to characterize co-expression replicability. *Genome Biol.* 17:101.

Della Gatta, G., Palomero, T., Perez-Garcia, A., Ambesi-Impiombato, A., Bansal, M., Carpenter, Z. W., et al. (2012). Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nat. Med.* 18, 436–440. doi: 10.1038/nm.2610

Dhingra, P., Martinez-Fundichely, A., Berger, A., Huang, F. W., Forbes, A. N., Liu, E. M., et al. (2017). Identification of novel prostate cancer drivers using RegNetDriver: a framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network. *Genome Biol.* 18:141.

Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* 2:38. doi: 10.3389/fcell.2014.00038

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Epsi, N. J., Panja, S., Pine, S. R., and Mitrofanova, A. (2019). pathCHEMO, a generalizable computational framework uncovers molecular pathways of chemoresistance in lung adenocarcinoma. *Commun. Biol.* 2:334.

Erho, N., Crisan, A., Vergara, I. A., Mitra, A. P., Ghadessi, M., Buerki, C., et al. (2013). Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* 8:e66855. doi: 10.1371/journal.pone.0066855

Eskandari, E., Mahjoubi, F., and Motalebzadeh, J. (2018). An integrated study on TFs and miRNAs in colorectal cancer metastasis and evaluation of three co-regulated candidate genes as prognostic markers. *Gene* 679, 150–159. doi: 10.1016/j.gene.2018.09.003

Fiers, M., Minnoye, L., Aibar, S., Bravo Gonzalez-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* 17, 246–254. doi: 10.1093/bfgp/elx046

Fletcher, M. N., Castro, M. A., Wang, X., de Santiago, I., O'Reilly, M., Chin, S. F., et al. (2013). Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* 4:2464.

Freeman, L. C. A. (1977). Set of measures of centrality based on betweenness. *Sociometry* 40, 35–41. doi: 10.2307/3033543

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Gabay, M., Li, Y., and Felsher, D. W. (2014). MYC activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harb. Perspect. Med.* 4:a014241. doi: 10.1101/cshperspect.a014241

Garzotto, M., Beer, T. M., Hudson, R. G., Peters, L., Hsieh, Y. C., Barrera, E., et al. (2005). Improved detection of prostate cancer using classification and regression tree analysis. *J. Clin. Oncol.* 23, 4322–4329. doi: 10.1200/jco.2005.11.136

Giulietti, M., Occhipinti, G., Principato, G., and Piva, F. (2016). Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development. *Cell Oncol.* 39, 379–388. doi: 10.1007/s13402-016-0283-7

Giulietti, M., Occhipinti, G., Principato, G., and Piva, F. (2017). Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene co-expression network analysis. *Cell Oncol.* 40, 181–192. doi: 10.1007/s13402-017-0315-y

Greber, B. J., and Nogales, E. (2019). The structures of eukaryotic transcription pre-initiation complexes and their functional implications. *Subcell. Biochem.* 93, 143–192. doi: 10.1007/978-3-030-28151-9_5

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. doi: 10.1038/ng.3259

Haagenson, K. K., and Wu, G. S. (2010). The role of MAP kinases and MAP kinase phosphatase-1 in resistance to breast cancer treatment. *Cancer Metastasis Rev.* 29, 143–149. doi: 10.1007/s10555-010-9208-5

Han, Y., Ye, X., Cheng, J., Zhang, S., Feng, W., Han, Z., et al. (2019). Integrative analysis based on survival associated co-expression gene modules for predicting Neuroblastoma patients' survival time. *Biol. Direct* 14:4.

Han, Z., Zhang, J., Sun, G., Liu, G., and Huang, K. (2016). A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules. *BMC Genomics* 17(Suppl. 7):519. doi: 10.1186/s12864-016-2912-y

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems* 96, 86–103.

Heng, Y. J., Lester, S. C., Tse, G. M., Factor, R. E., Allison, K. H., Collins, L. C., et al. (2017). The molecular basis of breast cancer pathological phenotypes. *J. Pathol.* 241, 375–391. doi: 10.1002/path.4847

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e6.

Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* 4:e1000117. doi: 10.1371/journal.pcbi.1000117

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. U.S.A.* 103, 17402–17407. doi: 10.1073/pnas.0608396103

Hu, X., Bao, M., Huang, J., Zhou, L., and Zheng, S. (2020). Identification and validation of novel biomarkers for diagnosis and prognosis of hepatocellular carcinoma. *Front. Oncol.* 10:541479. doi: 10.3389/fonc.2020.541479

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., et al. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* 6, 484–495.e5.

Huang, Z., Han, Z., Wang Resource, T., Shao, W., Xiang, S., Salama, P., et al. (2021). TSUNAMI: translational bioinformatics tool suite for network analysis and mining. *Genomics Proteomics Bioinformatics*.

Huo, T., Canepa, R., Sura, A., Modave, F., and Gong, Y. (2017). Colorectal cancer stages transcriptome analysis. *PLoS One* 12:e0188697. doi: 10.1371/journal.pone.0188697

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e12776. doi: 10.1371/journal.pone.0012776

Jackson, C. A., Castro, D. M., Saldi, G. A., Bonneau, R., and Gresham, D. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *Elife* 9:e51254.

Jain, R. K., Duda, D. G., Willett, C. G., Sahani, D. V., Zhu, A. X., Loeffler, J. S., et al. (2009). Biomarkers of response and resistance to antiangiogenic therapy. *Nat. Rev. Clin. Oncol.* 6, 327–338.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi: 10.1038/35075138

Jia, R., Zhao, H., and Jia, M. (2020). Identification of co-expression modules and potential biomarkers of breast cancer by WGCNA. *Gene* 750:144757. doi: 10.1016/j.gene.2020.144757

Jiramongkol, Y., and Lam, E. W. (2020). FOXO transcription factor family in cancer and metastasis. *Cancer Metastasis Rev.* 39, 681–709. doi: 10.1007/s10555-020-09883-w

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 49, D545–D551.

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 9, 770–780.

Kotlyar, M., Pastrello, C., Malik, Z., and Jurisica, I. (2019). IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* 47, D581–D589.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., et al. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 38, 672–684.e6.

Kuiper, R., Broyl, A., de Knegt, Y., van Vliet, M. H., van Beers, E. H., van der Holt, B., et al. (2012). A gene expression signature for high-risk multiple myeloma. *Leukemia* 26, 2406–2413.

Lamere, A. T., and Li, J. (2019). Inference of gene co-expression networks from single-cell RNA-sequencing data. *Methods Mol. Biol.* 1935, 141–153. doi: 10.1007/978-1-4939-9057-3_10

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Lee, W. P., and Tzou, W. S. (2009). Computational methods for discovering gene networks from expression data. *Brief. Bioinform.* 10, 408–423.

Lefebvre, C., Rajbhandari, P., Alvarez, M. J., Bandaru, P., Lim, W. K., Sato, M., et al. (2010). A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* 6:377. doi: 10.1038/msb.2010.31

Lim, W. K., Lyashenko, E., and Califano, A. (2009). Master regulators used as breast cancer metastasis classifier. *Pac. Symp. Biocomput.* 14, 504–515.

Liu, J., Zhou, S., Li, S., Jiang, Y., Wan, Y., Ma, X., et al. (2019). Eleven genes associated with progression and prognosis of endometrial cancer (EC) identified by comprehensive bioinformatics analysis. *Cancer Cell Int.* 19, 136.

Liu, R., Guo, C. X., and Zhou, H. H. (2015a). Network-based approach to identify prognostic biomarkers for estrogen receptor-positive breast cancer treatment with tamoxifen. *Cancer Biol. Ther.* 16, 317–324. doi: 10.1080/15384047.2014.1002360

Liu, R., Lv, Q. L., Yu, J., Hu, L., Zhang, L. H., Cheng, Y., et al. (2015b). Correlating transcriptional networks with pathological complete response following neoadjuvant chemotherapy for breast cancer. *Breast Cancer Res. Treat.* 151, 607–618. doi: 10.1007/s10549-015-3428-x

Ludwig, J. A., and Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* 5, 845–856. doi: 10.1038/nrc1739

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* 15, 290–298. doi: 10.1038/nmeth.4627

Mak, H. C., Daly, M., Gruebel, B., and Ideker, T. (2007). CellCircuits: a database of protein network models. *Nucleic Acids Res.* 35, D538–D545.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006a). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006b). Reverse engineering cellular networks. *Nat. Protoc.* 1, 662–671.

Markowetz, F., and Spang, R. (2007). Inferring cellular networks–a review. *BMC Bioinformatics* 8(Suppl. 6):S5. doi: 10.1186/1471-2105-8-S6-S5

McDermott, J. E., Wang, J., Mitchell, H., Webb-Robertson, B. J., Hafen, R., Ramey, J., et al. (2013). Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin. Med. Diagn.* 7, 37–51. doi: 10.1517/17530059.2012.718329

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365, 488–492. doi: 10.1016/s0140-6736(05)17866-0

Miryala, S. K., Anbarasu, A., and Ramaiah, S. (2018). Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. *Gene* 642, 84–94. doi: 10.1016/j.gene.2017.11.028

Mitrofanova, A., Aytes, A., Zou, M., Shen, M. M., Abate-Shen, C., and Califano, A. (2015). Predicting drug response in human prostate cancer from preclinical analysis of in vivo mouse models. *Cell Rep.* 12, 2060–2071. doi: 10.1016/j.celrep.2015.08.051

Nishimura, D. (2001). BioCarta. *Biotech. Softw. Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363.

Ou, Y., and Zhang, C.-Q. (2007). A new multimembership clustering method. *J. Ind. Manage. Optim.* 3, 619–624. doi: 10.3934/jimo.2007.3.619

Palomero, T., Lim, W. K., Odom, D. T., Sulis, M. L., Real, P. J., Margolin, A., et al. (2006). NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proc. Natl. Acad. Sci. U.S.A.* 103, 18261–18266. doi: 10.1073/pnas.0606108103

Panja, S., Hayati, S., Epsi, N. J., Parrott, J. S., and Mitrofanova, A. (2018). Integrative (epi) genomic analysis to predict response to androgen-deprivation therapy in prostate cancer. *EBioMedicine* 31, 110–121. doi: 10.1016/j.ebiom.2018.04.007

Panja, S., Rahem, S., Chu, C. J., and Mitrofavnova, A. (2020). Big data to knowledge: application of machine learning to predictive modeling of therapeutic response in cancer. *Curr. Genomics* 21, 1–25. doi: 10.1201/b11508-2

Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O., and Gunawan, R. (2018). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 34, 258–266. doi: 10.1093/bioinformatics/btx575

Petty, R. D., Samuel, L. M., Murray, G. I., MacDonald, G., O'Kelly, T., Loudon, M., et al. (2009). APRIL is a novel clinical chemo-resistance biomarker in colorectal adenocarcinoma identified by gene expression profiling. *BMC Cancer* 9:434. doi: 10.1186/1471-2407-9-434

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17, 147–154. doi: 10.1038/s41592-019-0690-6

Rahem, S. M., Epsi, N. J., Coffman, F. D., and Mitrofanova, A. (2020). Genome-wide analysis of therapeutic response uncovers molecular pathways governing tamoxifen resistance in ER+ breast cancer. *EBioMedicine* 61:103047. doi: 10.1016/j.ebiom.2020.103047

Remo, A., Simeone, I., Pancione, M., Parcesepe, P., Finetti, P., Cerulo, L., et al. (2015). Systems biology analysis reveals NFAT5 as a novel biomarker and master regulator of inflammatory breast cancer. *J. Transl. Med.* 13:138.

Robichaud, N., Sonenberg, N., Ruggero, D., and Schneider, R. J. (2019). Translational control in cancer. *Cold Spring Harb. Perspect. Biol.* 11:a032896.

Robinson, D., Van Allen, E. M., Wu, Y. M., Schultz, N., Lonigro, R. J., Mosquera, J. M., et al. (2015). Integrative clinical genomics of advanced prostate cancer. *Cell* 161, 1215–1228.

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnol.* 26, 462–469.

Sanz-Pamplona, R., Berenguer, A., Cordero, D., Mollevi, D. G., Crous-Bou, M., Sole, X., et al. (2014). Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol. Cancer* 13:46. doi: 10.1186/1476-4598-13-46

Sartor, I. T., Zeidan-Chulia, F., Albanus, R. D., Dalmolin, R. J., and Moreira, J. C. (2014). Computational analyses reveal a prognostic impact of TULP3 as a transcriptional master regulator in pancreatic ductal adenocarcinoma. *Mol. Biosyst.* 10, 1461–1468. doi: 10.1039/c3mb70590k

Sekula, M., Gaskins, J., and Datta, S. (2020). A sparse bayesian factor model for the construction of gene co-expression networks from single-cell RNA sequencing count data. *BMC Bioinformatics* 21:361. doi: 10.1186/s12859-020-03707-y

Shaughnessy, J. D. Jr., Qu, P., Usmani, S., Heuck, C. J., Zhang, Q., Zhou, Y., et al. (2011). Pharmacogenomics of bortezomib test-dosing identifies hyperexpression of proteasome genes, especially PSMD4, as novel high-risk feature in myeloma treated with total therapy 3. *Blood* 118, 3512–3524. doi: 10.1182/blood-2010-12-328252

Shaughnessy, J. D. Jr., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., et al. (2007). A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* 109, 2276–2284. doi: 10.1182/blood-2006-07-038430

Sonabend, A. M., Bansal, M., Guarnieri, P., Lei, L., Amendolara, B., Soderquist, C., et al. (2014). The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression. *Cancer Res.* 74, 1440–1451. doi: 10.1158/0008-5472.can-13-2150

Song, H., Ding, N., Li, S., Liao, J., Xie, A., Yu, Y., et al. (2020). Identification of hub genes associated with hepatocellular carcinoma using robust rank aggregation combined with weighted gene co-expression network analysis. *Front. Genet.* 11:895. doi: 10.3389/fgene.2020.00895

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10869–10874. doi: 10.1073/pnas.191367098

Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., et al. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10393–10398. doi: 10.1073/pnas.1732912100

Strimbu, K., and Tavel, J. A. (2010). What are biomarkers? *Curr. Opin. HIV AIDS* 5, 463–466.

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

Talos, F., Mitrofanova, A., Bergren, S. K., Califano, A., and Shen, M. M. (2017). A computational systems approach identifies synergistic specification genes that facilitate lineage conversion to prostate tissue. *Nat. Commun.* 8:14662.

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Front. Oncol.* 8:374. doi: 10.3389/fonc.2018.00374

Tang, X., Xu, P., Wang, B., Luo, J., Fu, R., Huang, K., et al. (2019). Identification of a specific gene module for predicting prognosis in glioblastoma patients. *Front. Oncol.* 9:812. doi: 10.3389/fonc.2019.00812

Tian, Z., He, W., Tang, J., Liao, X., Yang, Q., Wu, Y., et al. (2020). Identification of important modules and biomarkers in breast cancer based on WGCNA. *Onco Targets Ther.* 13, 6805–6817. doi: 10.2147/ott.s258439

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174, 716–729.e27.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.

Walsh, L. A., Alvarez, M. J., Sabio, E. Y., Reyngold, M., Makarov, V., Mukherjee, S., et al. (2017). An integrated systems biology approach identifies TRIM25 as a key determinant of breast cancer metastasis. *Cell Rep.* 20, 1623–1640. doi: 10.1016/j.celrep.2017.07.052

Wang, C. C. N., Li, C. Y., Cai, J. H., Sheu, P. C., Tsai, J. J. P., Wu, M. Y., et al. (2019). Identification of prognostic candidate genes in breast cancer by integrated bioinformatic analysis. *J. Clin. Med.* 8:1160. doi: 10.3390/jcm8081160

Wang, H. Q., Wong, H. S., Zhu, H., and Yip, T. T. (2009). A neural network-based biomarker association information extraction approach for cancer classification. *J. Biomed. Inform.* 42, 654–666. doi: 10.1016/j.jbi.2008.12.010

Wang, J., Sun, Y., Zheng, S., Zhang, X. S., Zhou, H., and Chen, L. (2013). APG: an active protein-gene network model to quantify regulatory signals in complex biological systems. *Sci. Rep.* 3:1097.

Wang, X. G., Peng, Y., Song, X. L., and Lan, J. P. (2016). Identification potential biomarkers and therapeutic agents in multiple myeloma based on bioinformatics analysis. *Eur. Rev. Med. Pharmacol. Sci.* 20, 810–817.

Wang, Y., Chen, L., Ju, L., Qian, K., Liu, X., Wang, X., et al. (2019). Novel biomarkers associated with progression and prognosis of bladder cancer identified by co-expression analysis. *Front. Oncol.* 9:1030. doi: 10.3389/fonc.2019.01030

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679. doi: 10.1016/s0140-6736(05)17947-1

Werhli, A. V., and Husmeier, D. (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* 6:15.

Wilson, D. N., and Doudna Cate, J. H. (2012). The structure and function of the eukaryotic ribosome. *Cold Spring Harb. Perspect. Biol.* 4:a011536. doi: 10.1101/cshperspect.a011536

Yan, Z., Li, J., Xiong, Y., Xu, W., and Zheng, G. (2012). Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data. *Oncol. Rep.* 28, 1036–1042. doi: 10.3892/or.2012.1891

Yang, Q., Wang, R., Wei, B., Peng, C., Wang, L., Hu, G., et al. (2018). Candidate biomarkers and molecular mechanism investigation for glioblastoma multiforme utilizing WGCNA. *Biomed Res. Int.* 2018:4246703.

Ye, X., Zhang, W., Futamura, Y., and Sakurai, T. (2020). Detecting interactive gene groups for single-cell RNA-seq data based on co-expression network analysis and subgraph learning. *Cells* 9:1938. doi: 10.3390/cells9091938

Ying, C. Y., Dominguez-Sola, D., Fabi, M., Lorenz, I. C., Hussein, S., Bansal, M., et al. (2013). MEF2B mutations lead to deregulated expression of the oncogene

BCL6 in diffuse large B cell lymphoma. *Nat. Immunol.* 14, 1084–1092. doi: 10.1038/ni.2688

Yu, C. Y., Xiang, S., Huang, Z., Johnson, T. S., Zhan, X., Han, Z., et al. (2019). Gene co-expression network and copy number variation analyses identify transcription factors associated with multiple myeloma progression. *Front. Genet.* 10:468. doi: 10.3389/fgene.2019.00468

Yu, M. K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J., et al. (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Syst.* 2, 77–88. doi: 10.1016/j.cels.2016.02.003

Yu, M. K., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., and Ideker, T. (2018). Visible machine learning for biomedicine. *Cell* 173, 1562–1565. doi: 10.1016/j.cell.2018.05.056

Yue, W., Wang, J.-P., Conaway, M., Masamura, S., Li, Y., and Santen, R. J. (2002). Activation of the MAPK pathway enhances sensitivity of MCF-7 breast cancer cells to the mitogenic effect of estradiol. *Endocrinology* 143, 3221–3229. doi: 10.1210/en.2002-220186

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., et al. (2002). Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood* 99, 1745–1757. doi: 10.1182/blood.v99.5.1745

Zhan, F., Huang, Y., Colla, S., Stewart, J. P., Hanamura, I., Gupta, S., et al. (2006). The molecular classification of multiple myeloma. *Blood* 108, 2020–2028.

Zhan, T., Rindtorff, N., and Boutros, M. (2017). Wnt signaling in cancer. *Oncogene* 36, 1461–1473.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17.

Zhang, F., Chen, J., Wang, M., and Drabier, R. (2013). A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer. *BMC Proc.* 7(Suppl. 7):S10. doi: 10.1186/1753-6561-7-S7-S10

Zhang, H., Yu, C. Y., Singer, B., and Xiong, M. (2001). Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl. Acad. Sci. U.S.A.* 98, 6730–6735. doi: 10.1073/pnas.111153698

Zhang, J., and Huang, K. (2014). Normalized lmQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. *Cancer Inform.* 13(Suppl. 3), 137–146.

Zhang, J., Wang, L., Xu, X., Li, X., Guan, W., Meng, T., et al. (2020). Transcriptome-based network analysis unveils eight immune-related genes as molecular signatures in the immunomodulatory subtype of triple-negative breast cancer. *Front. Oncol.* 10:1787. doi: 10.3389/fonc.2020.01787

Zhang, S., Jing, Y., Zhang, M., Zhang, Z., Ma, P., Peng, H., et al. (2015). Stroma-associated master regulators of molecular subtypes predict patient prognosis in ovarian cancer. *Sci. Rep.* 5:16066.

Zhang, X., Yang, L., Szeto, P., Abali, G. K., Zhang, Y., Kulkarni, A., et al. (2020). The hippo pathway oncoprotein YAP promotes melanoma cell invasion and spontaneous metastasis. *Oncogene* 39, 5267–5281. doi: 10.1038/s41388-020-1362-9

Zhao, L., Lee, B. Y., Brown, D. A., Molloy, M. P., Marx, G. M., Pavlakis, N., et al. (2009). Identification of candidate biomarkers of therapeutic response to docetaxel by proteomic profiling. *Cancer Res.* 69, 7696–7703. doi: 10.1158/0008-5472.can-08-4901

**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: 09 August 2021
doi: 10.3389/fgene.2021.695130

# A Novel Signature Constructed by Immune-Related LncRNA Predicts the Immune Landscape of Colorectal Cancer

*Mengyu Sun[1,3], Tongyue Zhang[1,3], Yijun Wang[1,3], Wenjie Huang[2,3]* and Limin Xia[1,3]**

[1] Department of Gastroenterology, Institute of Liver and Gastrointestinal Diseases, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2] Hepatic Surgery Center, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [3] Hubei Key Laboratory of Hepato-Pancreato-Biliary Diseases, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Colorectal cancer (CRC) has the characteristics of high morbidity and mortality. LncRNA not only participates in the progression of CRC through genes and transcription levels, but also regulates the tumor microenvironment and leads to the malignant phenotype of tumors. Therefore, we identified immune-related LncRNAs for the construction of clinical prognostic model. We searched The Cancer Genome Atlas (TCGA) database for original data. Then we identified differentially expressed irlncRNA (DEirlncRNA), which was paired and verified subsequently. Next, univariate analysis, Lasso and Cox regression analysis were performed on the DEirlncRNA pair. The ROC curve of the signature was drawn, and the optimal cut-off value was found. Then the cohort was divided into a high-risk and a low-risk group. Finally, we re-evaluated the signature from different perspectives. A total of 16 pairs of DEirlncRNA were included in the construction of the model. After regrouping according to the cut-off value of 1.275, the high-risk group showed adverse survival outcomes, progressive clinicopathological features, specific immune cell infiltration status, and high sensitivity to some chemotherapy drugs. In conclusion, we constructed a signature composed of immune-related LncRNA pair with no requirement of the specific expression level of genes, which shows promising clinical predictive value in CRC patients.

**Keywords: colorectal cancer, immune-related lncRNA, immunotherapy, signature, TCGA**

OPEN ACCESS

***Edited by:***
*Suman Ghosal,*
*National Institutes of Health (NIH),*
*United States*

***Reviewed by:***
*Arijita Sarkar,*
*University of Southern California,*
*United States*
*Sourish Ghosh,*
*National Institutes of Health (NIH),*
*United States*

****Correspondence:***
*Wenjie Huang*
*huangwenjie@tjh.tjmu.edu.cn*
*Limin Xia*
*xialimin@tjh.tjmu.edu.cn*

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

***Received:*** *14 April 2021*
***Accepted:*** *12 July 2021*
***Published:*** *09 August 2021*

***Citation:***
*Sun M, Zhang T, Wang Y, Huang W and Xia L (2021) A Novel Signature Constructed by Immune-Related LncRNA Predicts the Immune Landscape of Colorectal Cancer. Front. Genet. 12:695130. doi: 10.3389/fgene.2021.695130*

## INTRODUCTION

According to global cancer statistics in 2020, colorectal cancer ranks third in cancer incidence and becomes the second leading cause of cancer deaths (Sung et al., 2021). And its incidence has been steadily increasing in many countries in Eastern Europe, Southeast Asia, Central-south Asia, and South America (Arnold et al., 2017, 2020). Although the diagnosis and treatment of

**Abbreviations:** CRC, Colorectal cancer; TCGA, The Cancer Genome Atlas; irlncRNA, Immune-related LncRNA; DEirlncRNA, Differentially expressed irlncRNA; AIC, Akaike information criterion; mabs, Monoclonal antibodies; FPKM, Fragments per kilobase million; GTF, Gene transfer format; ir-genes, Immune-related genes; FC, Fold change; FDR, False discovery rate; ROC, Receiver operating characteristic; IC50, Half-inhibitory concentrations; AUC, Area under curve; OS, overall survival; GEO, Gene Expression Ominibus.

colorectal cancer continue to make progress, the prognosis of advanced patients is poor, mainly due to recurrence, metastasis, and drug resistance (Nishihara et al., 2013). Moreover, the prognosis of patients in the same disease stage is different, accompanied by different gene mutations (Sinicrope et al., 2017). Genetic and molecular changes play an essential role in these events and provide potential targets for treatment (Sadanandam et al., 2013).

Immunotherapy accelerates the development of oncology and shows encouraging anti-tumor efficacy in many types of solid cancers (Farkona et al., 2016). Among various immunotherapy methods, immunomodulatory monoclonal antibodies (mabs) that target immune checkpoints have produced promising and long-lasting therapeutic results in multiple cancers. However, in colorectal cancer, the tumor has a high proportion of resistance and ineffectiveness to these monoclonal antibodies, such as anti-PD-1 mabs (Guo et al., 2020). Therefore, more alternative therapies and biomarkers used to predict the prognosis and treatment response of CRC need to be further investigated to allow more patients to benefit from immunotherapy.

Long noncoding RNAs, defined as RNA that transcriptional length is more than 200 nucleotides, do not encode proteins (Chan and Tay, 2018). LncRNAs interact with DNA, mRNAs, ncRNAs, and proteins to regulate gene expression at different levels, and play an essential role in both normal development and tumor progression (Ørom et al., 2010; Wang and Chang, 2011). LncRNAs are frequently involved in different stages of CRC from precancerous polyps to distant metastasis, and are considered potentially effective diagnostic biomarkers (Ye et al., 2015; Saus et al., 2016). Studies have shown that lncRNAs can also lead to cancer's malignant progression by changing the tumor microenvironment (Atianand et al., 2017). LncRNAs regulate the gene-coded products involved in the immune response and affect the activation of immune cells, thus leading to the infiltration of immune cells in tumor (Chen et al., 2017).

Clinical predictive signatures focusing on immune-related markers have shown favorable diagnostic and predictive performance in various tumors. Shen et al. (2020) identified 11 lncRNAs associated with immune cell infiltration to construct the predictive signature of breast cancer. Wu et al. (2020) identified 8 immune-related LncRNAs and demonstrated their value in predicting the prognosis and immunotherapy response of bladder cancer. Hong et al. (2020) incorporated 12 pairs of immune-related LncRNAs into the model, which has good clinical predictive value in hepatocellular carcinoma.

As far as the accuracy of the cancer diagnosis signature is concerned, the combination of two markers is better than a simple single gene (Lv et al., 2020). For the simplicity and practicality of the signature, we tried to construct a reasonable model based on 2-lncRNA combinations (Hong et al., 2020; Chen et al., 2021; Ping et al., 2021). Compared with the single gene model that needs to detect the specific expression level of each marker, our model only needs to compare the expression level of each lncRNA pair and substitute 0 or 1 into the model. This will effectively avoid data correction during model application. We evaluated this signature's predictive value in CRC patients, including survival rate, clinical progression, immune cell infiltration, and chemotherapy effects.

# MATERIALS AND METHODS

## Retrieval and Preparation of Transcriptome and Clinical Data

The transcriptome data of colorectal cancer were downloaded from The Cancer Genome Atlas (TCGA, RRID:SCR_003193) database,[1] and the data type was FPKM (fragments per kilobase million). The dataset includes 44 normal tissues and 568 tumor tissues. We downloaded the GTF (gene transfer format) file from Ensembl[2] to distinguish mRNA and lncRNA. Clinical data of CRC patients were retrieved from the TCGA database. To extract valid data, duplicate data and data with a follow-up time fewer than 31 days were eliminated.

## Identification of Differentially Expressed Immune-Related LncRNAs (DEirlncRNAs)

The list of immune-related genes (ir-genes) was obtained from the Immport database,[3] and the co-expression analysis was performed to screen immune-related LncRNAs. We analyzed the expression correlation between ir-genes and all lncRNAs. The screening criteria for irlncRNAs were immune gene correlation coefficient $> 0.4$ and $p$-value $< 0.001$. To identify DEirlncRNAs between normal and cancer tissues, R package limma was used to analyze the differential expression of irlncRNAs. We set thresholds as log2 |fold change| $> 1$ and false discovery rate (FDR) $< 0.05$.

## Pairing of DEirlncRNA

DEirLncRNAs were periodically paired to construct a 0-or-1 matrix. Suppose that C is equivalent to a pair of DEirncRNA, such as LncRNAA and LncRNAB. If the expression level of LncRNAA is lower than that of LncRNAB, then C is defined as 0; otherwise, C is defined as 1. Next, the matrix was screened further. If the expression of the DEirLncRNA pair is counted as 0 or 1 in most samples, this pair will not be used for subsequent prognostic analysis since gene pairs without a certain level of difference cannot accurately predict patients' survival. When the number of DEirncRNA pairs of which expression quantity was 0 or 1 accounted for more than 20% and less than 80% of the total samples, the pair was considered to be an effective match.

## Construction of Risk Signature to Evaluate Risk Score

The DEirlncRNA pair was analyzed by univariate analysis, followed by lasso regression with 10 fold cross-validation, and the $p$-value was set to 0.05. Lasso regression ran 1,000 cycles to obtain the DEirlncRNA pair combination with the smallest

[1]https://tcga-data.nci.nih.gov/tcga/

[2]http://asia.ensembl.org

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

cross-validation error, and then Cox proportional hazards regression analysis and model construction were carried out. We determined the optimal model according to the Akaike information criterion (AIC) value. When the AIC value was minimum, the calculation process was terminated, and the model was regarded as the optimal candidate. The receiver operating characteristic (ROC) curves of 1-, 3-, and 5-year were drawn afterward. The following formula can calculate the risk score of all cases: RiskScore = $\sum_{i=1}^{N} Expi * Wi$, where exp is the expression value of every DEirlncRNA pair, and W is the multivariate cox regression analysis coefficient of each DEirlncRNA pair in the signature. The sum of sensitivity and specificity of each point in the 5-year ROC curve was calculated, and the risk score corresponding to the maximum point was taken as the cut-off value to distinguish the risk level.

The R packages used in the above steps were survival, surviviner, survivalroc, and glmnet.

## Verification of the Constructed Risk Signature

We first used the LncAR database to verify the differential expression of the irlncRNA contained in the signature. To verify the cut-off value, Kaplan-Meier analysis was performed to show the survival difference between the high-risk and low-risk groups by using the survival curve. Using R tool, we also visualized the specific risk score of each sample in the signature.

For the sake of verifying the clinical application value of the signature, the chi-square test was used to analyze the relationship between the signature and clinicopathological characteristics. Afterward a band diagram was drawn for visualization ($^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$). Wilcoxon signed rank sum test was applied to calculate the difference of risk score among groups with different clinicopathological characteristics. The analysis results were shown by box diagram. Univariate and multivariate Cox regression analyses were performed to evaluate the correlation between risk score, clinical variables, and prognosis of patients, so as to clarify whether the risk model can be used as an independent prognostic indicator of colorectal cancer. $P < 0.05$ was considered statistically significant. The results were demonstrated by forest map.

The R packages used for the above analysis steps are survival, surviviner, survivalroc, limma, ggpubr and complex Heatmap.

## Analysis of Tumor-Infiltrating Immune Cells

To explore the relationship between the risk score and the tumor-Infiltrating immune cells, we used various currently recognized methods to evaluate the immune cell infiltration status of colorectal cancer, including XCELL, TIMER, QUANTISEQ, MCPCOUNTER, EPIC, CIBERSORT-ABS, and CIBERSORT. The Wilcoxon signed rank sum test was implemented to compare the content of infiltrating immune cells between the high-risk and low-risk groups. Through Spearman correlation analysis, the relationship between the risk score and infiltrating immune cells was analyzed. The threshold was $P < 0.05$, and the results were displayed in the lollipop graph. The R packages limma, scales, ggplot2, and ggtext were used for the analysis.

## Evaluation of the Model's Role in Clinical Treatment

To evaluate whether the model has a certain application value in the clinical treatment of colorectal cancer, we calculated the half-inhibitory concentrations (IC50) of commonly used chemotherapy drugs in the TCGA data set. The anti-tumor medications used in the analysis included gemcitabine, Rapamycin, Imatinib, Lenalidomide, and Shikonin. Wilcoxon signed rank sum test was performed to compare the difference in IC50 of drugs between the high-risk and low-risk groups. The results were displayed in the form of the box plot.

The R packages used in this part include limma, ggpubr, pRRophetic and ggplot2.

# RESULTS

## Recognition of Differentially Expressed Immune-Related LncRNAs (DEirlncRNAs)

First, we downloaded the transcriptome data of colorectal cancer from the TCGA database. Then, the data were annotated based on the GTF file. Co-expression analysis between immune-related genes and lncRNA was performed. A total of 1017 immune-related lncRNAs were identified. Through differential expression analysis, 383 were classified as DEirlncRNAs, of which 339 were highly expressed and 44 were low expressed (**Figure 1B**). The expression of the DEirlncRNAs ranked in the top 200 based on fold change was displayed in the heat map (**Figure 1A**). The complete list of differentially expressed immune-related lncRNAs was shown in **Supplementary Table 1**.

## Construction of DEirlncRNA Pair and Risk Assessment Signature

Through the iterative loop and the construction and screening of a 0-or-1 matrix, 39528 valid DEirlncRNA pairs were obtained from 383 DEirlncRNAs. By univariate analysis, 4197 lncRNA pairs related to prognosis were identified. Subsequently, 26 DEirlncRNA pairs were extracted by LASSO regression analysis to prevent the model from over-fitting, of which 16 pairs were incorporated into the COX proportional hazard model by a stepwise method (**Figure 1C**).

The value of the area under curve (AUC) of the signature was 0.904 (**Figure 2A**), indicating an ideal predictive performance of the model. To verify the signature's superiority, we plotted the 1-, 3-, and 5-year ROC curves, and the results showed that the AUC values of all three curves were over 0.80 (**Figure 2B**). We also compared 5-year ROC curve with other clinical characteristics, and the risk score had the most considerable AUC value (**Figure 2C**).

By calculating the sum of sensitivity and specificity of each point of the ROC curve for 5 years, the risk score of 1.275 at the maximum end was taken as the cut-off value to distinguish

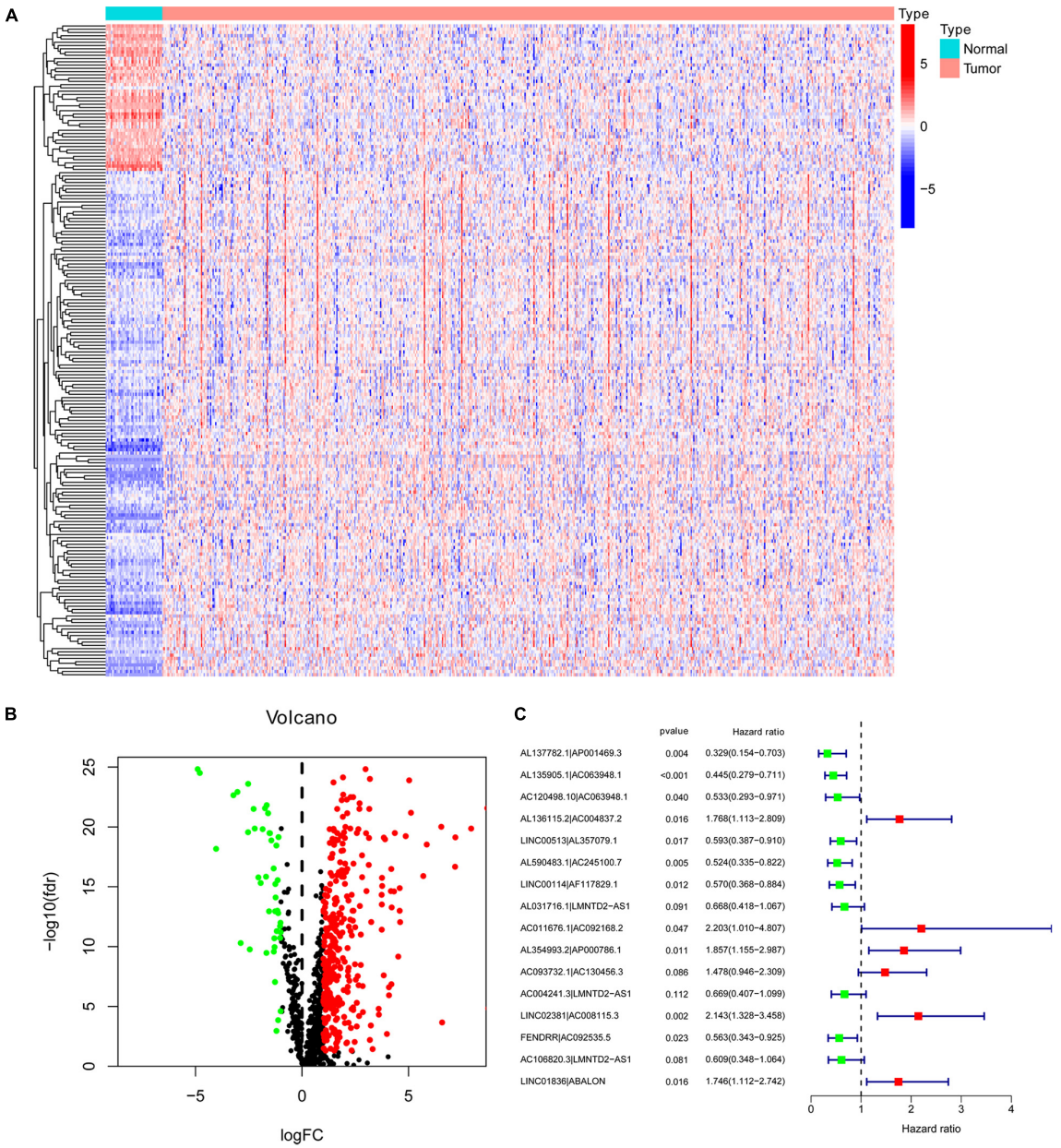



**FIGURE 1 |** Identification of differentially expressed immune-related LncRNA (DEirlncRNA) and the construction of a signature. The expression information of DEirlncRNA was displayed in the heatmap **(A)** and volcano plot **(B)**. **(C)** A forest map showed 16 DEirlncRNA pairs included in the signature.

the high and low risk of samples (**Figure 2D**). The signature was applied to 500 colorectal cancer samples available from the TCGA database, and the risk score of these patients was calculated. Whereafter, these samples were divided into a high-risk group and a low-risk group by the cut-off point identified above for further validation.

## Validation of Risk Assessment Model and Its Application in Clinical Evaluation

The external data validation of the expression of irlncRNA in the model was shown in **Supplementary Figures 1, 2**, and the detailed data sources were shown in **Supplementary Table 2**. Based on the cut-off value, 209 samples were classified into the high-risk group and 291 cases into the low-risk group. The risk scores and survival time of each case were shown in **Figures 3A,B**. The results showed that the survival rate and survival time decreased with the increase of the risk score. Survival analysis demonstrated that the survival time of the high-risk group was significantly shorter than that of the low-risk group ($p < 0.001$) (**Figure 3C**).

Subsequently, we applied the chi-square test to explore the relationship between the risk score and clinicopathological

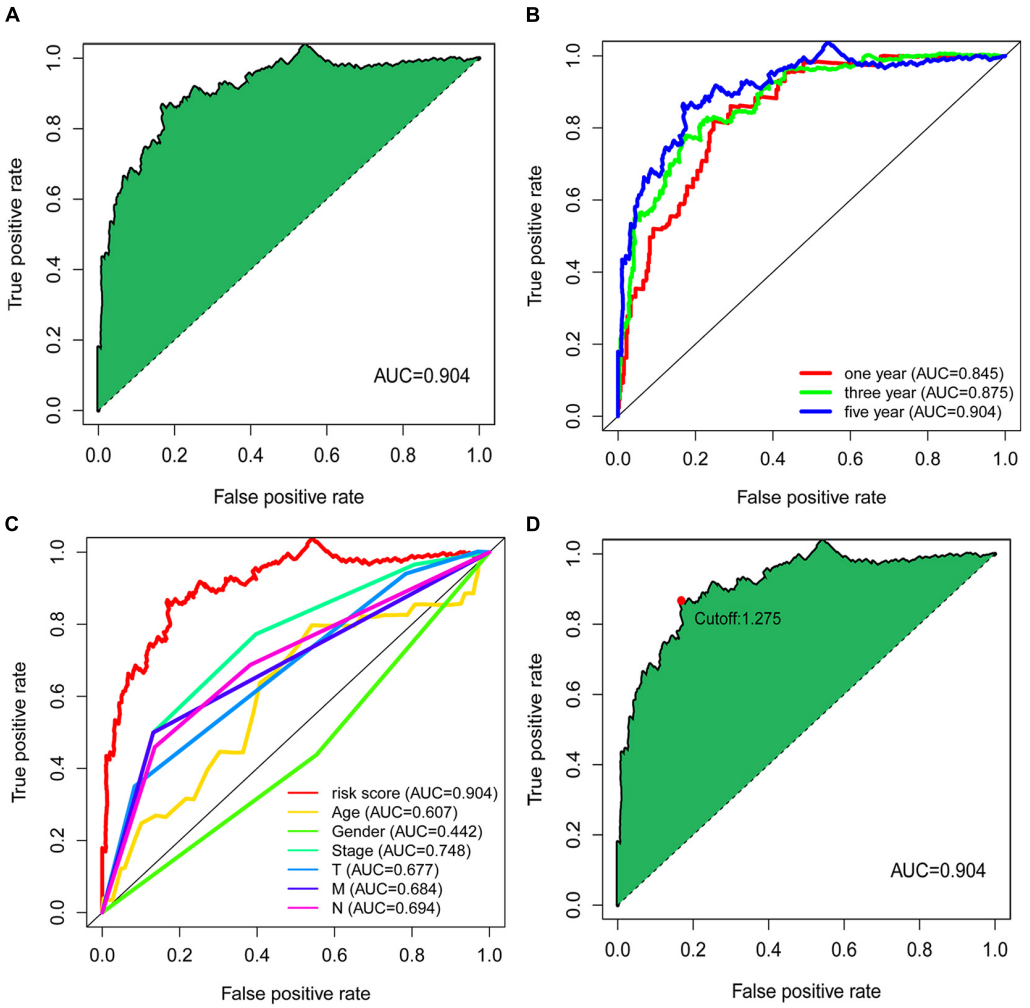



**FIGURE 2 |** Establishment of the signature based on DEirlncRNA pairs. **(A)** The 5-year ROC of the optimal signature. **(B)** The AUC values of the 1-year, 3-year, and 5-year ROC curves of the model. **(C)** The comparation of 5-year ROC curve with other clinical characteristics. **(D)** The risk score of 1.275 at the maximum end was taken as the cut-off value to distinguish the high and low risk of samples.

characteristics. The band diagram (**Figure 4A**) and scatter plots showed that clinical stage (**Figure 4B**), T stage (**Figure 4C**), M stage (**Figure 4D**), and N stage (**Figure 4E**) were significantly associated with risk score. In general, groups with advance stage were accompanied by higher risk scores. Univariate COX regression analysis showed that age ($p = 0.002$, HR = 1.032, 95% CI [1.011–1.053]), clinical stage ($p < 0.001$, HR = 2.501, 95% CI [1.951–3.206]), T stage ($p < 0.001$, HR = 3.245, 95% CI [2.119–4.969]), M stage ($p < 0.001$, HR = 5.070, 95% CI [3.269–7.862]), N stage ($p < 0.001$, HR = 2.241, 95% CI [1.739–2.888]), and RiskScore (p < 0.001, HR = 1.070, 95% CI [1.057–1.084]) were considered statistically significant (**Figure 4F**). Multivariate Cox regression analysis showed age ($p < 0.001$, HR = 1.047, 95% CI [1.026–1.069]), T stage ($p = 0.008$, HR = 1.940, 95% CI [1.186–3.171]), and RiskScore ($p < 0.001$, HR = 1.075, 95% CI [1.057–1.093]) were independent prognostic predictors (**Figure 4G**). The detailed information of univariate and multivariate Cox regression analysis was shown in **Supplementary Table 3**.

## Analysis of Immune Cell Infiltration Based on the Risk Score

Since the lncRNA identified by the co-expression method was related to immune genes, we explored whether the model was linked to the tumor microenvironment. The results showed that patients' high risk was positively correlated with tumor-infiltrating immune cells such as $CD4^+$ T cells, macrophage, and cancer-associated fibroblast, whereas negatively correlated with neutrophil (**Figures 5B–G**). Through Spearman correlation analysis, the relationship between risk score and immune infiltrating cells in multiple databases was displayed in **Figure 5A**.

## Correlation Between the Risk Model and Chemotherapy Drugs

We attempted to explore the relationship between the risk score and efficacy of common chemotherapy drugs for CRC in the TCGA dataset. The results showed that a high-risk score was

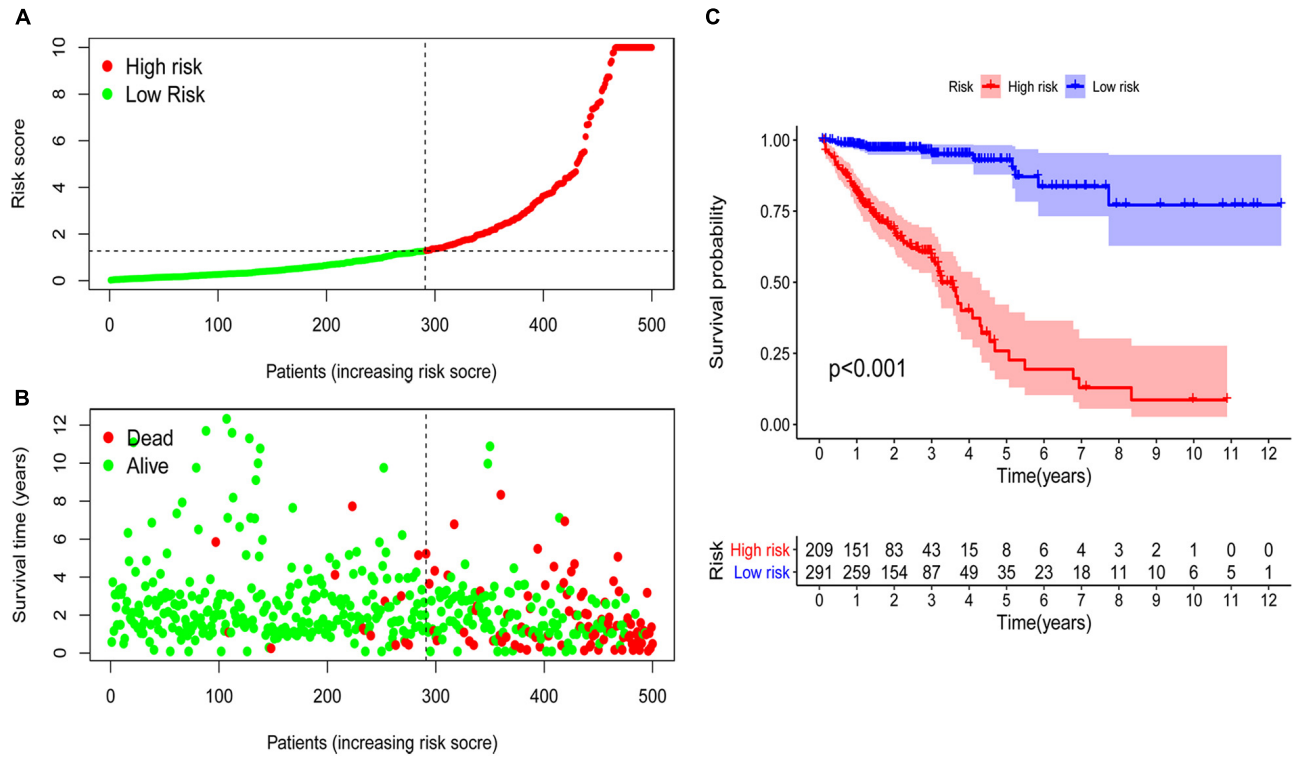



**FIGURE 3 |** Risk assessment model for survival prediction. The risk scores **(A)** and survival time **(B)** of each case were shown. **(C)** The survival time of the high-risk group was significantly shorter than that of the low-risk group.

related to lower IC50 of chemotherapeutics such as Rapamycin ($P$ = 0.00017), Imatinib ($P$ = 0.016), Lenalidomide ($P$ = 5.3e-07), and Shikonin ($P$ = 0.00075), suggesting that the model could be regarded as a potential predictor of chemotherapy sensitivity (**Figure 6**).

## DISCUSSION

Recently, an increasing number of studies showed that infiltrating immune cells play an essential role in tumor management and become an effective prognostic factor for colorectal cancer (Picard et al., 2020). There were data disclosed that there seem to be subtle differences in the composition of immune cells infiltrated in colorectal cancer, which may be a key determinant of treatment and prognosis (Xiong et al., 2018). Besides, Pagès et al. (2018) have proposed that immune scores based on tumor-infiltrating immune cells can reliably estimate the risk of recurrence of colorectal cancer patients. These results support the view that the immune score could work as a new part of tumor TNM-immune classification. In recent years, many studies have focused on establishing immune-related coding genes and non-coding RNA signatures to assess the prognosis of colorectal cancer. However, most of the prognostic models are based on the quantified expression level of the sample. In this study, inspired by the gene pairing strategy, we tried for the first time in colorectal cancer to construct a reasonable model composed of paired lncRNA, which does not require the exact expression of lncRNA.

First, we obtained the original transcriptome data of colorectal cancer from the TCGA database, performed co-expression analysis and differential expression analysis to identify DEirlncRNA, and verified the effective DEirlncRNA pair by loop pairing and a matrix of 0 or 1. Second, we performed univariate analysis, LASSO regression analysis, and COX regression analysis to determine the DEirlncRNA pair for inclusion in the signature. Third, we determined the optimal signature by calculating the AIC value, and calculated the sum of sensitivity and specificity of each point on the 5-year ROC curve to find the best cut-off value. Finally, we evaluated the model from several aspects, including survival time, clinicopathological progress, distribution of tumor-infiltrating immune cells, and chemosensitivity.

At present, a number of studies have identified predictive biomarkers for colorectal cancer and have shown good clinical utility, which also provides ideas for the construction of more clinical models (Clarke et al., 2017; Martinez-Romero et al., 2018; Ahluwalia et al., 2019). Many studies suggested that lncRNA plays a non-negligible role in the development of colorectal cancer, and may participate in the remodeling of the tumor microenvironment and affect the infiltration of immune cells in the tumor. Qin et al. (2021) constructed an independent model based on 7 immune-related lncRNAs, which may promote the accurate assessment of the prognosis of CRC patients. Lin et al. (2020) identified a model containing 9 immune-related

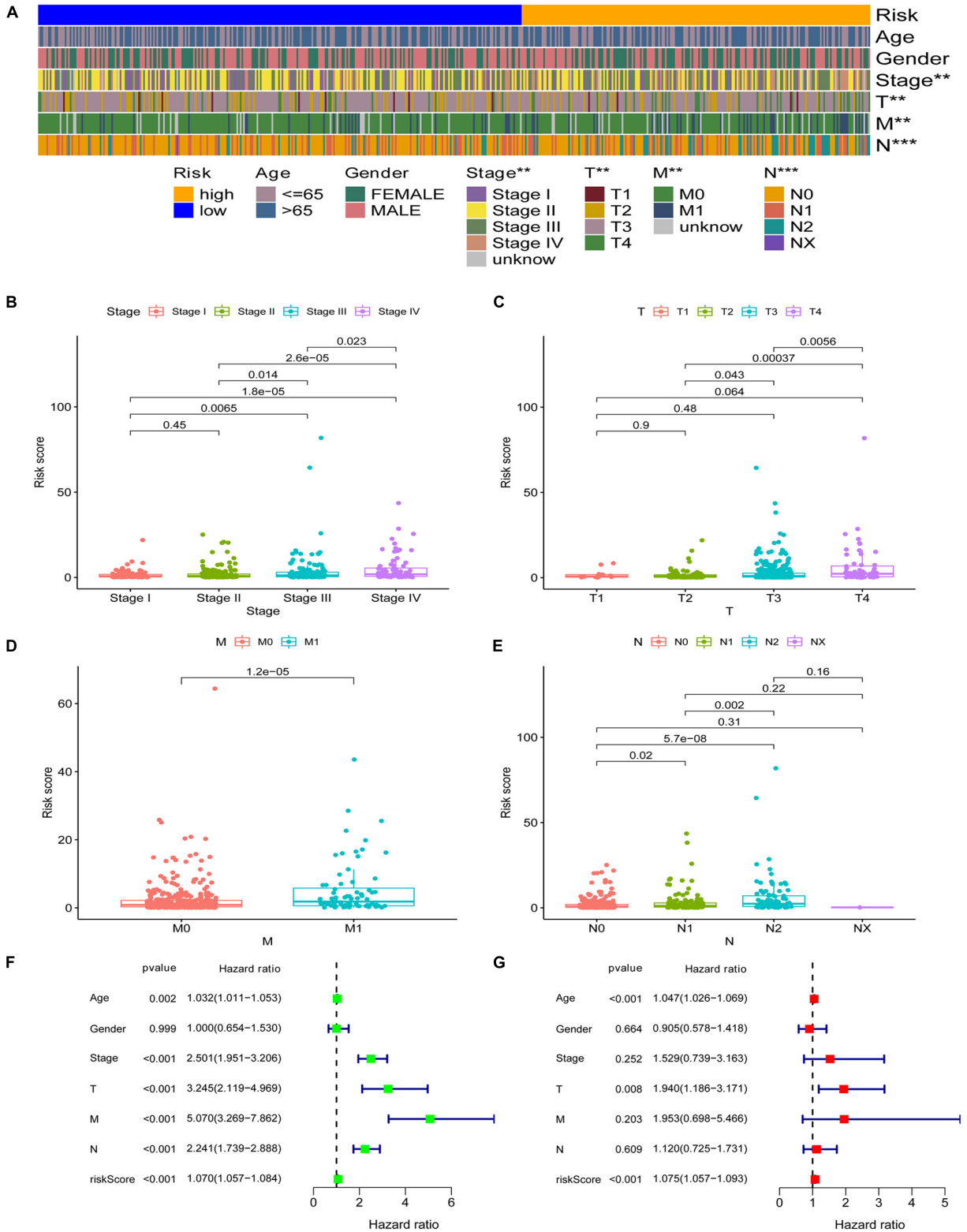



**FIGURE 4 |** Application of the signature in the clinical evaluation. The band diagram **(A)** and scatter plots showed that clinical stage **(B)**, T stage **(C)**, M stage **(D)**, and N stage **(E)** were significantly associated with risk score. Univariate COX regression analysis showed that age, clinical stage, T stage, M stage, N stage, and RiskScore were considered statistically significant **(F)**. **(G)** Multivariate Cox regression analysis showed age, T stage, and RiskScore were independent prognostic predictors. ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

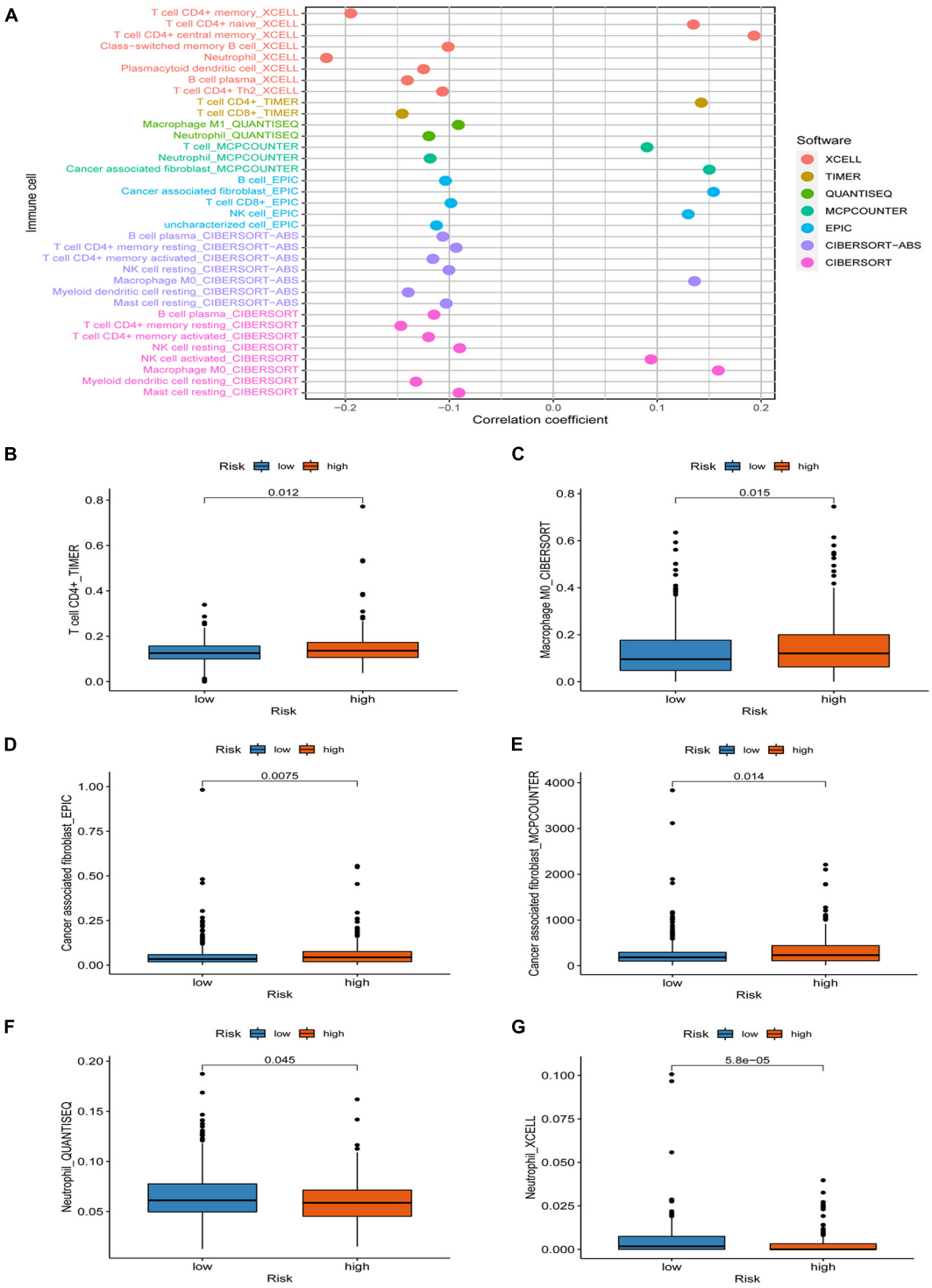



**FIGURE 5 |** Association between immune cell infiltration and the risk score. **(A)** The relationship between risk score and immune infiltrating cells in multiple databases. **(B–G)** High risk was positively correlated with tumor-infiltrating immune cells such as CD4$^+$ T cells, macrophage, and cancer-associated fibroblast, whereas negatively correlated with neutrophil.

lncRNAs, which may help to improve the prediction results of colon cancer patients and guide individualized treatment. Li et al. (2020) constructed a seven immune-related lncRNA signature, which showed promising clinical significance in colon adenocarcinoma. Our algorithm showed that we could identify DEirlncRNAs and construct the most important irlncRNA pair

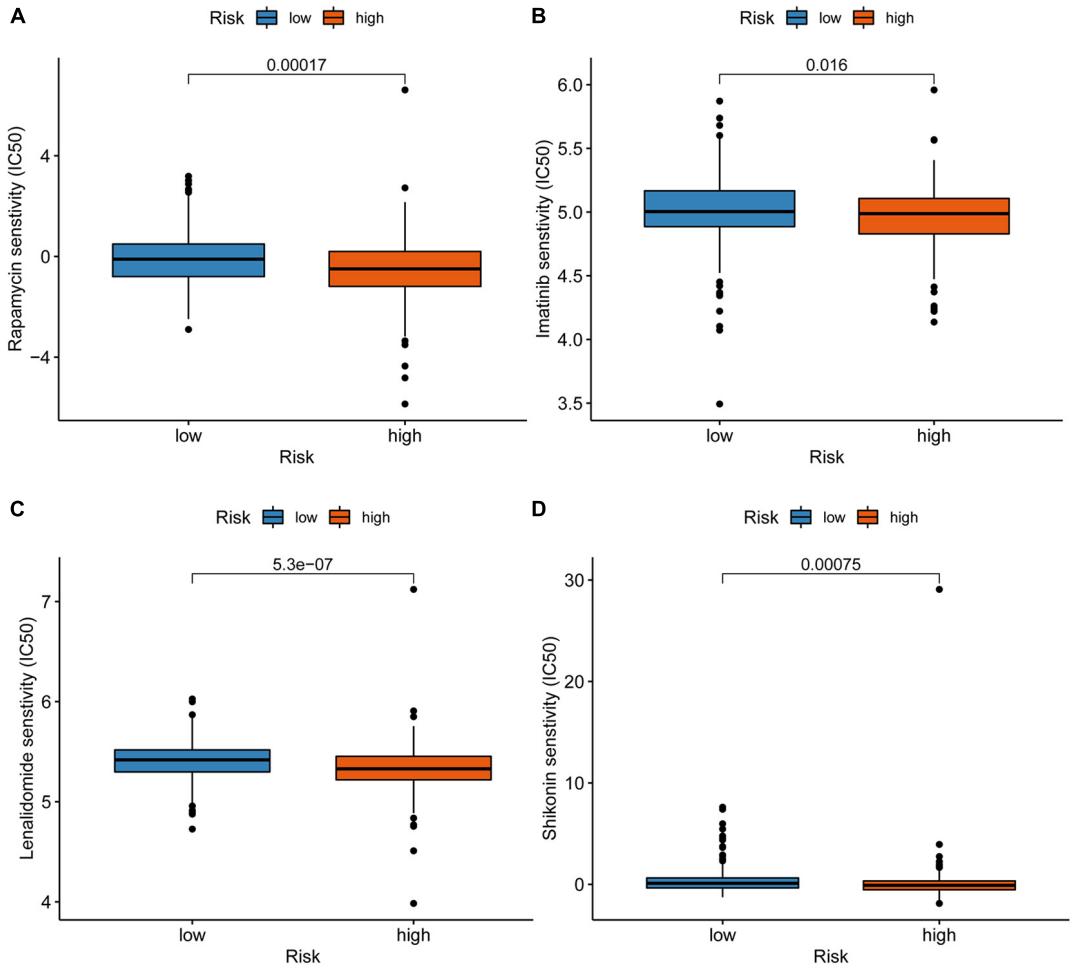



**FIGURE 6 |** Correlation between the risk model and chemotherapeutics. A high-risk score was associated with lower IC50 for chemotherapeutics such as Rapamycin **(A)**, Imatinib **(B)**, Lenalidomide **(C)**, and Shikonin **(D)**.

for the first time in colorectal cancer. The model showed good predictive performance. The AUC values of the 1-, 3-, and 5-year ROC curves of the model were all above 0.80. What's more, the most significant difference between our signature and the above-mentioned prognostic model is that the signature does not require each marker's exact expression, but only needs to compare the expression level in each pair of DEirlncRNA. This dramatically improves the model's clinical utility and largely avoids the error caused by differences in marker expression detection.

The DEirlncRNAs used to construct the signature in this study play critical roles in a variety of tumors. Dysregulation of FENDRR expression is associated with tumorigenesis, resistance to chemotherapy, fibrosis, and inflammatory diseases (Szafranski and Stankiewicz, 2021). In colorectal cancer, Yin et al. (2019) revealed that FENDRR could inhibit tumor aggressiveness by regulating the miR-18a-5p/ING4 axis. Data from Cheng et al. (2020) indicated that FENDRR could inhibit cell proliferation, migration, and invasion in CRC by targeting miR-424-5p. Liu and Du (2019) proved that FENDRR might work as a tumor suppressor gene in colon cancer by inhibiting SOX4. Also, FENDRR is closely related to immune regulation (Munteanu et al., 2020; Shen et al., 2021). Moreover, FENDRR takes effect in different cancers, such as hepatocellular carcinoma, cholangiocarcinoma, gastric cancer, cervical cancer, breast cancer, prostate cancer, endometrial cancer, and non-small cell lung cancer (He et al., 2018; Li et al., 2018; Qin et al., 2019; Yu et al., 2019; Zhang G. et al., 2019; Zhang Y.Q. et al., 2019; Zhu et al., 2020). It can be seen that FENDRR is closely involved in the process of tumors related to the digestive system and reproductive system. Lv et al. (2019) reported that LINC00114 promotes colorectal cancer by regulating the EZH2/DNMT1/miR-133bz axis, and Han et al. (2020) found that LINC00114 promotes the progression and radioresistance of nasopharyngeal carcinoma by targeting miR-203 to regulate the ERK/JNK signaling pathway. Jafarzadeh et al. (2020) indicated that LINC02381 might inhibit colorectal cancer tumorigenesis partly by regulating the PI3K signaling pathway. Besides, Jafarzadeh and Soltani (2020) also revealed that LINC02381 inhibits gastric cancer progression through the Wnt signaling

pathway. In addition, LINC02381 plays a cancer-promoting role in cervical cancer and osteosarcoma (Chen et al., 2020; Bian et al., 2021). Some identified DEirlncRNAs were also included in the signatures of other colorectal cancer studies, such as LINC02381, LINC00114, and AL590483.1 (Wang et al., 2018; Li et al., 2020; Liu et al., 2020; Sun et al., 2020; Zhou et al., 2020), which verified the effectiveness of our algorithm. Another part of the selected DEirlncRNAs was reported for the first time. Therefore, the novel biomarkers need to be further explored.

We improved the modeling process, calculated the AIC value in Cox regression analysis to determine the best model, and compared the model with other clinicopathological characteristics. Instead of using the median value of risk score to distinguish the high and low risk of patients, we calculated the sum of sensitivity and specificity of each point on the ROC curve to find the optimal cut-off value. Then we re-evaluated the signature, and the results showed that its application effect was pretty good.

The occurrence and development of colorectal cancer involve many aspects of immunodeficiency. Tumor-infiltrating immune cells may affect the therapeutic effect of immune checkpoint inhibitors (Guo et al., 2020). To explore the relationship between risk score and tumor-infiltrating immune cells, we performed various methods to estimate infiltrating immune cells in colorectal cancer. Our signature was closely related to $CD4^+$ T cells, $CD8^+$ T cells, macrophages, cancer-associated fibroblast, and neutrophil through a comprehensive analysis. According to our signature, the high risk was associated with the sensitivity of chemotherapy drugs such as Rapamycin, Imatinib, Lenalidomide, and Shikonin. Given the limited drug data in the database, the sensitivity of more first-line chemotherapy drugs for colorectal cancer needs to be analyzed to further improve the signature's practicality.

Our research also has shortcomings and limitations. First of all, we only obtained the original CRC data from the TCGA database, thus the number of samples may be relatively insufficient. We have not retrieved a useful dataset containing lncRNA expression levels and clinical information of CRC in other commonly used databases such as Gene Expression Ominibus (GEO). Since there were no data available, our model has not been externally verified. When a model based on the marker's expression is validated with an external data set, due to the possible differences in sequencing on different platforms, the model's effect may be affected. We constructed a 0-or-1 matrix to screen markers to minimize the errors caused by expression variations. Besides, we optimized the process of model construction and utilized various methods to verify the effectiveness of the signature. Based on the results, we believe that the signature we constructed is acceptable. However, the verification of the signature in a larger number of samples is still necessary. We will continue to collect samples in future clinical work and expand the verification scope for further evaluation.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://tcga-data.nci.nih.gov/tcga/.

## AUTHOR CONTRIBUTIONS

LX and MS designed the work. TZ and YW collected and integrated the data. MS analyzed the data and prepared the manuscript. TZ, YW, MS, and WH edited and revised the manuscript. All authors approved the final manuscript.

## FUNDING

This research was funded by the National Natural Science Foundation of China (No. 81871911 to WH and Nos. 81972237 and 81772623 to LX), and the National Key Research and Development Program of China (No. 2018YFC1312103 to LX).

## ACKNOWLEDGMENTS

We express gratitude to the public database TCGA.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.695130/full#supplementary-material

## REFERENCES

Ahluwalia, P., Mondal, A. K., Bloomer, C., Fulzele, S., Jones, K., Ananth, S., et al. (2019). Identification and clinical validation of a Novel 4 gene-signature with prognostic utility in colorectal cancer. *Int. J. Mol. Sci.* 20:3818. doi: 10.3390/ijms20153818

Arnold, M., Abnet, C. C., Neale, R. E., Vignat, J., Giovannucci, E. L., McGlynn, K. A., et al. (2020). Global burden of 5 major types of gastrointestinal cancer. *Gastroenterology* 159, 335–349.e15.

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683–691. doi: 10.1136/gutjnl-2015-310912

Atianand, M. K., Caffrey, D. R., and Fitzgerald, K. A. (2017). Immunobiology of Long Noncoding RNAs. *Annu. Rev. Immunol.* 35, 177–198.

Bian, X., Sun, Y. M., Wang, L. M., and Shang, Y. L. (2021). ELK1-induced upregulation lncRNA LINC02381 accelerates the osteosarcoma tumorigenesis through targeting CDCA4 via sponging miR-503-5p. *Biochem. Biophys. Res. Commun.* 548, 112–119. doi: 10.1016/j.bbrc.2021.02.072

Chan, J. J., and Tay, Y. (2018). Noncoding RNA:RNA regulatory networks in cancer. *Int. J. Mol. Sci.* 19:1310.

Chen, L., Cai, Z., Lyu, K., Cai, Z., and Lei, W. (2021). A novel immune-related long non-coding RNA signature improves the prognosis prediction in the context of head and neck squamous cell carcinoma. *Bioengineered* 12, 2311–2325. doi: 10.1080/21655979.2021.1943284

Chen, X., Zhang, Z., Ma, Y., Su, H., Xie, P., and Ran, J. (2020). LINC02381 promoted cell viability and migration via targeting miR-133b in cervical cancer cells. *Cancer Manag. Res.* 12, 3971–3979. doi: 10.2147/cmar.s237285

Chen, Y. G., Satpathy, A. T., and Chang, H. Y. (2017). Gene regulation in the immune system by long noncoding RNAs. *Nat. Immunol.* 18, 962–972.

Cheng, C., Li, H., Zheng, J., Xu, J., Gao, P., and Wang, J. (2020). FENDRR Sponges miR-424-5p to inhibit cell proliferation, migration and invasion in colorectal cancer. *Technol. Cancer Res. Treat.* 19:1533033820980102.

Clarke, C. N., Lee, M. S., Wei, W., Manyam, G., Jiang, Z. Q., Lu, Y., et al. (2017). Proteomic features of colorectal cancer identify tumor subtypes independent of oncogenic mutations and independently predict relapse-free survival. *Ann. Surg. Oncol.* 24, 4051–4058. doi: 10.1245/s10434-017-6054-5

Farkona, S., Diamandis, E. P., and Blasutig, I. M. (2016). Cancer immunotherapy: the beginning of the end of cancer? *BMC Med.* 14:73.

Guo, L., Wang, C., Qiu, X., Pu, X., and Chang, P. (2020). Colorectal cancer immune infiltrates: significance in patient prognosis and immunotherapeutic efficacy. *Front. Immunol.* 11:1052.

Han, Y. Y., Liu, K., Xie, J., Li, F., Wang, Y., and Yan, B. (2020). LINC00114 promoted nasopharyngeal carcinoma progression and radioresistance in vitro and in vivo through regulating ERK/JNK signaling pathway via targeting miR-203. *Eur. Rev. Med. Pharmacol. Sci.* 24, 2491–2504.

He, Z., Wang, X., Huang, C., Gao, Y., Yang, C., Zeng, P., et al. (2018). The FENDRR/miR-214-3P/TET2 axis affects cell malignant activity via RASSF1A methylation in gastric cancer. *Am. J. Transl. Res.* 10, 3211–3223.

Hong, W., Liang, L., Gu, Y., Qi, Z., Qiu, H., Yang, X., et al. (2020). Immune-Related lncRNA to construct novel signature and predict the immune landscape of human hepatocellular carcinoma. *Mol. Ther. Nucleic Acids* 22, 937–947. doi: 10.1016/j.omtn.2020.10.002

Jafarzadeh, M., and Soltani, B. M. (2020). Long noncoding RNA LOC400043 (LINC02381) inhibits gastric cancer progression through regulating wnt signaling pathway. *Front. Oncol.* 10:562253.

Jafarzadeh, M., Soltani, B. M., Soleimani, M., and Hosseinkhani, S. (2020). Epigenetically silenced LINC02381 functions as a tumor suppressor by regulating PI3K-Akt signaling pathway. *Biochimie* 17, 63–71. doi: 10.1016/j.biochi.2020.02.009

Li, Y., Zhang, W., Liu, P., Xu, Y., Tang, L., Chen, W., et al. (2018). Long non-coding RNA FENDRR inhibits cell proliferation and is associated with good prognosis in breast cancer. *Onco Targets Ther.* 11, 1403–1412. doi: 10.2147/ott.s149511

Li, Z., Wang, D., and Yin, H. (2020). A seven immune-related lncRNA signature predicts the survival of patients with colon adenocarcinoma. *Am. J. Transl. Res.* 12, 7060–7078.

Lin, Y., Pan, X., Chen, Z., Lin, S., and Chen, S. (2020). Identification of an immune-related Nine-lncRNA signature predictive of overall survival in colon cancer. *Front. Genet.* 11:318.

Liu, J., and Du, W. (2019). LncRNA FENDRR attenuates colon cancer progression by repression of SOX4 protein. *Onco Targets Ther.* 12, 4287–4295. doi: 10.2147/ott.s195853

Liu, S., Cao, Q., An, G., Yan, B., and Lei, L. (2020). Identification of the 3-lncRNA signature as a prognostic biomarker for colorectal cancer. *Int. J. Mol. Sci.* 21:9359. doi: 10.3390/ijms21249359

Lv, L., He, L., Chen, S., Yu, Y., Che, G., Tao, X., et al. (2019). Long Non-coding RNA LINC00114 facilitates colorectal cancer development through EZH2/DNMT1-Induced miR-133b suppression. *Front. Oncol.* 9:1383.

Lv, Y., Lin, S. Y., Hu, F. F., Ye, Z., Zhang, Q., Wang, Y., et al. (2020). Landscape of cancer diagnostic biomarkers from specifically expressed genes. *Brief. Bioinform.* 21, 2175–2184. doi: 10.1093/bib/bbz131

Martinez-Romero, J., Bueno-Fortes, S., Martín-Merino, M., Ramirez, D. M. A., and De Las, R. J. (2018). Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genom.* 19:857.

Munteanu, M. C., Huang, C., Liang, Y., Sathiaseelan, R., Zeng, X., and Liu, L. (2020). Long non-coding RNA FENDRR regulates IFNγ-induced M1 phenotype in macrophages. *Sci. Rep.* 10:13672.

Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z. R., et al. (2013). Long-term colorectal-cancer incidence and mortality after lower endoscopy. *N. Engl. J. Med.* 369, 1095–1105. doi: 10.1056/nejmoa1301969

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58. doi: 10.1016/j.cell.2010.09.001

Pagès, F., Mlecnik, B., Marliot, F., Bindea, G., Ou, F. S., Bifulco, C., et al. (2018). International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* 391, 2128–2139.

Picard, E., Verschoor, C. P., Ma, G. W., and Pawelec, G. (2020). Relationships between immune landscapes, genetic subtypes and responses to immunotherapy in Colorectal cancer. *Front. Immunol.* 11:369.

Ping, S., Wang, S., He, J., and Chen, J. (2021). Identification and validation of immune-related lncRNA signature as a prognostic model for skin Cutaneous Melanoma. *Pharmgenomics Pers. Med.* 14, 667–681. doi: 10.2147/pgpm.s310299

Qin, F., Xu, H., Wei, G., Ji, Y., Yu, J., Hu, C., et al. (2021). A Prognostic model based on the immune-related lncRNAs in Colorectal cancer. *Front. Genet.* 12:658736.

Qin, X., Lu, M., Zhou, Y., Li, G., and Liu, Z. (2019). LncRNA FENDRR represses proliferation, migration and invasion through suppression of survivin in cholangiocarcinoma cells. *Cell Cycle* 18, 889–897. doi: 10.1080/15384101.2019.1598726

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* 19, 619–625. doi: 10.1038/nm.3175

Saus, E., Brunet-Vega, A., Iraola-Guzmán, S., Pegueroles, C., Gabaldón, T., and Pericay, C. (2016). Long Non-Coding RNAs as potential novel prognostic biomarkers in colorectal cancer. *Front. Genet.* 7:54.

Shen, J., Feng, X. P., Hu, R. B., Wang, H., Wang, Y. L., Qian, J. H., et al. (2021). N-methyladenosine reader YTHDF2-mediated long noncoding RNA FENDRR degradation promotes cell proliferation in endometrioid endometrial carcinoma. *Lab. Invest.* 101, 775–784. doi: 10.1038/s41374-021-00543-3

Shen, Y., Peng, X., and Shen, C. (2020). Identification and validation of immune-related lncRNA prognostic signature for breast cancer. *Genomics* 112, 2640–2646. doi: 10.1016/j.ygeno.2020.02.015

Sinicrope, F. A., Shi, Q., Allegra, C. J., Smyrk, T. C., Thibodeau, S. N., Goldberg, R. M., et al. (2017). Association of DNA mismatch repair and mutations in BRAF and KRAS with survival after recurrence in stage III Colon cancers : a secondary analysis of 2 randomized clinical trials. *JAMA Oncol.* 3, 472–480. doi: 10.1001/jamaoncol.2016.5469

Sun, Y., Peng, P., He, L., and Gao, X. (2020). Identification of lnc RNAs related to prognosis of patients with Colorectal cancer. *Technol. Cancer Res. Treat.* 19:1533033820962120.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660

Szafranski, P., and Stankiewicz, P. (2021). Long non-coding RNA FENDRR: gene structure, expression, and biological relevance. *Genes* 12:177. doi: 10.3390/genes12020177

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018

Wang, X., Zhou, J., Xu, M., Yan, Y., Huang, L., Kuang, Y., et al. (2018). A 15-lncRNA signature predicts survival and functions as a ceRNA in patients with colorectal cancer. *Cancer Manag. Res.* 10, 5799–5806. doi: 10.2147/cmar.s178732

Wu, Y., Zhang, L., He, S., Guan, B., He, A., Yang, K., et al. (2020). Identification of immune-related LncRNA for predicting prognosis and immunotherapeutic response in bladder cancer. *Aging* 12, 23306–23325.

Xiong, Y., Wang, K., Zhou, H., Peng, L., You, W., and Fu, Z. (2018). Profiles of immune infiltration in colorectal cancer and their clinical significant: a gene expression-based study. *Cancer Med.* 7, 4496–4508. doi: 10.1002/cam4.1745

Ye, L. C., Zhu, X., Qiu, J. J., Xu, J., and Wei, Y. (2015). Involvement of long non-coding RNA in colorectal cancer: from benchtop to bedside (Review). *Oncol. Lett.* 9, 1039–1045. doi: 10.3892/ol.2015.2846

Yin, S. L., Xiao, F., Liu, Y. F., Chen, H., and Guo, G. C. (2019). Long non-coding RNA FENDRR restrains the aggressiveness of CRC via regulating miR-18a-5p/ING4 axis. *J. Cell Biochem.* doi: 10.1002/jcb.29555 Online ahead of print.

Yu, Z., Zhao, H., Feng, X., Li, H., Qiu, C., Yi, X., et al. (2019). Long Non-coding RNA FENDRR acts as a miR-423-5p sponge to suppress the treg-mediated

immune escape of hepatocellular Carcinoma cells. *Mol. Ther. Nucleic Acids* 17, 516–529. doi: 10.1016/j.omtn.2019.05.027

Zhang, G., Wang, Q., Zhang, X., Ding, Z., and Liu, R. (2019). LncRNA FENDRR suppresses the progression of NSCLC via regulating miR-761/TIMP2 axis. *Biomed. Pharmacother.* 118:109309. doi: 10.1016/j.biopha.2019.109309

Zhang, Y. Q., Chen, X., Fu, C. L., Zhang, W., Zhang, D. L., Pang, C., et al. (2019). FENDRR reduces tumor invasiveness in prostate cancer PC-3 cells by targeting CSNK1E. *Eur. Rev. Med. Pharmacol. Sci.* 23, 7327–7337.

Zhou, W., Zhang, S., Li, H. B., Cai, Z., Tang, S., Chen, L. X., et al. (2020). Development of prognostic indicator based on autophagy-related lncRNA analysis in Colon Adenocarcinoma. *Biomed. Res. Int.* 2020:9807918.

Zhu, Y., Zhang, X., Wang, L., Zhu, X., Xia, Z., Xu, L., et al. (2020). FENDRR suppresses cervical cancer proliferation and invasion by targeting miR-15a/b-5p and regulating TUBA1A expression. *Cancer Cell Int.* 20:152.

**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: 28 October 2021
doi: 10.3389/fgene.2021.709329

# Construction and Validation of a Novel Ferroptosis-Related lncRNA Signature to Predict Prognosis in Colorectal Cancer Patients

*Wenqi Zhang*[1], *Daoquan Fang*[1], *Shuhan Li*[1], *Xiaodong Bao*[1], *Lei Jiang*[1]* *and Xuecheng Sun*[2]*

[1]Central Laboratory, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China, [2]Department of Gastroenterology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

**Background:** Colorectal cancer (CRC) ranks as the third most common malignancy worldwide but a reliable prognostic biomarker of CRC is still lack. Thus, the purpose of our study was to explore whether ferroptosis - related lncRNAs could predict the prognosis of CRC.

**Methods:** The mRNA expression profiling of colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) patients in The Cancer Genome Atlas (TCGA) database were downloaded. Univariate Cox and multivariate Cox regression analyses was used to obtain prognostic differently expressed ferroptosis-related lncRNAs (DE-FLs) and a risk signature was developed. Quantitative polymerase chain reaction (q-PCR) was used to validated the different expressions of DE-FLs. The calibration curves, C-index and the receiver operating characteristic (ROC) curves were applied to evaluate the accuracy of nomogram. Gene set enrichment analyses (GSEA) were carried out to explore the biological mechanism between high- and low-risk group and the potential regulated pathway of prognostic DE-FLs in CRC.

**Results:** Forty-nine DE-FLs were identified between CRC and normal tissue. Then, a 4-DE-FLs (AC016027.1, AC099850.3, ELFN1-AS1, and VPS9D1-AS1) prognostic signature model was generated. AC016027.1 was downregulated in CRC tissue; VPS9D1-AS1 and ELFN1-AS1 were upregulated by q-PCR. The model had a better accuracy presenting by 1-, 3-, and 5-years ROC curve (AUC ≥0.6), and identified survival probability ($p < 0.05$) well. Moreover, the risk signature could play as an independent factor of CRC ($p < 0.05$). Further, a nomogram including age, pathologic stage, T stage, and risk score with good prognostic capability (C-index = 0.789) was constructed. In addition, we found biological pathways mainly related to metabolism and apoptosis were down-regulated in high-risk group who with poor outcome. Finally, the functional enrichment showed prognostic DE-FLs may significantly impact bile secretion in CRC.

**Conclusion:** A risk model and nomogram based on ferroptosis-related lncRNAs were created to predict 1-, 3-, and 5-years survival probability of CRC patients. Our data suggested that the prognostic lncRNAs could serve as valuable prognostic marker.

**Keywords: ferroptosis, lncRNAs, colorectal cancer, prognosis, risk signature, nomogram**

OPEN ACCESS

***Edited by:***
*Suman Ghosal,*
*National Institutes of Health (NIH),*
*United States*

***Reviewed by:***
*Balaji Banoth,*
*St. Jude Children's Research Hospital,*
*United States*
*Arijita Sarkar,*
*University of Southern California,*
*United States*

***Correspondence:**
*Lei Jiang*
*jiangleistone79@163.com*
*Xuecheng Sun*
*sxc1979@163.com*

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

***Received:*** *13 May 2021*
***Accepted:*** *01 October 2021*
***Published:*** *28 October 2021*

***Citation:***
*Zhang W, Fang D, Li S, Bao X, Jiang L and Sun X (2021) Construction and Validation of a Novel Ferroptosis-Related lncRNA Signature to Predict Prognosis in Colorectal Cancer Patients. Front. Genet. 12:709329. doi: 10.3389/fgene.2021.709329*

## INTRODUCTION

CRC is the third major cause of cancer mortality in industrialized countries, which seriously endangers human health (Siegel et al., 2018). The relevant data of the World Health Organization (WHO) shows that the incidence of CRC has begun to decline in some developed countries, but it still remains increase in the developing world (Zeuner et al., 2014). In new CRC diagnoses, 20% of patients have metastases at presentation and another 25% with localized disease will later develop metastases (Biller and Schrag, 2021), therefore early disease diagnosis is especially critical. Despite the encouraging amelioration in CRC diagnostic and therapeutic methods, a relatively high proportion of CRC patients suffering from poor survival outcomes still exists because of late disease detection and lacking availability of adequate risk-assessment biomarkers (Yang et al., 2021). Thus, it is urgently to identify novel and reliable biomarkers for the individualized diagnosis of CRC, which may potentially improve overall outcome of this disease.

Ferroptosis is considered a nonapoptotic, iron-dependent form of cell death with three hallmarks including oxidation of polyunsaturated fatty acid, redox active iron and lipid peroxide repair loss (Dixon et al., 2012; Jiang X. et al., 2021). Cancer cells are more vulnerable to ferroptosis due to their high demand of iron to support fast proliferation (Hassannia et al., 2019). Recently, evidence is emerging that ferroptosis has a tumor-suppressor effect that could be employed for tumor treatment (Stockwell et al., 2017). For example, BAP1 restrains tumor progression partly through SLC7A11 and ferroptosis (Zhang et al., 2018). Ferroptosis also has great potential to eliminate malignant cells which are resistant to conventional therapy. Shin et al. (2018) reported that inhibition of GPX4 made chemoresistance cancer cells more vulnerable to ferroptosis.

Long non-coding RNAs (lncRNAs) are transcripts with more than 200 nucleotides in length, but being not translated into proteins (Mercer et al., 2009). LncRNA possesses variously functional activity including RNA decay, gene expression and control, RNA splicing, miRNA regulation, protein folding (Chen et al., 2017). Some lncRNAs can also prevent oxidation and thus inhibit ferroptosis as rival endogenous RNAs (Jiang N. et al., 2021). A study found that LINC00336 acted as an oncogene, which bound ELAVL1 using nucleotides 1901-2107 of LINC00336 and the RRM interaction domain and key amino acids of ELAVL1 (aa 101-213), inhibiting ferroptosis (Wang M. et al., 2019). Moreover, recent evidences indicated that identification of lncRNAs could help early disease detection and advance therapy outcomes in CRC patients (Yang et al., 2021). Given their roles in malignant development and disorder of expression in CRC patients, lncRNA is undoubtedly potential to be a reliable diagnostic and prognostic biomarker for CRC.

In this study, we hypothesized that ferroptosis-related lncRNAs might be promising prognostic biomarkers for CRC patients. We analyzed the correlation between the expression of ferroptosis-related lncRNAs with survival and the clinicopathological parameter of CRC patients from TCGA database. Moreover, we constructed a prognostic signature based on four ferroptosis-related lncRNAs and assessed its ability to independently and accurately predict the prognosis of CRC patients. The work flow of this study is illustrated in **Figure 1**.

## MATERIAL AND METHODS

### Acquisition of CRC Data

The level 3 RNA-Sequencing (RNA-Seq) dataset and corresponding clinical information of COAD and READ cancer patients were downloaded from TCGA data portal (https://portal.gdc.cancer.gov/) which was updated at November 9, 2020. There were 673 samples in the dataset, including 622 CRC and 51 normal ones. After excluding the samples without survival information, a total of 582 samples were enrolled in our study and randomly assigned into training and validation set at a ratio of 7:3. In addition, 382 ferroptosis-related genes (FRGs) were obtained from the FerrDb database (http://www.zhounan.org/ferrdb/). GENCODE v22 was used for gene annotation (https://www.gencodegenes.org/human/release_22.html).

### Identification of Differently Expressed Ferroptosis-Related lncRNAs in CRC.

Differently expressed genes (DEGs) and lncRNAs (DLRs) between CRC and normal samples were identified by "limma" package in *R* language. DEGs and DLRs meeting $|\log2FC| > 1$ and $p$_value <0.05 were considered as significantly expressed. The "ggplot2" package were used to construct volcano plot of these DEGs and DLRs. Then we extracted differently expressed ferroptosis-related genes (DE-FGs) from the overlap of FRGs and DEGs, which analyzed by Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/). Pearson analysis were performed to screen the DE-FLs with criterion of $|cor| > 0.3$ and $p < 0.05$.

### Construction of Prognostic DE-FLs Signature

DE-FLs which significantly associated with CRC prognosis were identified by univariate Cox regression model (cut off <0.2). Then the candidate DE-FLs were entered into a stepwise multivariate Cox regression analysis and constructed a prognostic signature model according to Akaike Information Criterion (AIC). The expression levels of prognostic DE-FLs in normal and CRC tumor samples in TCGA database were checked by Wilcoxon test. And the survival curve of CRC patients based on DE-FLs expression were draw. Further, the risk score of individual patients was established based on the summation of coefficients and expression level of each prognostic DE-FLs according to the following formula:

$$\text{Risk score} = \sum (Coefi \text{ X } Expi)$$

Thus, the CRC patients in each set were classified into high- and low-risk group reference the median risk score.

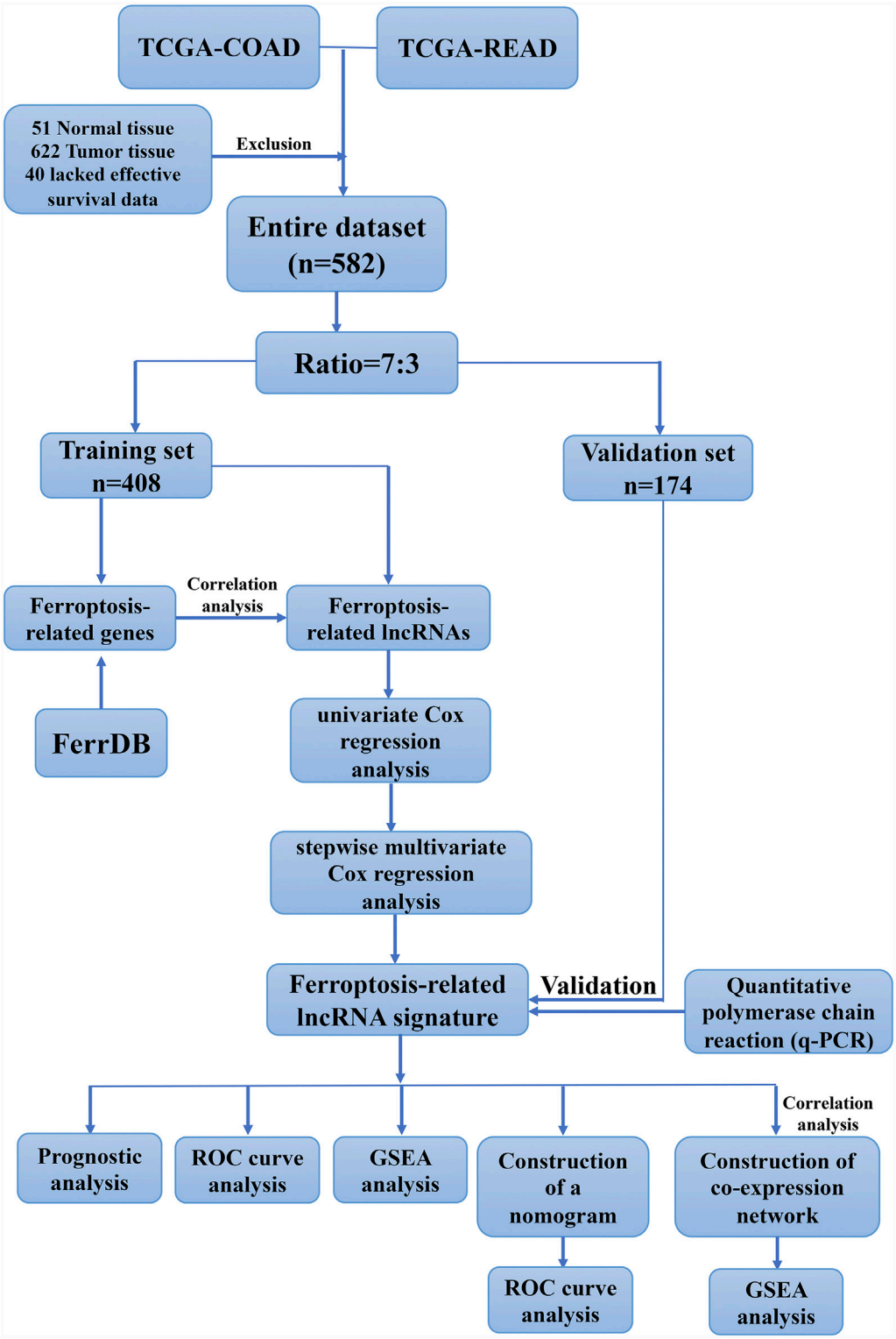



**FIGURE 1 |** Work flow for the construction of a risk signature in colorectal cancer.

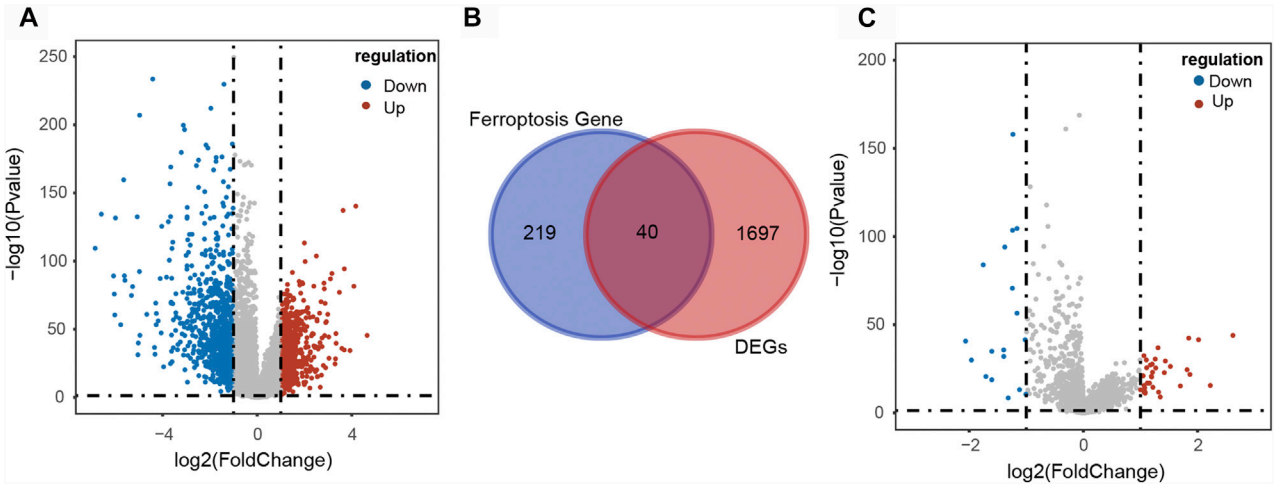



**FIGURE 2 |** Identification of differentially expressed ferroptosis-related lncRNAs in CRC patients. **(A)** Volcano plot of the DEGs in CRC patients. **(B)** Venn diagram showed the intersection of 40 differentially expressed ferroptosis-related genes. **(C)** Volcano plot of the differentially expressed lncRNAs in CRC patients.

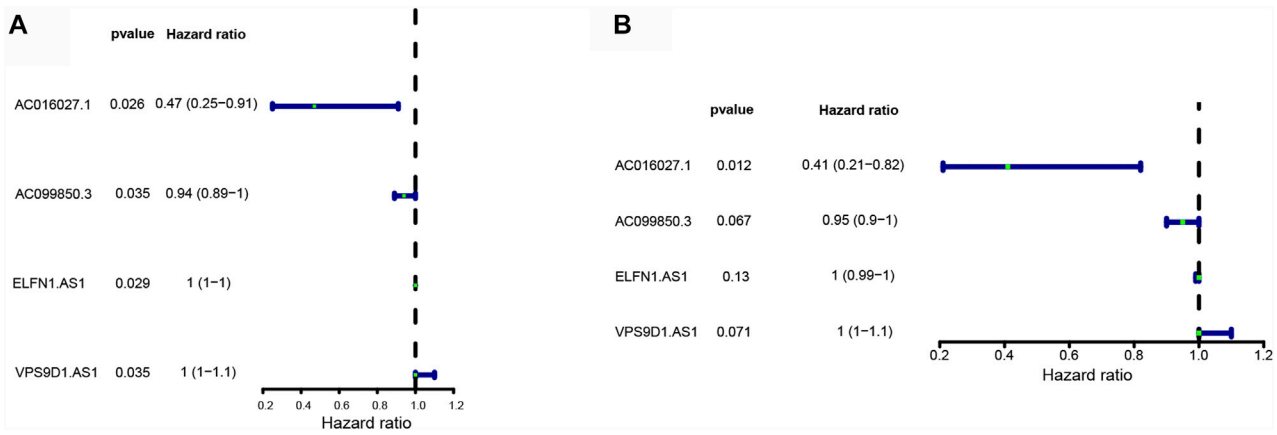



**FIGURE 3 |** Construction of ferroptosis-related lncRNA prognostic signature in CRC patients. **(A)** Forest plots showed 4 DE-FLs selected by the univariate Cox regression. **(B)** Forest plots showed four prognostic DE-FLs identified by the multivariate Cox regression.

## Evaluation of the Prognostic Signature

To evaluate the valuable of prognostic signature in each set, the differences in patients' survival between high- and low-risk group were evaluated by Kaplan-Meier curve analyses and log-rank test ($p < 0.05$). Then, the 1-, 3-, and 5-years ROC curves were employed to compare the specificity and sensitivity of the survival prediction based on the prognostic signature *via* "pROC" package. Moreover, the relationship between risk score and clinical characteristics were analyzed by *t-test*.

## Establishment of Nomogram for CRC Prognostic Prediction

To identify independent prognostic factors of CRC, the univariate and multivariate Cox regression analyses were performed to evaluate the risk score and other clinical variables such as age, gender, grade, M stage, T stage, and N stage. The factor with $p < 0.05$ was considered statistically significant. Then, we integrated all of the independent prognostic factors to build a nomogram by "rms" package for inspecting the probability of 1-, 3-, and 5-years overall survival (OS) of the CRC patients. The discrimination and predictive ability of the nomogram in CRC were assessed with calibration curve and a C-index indicate. We also plotted ROC curve and calculated the area under the ROC curve (AUC) values based on total points of nomogram.

## Functional Annotation

Functional enrichment analyses were performed to explore the underlying mechanism of prognostic signature. Firstly, the GSEA software was conducted for Gene Ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analyses in high- and low-risk group in CRC. Then, the co-expression of

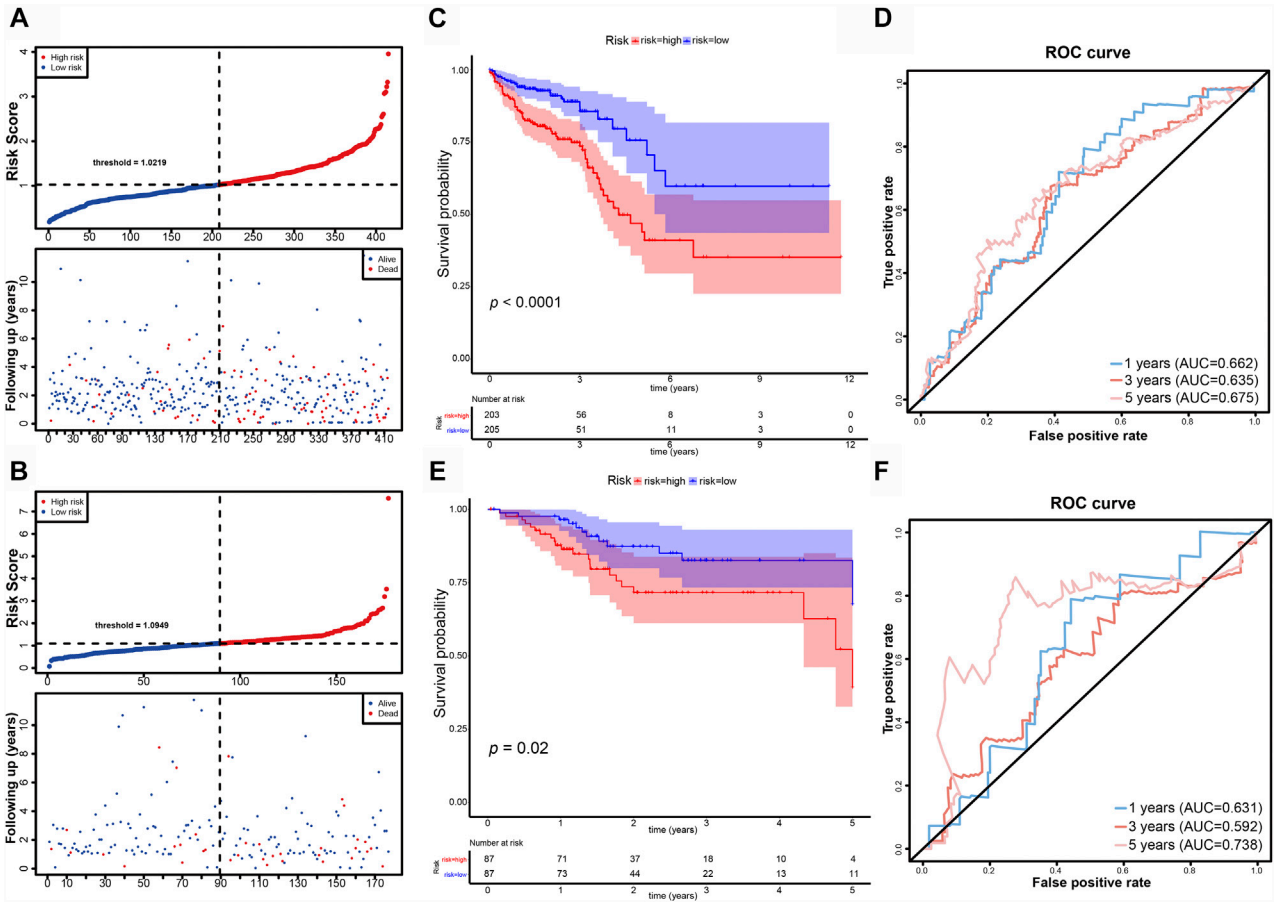



**FIGURE 4 |** Evaluation and validation of the DE-FLs prognostic signature in CRC patients. **(A, B)** Distribution and Survival status of CRC patients with different risk scores in training set and validation set. **(C)** Kaplan–Meier survival curve of high- and low-risk group in training set. **(D)** The 1-, 3-, and 5-years ROC curve for risk model in training set. **(E)** Kaplan–Meier survival curve of high- and low-risk group in testing set. **(F)** The 1-, 3-, and 5-years ROC curve for risk model in testing set.

the signature DE-FLs and mRNA was assessed with Pearson correlation analyses. The mRNAs in co-expression pairs whose correlation coefficient >0.7 and $p < 0.0.5$ were selected for functional annotation by "clusterProfiler" package. Biological processes were collected from the GSEA (https://www.gsea-msigdb.org/gsea/index.jsp). The *p*. adjust-value below 0.05 was considered significant.

## Quantitative Polymerase Chain Reaction

Thirty pairs of CRC tumor and adjacent tissues were collected from clinical patients at The First Affiliated Hospital of Wenzhou Medical University and preserved in -80 C refrigerator. All patients gave the written informed consent. All assay regimens gained the approval of the Ethics Committees in Clinical Research of the First Affiliated Hospital of Wenzhou Medical University and the corresponding ethical approval code was KY2021-R005. Total RNA was isolated from tissue samples using TRIzol reagent (Invitrogen, Carlsbad, CA). Then PrimeScript™ RT Master Mix kit (TaKaRa, Japan) kit was used to synthesize cDNA according to the instruction manual. q-PCR for cDNA amplification was performed with Green™ Premix Ex Taq™ II (TaKaRa) kit. GAPDH was used as endogenous control and primers were shown in **Supplementary Table S1**. The expression of signature DE-FLs were normalized using the relative quantification method of 2-ΔΔCt.

## Statistical Analysis

Statistical procedures were applied using R v.4.0.3 and Prism v.8.0.0. The Student's *t-test* or the Wilcoxon test was utilized for differences analysis. $p < 0.05$ was statistically significant.

# RESULTS

## Screening of DE-FLs in CRC

A total of 1737 DEGs, including 780 up-regulated and 957 down-regulated genes, between CRC and normal samples were obtained from TCGA (**Figure 2A**). And 382 FRGs were downloaded from the FerrDb database. Then 40 DE-FGs were identified by Venn diagram (**Figure 2B**) and the correlation among them was shown in **Supplementary Figure S1**. In addition, there were 51 DLRs selected in CRC, including 33 up-regulated and 18 down-regulated

**TABLE 1 |** The relationship of CRC patients clinical feature and the DE-FLs model.

| | Total (n = 408) | Expression | | *p*_value |
|---|---|---|---|---|
| | | High (n = 205) | Low (n = 203) | |
| Gender | | | | |
| female | 192 (47.1%) | 99 (48.3%) | 93 (45.8%) | 0.687 |
| male | 216 (52.9%) | 106 (51.7%) | 110 (54.2%) | |
| Age (years) | | | | 0.138 |
| ≥ 60 | 290 (71.1%) | 153 (74.6%) | 137 (67.5%) | |
| <60 | 118 (28.9%) | 52 (25.4%) | 66 (32.5%) | |
| Pathologic stage | | | | 0.592 |
| stage_I | 67 (16.4%) | 32 (15.6%) | 35 (17.2%) | |
| stage_II | 152 (37.3%) | 74 (36.1%) | 78 (38.4%) | |
| stage_III | 122 (29.9%) | 61 (29.8%) | 61 (30.0%) | |
| stage_IV | 53 (13.0%) | 32 (15.6%) | 21 (10.3%) | |
| unknown | 14 (3.4%) | 6 (2.9%) | 8 (3.8%) | |
| T stage | | | | |
| T1 | 9 (2.2%) | 6 (2.9%) | 3 (1.5%) | 0.568 |
| T2 | 71 (17.4%) | 32 (15.6%) | 39 (19.2%) | |
| T3 | 290 (71.1%) | 146 (71.2%) | 144 (70.9%) | |
| T4 | 37 (9.1%) | 20 (9.8%) | 17 (8.4%) | |
| unknown | 1 (0.2%) | 1 (0.5%) | 0 (0%) | |
| M stage | | | | 0.174 |
| M0 | 302 (74.0%) | 148 (72.2%) | 154 (75.9%) | |
| M1 | 52 (12.7%) | 32 (15.6%) | 20 (9.9%) | |
| MX | 45 (11.0%) | 19 (9.3%) | 26 (12.8%) | |
| unknown | 9 (2.2%) | 6 (2.9%) | 3 (1.5%) | |
| N stage | | | | 0.285 |
| N0 | 231 (56.6%) | 112 (54.6%) | 119 (58.6%) | |
| N1 | 92 (22.5%) | 42 (20.5%) | 50 (24.6%) | |
| N2 | 82 (20.1%) | 49 (23.9%) | 33 (16.3%) | |
| NX | 2 (0.5%) | 1 (0.5%) | 1 (0.5%) | |
| unknown | 1 (0.2%) | 1 (0.5%) | 0 (0%) | |

lncRNAs (**Figure 2C**). Finally, we performed Pearson correlation analysis to calculate the correlation between the DLRs and DE-FGs, and used $|cor| > 0.3$ and $p < 0.05$ as the selection criteria. As shown in **Supplementary Data S1**, 49 lncRNAs were acquired and termed as DE-FLs for further research.

## Construction of Ferroptosis-Related lncRNA Prognostic Signature in CRC

Univariate Cox regression analysis showed that 4 DE-FLs (AC016027.1, AC099850.3, ELFN1-AS1, and VPS9D1-AS1) were associated with OS in CRC, except VPS9D1-AS1 played a risk factor with HR > 1, the others acted as protectors with HR < 1 (**Figure 3A**). Multivariate Cox regression analysis further ascertained these 4 DE-FLs with prognostic significance (**Figure 3B**). Thus, they were employed to construct a prognostic signature model. **Supplementary Figure S2** showed the survival probability of CRC patients in high- and low-expression DE-FLs respectively.

## Evaluation and Validation of the Ferroptosis-Related lncRNA Prognostic Signature

CRC patients in the TCGA dataset were classified into high-risk (training set n = 205; validation set n = 87; total n = 292) and low-risk (training set n = 203; validation set n = 87; total n = 290) groups using the median risk-score as the cutoff point (**Figures 4A,B**). Kaplan-Meier survival curve analysis showed that the OS of high-risk group patients were significantly poorer than low-risk group in the training set (n = 408) (**Figure 4C**). The 5-years survival rates were approximately 40 and 75% of the high-risk and low-risk group respectively. Time-dependent ROC curve analysis showed an appropriate accuracy of the prognostic signature in predicting OS in CRC, and AUC values were 0.662 at 1 year, 0.635 at 3 years, and 0.657 at 5 years (**Figure 4D**). For further validation, we confirmed that the results in the validation set coincided with the outcomes in the training set. In the validation set (n = 174), the significant prognostic value was $p = 0.02$ (**Figure 4E**) and AUC values for 1-, three- and 5- year OS were 0.631, 0.592, and 0.738, respectively (**Figure 4F**).

Then, we conducted the correlation analysis between the risk scores and the clinical characteristics of the CRC patients in TCGA database. We found that none of the clinical feature associated with risk scores in training set (**Table 1** and **Figure 5**) and in validation set (**Supplementary Table S2**, **Supplementary Figure S3**).

## The Risk Score Is an Independent Prognostic Factor in CRC

In order to determine if the ferroptosis-related lncRNA prognostic signature was an independent prognostic factor for

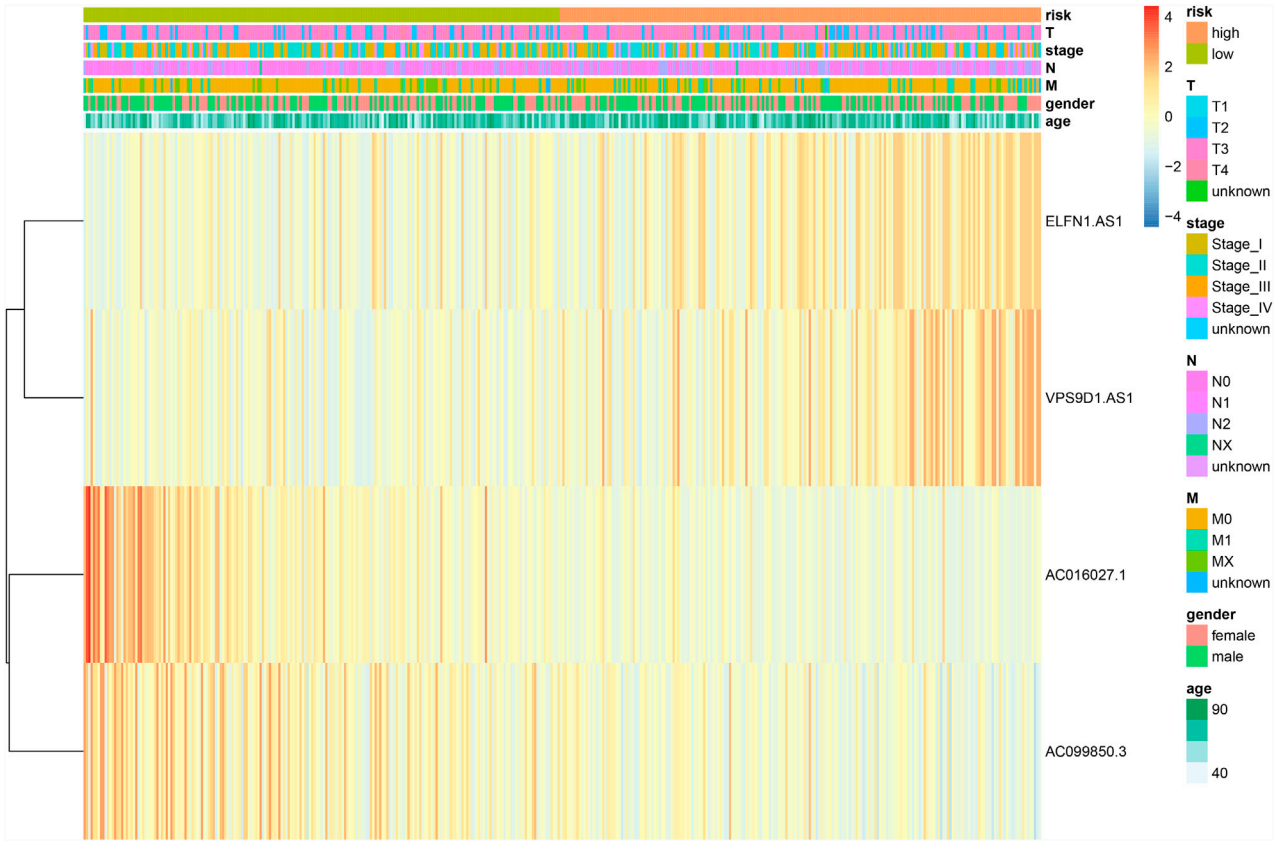



**FIGURE 5 |** Relationship between the risk score and clinical significance in training set.

CRC patients, we performed univariate and multivariate Cox regression analyses. Univariate analyses showed that age, MNT ($p < 0.001$), stage ($p < 0.001$) and the risk score were significantly associated with OS (**Figure 6A**). Multivariate analyses showed that pathologic stage ($p < 0.001$), age, T stage and risk score could act as independent prognostic factor (**Figure 6B**).

Subsequently, we developed a nomogram to predict 1-, 3-, and 5-years OS of CRC using the all of the independent prognostic factors (**Figure 6C**). The C-index for the nomogram was 0.789. The 1-year, 3-years and 5-years calibration curves showed the nomogram with an accurate prediction in CRC (**Figure 6D**). Finally, the AUC values at 1-year, 3-years and 5-years were 0.736, 0.710, 0.746, respectively (**Figure 6E**), also indicated the predictive capacity of the nomogram was reliable.

## Functional Analysis of Ferropotosis-Related lncRNA Signature

GSEA analyses were conducted to further explore the difference biological mechanism between low- and high-risk groups. In high-risk group, the biological process mainly related to chromosome (**Figure 7A**). And the pathways such as OLFACTORY_TRANSDUCTION, OOCYTE_MEIOSIS, ASCORBATE_AND_ALDARATE_METABOLISM, O_GLYCAN_BIOSYNTHESIS, STARCH_AND_ SUCROSE_METABOLISM, UBIQUITIN_MEDIATED_PROTEOLYSIS, APOPTOSIS and PROTEIN_EXPORT were significantly enriched in the high-risk group (**Figure 7B**). Except RIBOSOME, all of the others were down-regulated.

## Construction of Co-expression Network

We used Pearson correlation analyses and Cytoscape to construct lncRNA-mRNA co-expression network. When threshold parameter $|cor| > 0.7$ was set, 132 mRNA which significantly related to two prognostic DE-FLs (AC016027.1, AC099850.3) were involved in the network (**Figure 8**, **Supplementary Data S2**). Pearson correlation analysis also showed all the mRNAs ($|cor| > 0.5$) associated with these 4 DE-FLs (**Supplementary Data S3**). PEX26, SLC51B, TMEM236, CA4, and SLC26A3 ranked as top five genes high correlated with AC016027.1. PRR11, BRCA1, KPNA2, TOP2A and NCAPH were top five high correlated with AC099850.3.

## Functional Enrichment Analysis

To investigate the biological pathways regulated by the prognostic lncRNAs, we performed GO and KEGG enrichment analysis on the network-genes. Five KEGG pathways and 84 GO functional items were enriched. The top ten GO items were mainly related with small molecule catabolic process, lipid catabolic process, fatty acid metabolism process (**Figure 9A**). KEGG pathway analysis confirmed that bile secretion was the most significantly enriched pathway (**Figure 9B**).

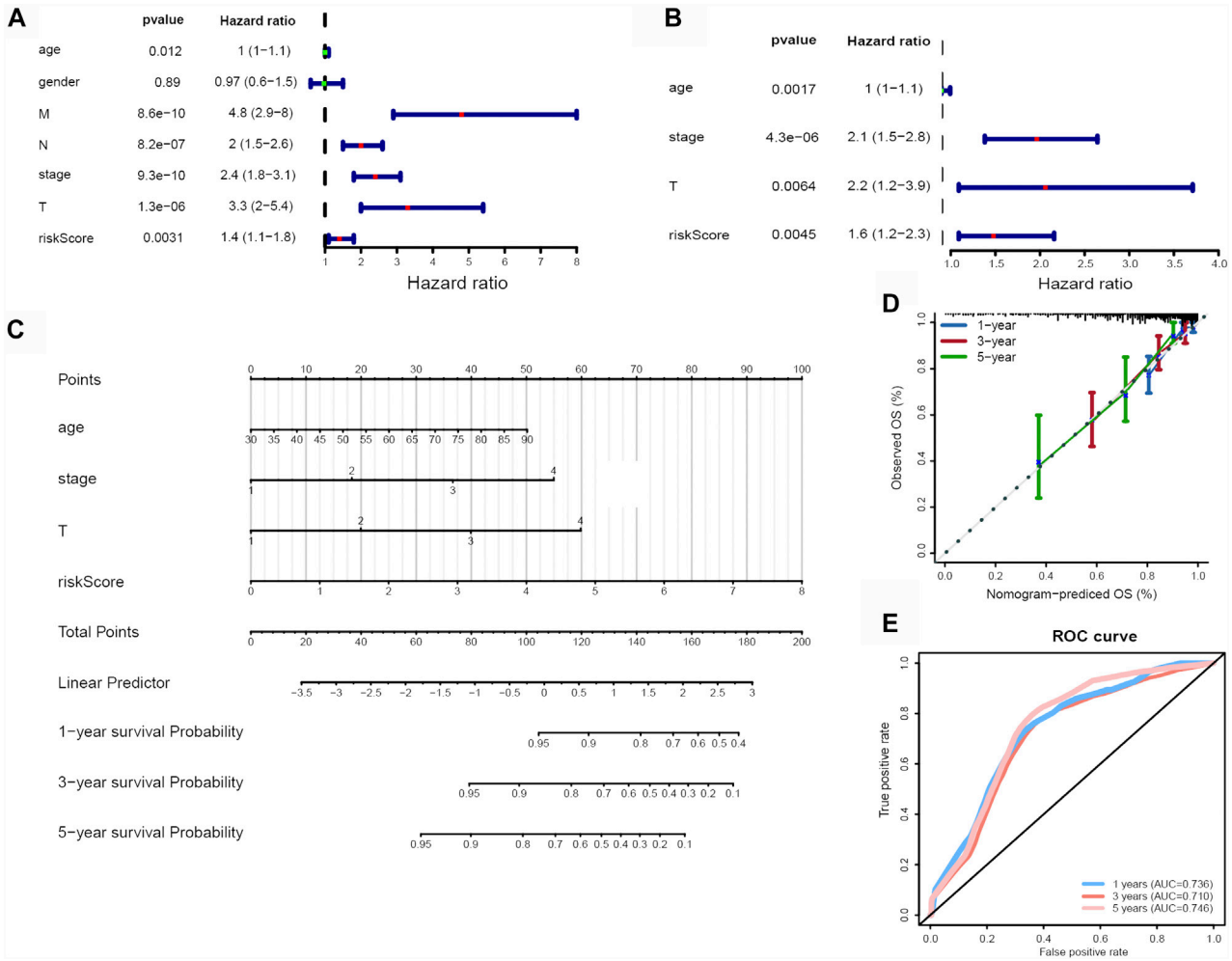



**FIGURE 6 |** Identification of independent risk factors and construction of a nomogram. **(A)** The univariate and **(B)** multivariate Cox regression analysis of risk score and clinical feature prognostic value. **(C)** A nomogram was constructed to predict 1-, 3-, and 5-years OS using risk score, stage, T and age. **(D)** Calibration curves of nomogram. **(E)** 1-, 3-, and 5-years ROC curve based on nomogram.

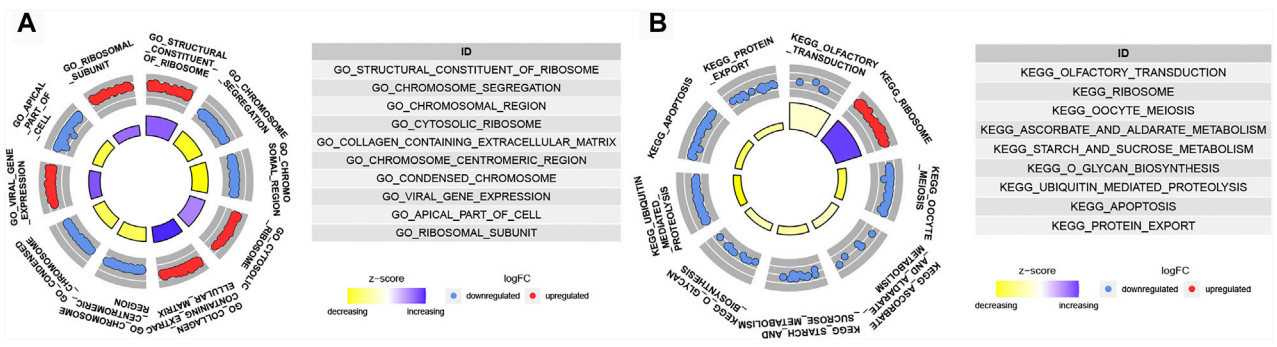



**FIGURE 7 |** Functional enrichment analysis. **(A)** GO items enriched by GSEA analysis in high risk group. **(B)** KEGG pathways enriched by GSEA analysis in high risk group.

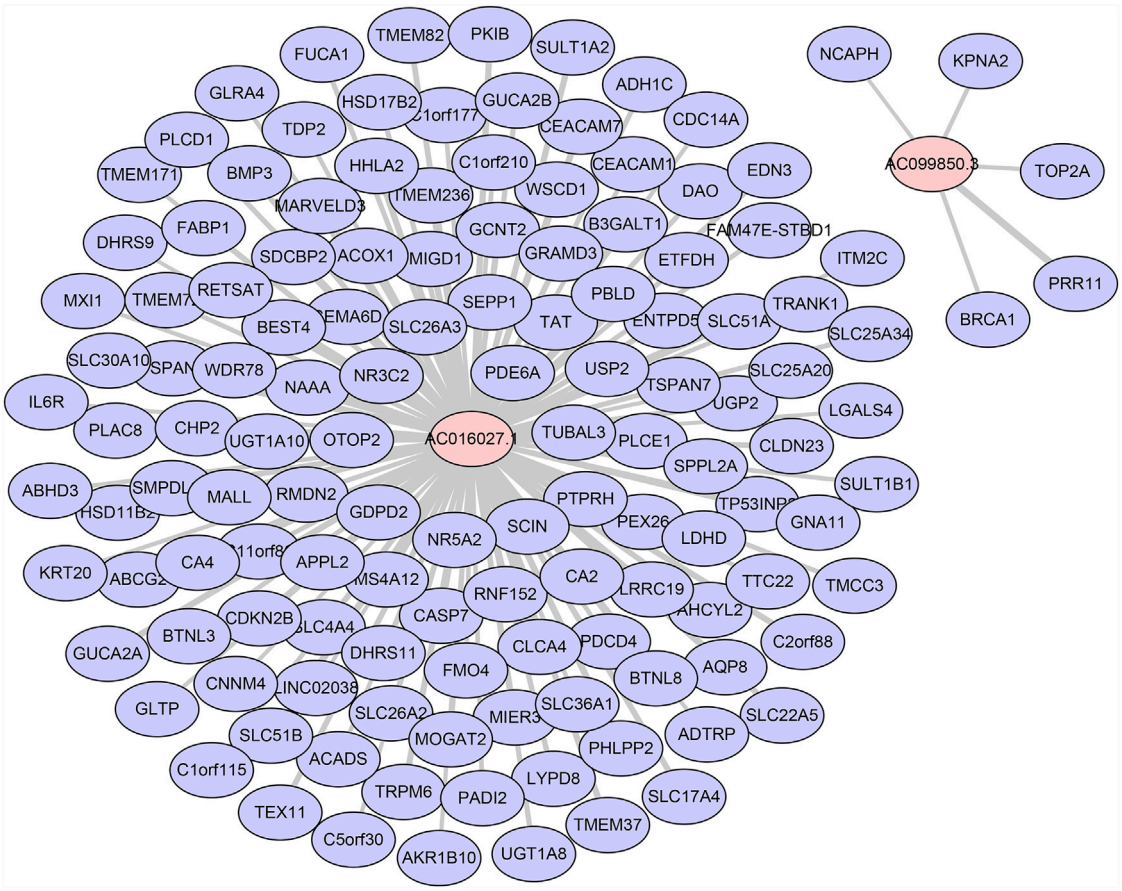



**FIGURE 8 |** Co-expression network of DE-FLs and mRNA.

## Validation of the Ferrptosis-Related Signature in Clinical CRC Tissues

The expression levels of prognostic DE-FLs were detected by q-PCR in thirty pairs of CRC tissues and adjacent tissues. As shown in **Figures 10A–D**, AC016027.1 was down-regulated in CRC from TCGA database whereas VPS9D1-AS1, ELFN1-AS1 and AC099850.3 were up-regulated. Q-PCR analyses showed that expression level of AC016027.1 was down-regulated in CRC tissues compared with paired normal tissues; VPS9D1-AS1 and ELFN1-AS1 were up-regulated (**Figures 10E–G**) which were consistent with the results from TCGA. There was no significant difference in the expression level of AC099850.3 between CRC and normal tissues (**Figure 10H**).

# DISCUSSION

A growing body of researches showed that aberrant expression of lncRNAs were associated with risk for CRC (Ni et al., 2019; Wang Y. et al., 2019; Yuan et al., 2021). Their results indicated the clinic potential of lncRNAs as stratification markers, diagnostic indicators, therapeutic targets for CRC (Lin and Yang, 2018). Ferroptosis, dependent on iron and reactive oxygen species (ROS) and characterized by lipid peroxidation, is a non-apoptotic regulated cell death (Kajarabille and Latunde-Dada, 2019). Much evidence has indicated that ferroptosis participated in multiple pathways and was critical for eradicating the carcinogenic cells (Dixon, 2017; Mou et al., 2019). Xu et al. (2020) demonstrated that knockout of SLC7A11 induced ferroptosis, thereby suppressed the progression of CRC stem cells. Previous studies which investigated the function of ferroptosis in CRC have largely concentrated on TP53 (Murphy, 2016; Xie et al., 2017). Xie et al. (2017) reported that TP53 suppressed ferroptosis triggered by erastin through the inhibition of dipeptidyl-peptidase-4 (DPP4) in a transcription-independent manner in CRC. Emerging evidences indicated that lncRNAs have been involved in tumorigenesis and tumor progression by targeting ferroptosis-related genes, such as p53RRA which restrains ferroptosis-modulated genes in a p53-dependent manner by interacting with G3BP1 (Mao et al., 2018). Mao C *et al.* validated a novel mechanism that p53RRA regulates apoptosis and ferroptosis in cancer. Also, p53RRA can be used as a prognostic marker in lung adenocarcinoma patients. Therefore, particular emphasis should be placed on ferroptosis-related lncRNAs for individualized diagnosis and treatment in CRC patients. In present study, we screened four ferroptosis-related lncRNAs and constructed a ferroptosis-

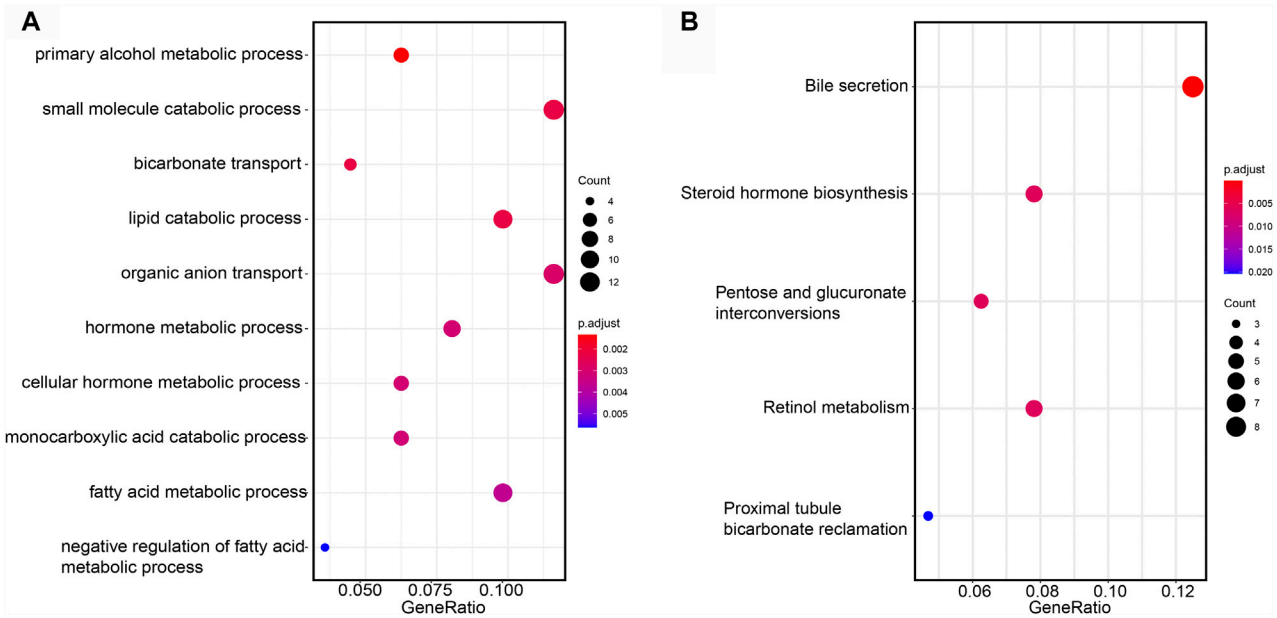



**FIGURE 9 |** Functional annotation of mRNA. **(A)** GO items enriched by GSEA analysis on the network-genes. **(B)** KEGG pathways enriched by GSEA analysis on the network-genes.

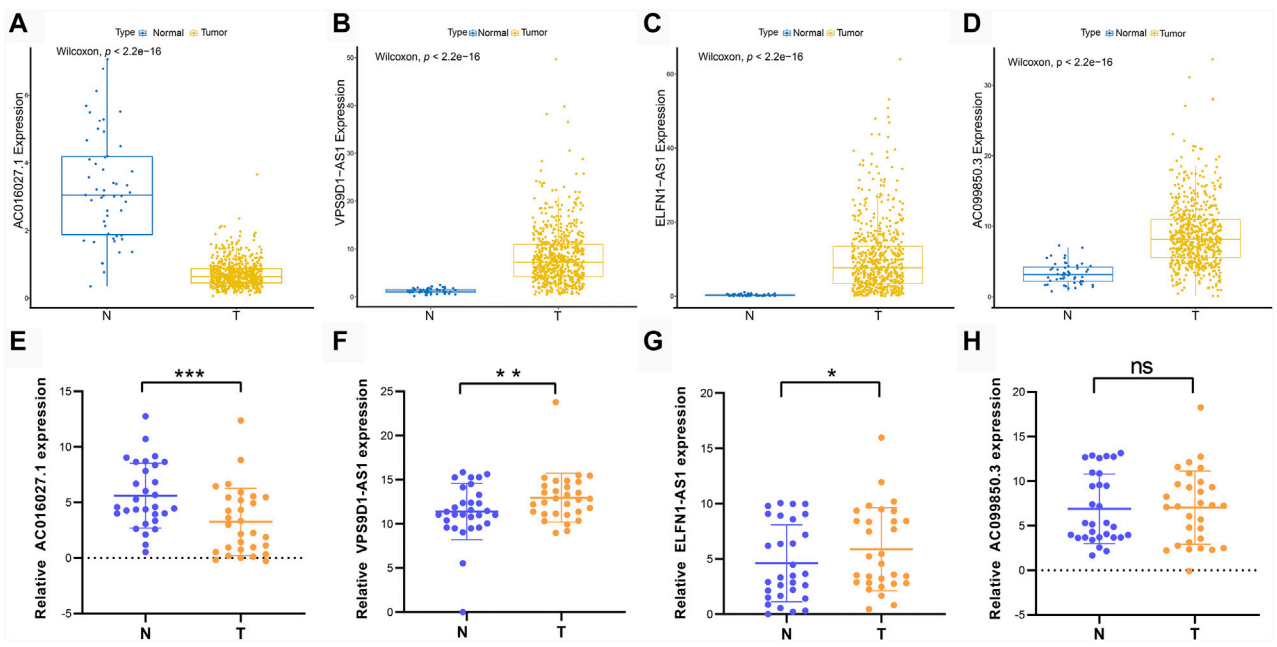



**FIGURE 10 |** Different expressions of prognostic DE-FLs in CRC tissues and TCGA. **(A–D)** The different expression levels of 4 DE-FLs in TCGA. **(E–G)** Q-PCR results demonstrated the down-regulated expression level of AC016027.1 and up-regulated level of VPS9D1-AS1 and ELFN1-AS1 in CRC tissues compared with paired normal tissues. **(H)** Q-PCR result of expression level of AC099850.3 in CRC tissues compared with paired normal tissues. *, $p < 0.05$. **, $p < 0.01$. ***, $p < 0.001$.

related risk model which might be a potential diagnostic biomarker, using bioinformatics and statistical tools.

Firstly, we examined 49 DE-FLs in CRC by analyzing TCGA-COADREAD data set (Smyth, 2004), then 4 DE-FLs (AC016027.1, AC099850.3, ELFN1-AS1, and VPS9D1-AS1) that significantly correlated with OS were found to construct the risk signature according to the univariate and multivariate Cox regression analysis. CRC patients in high-risk groups showed shorter OS compared to those in low-risk groups. Kaplan-Meier survival curve and ROC curve evaluated the predictive accuracy of the ferroptosis-related signature in CRC patients (Robin et al., 2011).

Previous research reported that AC099850.3 and ELFN1-AS1 participated in prognostic autophagy-related lncRNAs signature in HCG patients (Jia et al., 2020). Jiang Q. et al. (2021) reported AC099850.3 was selected by a prognostic autophagy-related lncRNAs signature in oral and oropharyngeal squamous cell carcinoma. The above results appeared to support an existing viewpoint that ferroptosis is a type of autophagy-dependent cell death because a lot of autophagy-related signal pathways contributed to ferroptosis, including BECN1-mediated system xc- inhibition and NCOA4-facilitated ferritinophagy (Song et al., 2018; Quiles Del Rey and Mancias, 2019; Zhou et al., 2020). ELFN1-AS1 was validated with high expression in colon cancer tissues and cells and was reported to promote proliferation and invasion of colon cancer cells by adjusting the miR-191-5p/SATB1 axis (Du et al., 2020). Liu et al. (2020) suggested that VPS9D1-AS1 was a competing endogenous RNA in CRC cells and increased the expression of HMGA1, thereby influenced CRC progression. However, to our knowledge, AC016027.1 and AC099850.3 have not been reported in CRC, which means our findings indicated further research is necessary.

In addition, the ferroptosis-related lncRNA signature is an independent prognostic factor. We constructed a robust nomogram integrating risk score and prognostic clinical features including age, pathological staging and T stage for predicting patient outcomes. ROC curve further demonstrated that this nomogram provided a personalized and accurate survival prediction. Collectively, in our study, there are plenty of evidences suggested that the ferroptosis-related lncRNA signature made accurately predictive prognosis of CRC patients and showed great potential for clinical individualized prognosis and therapy.

Ferroptosis-related GO terms and KEGG signaling pathways were enriched, in order to illustrate the specific mechanism behind the predictive signature. We identified APOPTOSIS pathway reported to be closely associated with ferroptosis and the expression levels of genes involved in the pathway were significantly upregulated in the high-risk group. Previous study revealed an existing cross talk between ferroptosis and apoptosis through ferroptosis-induced endoplasmic reticulum stress (Lee et al., 2018). C/EBP homologous protein (CHOP) signaling pathway-mediated p53 upregulated modulator of apoptosis (PUMA) expression participated in the cooperative interaction between ferroptosis and apoptosis, indicating combination of ferroptotic and apoptotic agent treatment could be considered as a new therapeutic strategy for cancer. Meanwhile, the interaction network between prognostic lncRNAs and DE-FGs was also constructed. The correlation analysis gave us clues about the regulatory relationship between these lncRNAs and mRNA, as well as the molecular mechanism of their role in colorectal cancer. Through the above analysis, we preliminarily inferred that these four ferroptosis-related lncRNAs may directly or indirectly regulate DE-FGs or genes participated in the pathways which were closely related to CRC and thus caused the difference in patient survival.

Nie et al. (2021) comprehensively reported construction of a ferroptosis related genes prognosis model in colon cancer, which aimed to predict survival probability of patients. Our study further screened for differently expressed lncRNAs in CRC which were associated with ferroptosis-related genes, and then constructed a ferroptosis-related lncRNA prognosis model of CRC by Cox regression analyses. There are some advantages in our study that not only did we conduct the mining and exploration of public databases, but also developed the biochemical experiment such as q-PCR to verify our findings in clinical CRC patient samples. However, the shortcoming is that the specific biological molecular mechanisms of these ferroptosis-related prognostic lncRNAs have not been studied in depth. Therefore, future studies are required to explore its exact molecular functions of these ferroptosis-related prognostic genes in CRC.

In conclusion, our study constructed and validated a ferroptosis-related lncRNA prognosis signature in CRC, which consist of AC016027.1, AC099850.3, ELFN1-AS1 and VPS9D1-AS1. The novel ferroptosis-related lncRNA prognosis signature accurately predicts the survival of CRC patients and differentiates them into high- and low-risk groups. Furthermore, the prediction model was independent of clinical features and a reliable nomogram was constructed. Our results might shed lights on promising biomarkers and targets for the individualized therapy of CRC.

## DATA AVAILABILITY STATEMENT

The three transcriptome datasets used in this study are all publicly available. The datasets TCGA-COAD and TCGA-READ for this study can be found in the National Cancer Institute GDC Data Portal [https://portal.gdc.cancer.gov]. The gene set FRGs was downloaded from the FerrDb database [http://www.zhounan.org/ferrdb]. All other data generated in this study are included in the article or the **Supplementary Material**.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee in Clinical Research of the First Affiliated Hospital of Wenzhou Medical University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

Conception and design of the research: LJ, XS; Acquisition of data: WZ; Analysis and interpretation of data: WZ, DF; Statistical analysis: SL, XB; Drafting manuscript: WZ; Obtaining funding: LJ, XS. All authors read and approved the final manuscript.

## FUNDING

This work was supported by the Natural Science Foundation of Zhejiang Province (Grant No LY19H160024), Wenzhou Science and Technological Project (Grant No Y20180081), Key funding for Wenzhou High-level Talent Innovation Technology Project, and the Medicine and Health Technology Program of Zhejiang Province (2021KY790).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.709329/full#supplementary-material

## REFERENCES


Biller, L. H., and Schrag, D. (2021). Diagnosis and Treatment of Metastatic Colorectal Cancer. *JAMA* 325 (7), 669–685. doi:10.1001/jama.2021.0106

Chen, Y. G., Satpathy, A. T., and Chang, H. Y. (2017). Gene Regulation in the immune system by long noncoding RNAs. *Nat. Immunol.* 18 (9), 962–972. doi:10.1038/ni.3771

Dixon, S. J. (2017). Ferroptosis: bug or feature. *Immunol. Rev.* 277 (1), 150–157. doi:10.1111/imr.12533

Dixon, S. J., Lemberg, K. M., Lamprecht, M. R., Skouta, R., Zaitsev, E. M., and Gleason, C. E. (2012). Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* 149 (5), 1060–1072. doi:10.1016/j.cell.2012.03.042

Du, Y., Hou, Y., Shi, Y., Liu, J., and Li, T. (2020). Long Non-Coding RNA ELFN1-AS1 Promoted Colon Cancer Cell Growth and Migration via the miR-191-5p/Special AT-Rich Sequence-Binding Protein 1 Axis. *Front. Oncol.* 10, 588360. doi:10.3389/fonc.2020.588360

Hassannia, B., Vandenabeele, P., and Vanden Berghe, T. (2019). Targeting Ferroptosis to Iron Out Cancer. *Cancer Cell* 35 (6), 830–849. doi:10.1016/j.ccell.2019.04.002

Jia, Y., Chen, Y., and Liu, J. (2020). Prognosis-Predictive Signature and Nomogram Based on Autophagy-Related Long Non-coding RNAs for Hepatocellular Carcinoma. *Front. Genet.* 11, 608668. doi:10.3389/fgene.2020.608668

Jiang, N., Zhang, X., Gu, X., Li, X., and Shang, L. (2021a). Progress in understanding the Role of lncRNA in programmed cell death. *Cell Death Discov* 7 (1), 30. doi:10.1038/s41420-021-00407-1

Jiang, Q., Xue, D., Shi, F., and Qiu, J. (2021b). Prognostic significance of an autophagy-Related long non-coding RNA signature in patients with oral and oropharyngeal squamous cell carcinoma. *Oncol. Lett.* 21 (1), 29. doi:10.3892/ol.2020.12290

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

Kajarabille, N., and Latunde-Dada, G. O. (2019). Programmed Cell-Death by Ferroptosis: Antioxidants as Mitigators. *Int. J. Mol. Sci.* 20 (19). doi:10.3390/ijms20194968

Lee, Y. S., Lee, D. H., Choudry, H. A., Bartlett, D. L., and Lee, Y. J. (2018). Ferroptosis-Induced Endoplasmic Reticulum Stress: Cross-talk between Ferroptosis and Apoptosis. *Mol. Cancer Res.* 16 (7), 1073–1076. doi:10.1158/1541-7786.MCR-18-0055

Lin, C., and Yang, L. (2018). Long Noncoding RNA in Cancer: Wiring Signaling Circuitry. *Trends Cel Biol* 28 (4), 287–301. doi:10.1016/j.tcb.2017.11.008

Liu, H., Zhang, X., Jin, X., Yang, Y., Liang, G., Ma, Y., et al. (2020). Long Noncoding RNA VPS9D1-AS1 Sequesters microRNA-525-5p to Promote the Oncogenicity of Colorectal Cancer Cells by Upregulating HMGA1. *Cancer Manag. Res.* 12, 9915–9928. doi:10.2147/CMAR.S273687

Mao, C., Wang, X., Liu, Y., Wang, M., Yan, B., Jiang, Y., et al. (2018). A G3BP1-Interacting lncRNA Promotes Ferroptosis and Apoptosis in Cancer via Nuclear Sequestration of p53. *Cancer Res.* 78 (13), 3484–3496. doi:10.1158/0008-5472.CAN-17-3454

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi:10.1038/nrg2521

Mou, Y., Wang, J., Wu, J., He, D., Zhang, C., Duan, C., et al. (2019). Ferroptosis, a new form of cell death: opportunities and challenges in cancer. *J. Hematol. Oncol.* 12 (1), 34. doi:10.1186/s13045-019-0720-y

Murphy, M. E. (2016). Ironing out how p53 Regulates ferroptosis. *Proc. Natl. Acad. Sci. U S A.* 113 (44), 12350–12352. doi:10.1073/pnas.1615159113

Ni, W., Yao, S., Zhou, Y., Liu, Y., Huang, P., Zhou, A., et al. (2019). Long noncoding RNA GAS5 inhibits progression of colorectal cancer by interacting with and triggering YAP phosphorylation and degradation and is negatively Regulated by the m(6)A Reader YTHDF3. *Mol. Cancer* 18 (1), 143. doi:10.1186/s12943-019-1079-y

Nie, J., Shan, D., Li, S., Zhang, S., Zi, X., Xing, F., et al. (2021). A Novel Ferroptosis Related Gene Signature for Prognosis Prediction in Patients with Colon Cancer. *Front. Oncol.* 11, 654076. doi:10.3389/fonc.2021.654076

Quiles Del Rey, M., and Mancias, J. D. (2019). NCOA4-Mediated Ferritinophagy: A Potential Link to Neurodegeneration. *Front. Neurosci.* 13, 238. doi:10.3389/fnins.2019.00238

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 77. doi:10.1186/1471-2105-12-77

Shin, D., Kim, E. H., Lee, J., and Roh, J. L. (2018). Nrf2 inhibition Reverses Resistance to GPX4 inhibitor-induced ferroptosis in head and neck cancer. *Free Radic. Biol. Med.* 129, 454–462. doi:10.1016/j.freeradbiomed.2018.10.426

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer J. Clinicians* 68 (1), 7–30. doi:10.3322/caac.21442

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3. doi:10.2202/1544-6115.1027Article3

Song, X., Zhu, S., Chen, P., Hou, W., Wen, Q., Liu, J., et al. (2018). AMPK-mediated BECN1 Phosphorylation Promotes Ferroptosis by Directly Blocking System Xc (-) Activity. *Curr. Biol.* 28 (15), 2388–2399 e2385. doi:10.1016/j.cub.2018.05.094

Stockwell, B. R., Friedmann Angeli, J. P., Bayir, H., Bush, A. I., Conrad, M., Dixon, S. J., et al. (2017). Ferroptosis: A Regulated Cell Death Nexus Linking Metabolism, Redox Biology, and Disease. *Cell* 171 (2), 273–285. doi:10.1016/j.cell.2017.09.021

Wang, M., Mao, C., Ouyang, L., Liu, Y., Lai, W., Liu, N., et al. (2019a). Long noncoding RNA LINC00336 inhibits ferroptosis in lung cancer by functioning as a competing endogenous RNA. *Cell Death Differ* 26 (11), 2329–2343. doi:10.1038/s41418-019-0304-y

Wang, Y., Lu, J. H., Wu, Q. N., Jin, Y., Wang, D. S., Chen, Y. X., et al. (2019b). LncRNA LINRIS stabilizes IGF2BP2 and promotes the aerobic glycolysis in colorectal cancer. *Mol. Cancer* 18 (1), 174. doi:10.1186/s12943-019-1105-0

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. *Cell Rep* 20 (7), 1692–1704. doi:10.1016/j.celrep.2017.07.055

Xu, X., Zhang, X., Wei, C., Zheng, D., Lu, X., Yang, Y., et al. (2020). Targeting SLC7A11 specifically suppresses the progression of colorectal cancer stem cells via inducing ferroptosis. *Eur. J. Pharm. Sci.* 152, 105450. doi:10.1016/j.ejps.2020.105450

Yang, Y., Yan, X., Li, X., Ma, Y., and Goel, A. (2021). Long non-coding RNAs in colorectal cancer: Novel oncogenic mechanisms and promising clinical applications. *Cancer Lett.* 504, 67–80. doi:10.1016/j.canlet.2021.01.009

Yuan, H., Tu, S., Ma, Y., and Sun, Y. (2021). Downregulation of lncRNA RPLP0P2 inhibits cell proliferation, invasion and migration, and promotes apoptosis in colorectal cancer. *Mol. Med. Rep.* 23 (5). doi:10.3892/mmr.2021.11948

Zeuner, A., Todaro, M., Stassi, G., and De Maria, R. (2014). Colorectal cancer stem cells: from the crypt to the clinic. *Cell Stem Cell* 15 (6), 692–705. doi:10.1016/j.stem.2014.11.012

Zhang, Y., Shi, J., Liu, X., Feng, L., Gong, Z., Koppula, P., et al. (2018). BAP1 links metabolic Regulation of ferroptosis to tumour suppression. *Nat. Cel Biol* 20 (10), 1181–1192. doi:10.1038/s41556-018-0178-0

Zhou, B., Liu, J., Kang, R., Klionsky, D. J., Kroemer, G., and Tang, D. (2020). Ferroptosis is a type of autophagy-dependent cell death. *Semin. Cancer Biol.* 66, 89–100. doi:10.1016/j.semcancer.2019.03.002


**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: 03 December 2021
doi: 10.3389/fgene.2021.721021

# Prognostic Implications and Immune Infiltration Analysis of ALDOA in Lung Adenocarcinoma

*Guojun Lu, Wen Shi and Yu Zhang**

*Department of Respiratory Medicine, Nanjing Chest Hospital, Affiliated Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China*

**Background:** aldolase A (*ALDOA*) has been reported to be involved in kinds of cancers. However, the role of *ALDOA* in lung adenocarcinoma has not been fully elucidated. In this study, we explored the prognostic value and correlation with immune infiltration of *ALDOA* in lung adenocarcinoma.

**Methods:** The expression of *ALDOA* was analyzed with the Oncomine database, the Cancer Genome Atlas (TCGA), and the Human Protein Atlas (HPA). Mann-Whitney *U* test was performed to examine the relationship between clinicopathological characteristics and *ALDOA* expression. The receiver operating characteristic (ROC) curve and Kaplan-Meier method were conducted to describe the diagnostic and prognostic importance of *ALDOA*. The Search Tool for the Retrieval of Interacting Genes (STRING) and Cytoscape were used to construct PPI networks and identify hub genes. Functional annotations and immune infiltration were conducted.

**Results:** The mRNA and protein expression of *ALDOA* were higher in lung adenocarcinoma than those in normal tissues. The overexpression of *ALDOA* was significantly correlated with the high T stage, N stage, M stage, and TNM stage. Kaplan-Meier showed that high expression of *ALDOA* was correlated with short overall survival (38.9 vs 72.5 months, $p < 0.001$). Multivariate analysis revealed that *ALDOA* (HR 1.435, 95%CI, 1.013–2.032, $p = 0.042$) was an independent poor prognostic factor for overall survival. Functional enrichment analysis showed that positively co-expressed genes of *ALDOA* were involved in the biological progress of mitochondrial translation, mitochondrial translational elongation, and negative regulation of cell cycle progression. KEGG pathway analysis showed enrichment function in carbon metabolism, the HIF-1 signaling pathway, and glycolysis/gluconeogenesis. The "SCNA" module analysis indicated that the copy number alterations of *ALDOA* were correlated with three immune cell infiltration levels, including B cells, $CD8^+$ T cells, and $CD4^+$ T cells. The "Gene" module analysis indicated that *ALDOA* gene expression was negatively correlated with infiltrating levels of B cells, $CD8^+$ T cells, $CD4^+$ T cells, and macrophages.

**Conclusion:** Our study suggested that upregulated *ALDOA* was significantly correlated with tumor progression, poor survival, and immune infiltrations in lung adenocarcinoma. These results suggest that *ALDOA* is a potential prognostic biomarker and therapeutic target in lung adenocarcinoma.

**Keywords: lung adenocarcinoma, AldoA, biomarker, prognosis, immune infiltration**

OPEN ACCESS

***Edited by:***
*Rosalba Giugno,*
*University of Verona, Italy*

***Reviewed by:***
*Cheng Liang,*
*Shandong Normal University, China*
*Khanh N. Q. Le,*
*Taipei Medical University, Taiwan*

****Correspondence:***
*Yu Zhang*
*zhangyu2113_nj@163.com*

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

***Received:*** *05 June 2021*
***Accepted:*** *28 October 2021*
***Published:*** *03 December 2021*

***Citation:***
*Lu G, Shi W and Zhang Y (2021) Prognostic Implications and Immune Infiltration Analysis of ALDOA in Lung Adenocarcinoma. Front. Genet. 12:721021. doi: 10.3389/fgene.2021.721021*

## INTRODUCTION

According to the latest data from global cancer statistics, lung cancer is the most commonly diagnosed cancer and the leading cause of cancer-related death around the whole world (Bray et al., 2018). Lung adenocarcinoma is the most common pathological type and accounts for more than 40% of all lung cancers (Travis et al., 2015; Denisenko et al., 2018). Despite advances that have been made in early diagnosis and treatment for lung adenocarcinoma in the past years, including targeted therapy and immunotherapy (Zhou and Yao, 2016; Hanna et al., 2017; Xu et al., 2018), the prognosis of lung adenocarcinoma patients remains bleak (Zhang et al., 2019). Therefore, it is imperative to search for novel prognostic markers and therapeutic targets for lung adenocarcinoma.

Aldolase A (*ALDOA*), also called muscle-type aldolase, is mainly expressed in muscle tissues (Tochio et al., 2010). *ALDOA* encodes a glycolytic enzyme that catalyzes the reversible conversion of fructose-1,6-bisphosphate to glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. Ectopic expression of *ALDOA* is important in the development of cardiac hypertrophy, heart failure, and many cardio-cerebrovascular diseases (Hu et al., 2013). Furthermore, *ALDOA* has been reported to be involved in gluconeogenesis and glycolysis (Zeng et al., 2016). Based on both gluconeogenesis and glycolysis can provide energy for tumor proliferation, accumulating evidence has indicated that *ALDOA* plays an important role in the pathological progress of several cancers. A paper from Saito *et al.* indicated that upregulated *ALDOA* in cervical adenocarcinoma can increase the metastasis and invasion of cervical adenocarcinoma cells via promoting epithelial-mesenchymal transition (EMT) (Saito et al., 2020). Concerning non-small cell lung cancer, Fu et al. reported that *ALDOA* can activate the EGFR/MAPK pathway to promote cyclin D1 expression, enhance proliferation and G1/G transition, and facilitate aerobic glycolysis (Fu et al., 2018). These findings indicate that *ALDOA* plays an important role in tumor progression.

In the present study, we conducted bioinformatics analyses on *ALDOA* in lung adenocarcinoma patients, including transcriptional expression and mutation analysis, survival analysis, functional enrichment analysis. We also performed co-expression analysis, constructed the predicted protein-protein interaction (PPI) networks, and identified hub genes of co-expressed genes with *ALDOA*. Moreover, we determined the relationship between *ALDOA* expression and immune cell infiltration in lung adenocarcinoma. Our results link the expression of *ALDOA* with a poor prognosis and provide a potential therapeutic target for lung adenocarcinoma.

## MATERIALS AND METHODS

### Oncomine Database

Oncomine (https://www.oncomine.org/) is an online platform that provides solutions to compute gene expression signatures, clusters, and gene-set modules, automatically extracting biological insights (Rhodes et al., 2007). In this study, we conducted Oncomine to evaluate the mRNA expression of *ALDOA* in lung adenocarcinoma. The results drew from a series of lung adenocarcinoma studies, including Selamat lung, Landi lung, Hou lung, Okayama lung, Stearman lung, Su lung, and Garber lung (Garber et al., 2001; Stearman et al., 2005; Su et al., 2007; Landi et al., 2008; Hou et al., 2010; Okayama et al., 2012; Selamat et al., 2012).

### The Cancer Genome Atlas (TCGA)

TCGA (https://portal.gdc.cancer.gov/) is a genomics data resource that characterized, and analyzed cancer samples (Tomczak et al., 2015). In this study, we analyzed the transcription level of *ALDOA* in multiple cancers from TCGA. The mRNA expression and associated clinical data of *ALDOA* in lung adenocarcinoma were also downloaded from TCGA. The mRNA data of FPKM format has been converted into TPM.

### Tumor Immune Estimation Resource (TIMER)

TIMER (http://timer.cistrome.org/) is an online database and allows users to analyze the differential expression between tumor and normal tissues across all TCGA tumors, and study the correlation between gene expression and immune infiltration level (Li et al., 2020). In the present study, we conducted TIMER to determine the expression of *ALDOA* in diverse cancer types. Moreover, we applied TIMER to explore the correlation between *ALDOA* expression and the abundance of tumor-infiltrating immune cells (B cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils, macrophages, and dendritic cells). In addition, TIMER was used to study the correlation between *ALDOA* expression and gene markers of tumor-infiltrating immune cells.

### UALCAN

The UALCAN (http://ualcan.path.uab.edu/) is a comprehensive online web resource that provides easy access to analyze publicly available cancer omics data (Chandrashekar et al., 2017). In this study, we performed UALCAN to compare the mRNA and protein expression of *ALDOA* from TCGA and Clinical Proteomic Tumor Analysis Consortium (CPTAC, https://proteomics.cancer.gov/programs/cptac) (Edwards et al., 2015).

### The Human Protein Atlas (HPA)

The HPA database (https://proteinatlas.org/) is aimed to map all the human proteins with an integration of various omics technologies (Uhlén et al., 2015; Uhlen et al., 2017). All the data of human proteins includes expression profiles in cells, tumor tissues, and normal tissues. In this study, we performed HPA to confirm the protein expression of *ALDOA* in lung adenocarcinoma.

### Gene Expression Profiling Interactive Analysis (GEPIA2)

GEPIA 2 (http://gepia2.cancer-pku.cn/) is a web-based tool to provide interactive and customizable functions, including gene expression analysis, correlation analysis, survival analysis, similar

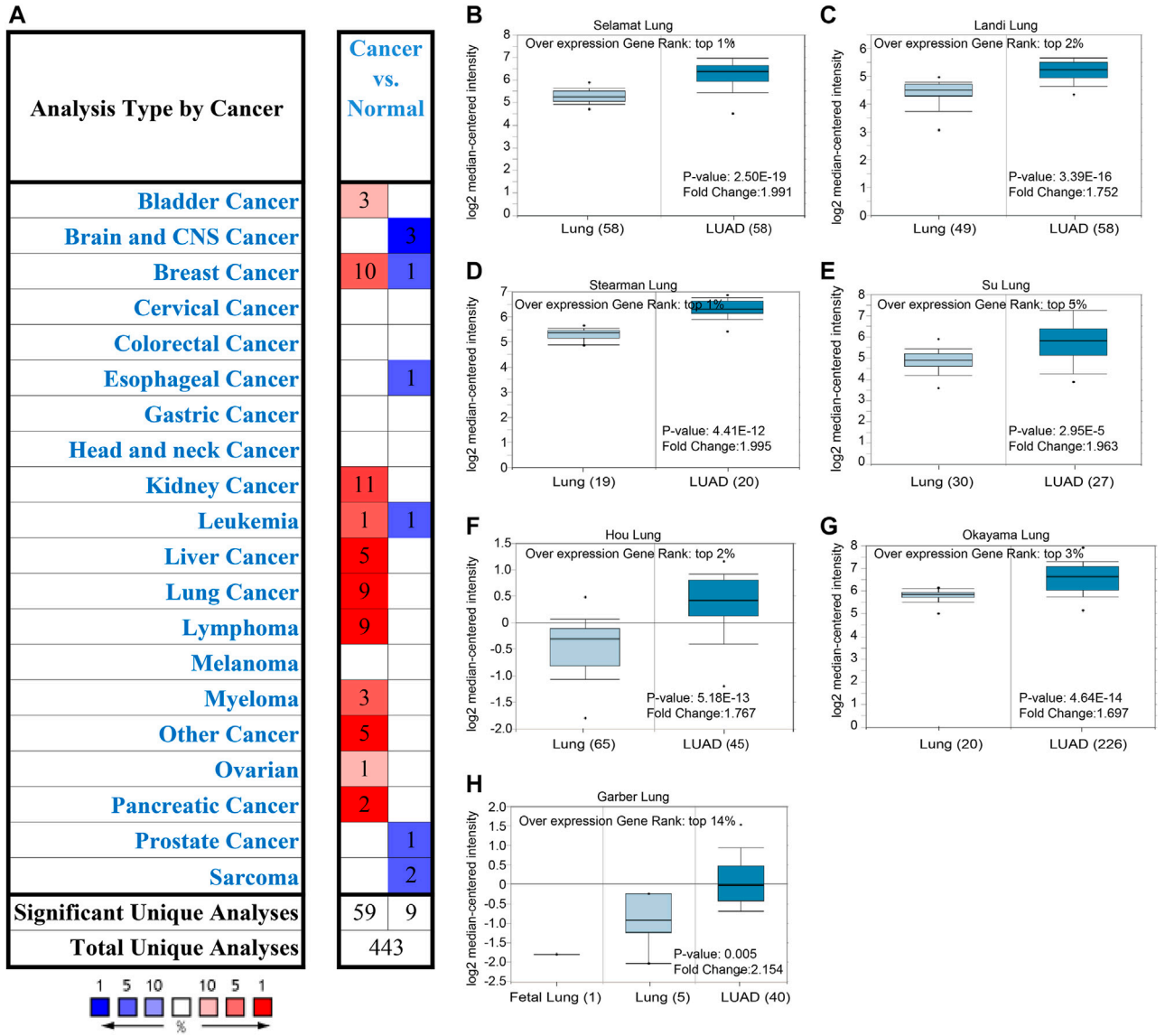



**FIGURE 1 |** Expression of ALDOA from Oncomine **(A)** *ALDOA* expression in different types of cancers. Red means up-regulated and blue means down-regulated **(B-H)** BOX plot showing the mRNA expression levels of *ALDOA* in Selamat lung, Landi lung, Hou lung, Okayama lung, Stearman lung, Su lung, and Garber lung datasets, respectively.

genes detection, and dimensionality reduction analysis (Tang et al., 2019). In this study, we conducted GEPIA2 to examine the correlation between *ALDOA* expression and overall survival. In addition, GEPIA2 was used to assess the correlation between *ALDOA* expression and gene markers of tumor-infiltrating immune cells.

## The Kaplan Meier Plotter

The Kaplan Meier plotter (http://www.kmplot.com/analysis/) is an online tool to assess the effect of 54k genes on survival across cancers including breast, ovarian, lung, and gastric cancer (Nagy et al., 2021). Gene expression data and information of relapse-free and overall survival are downloaded from Gene Expression Omnibus (GEO), European Genome-phenome Archive (EGA), and TCGA. In this study, we performed a Kaplan Meier plotter to validate the prognostic value of *ALDOA* in lung adenocarcinoma.

## c-BioPortal Database

The c-Bio Cancer Genomics Portal (https://www.cbioportal.org/) is an open-access online resource for interactive exploration of many cancer genomics databases (Cerami et al., 2012). The cancer research community can utilize genomic data easily and directly with c-BioPortal. In this study, we performed c-BioPortal databases to explore mutation data of *ALDOA* in lung adenocarcinoma, obtain its prognostic value in altered lung adenocarcinoma patients, acquire co-expressed genes of *ALDOA*,

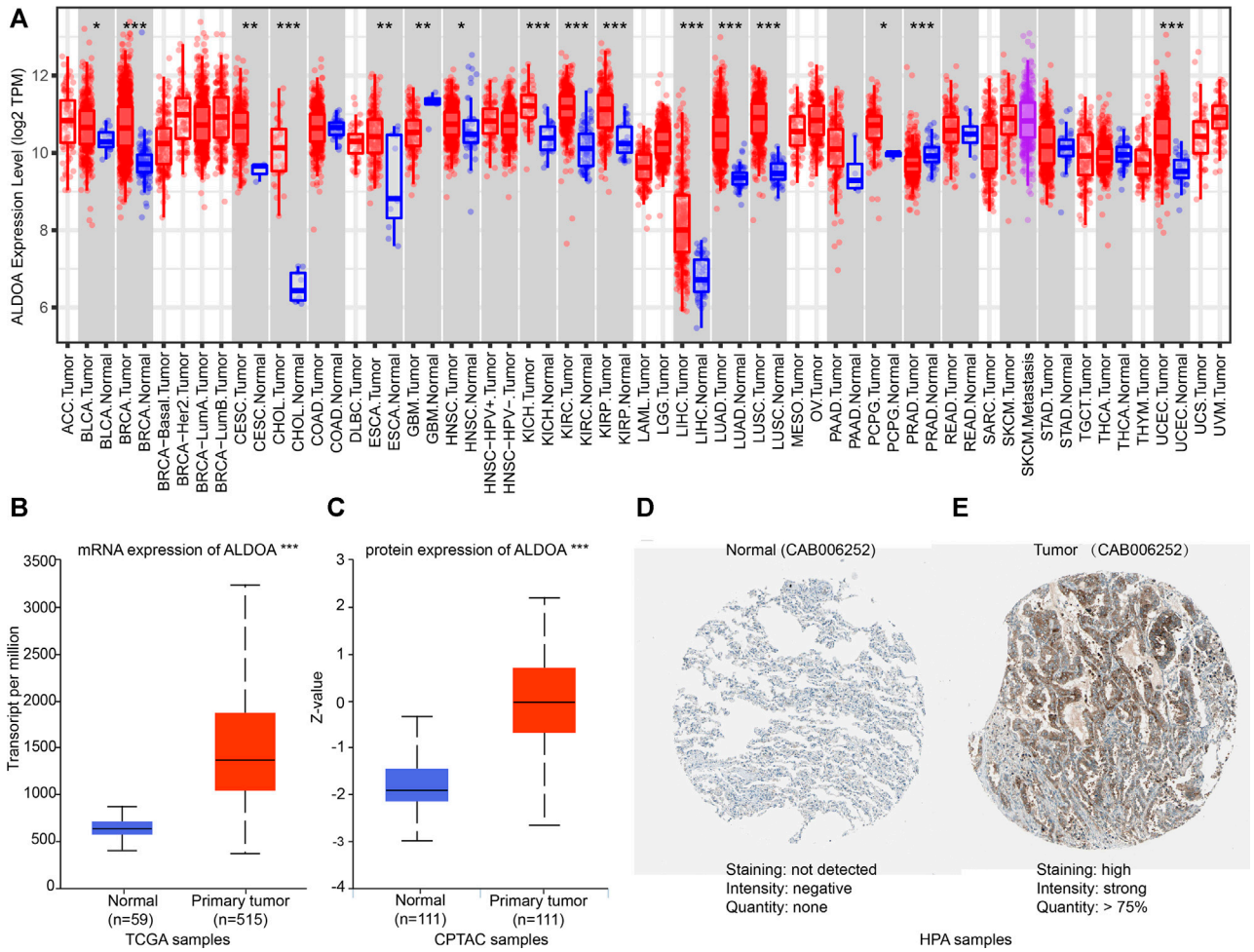



**FIGURE 2 |** Expression of ALDOA from TCGA, CPTAC, and HPA **(A)** *ALDOA* was upregulated in 14 cancer types, including BLCA, BRCA, CESC, CHOL, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PCPG, and UCEC. *ALDOA* was downregulated in GBM and PRAD **(B)** The mRNA expression of *ALDOA* in 59 normal tissues and 515 lung adenocarcinoma tissues **(C)** The protein expression of *ALDOA* in 111 lung normal tissues and 111 lung adenocarcinoma tissues **(D)** The protein expression of *ALDOA* in normal tissue. Staining was not detected, intensity was negative, and quantity was none **(E)** The protein expression of *ALDOA* in lung adenocarcinoma tissue. Staining was high, intensity was strong, and quantity >75%. HPA normal: https://www.proteinatlas.org/ENSG00000149925-ALDOA/tissue/lung#img,HPA tumor: https://www.proteinatlas.org/ENSG00000149925-ALDOA/pathology/lung+cancer#img (*, $p < 0.05$; **, $p < 0.01$, ***, $p < 0.001$).

and determine the correlation between *ALDOA* and mRNA expression of 10 hub genes.

## STRING Database and Cytoscape Platform

The Search Tool for the Retrieval of Interacting Genes (STRING, http://string-db.org, Version 11.0) is an online database to analyze functional enrichment and PPI networks (Szklarczyk et al., 2019). Cytoscape (Version 3.6.1) is an open-source software platform for integrating and visualizing complex networks (Shannon et al., 2003). In this study, to construct PPI networks of co-expressed genes and identify hub genes, we imported the co-expressed genes into STRING and then explored the degree scores with cytoHubba tool kits in Cytoscape.

## Statistical Analyses

Statistical analyses and visualization of expression differences were performed with R (V 3.6.3, https://www.r-project.org/) and R package ggplot2. Mann-Whitney *U* test was conducted to observe the differences between lung adenocarcinoma tissues and adjacent normal tissues. R package pROC (Robin et al., 2011) and clusterProfiler (Yu et al., 2012) were conducted to explore the diagnostic importance and functional enrichment analysis of co-expressed genes in lung adenocarcinoma.

# RESULTS

## Expression of *ALDOA* in Lung Adenocarcinoma From Oncomine

To evaluate the transcription level of *ALDOA* in multiple lung adenocarcinoma studies, we performed an analysis on Oncomine. As shown in **Figures 1A–H**, the transcription level of *ALDOA* was upregulated in lung adenocarcinoma tissues than in normal tissues. The fold change of *ALDOA* differed from 1.697 to 2.154,

**TABLE 1 |** The clinicopathological characteristics of lung adenocarcinoma patients.

| Characteristics | Total | Low expression | High expression | *p*-value |
|---|---|---|---|---|
| | N (%) | N (%) | N (%) | |
| T stage | — | — | — | 0.041* |
| T1 | 175 (32.9) | 101 (19.0) | 74 (13.9) | — |
| T2 | 289 (54.3) | 136 (25.6) | 153 (28.8) | — |
| T3 | 49 (9.2) | 22 (4.1) | 27 (5.1) | — |
| T4 | 19 (3.6) | 6 (1.1) | 13 (2.4) | — |
| N stage | — | — | — | <0.001*** |
| N0 | 348 (67.0) | 194 (37.4) | 154 (29.7) | — |
| N1 | 95 (18.3) | 31 (6.0) | 64 (12.3) | — |
| N2 | 74 (14.3) | 30 (5.8) | 44 (8.5) | — |
| N3 | 2 (0.4) | 0 (0) | 2 (0.4) | — |
| M stage | — | — | — | 0.003** |
| M0 | 361 (93.5) | 175 (45.3) | 186 (48.2) | — |
| M1 | 25 (6.5) | 4 (1.0) | 21 (5.4) | — |
| Pathologic stage | — | — | — | <0.001*** |
| Stage I | 294 (55.8) | 172 (32.6) | 122 (23.1) | — |
| Stage II | 123 (23.3) | 50 (9.5) | 73 (13.9) | — |
| Stage III | 84 (16.0) | 33 (6.3) | 51 (9.7) | — |
| Stage IV | 26 (4.9) | 5 (0.9) | 21 (4.0) | — |
| Gender | — | — | — | 0.178 |
| Female | 286 (53.5) | 151 (28.2) | 135 (25.2) | — |
| Male | 249 (46.5) | 116 (21.7) | 133 (24.9) | — |
| Age | — | — | — | 0.860 |
| ≤65 | 255 (49.4) | 126 (24.4) | 129 (25) | — |
| >65 | 261 (50.6) | 132 (25.6) | 129 (25) | — |
| Smoker | — | — | — | 1.000 |
| No | 75 (14.4) | 37 (7.1) | 38 (7.3) | — |
| Yes | 446 (85.6) | 223 (42.8) | 223 (42.8) | — |
| Anatomic neoplasm subdivision | — | — | — | 0.282 |
| Left | 205 (39.4) | 109 (21) | 96 (18.5) | — |
| Right | 315 (60.6) | 151 (29) | 164 (31.5) | — |
| Anatomic neoplasm subdivision2 | — | — | — | 0.816 |
| Central Lung | 62 (32.8) | 25 (13.2) | 37 (19.6) | — |
| Peripheral Lung | 127 (67.2) | 55 (29.1) | 72 (38.1) | — |

*p < 0.05; **p < 0.01; *** p < 0.001.
*CI, confidence interval.*

and mRNA expression was up to the top 14%. These data indicated transcription level of *ALDOA* is increased in lung adenocarcinoma tissues.

## Expression of *ALDOA* in Pan-Cancer Perspective and Lung Adenocarcinoma From TCGA, UALCAN, and HPA

To further evaluate the expression of *ALDOA* in multiple cancers, we performed an analysis on TCGA with TIMER. As shown in **Figure 2A**, the mRNA expression of *ALDOA* was upregulated in 14 cancer types, including lung adenocarcinoma and lung squamous cell carcinoma. *ALDOA* was downregulated in two cancer types, including GBM and PRAD. The result from UALCAN indicated that both mRNA and protein expression of *ALDOA* in lung adenocarcinoma tissues were significantly higher than that in normal tissues ($p$ = 1.62E-12, $p$ = 7.75E-31, respectively) (**Figures 2B,C**). Consistent with the results of UALCAN, HPA showed protein expression of *ALDOA* in lung adenocarcinoma tissue was higher than that in normal lung tissue (**Figures 2D,E**). All these results indicate that *ALDOA* is upregulated in lung adenocarcinoma tissues.

## The Relationship Between Clinicopathological Characteristics and *ALDOA* Expression in Lung Adenocarcinoma Patients From TCGA

We downloaded the mRNA expression and associated clinical data of *ALDOA* in lung adenocarcinoma from TCGA. The clinicopathological characteristics of lung adenocarcinoma patients were shown in **Table 1**. To examine the relationship between clinicopathological characteristics and *ALDOA* expression in TCGA cohorts, we conducted the Mann-Whitney *U* test. As shown in **Figure 3**, *ALDOA* mRNA expression in lung adenocarcinoma patients was significantly increased with high T stage ($p$ = 0.025), N stage ($p$ = 0.013), M stage ($p$ = 0.002), and TNM stage ($p$ < 0.001). However, no significantly differences were found between *ALDOA* mRNA expression and other characteristics, such as gender ($p$ = 0.329), age ($p$ = 0.594), smoke condition ($p$ = 0.754), and anatomic subdivision (right vs left, $p$ = 0.456; central vs peripheral, $p$ = 0.682). To sum up, these data suggest that *ALDOA* might play an important role in tumorigenesis and metastasis of lung adenocarcinoma.

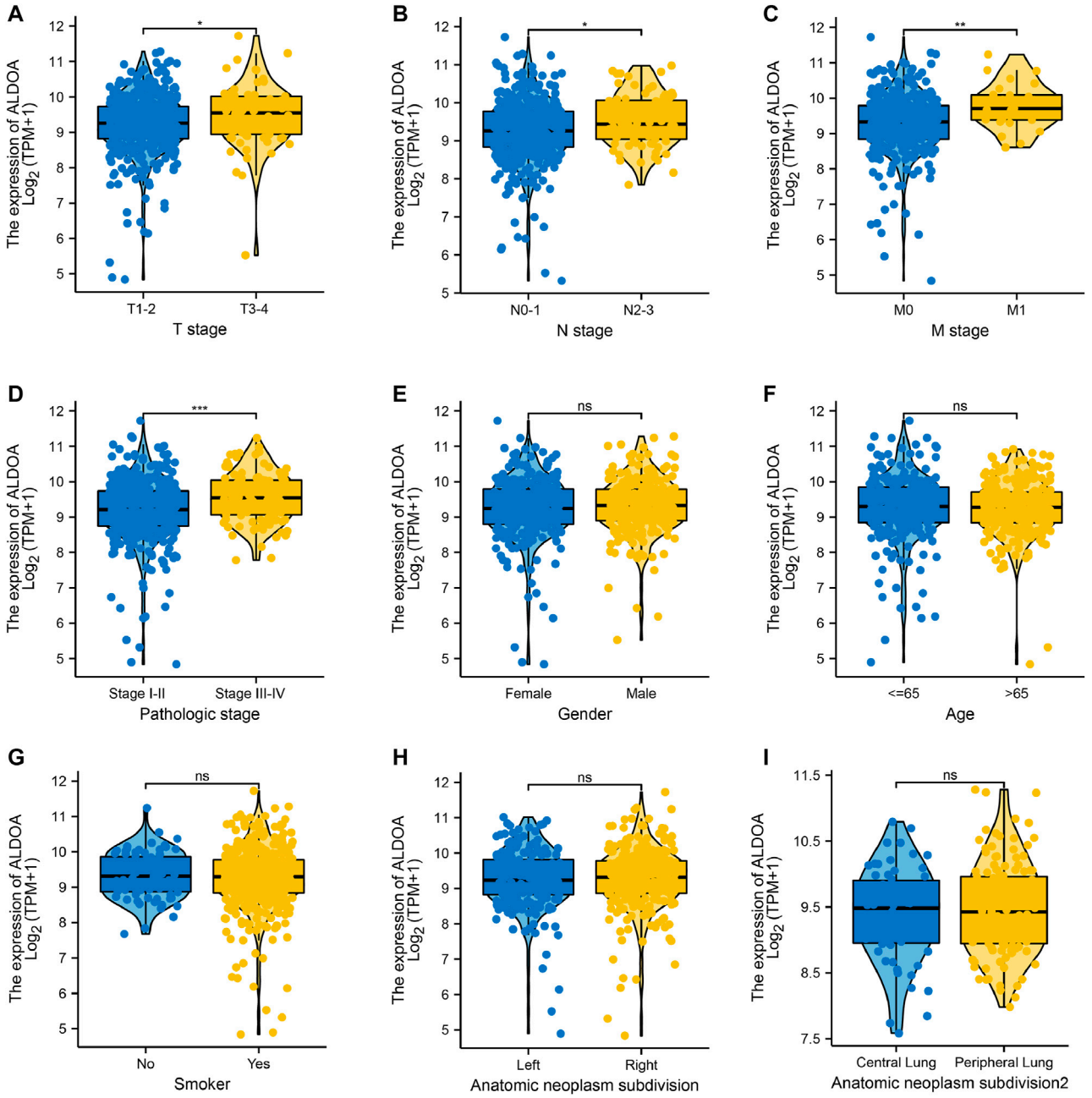



**FIGURE 3 |** The relationship between clinicopathological characteristics and ALDOA expression in lung adenocarcinoma patients from TCGA **(A-D)** *ALDOA* mRNA expression was significantly increased with high T stage, N stage, M stage, and TNM stage **(E-I)** No significant differences were found in gender, age, smoke condition, and anatomic subdivision (right vs left, central vs peripheral) (ns, no significant; *, $p < 0.05$; **, $p < 0.01$, ***, $p < 0.001$).

## Diagnostic Value of *ALDOA* for Distinguishing Lung Adenocarcinoma Tissues From Normal Tissues

To study the diagnostic value of *ALDOA* for distinguishing adenocarcinoma tissues from normal tissues, we performed ROC curve analysis with R package pROC. As shown in **Figure 4**. *ALDOA* had an AUC value of 0.909 (95% CI: 0.883–0.935). With a cutoff of 8.598, *ALDOA* had a sensitivity, specificity, and accuracy of 84.7, 93.2, and 85.5%, respectively. This result indicates that *ALDOA* might be used as a diagnostic biomarker for distinguishing lung adenocarcinoma tissues from normal tissues.

## Correlation Between *ALDOA* Expression and Overall Survival

To explore the correlation between *ALDOA* expression and overall survival and disease-free survival in lung

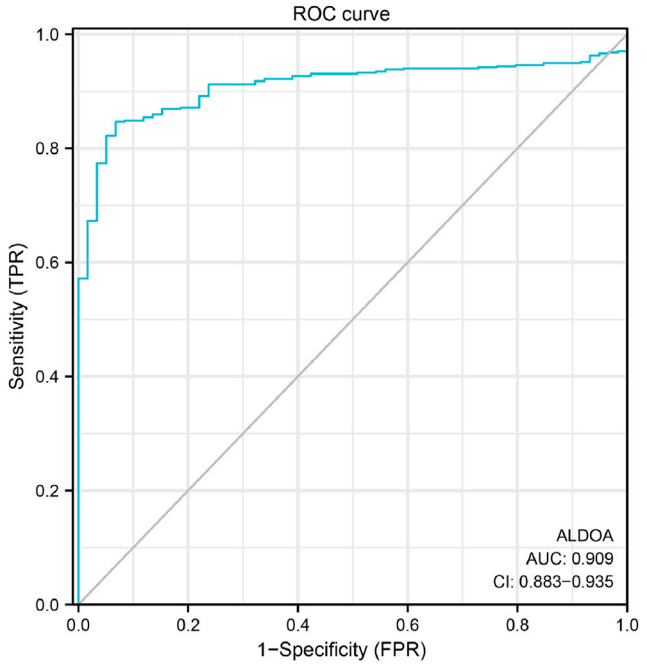



**FIGURE 4 |** ROC curve for distinguishing lung adenocarcinoma tissues from normal tissues. With a cutoff of 8.598, *ALDOA* had a sensitivity, specificity, and accuracy of 84.7, 93.2, and 85.5%, respectively.

adenocarcinoma patients, the GEPIA2 server and Kaplan Meier plotter were performed. As shown in **Figures 5A,B**, the GEPIA2 server indicated that the overall survival rate of lung adenocarcinoma patients with high *ALDOA* expression was significantly lower than that of patients with low *ALDOA* expression ($p$ = 0.00021). However, the correlation of *ALDOA* expression with disease-free survival rate was not statistically significant (**Figure 5B**). Consistently, the Kaplan Meier plotter showed that lung adenocarcinoma patients with high *ALDOA* were correlated with short overall survival (41.0 vs 55.1 months, $p$ = 0.0022) compared to low *ALDOA* mRNA expression (**Figure 5C**). There was no statistically significant between high/low expression of *ALDOA* for recurrence-free survival (68.2 vs 101.5 months, $p$ = 0.38) (**Figure 5D**). These data indicated that high mRNA expression of *ALDOA* is correlated with short overall survival in lung adenocarcinoma.

## Prognostic Importance of *ALDOA* mRNA in Lung Adenocarcinoma Patients

To further determine the prognostic importance of *ALDOA* mRNA, we conducted univariate and multivariate analyses with R package survival. Univariate analysis in **Table 2** showed that the overall survival of lung adenocarcinoma patients was correlated with T stage, N stage, M stage, TNM stage, as well as the mRNA expression of *ALDOA* ((HR 1.799, 95%CI, 1.342–2.413, $p$ = 0.011). Furthermore, we performed a multivariate analysis of five prognostic factors with the Cox proportional hazards model. As shown in **Table 2**, multivariate analysis revealed that T stage (HR 1.652, 95%CI, 1.020–2.673, $p$ = 0.041), and mRNA expression of *ALDOA* (HR 1.435, 95%CI, 1.013–2.032, $p$ = 0.042) were independent poor prognostic factors for overall survival. Moreover, a nomogram was constructed to predict the 1-, 3-, and 5-years survival probability of lung adenocarcinoma patients by combining the mRNA expression of *ALDOA* and clinical characteristics (**Figure 6**). Our data reveal that *ALDOA* is an independent poor prognostic factor for lung adenocarcinoma.

## Genetic Mutation of *ALDOA* and Its Correlation With Poor Survival

To determine the mutation characteristics of *ALDOA* and its correlation with survival in lung adenocarcinoma, we performed an analysis on c-BioPortal databases. As shown in **Figure 7A**, *ALDOA* had a high mutation frequency of 10% in lung adenocarcinoma (TCGA, PanCancer Atlas). The main genetic mutations of *ALDOA* were DNA copy number amplifications and mRNA upregulation (**Figure 7B**). Furthermore, compared with the unaltered group (n = 449), survival analysis revealed that the altered group (n = 56) was associated with poor overall survival (**Figure 7C**).

## Functional Enrichment Analysis of Positively Co-expressed Genes of *ALDOA*

The top 300 co-expressed genes of *ALDOA* were downloaded from the c-BioPortal. As shown in **Supplementary Table S1**, *ALDOA* had 165 positively co-expressed genes. To further demonstrate the enrichment function of these positively co-expressed genes, we conducted the analyses of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway with R package clusterProfiler. GO analysis showed that positively co-expressed genes of *ALDOA* were involved in the biological progress of mitochondrial translation, mitochondrial translational elongation, and negative regulation of cell cycle progress (**Figure 8A**). They acted as structural constituents in the mitochondrial inner membrane, mitochondrial matrix, and protein complex (**Figure 8B**), and played an important part in the structural constituent of ribosome, isomerase activity, and monosaccharide binding (**Figure 8C**). KEGG pathway analysis in **Figure 8D** showed enrichment function in carbon metabolism, HIF-1 signaling pathway, and glycolysis/gluconeogenesis.

## Construction of PPI Networks, Identification, and Enrichment Function of Hub Genes Among Co-expressed Genes of *ALDOA*

To construct PPI networks and identify hub genes of *ALDOA*, we imported the 165 positively co-expressed genes into STRING and then explored the degree scores with cytoHubba tool kits in Cytoscape. The PPI networks of co-expressed genes were shown in **Figure 9A**. As shown in **Figure 9B**, *GADD45GIP1*, *MRPL22*, *MRPL28*, *MRPL21*, *MRPL12*, *MRPS12*, *MRPL52*, *MRPL17*, *TUFM*, and *MRPL53* were the top 10 hub genes of *ALDOA*.

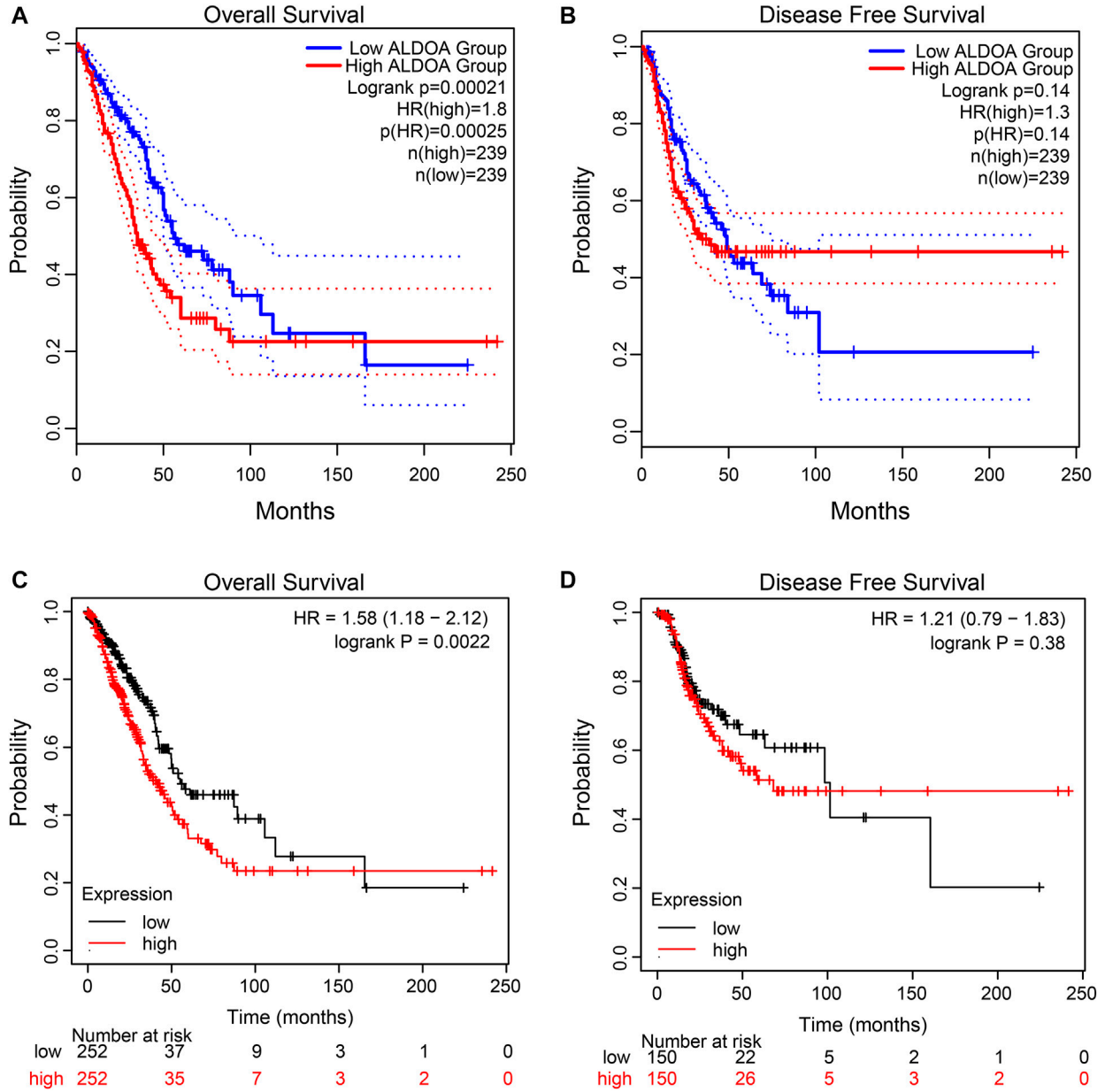



**FIGURE 5 |** Kaplan-Meier curves for ALDOA in lung adenocarcinoma **(A)** Higher ALDOA led to shorter overall survival from GEPIA2 **(B)** ALDOA had no correlation with disease-free survival from GEPIA2 **(C)** High ALDOA was correlated with short overall survival from the Kaplan Meier plotter **(D)** ALDOA had no correlated with recurrence-free survival from the Kaplan Meier plotter.

To further evaluate their prognostic values and correlation with ALDOA in lung adenocarcinoma, we performed analyses on GEPIA and c-BioPortal. As shown in **Figure 9C**, upregulation of *MRPL22* (HR = 1.5, $p$ = 0.0048), *MRPL28* (HR = 1.5, $p$ = 0.0098), *MRPL21* (HR = 1.7, $p$ = 0.001), *MRPL12* (HR = 1.7, $p$ = 0.00059), *MRPS12* (HR = 1.6, $p$ = 0.002), and *MRPL17* (HR = 1.5, $p$ = 0.0051) were correlated with poor overall survival in lung adenocarcinoma. Based on the R-value of Spearman correlation were all more than 0.4, the genes were considered as the most potential hub genes of *ALDOA* (**Figure 9C**). We also conducted enrichment function of these top 10 hub genes with R package clusterProfiler. As shown in **Figure 10**, GO analysis showed they were involved in the biological progress of mitochondrial translational elongation, translational elongation, and mitochondrial translation. They may be associated with a molecular function of the structural constituent of the ribosome and acted as structural constituents in the organellar ribosome, mitochondrial ribosome, and mitochondrial matrix. KEGG pathway analysis showed enrichment function in the ribosome. Taken together, all these data suggest that these hub genes may play an important role in lung adenocarcinoma by cooperating with *ALDOA*.

**TABLE 2 |** Univariate and multivariate analyses of prognostic variables for overall survival.

| Characteristics | Total(N) | Univariate analysis | | Multivariate analysis | |
|---|---|---|---|---|---|
| | | Hazard ratio (95% CI) | *p* Value | Hazard ratio (95% CI) | *p* Value |
| T stage (T3-4 vs T1-2) | 523 | 2.317 (1.591–3.375) | <0.001*** | 1.652 (1.020–2.673) | 0.041* |
| N stage (N2-3 vs N0-1) | 510 | 2.321 (1.631–3.303) | <0.001*** | 1.476 (0.725–3.003) | 0.283 |
| M stage (M1 vs M0) | 377 | 2.136 (1.248–3.653) | 0.006** | 1.158 (0.531–2.526) | 0.712 |
| Pathologic stage (Stage III- IV vs Stage I- II) | 518 | 2.664 (1.960–3.621) | <0.001*** | 1.651 (0.771–3.534) | 0.197 |
| Gender (Male vs Female) | 526 | 1.070 (0.803–1.426) | 0.642 | — | — |
| Age (>65 vs ≤65) | 516 | 1.223 (0.916–1.635) | 0.172 | — | — |
| Smoker (Yes vs No) | 512 | 0.894 (0.592–1.348) | 0.591 | — | — |
| ALDOA (High vs Low) | 526 | 1.799 (1.342–2.413) | <0.001*** | 1.435 (1.013–2.032) | 0.042* |

$p < 0.05$; $p < 0.01$; $p < 0.001$.

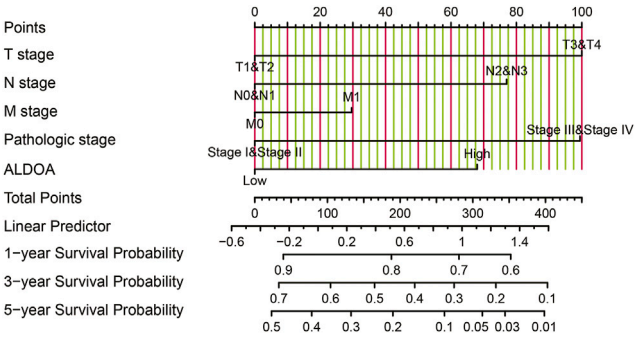



**FIGURE 6 |** The nomogram to predict 1-, 3- and 5-years overall survival probability.

## Relationship Between *ALDOA* Expression and Immune Infiltration in Lung Adenocarcinoma

To determine the potential relationship between *ALDOA* expression and immune infiltration levels in lung adenocarcinoma, we conducted a series of analyses by using TIMER. First, as shown in **Figure 11A**, the "SCNA" module analysis indicated that the copy number alterations of *ALDOA* were correlated with three immune cell infiltration levels, including B cells, $CD8^+$ T cells, and $CD4^+$ T cells in lung adenocarcinoma. Second, as shown in **Figure 11B**, the "Gene" module analysis indicated that there was no correlation between *ALDOA* expression and tumor purity. However, *ALDOA* expression was negatively correlated with infiltrating levels of B cells, $CD8^+$ T cells, $CD4^+$ T cells, and macrophages in lung adenocarcinoma. Third, to evaluate the impact of immune infiltration and *ALDOA* on the survival differences of lung adenocarcinoma patients, we used TIMER to draw Kaplan-Meier plots for immune infiltration and *ALDOA*. The results in **Figure 11C** showed that low levels of B cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells were associated with poor prognosis of lung adenocarcinoma patients. On the contrary, high levels of *ALDOA* were correlated with the poor prognosis of lung adenocarcinoma patients. Taken together, these results suggest that *ALDOA* may regulate the expression level of tumor-infiltrating immune cells to affect lung adenocarcinoma and clinical prognosis.

## Correlation Between *ALDOA* Expression and Gene Markers of Tumor-Infiltrating Immune Cells

To further evaluate the relationship between *ALDOA* and tumor-infiltrating immune cells, we next explored the correlation between *ALDOA* expression and immunological markers in lung adenocarcinoma using the TIMER database. We determined *ALDOA* expression and immunological markers of various immune cells, including B cell, $CD8^+$ T cell, T cell (general), M1 and M2 macrophage, neutrophils, and dendritic cell. After adjusting the correlation by tumor purity, these results revealed that there was a correlation between *ALDOA* expression and most immune marker sets (**Table 3**). In particular, *ALDOA* was significantly correlated with T cell markers (CD3E, CD2), neutrophils markers (CD66b, CCR7), and dendritic cell markers (HLA-DPB1, BDCA-1). We also assessed the correlation between *ALDOA* and these markers in lung adenocarcinoma using the GEPIA2 database, and the results were similar to those in TIMER (**Supplementary Table S1**).

# DISCUSSION

Many studies about the dysregulation of the *ALDOA* gene have emerged in recent years, including colorectal cancer (Dai et al., 2018), gastric cancer (Jiang et al., 2018), and renal cell carcinoma (Huang et al., 2018). Previous bioinformatics results also indicated that *ALDOA* expression was correlated with prognosis in bladder cancer (Li et al., 2019), hepatocellular cancer (Tang et al., 2021). In lung cancer, overexpression of *ALDOA* is reported to promote lung cancer cell proliferation and metastasis (Chang et al., 2019). Moreover, Zhang *et al.* reported that upregulated transcriptional levels of *ALDOA* were correlated with cell cycle-related genes and could regulate progress in non-small cell lung cancer (Zhang et al., 2017). However, the relationship between the expression level of *ALDOA* and prognostic value

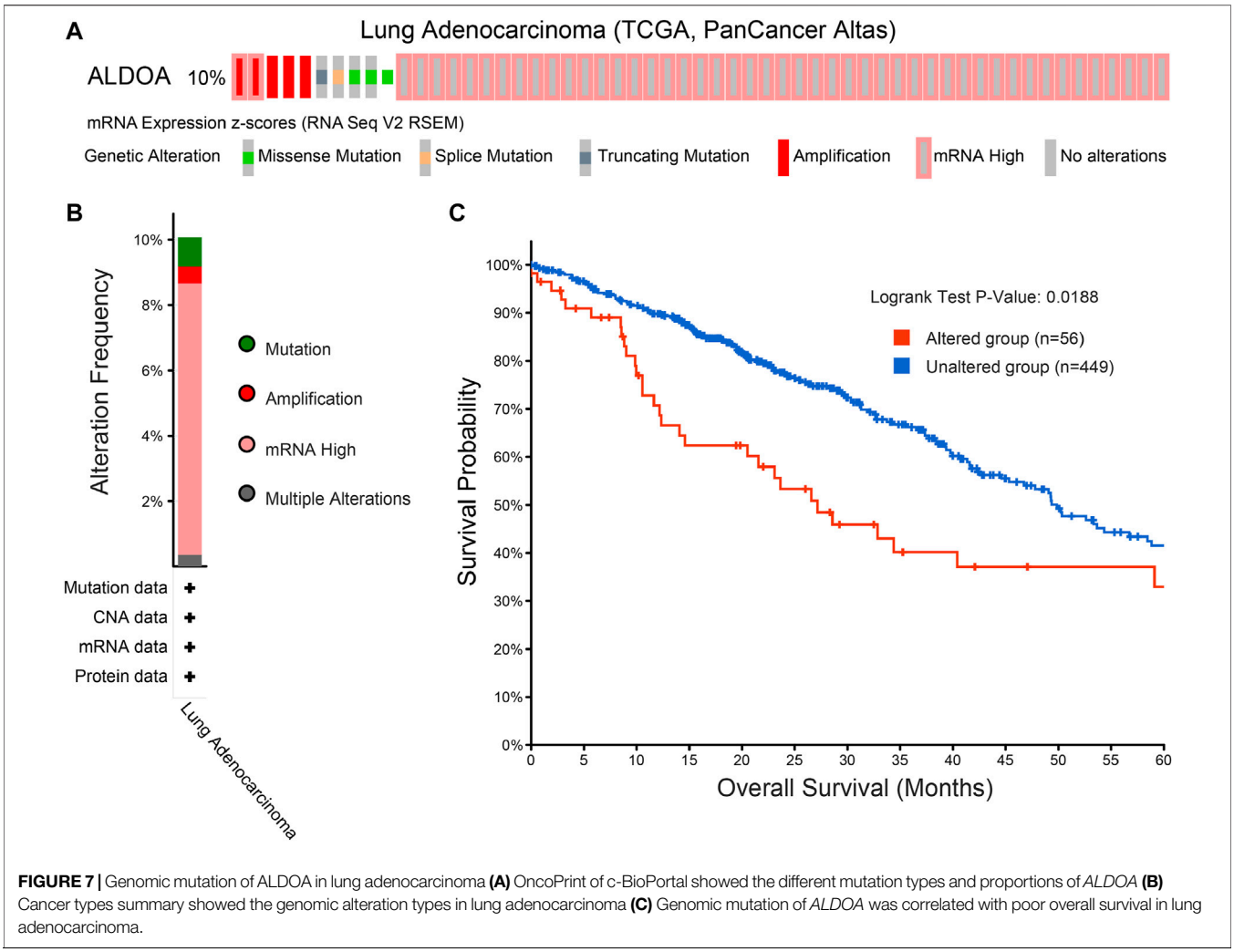



**FIGURE 7 |** Genomic mutation of ALDOA in lung adenocarcinoma **(A)** OncoPrint of c-BioPortal showed the different mutation types and proportions of *ALDOA* **(B)** Cancer types summary showed the genomic alteration types in lung adenocarcinoma **(C)** Genomic mutation of *ALDOA* was correlated with poor overall survival in lung adenocarcinoma.

and immune infiltration of lung adenocarcinoma has not been studied. To the best of our knowledge, for the first time, our study explored the prognostic value and correlation with immune infiltration of *ALDOA* in lung adenocarcinoma.

In this study, based on the data from Oncomine, TCGA, UALCAN, and HPA, we revealed that the mRNA and protein expression of *ALDOA* is upregulated in lung adenocarcinoma tissues. Given that there are significant differences between lung adenocarcinoma and normal tissues grouped by T stage, N stage, M stage, and TNM stage, we conclude that *ALDOA* might promote tumorigenesis and metastasis in lung adenocarcinoma. A paper from Marcišauskas *et al.* reported that *ALDOA* in cyst fluids and serum can be used as a diagnostic biomarker to separate stage I type 1 and type 2 ovarian cancers from benign serous adenoma (Marcišauskas et al., 2019). ROC curve can be used to examine the diagnostic value of biomarkers (Do and Le, 2021; Le et al., 2021). In the current study, ROC curve analysis suggested that *ALDOA* can act as a prospective non-invasive diagnostic biomarker to differentiate lung adenocarcinoma tissues from adjacent normal tissues. Previous studies reported that upregulation of *ALDOA* is correlated with poor prognosis in colorectal cancer (Dai et al., 2018), gastric cancer (Jiang et al., 2018), and hepatocellular carcinoma (Tang et al., 2021). Our data on survival analysis with GEPIA2 and the Kaplan Meier plotter indicated that lung adenocarcinoma patients with high *ALDOA* expression or genetic alteration have a poor overall survival prognosis. Univariate and multivariate analysis revealed that *ALDOA* is an independent poor prognostic factor for overall survival in lung adenocarcinoma. The major strength of this study is our findings raise the possibility that the upregulation of *ALDOA* could be a potential prognostic marker in lung adenocarcinoma.

Functional enrichment analysis was carried out to further explore the role of these positively co-expressed genes with *ALDOA* in lung adenocarcinoma. GO enrichment analysis showed that these positively co-expressed genes of *ALDOA* were involved in the biological progress of mitochondrial translation and negative regulation of cell cycle progression. KEGG pathway enrichment analysis showed enrichment function in carbon metabolism, HIF-1 signaling pathway,

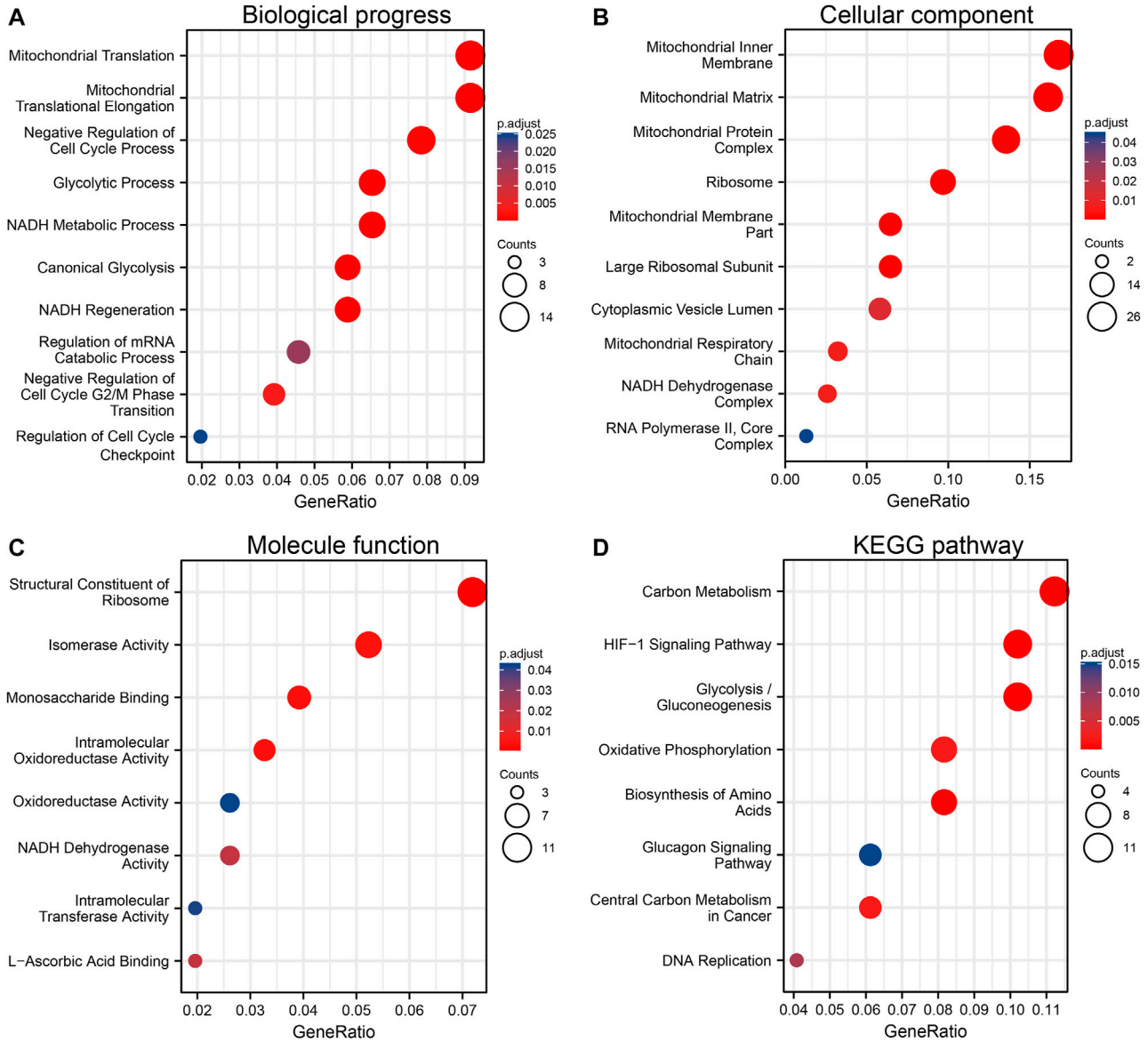



**FIGURE 8 |** Functional enrichment analysis of positively co-expressed genes of *ALDOA*. **(A)** Biological progress of mitochondrial translation, mitochondrial translational elongation, and negative regulation of cell cycle progress **(B)** Cellular component of the mitochondrial inner membrane, mitochondrial matrix, and protein complex **(C)** Molecule function of the structural constituent of ribosome, isomerase activity, and monosaccharide binding **(D)** KEGG pathway analysis showed enrichment function in carbon metabolism, HIF-1 signaling pathway, and glycolysis/gluconeogenesis.

and glycolysis/gluconeogenesis. It is well known that the HIF-1 signaling pathway and glycolysis/gluconeogenesis play an important role in tumor invasion and metastasis (Chen et al., 2011; Rankin and Giaccia, 2016; Peng et al., 2020). Based on our results, we speculate that *ALDOA* may be involved in the progress of invasion and metastasis in lung adenocarcinoma. However, this should be tested in other experiments. Moreover, we also imported the positively co-expressed genes of *ALDOA* into the STRING database and Cytoscape to obtain the PPI network and identify hub genes. In light of STRING database analysis, we conducted a PPI network and interactions among these positively co-expressed genes. The Cytoscape with cytoHubba tool kits, GEPIA, and c-BioPortal analysis suggested that upregulated hub genes of *MRPL22*, *MRPL28*, *MRPL21*, *MRPL12*, *MRPS12*, and *MRPL17* are correlated with poor overall survival and may play a key role by cooperating with *ALDOA* in lung adenocarcinoma.

Many studies about the possible role of immune infiltration have emerged in recent years. It is reported that immune infiltration is correlated with prognosis in human tumors (Pagès et al., 2010; Lei et al., 2020). However, the relationship between *ALDOA* expression and immune infiltration has not been investigated. In the present study, we reported that *ALDOA* copy number alterations were

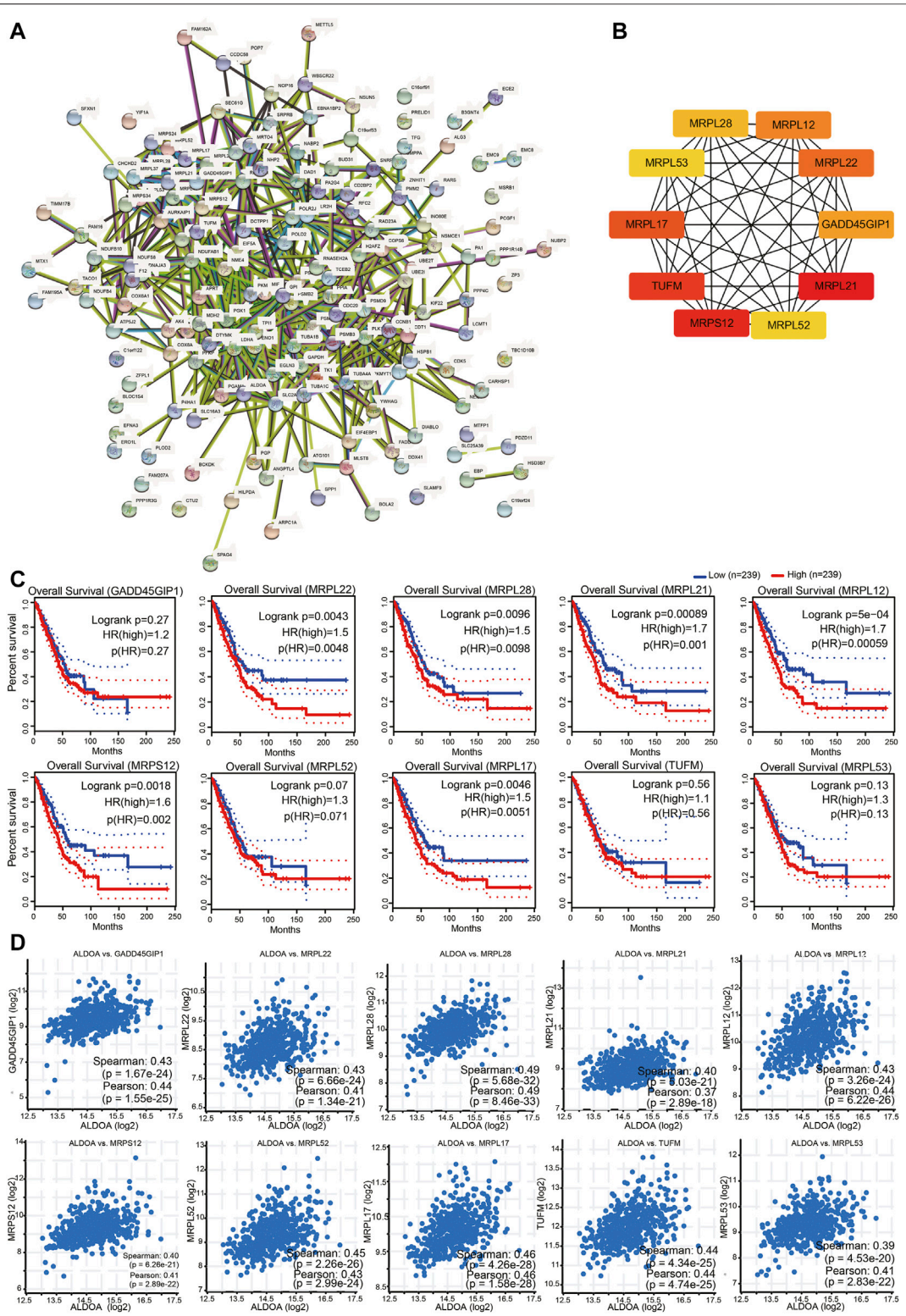



**FIGURE 9 |** Construction of PPI networks and identification of hub genes among co-expressed genes of ALDOA **(A)** Construction of PPI networks by STRING with 165 positively co-expressed genes of *ALDOA* **(B)** Identification of ten hub genes of *ALDOA* with cytoHubba tool in Cytoscape **(C)** Prognostic importance analyses of 10 hub genes with GEPIA **(D)** Correlation between *ALDOA* and mRNA expression of 10 hub genes determined with c-BioPortal.

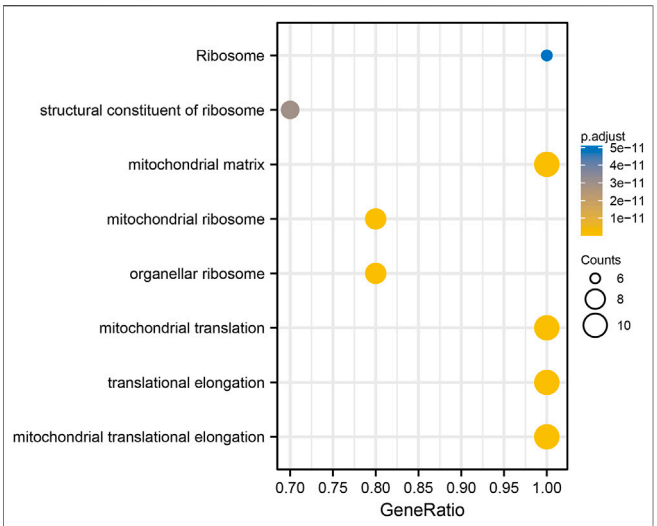



**FIGURE 10 |** Functional enrichment analysis of the top 10 hub genes.

correlated with immune infiltration levels of B cells, $CD8^+$ T cells, and $CD4^+$ T cells in lung adenocarcinoma by TIMER. We also confirmed that *ALDOA* gene expression was inversely correlated with infiltrating levels of B cells, $CD8^+$ T cells, $CD4^+$ T cells, and macrophages in lung adenocarcinoma. Moreover, previous studies showed that there was a correlation between tumor-infiltrating immune cell expression and the prognosis of lung cancer patients (Liu et al., 2017; Wang et al., 2019; Pan et al., 2020). In this study, our results showed low levels of B cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells were associated with poor prognosis of lung adenocarcinoma patients, while high *ALDOA* expression was correlated with poor prognosis in lung adenocarcinoma patients. Based on our data, we conclude for the first time that *ALDOA* is correlated with immune infiltration in lung adenocarcinoma. We further speculate that *ALDOA* can regulate the expression level of tumor-infiltrating immune cells to affect the clinical prognosis of lung adenocarcinoma patients.

In conclusion, our research suggests that the upregulation of *ALDOA* is correlated with tumorigenesis and metastasis in lung adenocarcinoma. Our results show high expression of *ALDOA* predicts poor prognosis and *ALDOA* is an independent poor prognostic factor for overall survival. *ALDOA* may regulate tumor-infiltrating immune cells to affect the clinical prognosis of lung adenocarcinoma patients. Our data provide a potential prognostic biomarker and therapeutic target for lung adenocarcinoma.

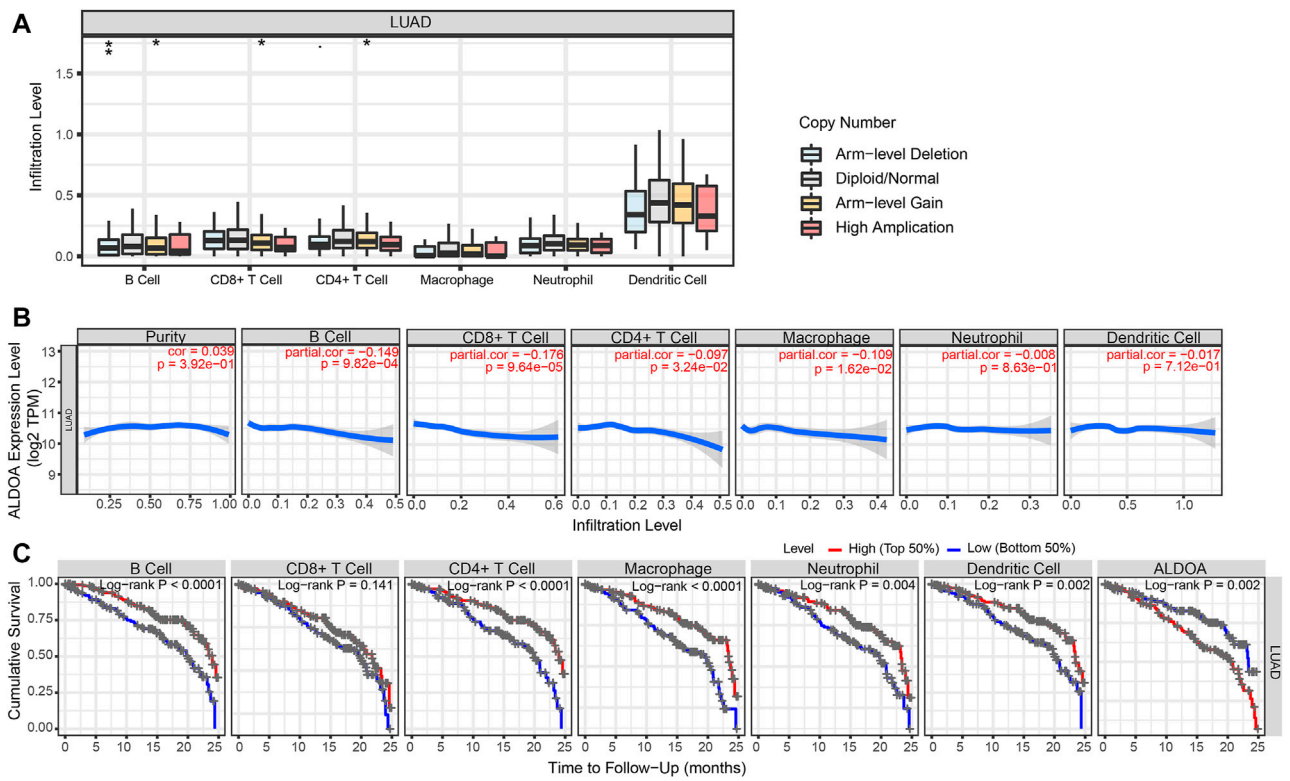



**FIGURE 11 |** Immune cell infiltration of *ALDOA* in lung adenocarcinoma (**A**) The "SCNA" module analysis indicated that the copy number alterations of *ALDOA* were correlated with three immune cell infiltration levels **(B)** The "Gene" module analysis indicated that there was no correlation between *ALDOA* expression and tumor purity. However, *ALDOA* expression was negatively correlated with infiltrating levels of B cells, $CD8^+$ T cells, $CD4^+$ T cells, and macrophages (**C**) Low levels of B cells, $CD4^+$ T cells, macrophages, neutrophils, and dendritic cells were associated with poor prognosis of lung adenocarcinoma patients.

**TABLE 3 |** Correlation analysis between ALDOA and immune infiltration markers in TIMER.

| Description | Gene markers | Lung adenocarcinoma | | | |
|---|---|---|---|---|---|
| | | None | | Purity | |
| | | Cor | *p* | Cor | *p* |
| B cell | CD19 | −0.157 | ** | 0.039 | 3.92e-01 |
| | CD79A | −0.106 | 1.59e-02 | −0.104 | 2.04e-02 |
| $CD8^+$ T cell | CD8A | −0.133 | * | −0.133 | * |
| | CD8B | −0.150 | ** | −0.147 | * |
| T cell (general) | CD3D | −0.130 | * | −0.125 | * |
| | CD3E | −0.169 | ** | −0.176 | *** |
| | CD2 | −0.162 | ** | −0.166 | ** |
| M1 Macrophage | INOS (NOS2) | −0.012 | 7.79e-01 | −0.010 | 8.24e-01 |
| | IRF5 | −0.023 | 6.08e-01 | 0.037 | 4.15e-01 |
| | COX2 (PTGS2) | 0.005 | 9.03e-01 | −0.001 | 9.74e-01 |
| M2 Macrophage | CD163 | −0.030 | 5.03e-01 | −0.004 | 9.22e-01 |
| | VSIG4 | −0.042 | 3.36e-01 | −0.028 | 5.38e-01 |
| | MS4A4A | −0.154 | ** | 0.039 | 3.92e-01 |
| Neutrophils | CD66b (CEACAM8) | −0.180 | *** | −0.183 | ** |
| | CD11b (ITGAM) | 0.023 | 6.06e-01 | 0.123 | 3.00e-01 |
| | CCR7 | −0.199 | *** | −0.203 | *** |
| Dendritic cell | HLA-DPB1 | −0.164 | ** | −0.156 | ** |
| | HLA-DQB1 | −0.101 | 2.15e-02 | −0.085 | 5.84e-02 |
| | HLA-DRA | −0.157 | ** | −-0.144 | * |
| | HLA-DPA1 | −0.122 | * | −0.111 | 1.34e-02 |
| | BDCA-1 (CD1C) | −0.249 | *** | −0.240 | *** |
| | BDCA-4 (NRP1) | 0.033 | 4.59e-01 | 0.033 | 4.59e-01 |
| | CD11c (ITGAX) | −0.002 | 9.69e-01 | 0.015 | 7.33e-01 |

*Cor, correlation of Spearman's R value; None, correlation with no adjustment; Purity, correlation with adjusted by purity.*
$p < 0.01$; $p < 0.001$; $p < 0.0001$.

## 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

YZ designed the study. GL contributed analysis tools and wrote the manuscript. WS performed data analysis. All authors reviewed the manuscript and declared that they have no conflict of interest.

## FUNDING

This study was funded by Wu JiePing Medical Foundation (320.6750.19059), "The 13 TH Five-Year Plan" the Major Program of Nanjing Medical Science and Technique Development Foundation (ZDX16012).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.721021/full#supplementary-material

## REFERENCES

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.cd-12-0095

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* 19 (8), 649–658. doi:10.1016/j.neo.2017.05.002

Chang, Y. C., Chiou, J., Yang, Y. F., Su, C. Y., Lin, Y. F., Yang, C. N., et al. (2019). Therapeutic Targeting of Aldolase A Interactions Inhibits Lung Cancer Metastasis and Prolongs Survival. *Cancer Res.* 79 (18), 4754–4766. doi:10.1158/0008-5472.CAN-18-4080

Chen, C., Fu, X., Zhang, D., Li, Y., Xie, Y., Li, Y., et al. (2011). Varied Pathways of Stage IA Lung Adenocarcinomas Discovered by Integrated Gene Expression Analysis. *Int. J. Biol. Sci.* 7 (5), 551–566. doi:10.7150/ijbs.7.551

Dai, L., Pan, G., Liu, X., Huang, J., Jiang, Z., Zhu, X., et al. (2018). High Expression of ALDOA and DDX5 Are Associated with Poor Prognosis in Human Colorectal Cancer. *Cmar* Vol. 10, 1799–1806. doi:10.2147/cmar.s157925

Denisenko, T. V., Budkevich, I. N., and Zhivotovsky, B. (2018). Cell Death-Based Treatment of Lung Adenocarcinoma. *Cell Death Dis* 9 (2), 117. doi:10.1038/s41419-017-0063-y

Do, D. T., and Le, N. Q. K. (2021). Using Extreme Gradient Boosting to Identify Origin of Replication in *Saccharomyces cerevisiae* via Hybrid Features. *Genomics.* 112, 2445–2451. doi:10.1016/j.ygeno.2020.01.017

Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., et al. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* 14 (6), 2707–2713. doi:10.1021/pr501254j

Fu, H., Gao, H., Qi, X., Zhao, L., Wu, D., Bai, Y., et al. (2018). Aldolase A Promotes Proliferation and G1/S Transition via the EGFR/MAPK Pathway in Non-small Cell Lung Cancer. *Cancer Commun.* 38 (1), 18. doi:10.1186/s40880-018-0290-3

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., et al. (2001). Diversity of Gene Expression in Adenocarcinoma of the Lung. *Proc. Natl. Acad. Sci.* 98 (24), 13784–13789. doi:10.1073/pnas.241500798

Hanna, N., Johnson, D., Temin, S., Baker, S., Brahmer, J., Ellis, P. M., et al. (2017). Systemic Therapy for Stage IV Non-small-cell Lung Cancer: American Society of Clinical Oncology Clinical Practice Guideline Update. *Jco* 35 (30), 3484–3515. doi:10.1200/jco.2017.74.6065

Hou, J., Aerts, J., den Hamer, B., van IJcken, W., den Bakker, M., Riegman, P., et al. (2010). Gene Expression-Based Classification of Non-small Cell Lung Carcinomas and Survival Prediction. *PLoS One* 5 (4), e10312. doi:10.1371/journal.pone.0010312

Hu, L.-j., Chen, Y.-q., Deng, S.-b., Du, J.-l., and She, Q. (2013). Additional Use of an Aldosterone Antagonist in Patients with Mild to Moderate Chronic Heart Failure: a Systematic Review and Meta-Analysis. *Br. J. Clin. Pharmacol.* 75 (5), 1202–1212. doi:10.1111/bcp.12012

Huang, Z., Hua, Y., Tian, Y., Qin, C., Qian, J., Bao, M., et al. (2018). High Expression of Fructose-Bisphosphate Aldolase A Induces Progression of Renal Cell Carcinoma. *Oncol. Rep.* 39 (6), 2996–3006. doi:10.3892/or.2018.6378

Jiang, Z., Wang, X., Li, J., Yang, H., and Lin, X. (2018). Aldolase A as a Prognostic Factor and Mediator of Progression via Inducing Epithelial-Mesenchymal Transition in Gastric Cancer. *J. Cel. Mol. Med.* 22 (9), 4377–4386. doi:10.1111/jcmm.13732

Landi, M. T., Dracheva, T., Rotunno, M., Figueroa, J. D., Liu, H., Dasgupta, A., et al. (2008). Gene Expression Signature of Cigarette Smoking and its Role in Lung Adenocarcinoma Development and Survival. *PLoS One* 3 (2), e1651. doi:10.1371/journal.pone.0001651

Le, N. Q. K., Hung, T. N. K., Do, D. T., Lam, L. H. T., Dang, L. H., and Huynh, T.-T. (2021). Radiomics-based Machine Learning Model for Efficiently Classifying Transcriptome Subtypes in Glioblastoma Patients from MRI. *Comput. Biol. Med.* 32, 104320, 1879-0534 (Electronic). doi:10.1016/j.compbiomed.2021.104320

Lei, X., Lei, Y., Li, J.-K., Du, W.-X., Li, R.-G., Yang, J., et al. (2020). Immune Cells within the Tumor Microenvironment: Biological Functions and Roles in Cancer Immunotherapy. *Cancer Lett.* 470, 126–133. doi:10.1016/j.canlet.2019.11.009

Li, J., Wang, F., Gao, H., Huang, S., Cai, F., and Sun, J. (2019). ALDOLASE A Regulates Invasion of Bladder Cancer Cells via E-cadherin-EGFR Signaling. *J. Cel Biochem* 120 (8), 13694–13705. doi:10.1002/jcb.28642

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48 (W1), W509–w514. doi:10.1093/nar/gkaa407

Liu, X., Wu, S., Yang, Y., Zhao, M., Zhu, G., and Hou, Z. (2017). The Prognostic Landscape of Tumor-Infiltrating Immune Cell and Immunomodulators in Lung Cancer. *Biomed. Pharmacother.* 95, 55–61. doi:10.1016/j.biopha.2017.08.003

Marcišauskas, S., Ulfenborg, B., Kristjansdottir, B., Waldemarson, S., and Sundfeldt, K. (2019). Univariate and Classification Analysis Reveals Potential Diagnostic Biomarkers for Early Stage Ovarian Cancer Type 1 and Type 2. *J. Proteomics* 196, 57–68. doi:10.1016/j.jprot.2019.01.017

Nagy, Á., Munkácsy, G., and Győrffy, B. (2021). Pancancer Survival Analysis of Cancer Hallmark Genes. *Sci. Rep.* 11 (1), 6047. doi:10.1038/s41598-021-84787-5

Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., et al. (2012). Identification of Genes Upregulated in ALK-Positive and EGFR/KRAS/ALK-negative Lung Adenocarcinomas. *Cancer Res.* 72 (1), 100–111. doi:10.1158/0008-5472.can-11-1403

Pagès, F., Galon, J., Dieu-Nosjean, M.-C., Tartour, E., Sautès-Fridman, C., and Fridman, W.-H. (2010). Immune Infiltration in Human Tumors: a Prognostic Factor that Should Not Be Ignored. *Oncogene* 29 (8), 1093–1102. doi:10.1038/onc.2009.416

Pan, Y., Sha, Y., Wang, H., Zhuang, H., Ren, X., Zhu, X., et al. (2020). Comprehensive Analysis of the Association between Tumor-Infiltrating Immune Cells and the Prognosis of Lung Adenocarcinoma. *J. Cancer Res. Ther.* 16 (2), 320–326. doi:10.4103/jcrt.JCRT_954_19

Peng, K., Zhuo, M., Li, M., Chen, Q., Mo, P., and Yu, C. (2020). Histone Demethylase JMJD2D Activates HIF1 Signaling Pathway via Multiple Mechanisms to Promote Colorectal Cancer Glycolysis and Progression. *Oncogene* 39 (47), 7076–7091. doi:10.1038/s41388-020-01483-w

Rankin, E. B., and Giaccia, A. J. (2016). Hypoxic Control of Metastasis. *Science* 352 (6282), 175–180. doi:10.1126/science.aaf4405

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia* 9 (2), 166–180. doi:10.1593/neo.07112

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics* 12, 77. doi:10.1186/1471-2105-12-77

Saito, Y., Takasawa, A., Takasawa, K., Aoyama, T., Akimoto, T., Ota, M., et al. (2020). Aldolase A Promotes Epithelial-mesenchymal Transition to Increase Malignant Potentials of Cervical Adenocarcinoma. *Cancer Sci.* 111 (8), 3071–3081. doi:10.1111/cas.14524

Selamat, S. A., Chung, B. S., Girard, L., Zhang, W., Zhang, Y., Campan, M., et al. (2012). Genome-scale Analysis of DNA Methylation in Lung Adenocarcinoma and Integration with mRNA Expression. *Genome Res.* 22 (7), 1197–1211. doi:10.1101/gr.132662.111

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

Stearman, R. S., Dwyer-Nield, L., Zerbe, L., Blaine, S. A., Chan, Z., Bunn, P. A., et al. (2005). Analysis of Orthologous Gene Expression between Human Pulmonary Adenocarcinoma and a Carcinogen-Induced Murine Model. *Am. J. Pathol.* 167 (6), 1763–1775. doi:10.1016/s0002-9440(10)61257-6

Su, L.-J., Chang, C.-W., Wu, Y.-C., Chen, K.-C., Lin, C.-J., Liang, S.-C., et al. (2007). Selection of DDX5 as a Novel Internal Control for Q-RT-PCR from Microarray Data Using a Block Bootstrap Re-sampling Scheme. *BMC Genomics* 8, 140. doi:10.1186/1471-2164-8-140

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–d613. doi:10.1093/nar/gky1131

Tang, Y., Yang, X., Feng, K., Hu, C., and Li, S. (2021). High Expression of Aldolase A Is Associated with Tumor Progression and Poor Prognosis in Hepatocellular Carcinoma. *J. Gastrointest. Oncol.* 12 (1), 174–183. doi:10.21037/jgo-20-534

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis. *Nucleic Acids Res.* 47 (W1), W556–W560. doi:10.1093/nar/gkz430

Tochio, T., Tanaka, H., Nakata, S., and Hosoya, H. (2010). Fructose-1,6-bisphosphate Aldolase A Is Involved in HaCaT Cell Migration by Inducing Lamellipodia Formation. *J. Dermatol. Sci.* 58 (2), 123–129. doi:10.1016/j.jdermsci.2010.02.012

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an Immeasurable Source of Knowledge. *Contemp. Oncol. (Pozn)* 19 (1a), A68–A77. doi:10.5114/wo.2014.47136

Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., et al. (2015). The 2015 World Health Organization Classification of Lung Tumors. *J. Thorac. Oncol.* 10 (9), 1243–1260. doi:10.1097/jto.0000000000000630

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-Based Map of the Human Proteome. *Science* 347 (6220), 1260419. doi:10.1126/science.1260419

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A Pathology Atlas of the Human Cancer Transcriptome. *Science* 357 (6352). doi:10.1126/science.aan2507

Wang, S.-s., Liu, W., Ly, D., Xu, H., Qu, L., and Zhang, L. (2019). Tumor-infiltrating B Cells: Their Role and Application in Anti-tumor Immunity in Lung Cancer. *Cell Mol Immunol* 16 (1), 6–18. doi:10.1038/s41423-018-0027-x

Xu, X., Huang, Z., Zheng, L., and Fan, Y. (2018). The Efficacy and Safety of Anti-PD-1/pd-L1 Antibodies Combined with Chemotherapy or CTLA4 Antibody as a First-Line Treatment for Advanced Lung Cancer. *Int. J. Cancer* 142 (11), 2344–2354. doi:10.1002/ijc.31252

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

Zeng, Y., Lv, Y., Tao, L., Ma, J., Zhang, H., Xu, H., et al. (2016). G6PC3, ALDOA and CS Induction Accompanies Mir-122 Down-Regulation in the Mechanical Asphyxia and Can Serve as Hypoxia Biomarkers. *Oncotarget* 7 (46), 74526–74536. doi:10.18632/oncotarget.12931

Zhang, F., Lin, J.-D., Zuo, X.-Y., Zhuang, Y.-X., Hong, C.-Q., Zhang, G.-J., et al. (2017). Elevated Transcriptional Levels of Aldolase A (ALDOA) Associates with Cell Cycle-Related Genes in Patients with NSCLC and Several Solid Tumors. *Bio Data Mining* 10. doi:10.1186/s13040-016-0122-4

Zhang, L., Zhang, Z., and Yu, Z. (2019). Identification of a Novel Glycolysis-Related Gene Signature for Predicting Metastasis and Survival in Patients with Lung Adenocarcinoma. *J. Transl Med.* 17 (1), 423. doi:10.1186/s12967-019-02173-2

Zhou, C., and Yao, L. D. (2016). Strategies to Improve Outcomes of Patients with EGFR-Mutant Non-small Cell Lung Cancer: Review of the Literature. *J. Thorac. Oncol.* 11 (2), 174–186. doi:10.1016/j.jtho.2015.10.002

**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: 11 January 2022
doi: 10.3389/fgene.2021.688207

# Development and Validation of a Tumor Mutation Burden-Related Immune Prognostic Signature for Ovarian Cancers

*Mengjing Cui[1†], Qianqian Xia[1†], Xing Zhang[1], Wenjing Yan[1], Dan Meng[1], Shuqian Xie[1], Siyuan Shen[1], Hua Jin[2]* and Shizhi Wang[1]**

[1]Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing, China, [2]Clinical Laboratory, Affiliated Tumor Hospital of Nantong University (Nantong Tumor Hospital), Nantong, China

Ovarian cancer (OC), one of the most common malignancies of the female reproductive system, is characterized by high incidence and poor prognosis. Tumor mutation burden (TMB), as an important biomarker that can represent the degree of tumor mutation, is emerging as a key indicator for predicting the efficacy of tumor immunotherapy. In our study, the gene expression profiles of OC were downloaded from TCGA and GEO databases. Subsequently, we analyzed the prognostic value of TMB in OC and found that a higher TMB score was significantly associated with a better prognosis ($p$ = 0.004). According to the median score of TMB, 9 key TMB related immune prognostic genes were selected by LASSO regression for constructing a TMB associated immune risk score (TMB-IRS) signature, which can effectively predict the prognosis of OC patients (HR = 2.32, 95% CI = 1.68–3.32; AUC = 0.754). Interestingly, TMB-IRS is also closely related to the level of immune cell infiltration and immune checkpoint molecules (PD1, PD-L1, CTLA4, PD-L2) in OC. Furthermore, the nomogram combined with TMB-IRS and a variety of clinicopathological features can more comprehensively evaluate the prognosis of patients. In conclusion, we explored the relationship between TMB and prognosis and validated the TMB-IRS signature based on TMB score in an independent database (HR = 1.60, 95% CI = 1.13–2.27; AUC = 0.639), which may serve as a novel biomarker for predicting OC prognosis as well as possible therapeutic targets.

**Keywords: ovarian cancer, tumor mutation burden, immune risk score, prognostic biomarkers, immune checkpoint**

OPEN ACCESS

***Edited by:***
*Rosalba Giugno,*
*University of Verona, Italy*

***Reviewed by:***
*Jinhui Liu,*
*Nanjing Medical University, China*
*Marco Beccuti,*
*University of Turin, Italy*

****Correspondence:***
*Shizhi Wang*
*shizhiwang2009@seu.edu.cn*
*Hua Jin*
*ntmgjh@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:*** *30 March 2021*
***Accepted:*** *22 December 2021*
***Published:*** *11 January 2022*

***Citation:***
*Cui M, Xia Q, Zhang X, Yan W, Meng D, Xie S, Shen S, Jin H and Wang S (2022) Development and Validation of a Tumor Mutation Burden-Related Immune Prognostic Signature for Ovarian Cancers. Front. Genet. 12:688207. doi: 10.3389/fgene.2021.688207*

## INTRODUCTION

Ovarian cancer (OC) is one of the most common malignancies of the female reproductive system, with worldwide incidence second only to cervical cancer, ranking first in the number of deaths from female reproductive system-related tumors (Webb and Jordan, 2017). The low efficiency of early diagnosis and screening of OC is due to the location of ovaries deep in the pelvic cavity, nonpalpable body surface, and lack of typical symptoms at onset (Stewart et al., 2019). In addition, the tumor grows rapidly, and most patients already have disseminated lesions at the time of diagnosis (Orr and Edwards, 2018). First-line conventional treatments for OC are mainly surgery and chemotherapy (Wang et al., 2016; Wang et al., 2019). Since many OC patients exhibit primary or secondary resistance to chemotherapeutic agents, new

therapeutic approaches need to be discovered to improve the prognosis of OC patients (Valmiki et al., 2021).

The tumor microenvironment (TME) plays an important role in tumor growth and therapy. As a critical part of the TME, immune cell infiltration can orchestrate innate and adaptive immune responses (Hinshaw and Shevde, 2019). With a deeper understanding of the tumor microenvironment, immunotherapy has been approved for the treatment of various types of advanced or recurrent cancers due to its long-term anti-tumor effects (Kruger et al., 2019). OC expresses highly immunogenic tissue-specific antigens, and immune infiltration is the main prognostic factor (Le Saux et al., 2020). Therefore, there is a strong biological basis for the development of immunotherapy for OC (Hao et al., 2018). Currently, checkpoint blockade is the most promising immunotherapy in OC (Ghisoni et al., 2019). However, the objective response rate of immunotherapy alone is not optimal (Wang et al., 2019). The combination of PD(L)-1 antibody and poly (ATP-ribose) polymerases (PARP) inhibitors or conventional chemotherapy has obtained a good response in clinical trials (Wang et al., 2019). Therefore, it is urgent to find molecular markers that can effectively predict the efficacy of OC immunotherapy and screen the appropriate immunotherapy population.

Tumor mutation burden (TMB) is defined as the total number of gene somatic mutations, base substitutions, gene insertion or deletion detected per million bases (Huo et al., 2020). TMB, as an important biomarker that can represent the degree of tumor mutation(Bi et al., 2020), is becoming an emerging biomarker that predicts prognosis and is sensitive to immune checkpoint inhibitors (ICIs) (Merino et al., 2020). Data from retrospective studies indicate that cancers with higher TMB are more likely to respond to ICIs(Snyder et al., 2014; Rizvi et al., 2015). For instance, Killock et al. found that higher TMB was significantly associated with improved survival in melanoma treated with programmed cell death protein 1 (PD-1) immune checkpoint blockade (Killock, 2020). Chalmers et al. reported that TMB could be accurately assessed using comprehensive genomic profiling (CGP) analysis, by which a large proportion of patients with high TMB across tumor types can benefit from immunotherapy (Chalmers et al., 2017).

However, the role of TMB associated immune genes in OC prognosis and the relationship between TMB associated immune genes and OC immune cell infiltration need further investigation. In the present study, somatic mutations and RNA-seq data of OC patients were obtained from TCGA. Subsequently, we analyzed the TMB prognostic value in OC and found that the higher TMB score group had a significantly better prognosis. According to the TMB grouping, 9 key TMB-related immune prognostic genes were selected out and used to construct a TMB-related immune risk score (TMB-IRS) signature that could effectively predict the outcome of ovarian cancer patients Finally, we explored the relationship between TMB and prognosis and validated the TMB-IRS signature based on TMB score in an independent database, which may serve as a novel biomarker and potential therapeutic target for predicting OC prognosis.

# MATERIALS AND METHODS

## Data Collection and Preprocessing

A total of 436 patients with OC were collected from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) database, including somatic mutation, clinical information, survival information and gene-expression data (FPKM normalized). Based on the following inclusion criteria: 1) The patient's pathological diagnosis is OC; 2) Complete mRNA expression profile; 3) Complete clinical information. Exclusion criteria: 1) Non-primary OC; 2) patients with missing mutation information and survival information; 3) patients who Relapsed OC. In all, we selected 271 OC samples as a training set, including corresponding clinical characteristics, such as age, cancer status, grade, stage, and race (**Supplementary Table S1**).

The gene names of all immune genes were downloaded directly from the website. From the Immunology database and Analysis Portal (ImmPort) database (https://immport.niaid.nih.gov) we downloaded the complete list of immune-related genes, including a total of 2483 immune-related genes (**Supplementary Table S2**).

## Calculation of TMB Scores and Prognostic Analysis

To evaluate the prognostic differences between different TMBs in OC patients, we performed the following analysis. In our study, the TMB score of each individual was calculated by, the number of mutations divided by exon length (30 MB). Then, OC samples were divided into high and low-TMB groups according to the median number. And further, Kaplan-Meier analysis was implemented for the comparison of differences in overall survival (OS) between the two groups. Visualization of the somatic mutation landscape of OC patients was done by using the "maptools" package in R. The version number of the R software used in this study is v 3.6.1.

## Differential Analysis

Based on TMB grouping, we first performed differential analysis to identify genes differentially expressed in the high- and low-TMB groups. Specifically, differentially expressed genes (DEGs) were obtained using the "limma" package in R. Among them, $\log2 \mid FC \mid > 0.58$ (FC, fold change) and $p < 0.05$ are criteria. Visualization of DEGs was implemented by plotting volcano plots *via* the "ggplot2", "Cairo" and "ggrepel" software packages in R.

## Construction and Validation of TMB-Related Immune Risk Score (TMB-IRS) Signature

TMB related immune prognostic genes in OC were screened out by stepwise analysis, as a way to construct a TMB-IRS signature that could effectively predict the prognosis of OC. Differential expression analysis was first performed to obtain

TMB-associated genes. The above gene and immune gene sets were intersected so that differentially expressed immune genes were obtained. Further, genes with an expression level of 0 in more than 50% of the samples were removed from differentially expressed immune genes. Subsequently, Cox regression and LASSO regression were performed to obtain independent immune genes related to prognosis using the "glmnet" R package. Based on the corresponding regression coefficient $\beta$ value, the risk score value of each sample was calculated by, TMB-IRS = $\Sigma$ Cox coefficient of gene Xi × scale expression value of gene Xi.

Each sample was ranked according to the risk score and grouped by the median, and patients were therefore divided into low- and high-risk groups. The prognostic value of the signature was assessed by performing Kaplan-Meier (KM) analysis with a log-rank test, using the "survminer" R package. Using the "survival" ROC R software package, we plotted the receiver operating characteristic (ROC) curve over time to evaluate the accuracy of the signature.

Search the GEO database for OC cohorts with gene expression and prognostic information, and finally select the GSE26712 cohort as a reasonable validation set, $n$ = 148. For comparability of data from different sources, gene expression from geo data were further log transformed.

## Relationship Between Clinicopathological Factors and TMB-IRS Signature

To evaluate whether the TMB-IRS could serve as an independent predictor of prognosis, we first employed univariate Cox regression analysis to look for clinical features associated with prognosis and then performed multivariate Cox regression analysis to look for independent factors. Besides, in order to comprehensively evaluate the prognosis of OC patients, we plan to establish a comprehensive assessment model that combines clinical information with the TMB-IRS signature. In brief, using the "rms" package in R, we constructed a nomogram that could predict 2-, 3-, and 5-years patient survival. To compare the consistency of the actual OS of OC with the predicted effect, calibration curves (2-, 3-, and 5-years survival prediction) were plotted, and the curve at 45 represented the nomogram with better prediction accuracy.

Further, we used the R survival package to calculate the concordance index (C-index) of TNM stage, TMB-IRS and nomogram for comparing the predictive ability of the three for the prognosis of OC patients. Meanwhile, decision curve analysis (DCA) at 2, 3 and 5 years were calculated to measure the clinical utility of our established nomogram. The $x$-axis represents the percentage of threshold probability, and the $y$-axis represents net income.

## Cibersort Database Analysis

In order to estimate the infiltration of immune cells, we used CIBERSORT online immune cell infiltration estimation analysis tool (http://cibersort.stanford.edu/). It is a tool to deconvolute immune cell subtype expression matrices based on linear support vector regression principles. In the present study, the tool was suitably employed to compare the proportions of 22 immune cells in the high- and low-TMB-IRS groups. The 22 types of immune cells included: 7 types of T cells ($CD8^+$ T cells, naive $CD4^+$ T cells, resting memory $CD4^+$ T cells, activated memory $CD4^+$ T cells, follicle-assisted T cells, regulatory T cells, and γδT cells), 3 types of B cells (naive B cells, memory B cells, and plasma cells) NK cells (resting NK cells and activated NK cells), and various myeloid cells (monocytes, M0 macrophages, M1 macrophages, M2 macrophages, resting dendritic cells, activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils). $p$ less than 0.05 was set as the criterion for statistical significance.

## Statistical Analysis

The SPSS 20.0 was adopted for multivariate Cox regression analysis, with a probability of a stepwise entry of 0.05 and removal of 0.1. And the simple mathematical operation processes and all table making were completed by the software Excel. Univariate and multivariate Cox regression was carried out to analyze the relationship among gene expression, clinical features and prognosis. Additionally, the "survival ROC" package was used to plot the survival ROC in R (v 3.6.1). All analyses associated with prognosis were performed with the "survival" package. The probability threshold with a significant difference was set as $p < 0.05$.

# RESULTS

## Landscape of the OC Mutation Profiles

In total, we analyzed the somatic mutation profiles of 271 patients. As shown in **Figure 1A**, there were 263 samples with somatic mutation data, accounting for 97.05%. *TP53*, *TTN*, and *CSMD3* mutations are the top three mutated genes in OC samples, and *TP53* mutations are found in more than 92% of OC samples. Moreover, missense mutations were the most common mutation classification, single nucleotide polymorphisms (SNPs) showed a higher fraction in the variant type than insertion or deletion, and C > T was the most common single nucleotide variant (SNV) in OC (**Figure 1B**). Furthermore, the number of variants in each sample was calculated, and the mutation types were also shown in **Figure 1B** with different colors for OC. The co-occurrence and exclusive associations between mutated genes are shown in **Figure 1C**.

After calculating the TMB value of each sample (**Supplementary Table S3**), all patients were divided into high- and low-TMB groups according to the median and interquartile range [M(IRQ) = 1.947 (1.316, 2.684)]. Interestingly, patients in the low-TMB group have an obviously shorter OS than those in the high-TMB group with $p$ = 0.004 (**Figure 1D**).

## Establishment and Evaluation of TMB-IRS Signature

To establish a TMB-IRS signature in the TCGA-OV cohort, multivariate Cox and LASSO analyses were employed to screen

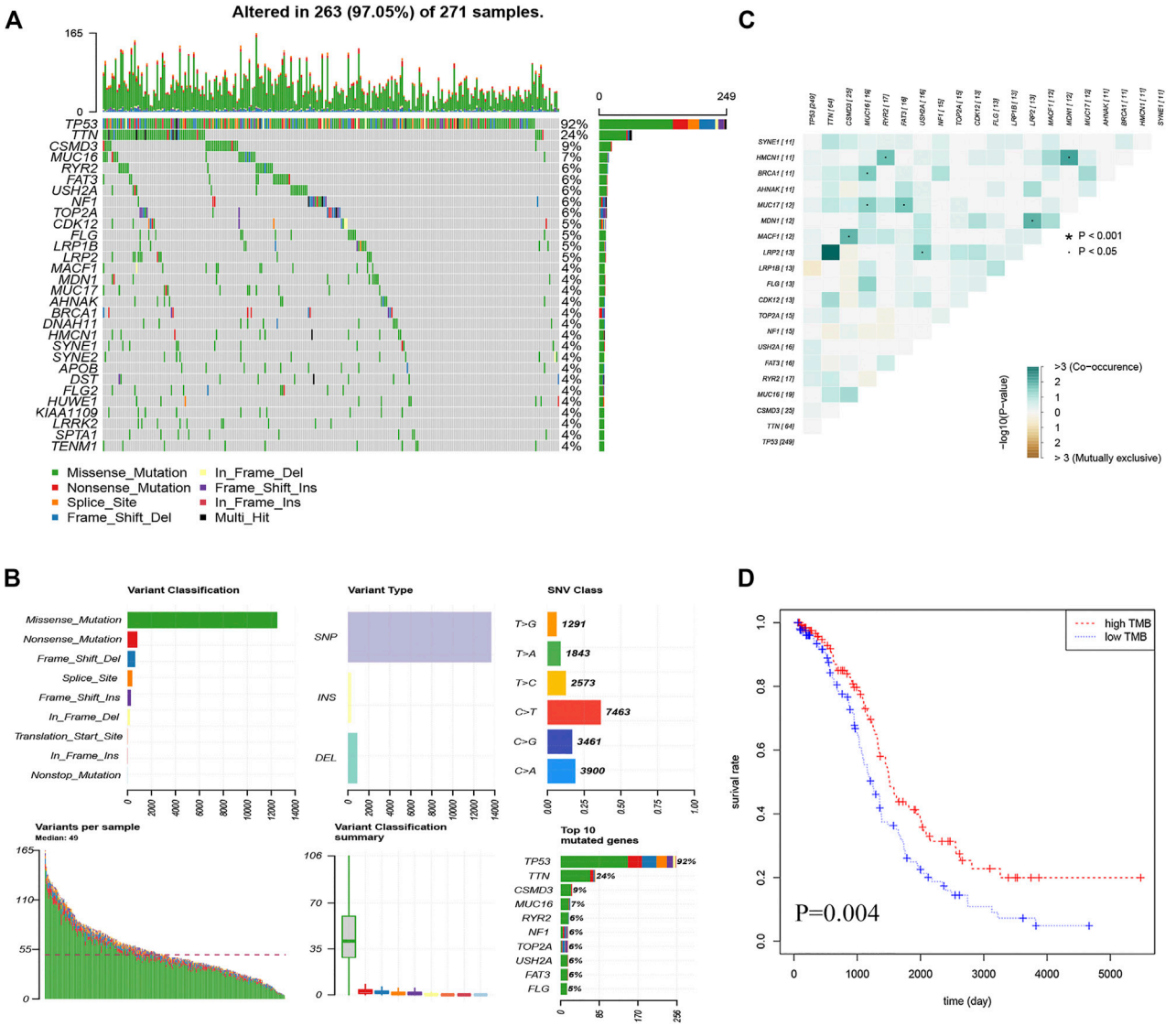



**FIGURE 1 |** Landscape of the OC mutation profiles. **(A)** Waterfall plots present the landscape of the top 30 genes somatically mutated in 271 OC patients. **(B)** Various bar graphs show somatic mutations data. Variant classification reveal that missense mutations are the most; Variant type show the highest number of SNP (single nucleotide polymorphism); SNV(single nucleotide variants) class show the highest number of C > T; Variant per sample, with a median of 49; Summary of variant type; Top 10 mutated genes **(C)** Correlation Heatmap of the top 20 mutated genes **(D)** Kaplan-Meier curves show that OS in the high-TMB was significantly higher than in the low TMB group.

out independent immune genes related to prognosis. Specifically, a total of 892 differentially expressed genes were first differentially analyzed between the high- and low-TMB groups (**Figure 2A**, **Supplementary Table S3**). The DEGs above intersected with 1793 immune genes to obtain 99 differentially expressed immune genes (**Figure 2B**). Further, univariate Cox regression analysis obtained 12 immune genes related to disease prognosis (**Supplementary Table S5**). After eliminating two genes with 0 expressions in more than 50% samples, LASSO regression analysis was performed, resulting in 9 independent prognostic immune genes (**Figures 2C,D**), namely *CSPG5*, *CXCL10*, *CXCL11*, *DKK1*, *PI3*, *TNFRSF17*, *DUOX1*, *TNFRSF13B* and *PAEP*. Finally, based on the regression coefficients and gene expression of the above 9 genes, and TMB-IRS was calculated for each patient with the following formula:

$$\begin{aligned}\text{TMB} - \text{IRS} = {} & 0.417 * \exp DKK1 + 0.091 * \exp PI3 \\ & + 0.166 * \exp DUOX1 + 0.013 * \exp PAEP \\ & + 0.184 * \exp CXCL10 - 0.254 * \exp CSPG5 \\ & - 0.392 * \exp CXCL11 - 0.219 * \exp TNFRSF17 \\ & - 0.428 * \exp TNFRSF13B\end{aligned}$$

Then, the risk score of each individual was calculated and ranked among OC patients, and then divided into high- and

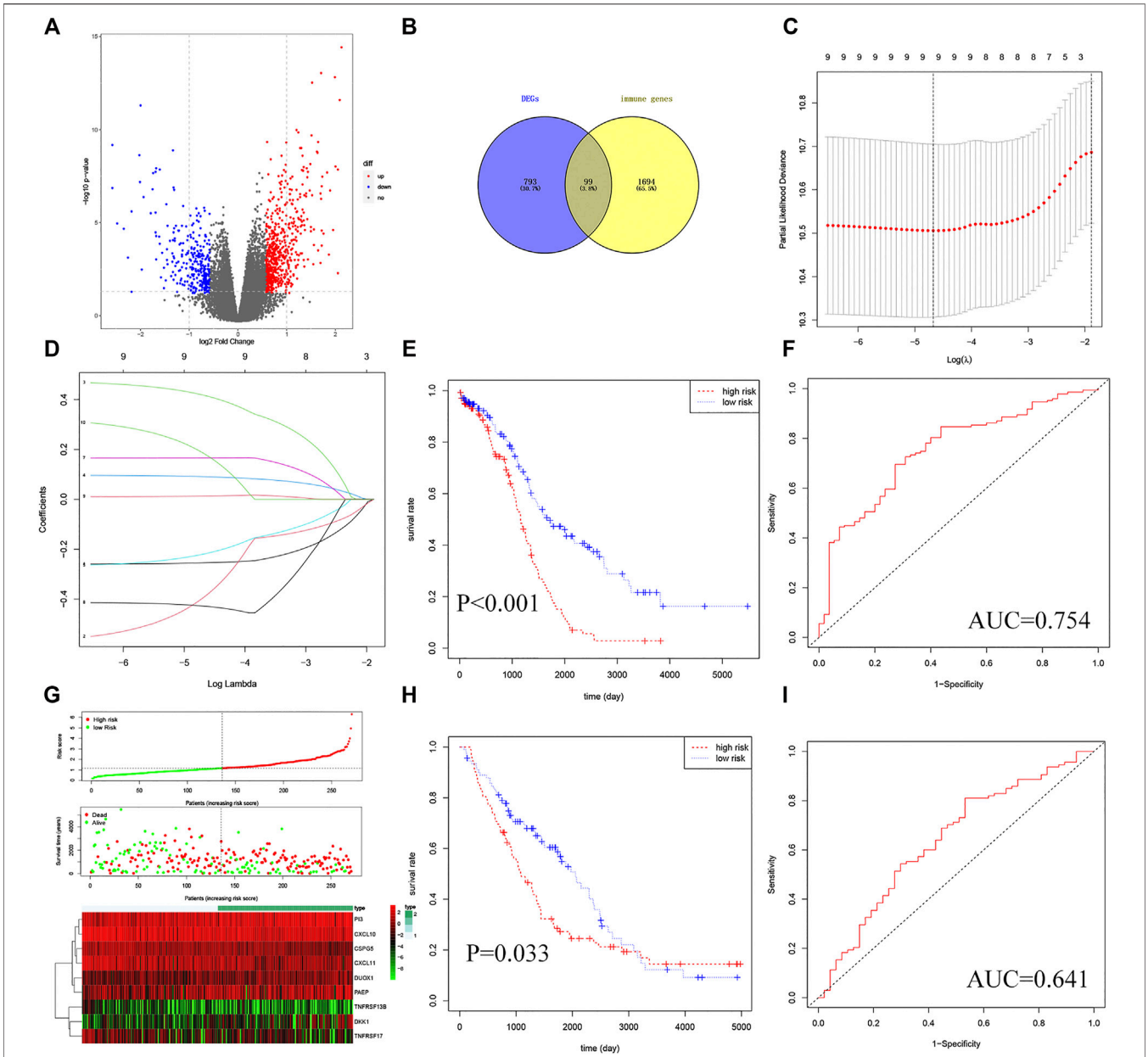



**FIGURE 2 |** Establishment and validation of TMB-IRS signature. **(A)** Identification of 892 DEGs based on TMB score. **(B)** DEGs intersected with immune genes in a Venn diagram with 99 genes available. **(C)** Ten-time cross-validation for tuning parameter selection in the LASSO model. **(D)** LASSO coefficient profiles. **(E)** Kaplan-Meier curves show that OS was significantly different between the high- and low-risk groups in TCGA-OC. **(F)** The TMB-IRS signature is shown by the time-dependent ROC curve for predicting 2, 3, 5-years survival. **(G)** Risk score, survival status, and heatmap of 9 immune genes in patients with OC. **(H)** Kaplan-Meier curves show that OS in the low-risk was significantly higher than in the high-risk group in GSE26712. **(I)** Time-dependent ROC curve analysis of the TMB-IRS signature at 2, 3, 5 years in GSE26712.

low-risk groups according to the median [M(IRQ) = 1.173 (0.798, 1.718)]. KM analysis indicated that patients in the high-risk group ($n$ = 135) tended to have a worse prognosis compared to those in the low-risk group ($n$ = 136) (**Figure 2E**, HR = 2.32, 95% CI = 1.68–3.32; $p$ < 0.001). In addition, the survival ROC curve results showed that the TMB-IRS signature had relative accuracy in predicting the prognosis of OC (**Figure 2F**, 5-years AUC = 0.754). The risk curve and heatmap (**Figure 2G**) showed the patient risk score for each individual as well as the expression levels of the 9 genes.

## Validation of the TMB-IRS Signature

In order to verify the universal applicability of the TMB-IRS signature, the OC cohort downloaded in the GEO database was

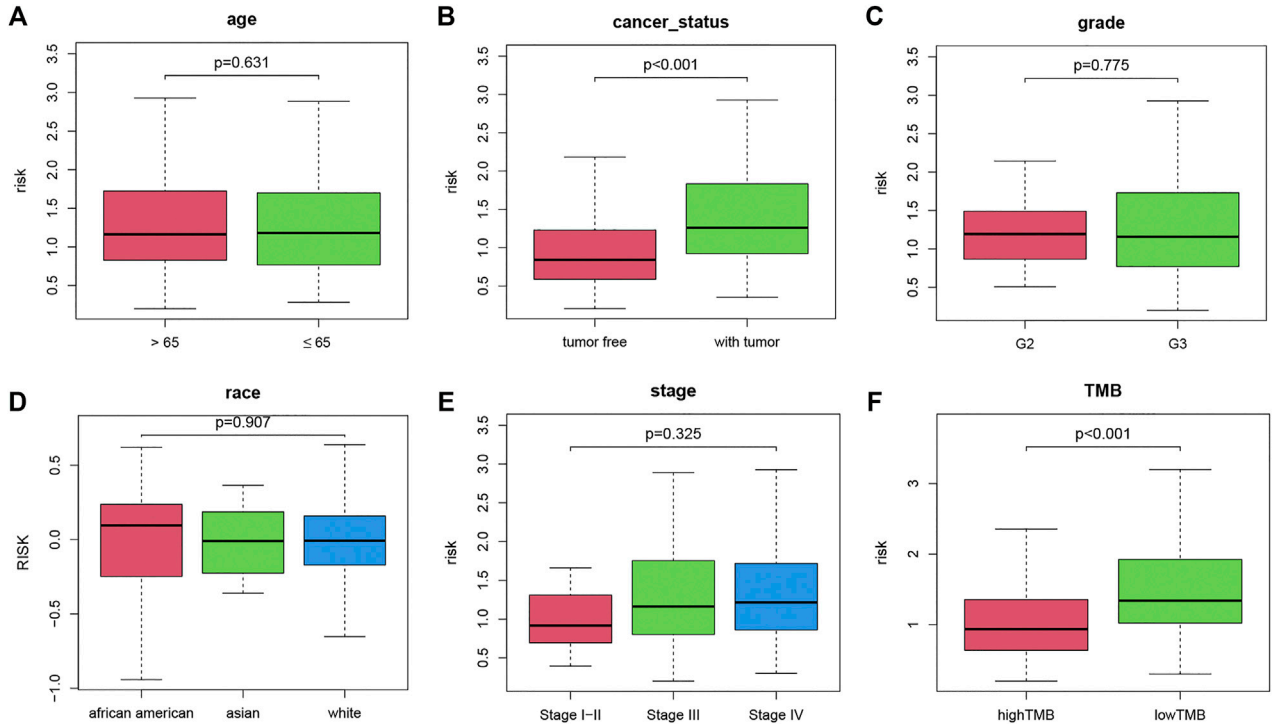



**FIGURE 3 |** Boxplot of the relationship between clinicopathological factors and TMB-IRS. **(A–F)** Relationship between age, tumor status, grade, race, stage, TMB and TMB-IRS, respectively.

used as a validation set, and patients with missing mutation information and survival time less than 30 days were excluded. A total of 148 patients were analyzed for prognosis. According to the TMB-IRS formula established by the OC cohort in the TCGA database, the risk score of each patient in the validation set was calculated. According to the median TMB-IRS calculated by the TCGA database cohort, the validation set was divided into low-risk group and high-risk group. The results of KM analysis showed that TMB-IRS was significantly related to the prognosis (**Figure 2H**, HR = 1.46, 95% CI = 1.03–2.08; $p$ = 0.033). The low-risk group had a better prognosis, while the high-risk group had a worse prognosis, which was consistent with the results of the TCGA database cohort. The ROC curve shows that the model has a good agreement between the predicted probability of OS and the actual probability (**Figure 2I**; 5-years AUC = 0.641).

## Correlations Between TMB-IRS and Clinical Variables

To investigate the correlation between clinical variables and the TMB-IRS, boxplots were drawn to visualize the immune risk profile across clinical subgroups. As shown in **Figure 3**, the immune risk score was significantly positively correlated with cancer status but negatively correlated with TMB. The risk score was significantly higher in the with-tumor group compared with the tumor-free group. In contrast, among the TMB subgroups, low-TMB tended to have a lower risk score. However, risk scores did not differ significantly between subgroups in other clinical characteristics (age, grade, stage, and race).

To demonstrate the prognostic predictive independence of the TMB-IRS signature in multiple clinical features, we employed univariate and multivariate Cox proportional hazards regression for analysis. As shown in **Table 1**, univariate Cox analysis results showed that age, cancer status, TMB and OS were significantly associated with OC patients. Furthermore, multivariate regression analysis demonstrated that the TMB-IRS signature could serve as an independent predictor for evaluating the prognosis of OC patients.

## DEVELOPMENT AND EVALUATION OF THE NOMOGRAM

To systematically predict the prognosis of OC, we constructed a nomogram model based on the risk score and clinical information in the TCGA dataset (**Figure 4A**). The calibration curve results showed that the prediction of prognostic survival probability of OC patients by the nomogram had good agreement with the actual probability (**Figure 4B**). Meanwhile, the C-index (95% confidence interval) of the nomogram, TNM stage, and TMB-IRS was 0.739 (0.717, 0.716), 0.643 (0.618, 0.668), and 0.537 (0.517, 0.557), respectively, and this result also demonstrated that the nomogram had better predictive accuracy. Consistent with this result, DCA plots (**Figures 4C,D**) also proved that TMB-IRS performed better than traditional TNM-stage for prediction,

**TABLE 1 |** Univariate/multivariate Cox regression analysis of OC clinicopathological characteristics associated with OS.

| Variables | | Patient N (271) | Univariate analysis | | Multivariate analysis | |
|---|---|---|---|---|---|---|
| | | | HR[a] (95% CI[b]) | *p* | HR (95% CI) | *p* |
| Age | <65 | 95 | 1 (reference) | — | 1 (reference) | — |
| | ≥65 | 176 | 0.715(0.521,0.982) | 0.038[c] | 0.656(0.465,0.926) | 0.016[c] |
| Stage | Stage I | 18 | 1 (reference) | — | — | — |
| Stage | Stage I-II | 18 | 1 (reference) | 0.504 | — | — |
| | Stage III | 204 | 1.429(0.628,3.251) | 0.395 | — | — |
| | Stage IV | 46 | 1.647(0.687,3.950) | 0.263 | — | — |
| Grade | G2 | 32 | 1 (reference) | — | — | — |
| | G3 | 229 | 1.001(0.636,1.574) | 0.997 | — | — |
| Cancer_status | Tumor free | 71 | 1 (reference) | — | 1 (reference) | — |
| | With tumor | 165 | 8.343(4.228,16.464) | <0.001[c] | 6.609(3.318,13.165) | <0.001[c] |
| TMB | low TMB | 135 | 1 (reference) | — | 1 (reference) | — |
| | high TMB | 136 | 0.654(0.479,0.892) | 0.007[c] | 0.816(0.578,1.151) | 0.258 |
| TMB-IRS | — | 271 | 1.944(1.638,2.307) | <0.001[c] | 1.758(1.425,2.168) | <0.001[c] |

[a]*HR, hazard ratio.*
[b]*CI, confidence interval.*
[c]p < *0.05.*

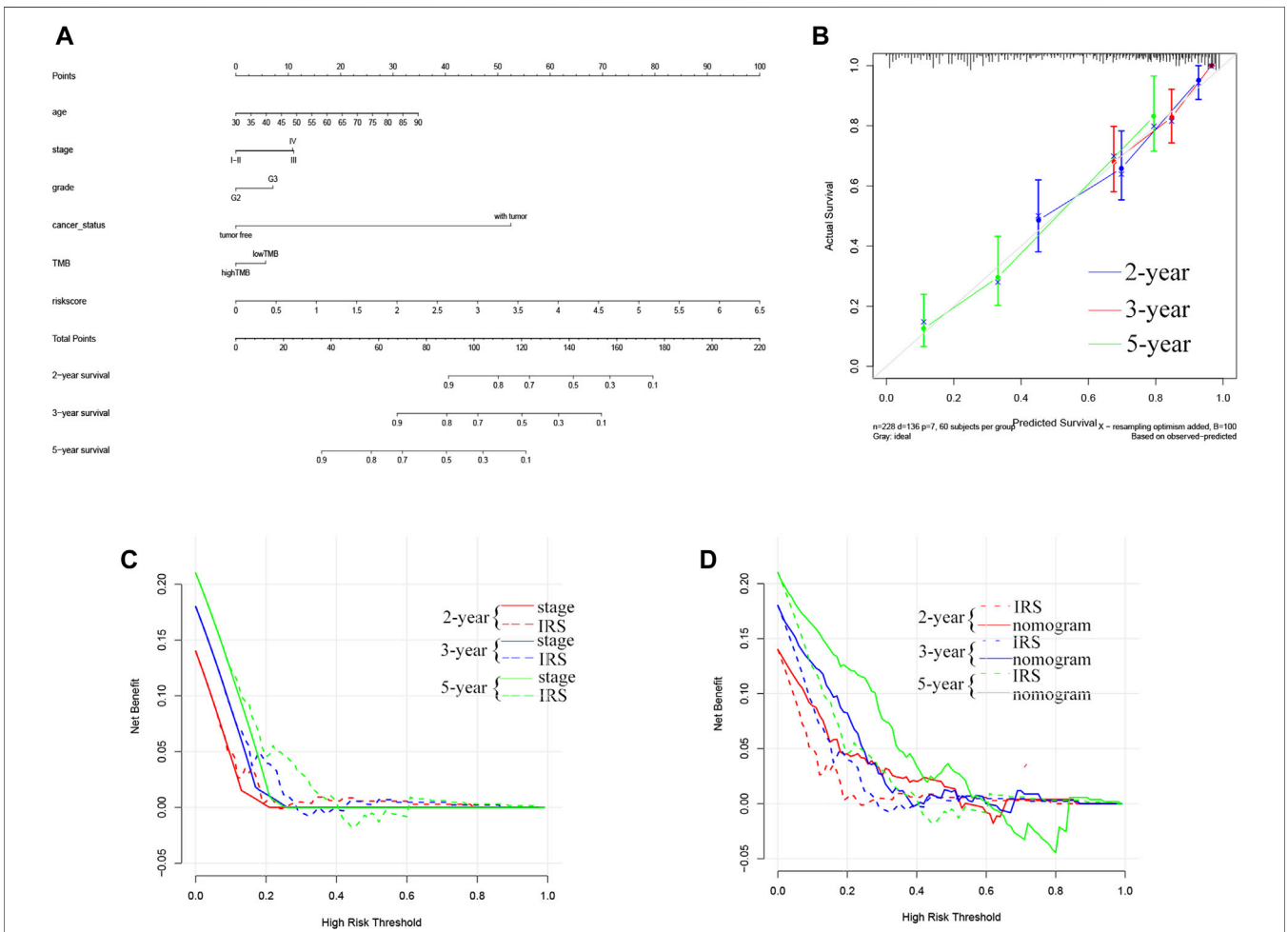



**FIGURE 4 |** Establishment of the OS nomogram for OC patients. **(A)** Nomogram for predicting OS of OC. There are eight components in this nomogram: age, stage, grade, cancer status, and risk score. **(B)** The Calibration curve of the nomogram for predicting OS rate at 2, 3, 5 years **(C,D)** Decision curve analysis for the evaluation of the net benefits of TNM-stage, IRS and nomogram at 2, 3, 5 years.

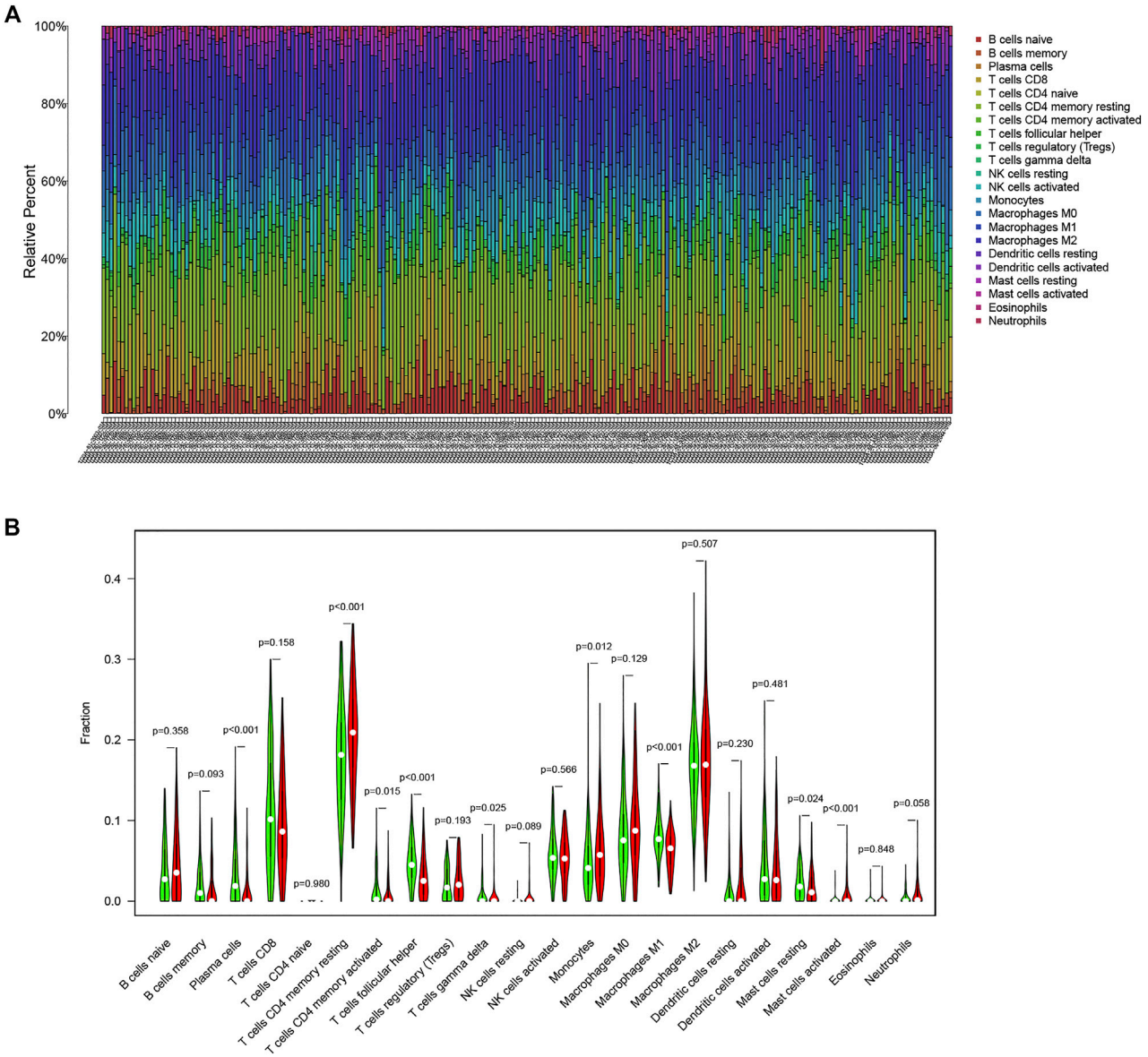



**FIGURE 5 |** Immune infiltration in OC. **(A)** Bar plot of the proportion of 22 immune cells in the TCGA-OC patients **(B)** Comparison of 22 immune cell infiltration between the two groups in the high-risk and low-risk groups (red means the high-risk group; green means the low-risk group).

however, nomograms combining multiple clinical features had the best clinical application value.

## TUMOR IMMUNE INFILTRATION IN OC

To explore the potential relationship between our risk score system and the immune infiltration microenvironment, we analyzed the correlation between the TMB-IRS and infiltrating immune cells using the "CIBERSORT" tool. The landscape of 22 immune cell infiltrates from each OC sample in TCGA was shown in **Figure 5A**. **Figure 5B** showed that Plasma cells, T cells CD4 memory activated, T cells follicular helper, Monocytes, Macrophages M1, and Mast cells resting were higher infiltrating in low-risk groups, while T cells CD4 memory resting, T cells gamma delta and Mast cells activated was higher infiltrating in high-risk groups. **Supplementary Figure S1** showed that CD4 memory resting, NK cells resting, Macrophages M0, Mast cells activated and Neutrophils were positively correlated with the risk score, while Plasma cells, CD4 memory activated, T cells follicular helper, T cells gamma delta, Macrophages M1, Mast cells resting were

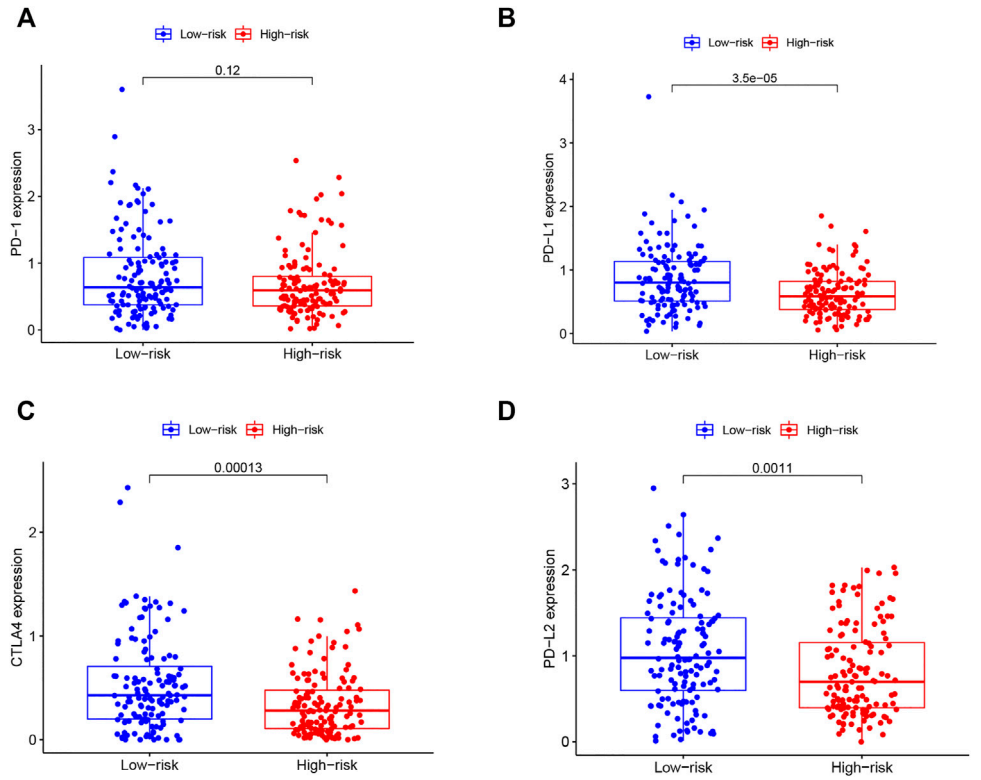



**FIGURE 6 |** Correlation of immune checkpoint molecules with risk score **(A–D)** boxplots of PD1, PDL1, CTLA4, PD-L2 expression of OC patients in high and low-risk groups.

negatively correlated with the risk score. Patients in the low-risk group had higher proportions of immune cell infiltration, with a $p < 0.05$.

## RELATIONSHIP BETWEEN IMMUNE CHECKPOINTS AND TMB-IRS

In recent years, cancer immunotherapy utilizing ICIs has shown promising efficacy in a proportion of cancer patients (O'Donnell et al., 2019). To explore the application value of our established TMB based model in immunotherapy, we plotted boxplots for the comparison between the expression levels of immune checkpoint molecules (PD1, PD-L1, PD-L2, CTLA4) between high-IRS and low-IRS. The results (**Figures 6A–D**) showed a significant negative correlation between the expression of PD-L1 ($p < 0.001$), PD-L2 ($p = 0.001$) as well as CTLA4 ($p < 0.001$) and the TMB-IRS. Specifically, the high-IRS group, with relatively lower expression of immune checkpoint genes, whereas in the low-IRS, gene expression was higher. Interestingly, there was no statistical difference in the expression of the PD-1 gene between the two groups ($p = 0.120$).

## DISCUSSION

OC is one of the common gynecological malignancies, with 14,070 patients dying of OC in 2018 in the United States alone, and most patients are already at an advanced stage at the time of diagnosis with a poor prognosis (Torre et al., 2018). Immunotherapy has become a promising personalized therapy for OC, but there is still a lack of reliable molecular biomarkers to distinguish patients with potential sensitivity to immunotherapy (Finkelmeier et al., 2018). Therefore, it is particularly important to identify more immune-related prognostic biomarkers, which can be used as potential therapeutic targets or can be used to screen patients sensitive to immunotherapy (Odunsi, 2017). TMB is a new type of biomarker that predicts the response of cancer immunotherapy. The findings of Wang et al. indicated that high TMB could promote antigen expression and inflammatory response of testicular tumors, and patients with high TMB might achieve a better prognosis if treated with immunotherapy (Wang and Li, 2019; Yan et al., 2020). However, few studies have focused on the prognostic role of TMB and the association between TMB and OC immune cell infiltration. Therefore, in this study, we aimed to explore the

prognostic role of TMB-related immune genes and their potential association with immune infiltration.

It is well known that cancer is a genetic disease and that neoplastic transformation results from the accumulation of somatic mutations in the DNA of diseased cells (Chan et al., 2019). In our study, missense mutations are the most common type of mutation in OC, and *TP53* mutations are the most frequently mutated gene, which can be identified in more than 90% of OC samples. The tumor suppressor gene *TP53* encodes the tumor suppressor protein p53, and its mutations are abundantly reported to be associated with poor prognosis in a variety of cancers (Luo et al., 2018; Li et al., 2019). In the current study, the high-TMB group has a more favorable prognosis, and conversely, the low-TMB group has a significantly poorer outcome. Yin et al.(Yin et al., 2020) suggested that high TMB could induce immune responses in humans, resulting in inhibition of tumor growth, followed by a relatively high survival rate of patients.

To screen out immune genes related to the prognosis of OC, immune-related genes were selected from the DEGs for univariate Cox and LASSO regression analysis. Nine independent prognostic immune genes associated with TMB were screened out and established a prognostic TMB-IRS signature. Among them, *DKK1*, *PI3*, *DUOX1*, *PAEP* and *CXCL10* genes are positively correlated with OS, while *CSPG5*, *CXCL11*, *TNFRSF17* and *TNFRSF13B* genes are negatively correlated with OS. Chondroitin sulfate proteoglycan 5 (*CSPG5*) encodes human chondroitin GSPG5, which is related to immune-related genes that are prognostic indicators of breast cancer and liver cancer patients (Shi et al., 2020). *CXCL10* and *CXCL11* are ligands of chemokine CXCR3, which can regulate the migration, differentiation and activation of immune cells, and are related to the selective migration and linear development of CD4 + and CD8 + T cells (Karin and Razon, 2018), thereby affecting the therapeutic effect of cancer (Tokunaga et al., 2018). DKK1, a regulator of Wnt signaling, is found to affect the tumor microenvironment by suppressing tumor immunity and can be used as an immunotherapeutic target for OC (Betella et al., 2020), which is consistent with our research results. *TNFRSF17* and *TNFRSF13B*, members of the tumor necrosis factor receptor superfamily, are primarily involved in the maturation of B lymphocytes and are associated with tumor growth and invasiveness and may serve as therapeutic targets in breast cancer (Pelekanou et al., 2018). Previous studies have shown that dual oxidase 1 (DUOX1) is commonly downregulated in lung, liver, and breast cancers, suggesting that it may have a tumor suppressor role (Little et al., 2016; Fortunato et al., 2018). Progesterone-associated endometrial protein (PAEP) can be used as a non-invasive biomarker to break down endometriosis (Irungu et al., 2019). Studies have reported its utility as a biomarker and immune system modulator in non-small cell lung cancer (Weber et al., 2019) and its association with prognosis in bladder cancer (Liu L et al., 2020). However, the prognostic relevance of PI3 in cancer has been less frequently reported, which may shed light on the mechanistic investigation of a novel immune gene in cancer.

In our study, KM analysis and ROC curve results confirm the favorable prognostic predictive value and accuracy of our established TMB-IRS signature. Specifically, the TMB-IRS signature can stratify patients into high- and low-risk groups with different outcomes and immunophenotypes, and the high-risk group is significantly associated with poor prognosis. Further, we have determined the relationship between the established model and multiple clinicopathological factors (age, cancer status, grade, stage, and ethnicity). TMB-IRS is significantly positively correlated with cancer status but negatively correlated with TMB, which is consistent with previous studies (Chan et al., 2019). Univariate and multivariate Cox regression results indicate that TMB-IRS, tumor status and age are independent prognostic predictors for the prognosis of OC patients. To comprehensively evaluate the prognosis of patients, we also establish a novel comprehensive nomogram risk assessment model based on clinical information. DCA and C-index results show that the predictive accuracy of TMB-IRS is higher than traditional TNM staging, while the nomogram containing multiple clinical information has the best prognostic predictive accuracy.

Accumulating evidence suggests that the immune component of the TME may be highly involved in tumor progression, as an immunosuppressive TME is associated with a worse patient prognosis (Tsogas et al., 2021). Immune cell infiltration in the tumor microenvironment can affect the treatment response and outcome of OC (Chalmers et al., 2017). Our research results show that Plasma cells, T cells CD4 memory activated, T cells follicular helper (Tfh), Monocytes, Macrophages M1, and Mast cells resting are higher infiltrating in low-risk groups, while T cells CD4 memory resting, T cells gamma delta and Mast cells activated is higher infiltrating in high-risk groups. This indirectly proves that the high immune response can inhibit the growth of OC tumors and improve the prognosis. Hollern et al. found that immune checkpoint therapy could induce the activation of Tfh of B cells, thereby promoting the anti-tumor response in a mouse model of triple-negative breast cancer (Hollern et al., 2019). In this study, 12 cells out of 22 immune cells were significantly correlated with TMB-IRS, and three of these cells (macrophage M1 T cell follicular helper plasma cells) were highly correlated with TMB-IRS ($R > 0.3$). High-affinity antibodies secreted by B cells and plasma cells are essential for the organism to fight and clear pathogen infections, whereas germinal center formation, B cell differentiation, and antibody affinity maturation are all independent of follicular helper T cell help, and macrophage M1, a macrophage that can produce proinflammatory cytokines, has strong microbial killing properties (He et al., 2018). In our study, the lower these three cell levels were when TMB-IRS was higher, which explained the potentially threatening and poor prognosis of tumors to some extent.

Currently, to effectively predict the prognosis of tumor patients, a large number of models matching the prognosis of tumor patients have been established and validated. For example, Shen et al. developed a promising biomarker based on immune genes that could predict overall survival in OC through the

Immport database (Shen et al., 2019). Using a TMB-associated signature to predict OS in OC, Bi et al. concluded that TMBB plays a critical role in the prognosis of OC and guides immunotherapy (Bi et al., 2020). In the study of fan et al.(Fan et al., 2020), the TMB-related genes were obtained by constructing the WGCNA network, and we were the DEGs obtained by differential analysis. Liu et al.'s (Liu J et al., 2020) study constructed a prognostic risk score for EOC (epithelial ovarian cancer) by obtaining all genes associated with TMB, while our study focused on the prognostic predictive role played by immune genes in OC. However, in our study, based on TMB high and low grouping, a signature constituted by 9 immune genes was established, which could more accurately predict the prognosis of OC, suggesting the level of immune cell infiltration, and thus guide immunotherapy.

ICIs with blocking antibodies targeting cytotoxic T lymphocyte antigen-4 (CTLA-4) as well as the programmed cell death protein 1 (PD-1) pathway and programmed death-1/programmed death-ligand 1 (PD-L1) has shown promising results in a variety of malignancies including OC (Odunsi, 2017; Memon and Patel, 2019). In our study, the expression of these immune checkpoint molecules was inversely correlated with that of TMB-IRS, suggesting a potential predictive role of our model for individual response to immunotherapy.

In the current study, we first explore the correlation between TMB and the prognosis of OC, and the results show that higher TMB levels are significantly associated with a better prognosis of OC. Based on the TMB score, nine TMB associated immune genes are identified, from which a biomarker TMB-IRS is constructed that can also effectively predict the prognosis of OC. We find that the TMB-IRS signature is negatively correlated with infiltrating immune cells, a new robust TMB-IRS signature, to help clinicians determine the most likely benefit from immunotherapy. The TMB-IRS signature, based on its strong prognostic predictive value and its association with immunotherapy, may serve as a novel biomarker and potential therapeutic target for predicting OC prognosis. The present study is a retrospective study, which is a limitation, so further prospective studies and clinical validation of its analytical accuracy and testing its clinical utility are warranted.

## 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

MC and QX conceived the study, performed data analysis and wrote the article. XZ downloaded gene expression data of OC. HJ and SW critically revised the article for research content and administrative support. The final manuscript was read and approved by all authors.

## FUNDING

This study was supported by Natural Science Foundation of China (81872684), the Fundamental Research Funds for the Central Universities, Southeast University "Zhongying Young Scholars" Project, the Six Talent Peaks Project in Jiangsu Province (wsw-201), the "SIX ONE" Talent Research Project for the High-level Health Personnel of Jiangsu Province (LGY2020050), the Fifth Scientific Research Project of Nantong ("226 Project"), Research Project from Nantong Commission of Health (MB2020018), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0162).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.688207/full#supplementary-material

## REFERENCES

Betella, I., Turbitt, W. J., Szul, T., Wu, B., Martinez, A., Katre, A., et al. (2020). Wnt Signaling Modulator DKK1 as an Immunotherapeutic Target in Ovarian Cancer. *Gynecol. Oncol.* 157 (3), 765–774. doi:10.1016/j.ygyno.2020.03.010

Bi, F., Chen, Y., and Yang, Q. (2020). Significance of Tumor Mutation burden Combined with Immune Infiltrates in the Progression and Prognosis of Ovarian Cancer. *Cancer Cel Int* 20, 373. doi:10.1186/s12935-020-01472-9

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., et al. (2017). Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational burden. *Genome Med.* 9 (1), 34. doi:10.1186/s13073-017-0424-2

Chan, T. A., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S. A., Stenzinger, A., et al. (2019). Development of Tumor Mutation burden as an Immunotherapy Biomarker: Utility for the Oncology Clinic. *Ann. Oncol.* 30 (1), 44–56. doi:10.1093/annonc/mdy495

Fan, S., Gao, X., Qin, Q., Li, H., Yuan, Z., and Zhao, S. (2020). Association between Tumor Mutation burden and Immune Infiltration in Ovarian Cancer. *Int. Immunopharmacol.* 89, 107126. doi:10.1016/j.intimp.2020.107126

Finkelmeier, F., Waidmann, O., and Trojan, J. (2018). Nivolumab for the Treatment of Hepatocellular Carcinoma. *Expert Rev. Anticancer Ther.* 18 (12), 1169–1175. doi:10.1080/14737140.2018.1535315

Fortunato, R. S., Gomes, L. R., Munford, V., Pessoa, C. F., Quinet, A., Hecht, F., et al. (2018). DUOX1 Silencing in Mammary Cell Alters the Response to Genotoxic Stress. *Oxid. Med. Cell Longev.* 2018, 1–9. doi:10.1155/2018/3570526

Ghisoni, E., Imbimbo, M., Zimmermann, S., and Valabrega, G. (2019). Ovarian Cancer Immunotherapy: Turning up the Heat. *Int. J. Mol. Sci.* 20 (12), 2927. doi:10.3390/ijms20122927

Hao, D., Liu, J., Chen, M., Li, J., Wang, L., Li, X., et al. (2018). Immunogenomic Analyses of Advanced Serous Ovarian Cancer Reveal Immune Score Is a Strong Prognostic Factor and an Indicator of Chemosensitivity. *Clin. Cancer Res.* 24 (15), 3560–3571. doi:10.1158/1078-0432.CCR-17-3862

He, L., Gu, W., Wang, M., Chang, X., Sun, X., Zhang, Y., et al. (2018). Extracellular Matrix Protein 1 Promotes Follicular Helper T Cell Differentiation and

Antibody Production. *Proc. Natl. Acad. Sci. USA* 115 (34), 8621–8626. doi:10.1073/pnas.1801196115

Hinshaw, D. C., and Shevde, L. A. (2019). The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Res.* 79 (18), 4557–4566. doi:10.1158/0008-5472.CAN-18-3962

Hollern, D. P., Xu, N., Thennavan, A., Glodowski, C., Garcia-Recio, S., Mott, K. R., et al. (2019). B Cells and T Follicular Helper Cells Mediate Response to Checkpoint Inhibitors in High Mutation Burden Mouse Models of Breast Cancer. *Cell* 179 (5), 1191–1206. doi:10.1016/j.cell.2019.10.028

Huo, J., Wu, L., and Zang, Y. (2020). A Prognostic Model of 15 Immune-Related Gene Pairs Associated with Tumor Mutation Burden for Hepatocellular Carcinoma. *Front. Mol. Biosci.* 7, 581354. doi:10.3389/fmolb.2020.581354

Irungu, S., Mavrelos, D., Worthington, J., Blyuss, O., Saridogan, E., and Timms, J. F. (2019). Discovery of Non-invasive Biomarkers for the Diagnosis of Endometriosis. *Clin. Proteom* 16, 14. doi:10.1186/s12014-019-9235-3

Karin, N., and Razon, H. (2018). Chemokines beyond Chemo-Attraction: CXCL10 and its Significant Role in Cancer and Autoimmunity. *Cytokine* 109, 24–28. doi:10.1016/j.cyto.2018.02.012

Killock, D. (2020). TMB - a Histology-Agnostic Predictor of the Efficacy of ICIs? *Nat. Rev. Clin. Oncol.* 17 (12), 718. doi:10.1038/s41571-020-00438-0

Kruger, S., Ilmer, M., Kobold, S., Cadilha, B. L., Endres, S., Ormanns, S., et al. (2019). Advances in Cancer Immunotherapy 2019 - Latest Trends. *J. Exp. Clin. Cancer Res.* 38 (1), 268. doi:10.1186/s13046-019-1266-0

Le Saux, O., Dubois, B., Stern, M.-H., Terme, M., Tartour, E., Classe, J.-M., et al. (2020). Les avancées actuelles de l'immunothérapie dans le cancer de l'ovaire. *Bull. Du Cancer* 107 (4), 465–473. doi:10.1016/j.bulcan.2019.11.015

Li, V. D., Li, K. H., and Li, J. T. (2019). TP53 Mutations as Potential Prognostic Markers for Specific Cancers: Analysis of Data from the Cancer Genome Atlas and the International Agency for Research on Cancer TP53 Database. *J. Cancer Res. Clin. Oncol.* 145 (3), 625–636. doi:10.1007/s00432-018-2817-z

Little, A. C., Sham, D., Hristova, M., Danyal, K., Heppner, D. E., Bauer, R. A., et al. (2016). DUOX1 Silencing in Lung Cancer Promotes EMT, Cancer Stem Cell Characteristics and Invasive Properties. *Oncogenesis* 5 (10), e261. doi:10.1038/oncsis.2016.61

Liu L, L., Hu, J., Wang, Y., Sun, T., Zhou, X., Li, X., et al. (2020). Establishment of a Novel Risk Score Model by Comprehensively Analyzing the Immunogen Database of Bladder Cancer to Indicate Clinical Significance and Predict Prognosis. *Aging* 12 (12), 11967–11989. doi:10.18632/aging.103364

Liu J, J., Xu, W., Li, S., Sun, R., and Cheng, W. (2020). Multi-omics Analysis of Tumor Mutational burden Combined with Prognostic Assessment in Epithelial Ovarian Cancer Based on TCGA Database. *Int. J. Med. Sci.* 17 (18), 3200–3213. doi:10.7150/ijms.50491

Luo, Y., Huang, W., Zhang, H., and Liu, G. (2018). Prognostic Significance of CD117 Expression and TP53 Missense Mutations in Triple-Negative Breast Cancer. *Oncol. Lett.* 15 (5), 6161–6170. doi:10.3892/ol.2018.8104

Memon, H., and Patel, B. M. (2019). Immune Checkpoint Inhibitors in Non-small Cell Lung Cancer: A Bird's Eye View. *Life Sci.* 233, 116713. doi:10.1016/j.lfs.2019.116713

Merino, D. M., McShane, L. M., Fabrizio, D., Funari, V., Chen, S.-J., White, J. R., et al. (2020). Establishing Guidelines to Harmonize Tumor Mutational burden (TMB): In Silico Assessment of Variation in TMB Quantification across Diagnostic Platforms: Phase I of the Friends of Cancer Research TMB Harmonization Project. *J. Immunother. Cancer* 8 (1), e000147. doi:10.1136/jitc-2019-000147

O'Donnell, J. S., Hoefsmit, E. P., Smyth, M. J., Blank, C. U., and Teng, M. W. L. (2019). The Promise of Neoadjuvant Immunotherapy and Surgery for Cancer Treatment. *Clin. Cancer Res.* 25 (19), 5743–5751. doi:10.1158/1078-0432.CCR-18-2641

Odunsi, K. (2017). Immunotherapy in Ovarian Cancer. *Ann. Oncol.* 28, viii1–viii7. doi:10.1093/annonc/mdx444

Orr, B., and Edwards, R. P. (2018). Diagnosis and Treatment of Ovarian Cancer. *Hematol. Oncol. Clin. North Am.* 32 (6), 943–964. doi:10.1016/j.hoc.2018.07.010

Pelekanou, V., Notas, G., Athanasouli, P., Alexakis, K., Kiagiadaki, F., Peroulis, N., et al. (2018). BCMA (TNFRSF17) Induces APRIL and BAFF Mediated Breast Cancer Cell Stemness. *Front. Oncol.* 8, 301. doi:10.3389/fonc.2018.00301

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., et al. (2015). Mutational Landscape Determines Sensitivity to PD-1 Blockade in Non-small Cell Lung Cancer. *Science* 348 (6230), 124–128. doi:10.1126/science.aaa1348

Shen, S., Wang, G., Zhang, R., Zhao, Y., Yu, H., Wei, Y., et al. (2019). Development and Validation of an Immune Gene-Set Based Prognostic Signature in Ovarian Cancer. *EBioMedicine* 40, 318–326. doi:10.1016/j.ebiom.2018.12.054

Shi, W., Feng, L., Dong, S., Ning, Z., Hua, Y., Liu, L., et al. (2020). Exploration of Prognostic index Based on Immune-Related Genes in Patients with Liver Hepatocellular Carcinoma. *Biosci. Rep.* 40 (7). doi:10.1042/BSR20194240

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* 371 (23), 2189–2199. doi:10.1056/NEJMoa1406498

Stewart, C., Ralyea, C., and Lockwood, S. (2019). Ovarian Cancer: An Integrated Review. *Semin. Oncol. Nurs.* 35 (2), 151–156. doi:10.1016/j.soncn.2019.02.001

Tokunaga, R., Zhang, W., Naseem, M., Puccini, A., Berger, M. D., Soni, S., et al. (2018). CXCL9, CXCL10, CXCL11/CXCR3 axis for Immune Activation - A Target for Novel Cancer Therapy. *Cancer Treat. Rev.* 63, 40–47. doi:10.1016/j.ctrv.2017.11.007

Torre, L. A., Trabert, B., DeSantis, C. E., Miller, K. D., Samimi, G., Runowicz, C. D., et al. (2018). Ovarian Cancer Statistics, 2018. *CA: A Cancer J. Clin.* 68 (4), 284–296. doi:10.3322/caac.21456

Tsogas, F. K., Majerczyk, D., and Hart, P. C. (2021). Possible Role of Metformin as an Immune Modulator in the Tumor Microenvironment of Ovarian Cancer. *Int. J. Mol. Sci.* 22 (2), 867. doi:10.3390/ijms22020867

Valmiki, S., Aid, M. A., Chaitou, A. R., Zahid, M., Valmiki, M., Fawzy, P., et al. (2021). Extracellular Matrix: A Treasure Trove in Ovarian Cancer Dissemination and Chemotherapeutic Resistance. *Cureus* 13 (3), e13864. doi:10.7759/cureus.13864

Wang, X., and Li, M. (2019). Correlate Tumor Mutation burden with Immune Signatures in Human Cancers. *BMC Immunol.* 20 (1), 4. doi:10.1186/s12865-018-0285-5

Wang, W., Kryczek, I., Dostál, L., Lin, H., Tan, L., Zhao, L., et al. (2016). Effector T Cells Abrogate Stroma-Mediated Chemoresistance in Ovarian Cancer. *Cell* 165 (5), 1092–1105. doi:10.1016/j.cell.2016.04.009

Wang, W., Liu, J. R., and Zou, W. (2019). Immunotherapy in Ovarian Cancer. *Surg. Oncol. Clin. North Am.* 28 (3), 447–464. doi:10.1016/j.soc.2019.02.002

Webb, P. M., and Jordan, S. J. (2017). Epidemiology of Epithelial Ovarian Cancer. *Best Pract. Res. Clin. Obstet. Gynaecol.* 41, 3–14. doi:10.1016/j.bpobgyn.2016.08.006

Weber, R., Meister, M., Muley, T., Thomas, M., Sültmann, H., Warth, A., et al. (2019). Pathways Regulating the Expression of the Immunomodulatory Protein Glycodelin in Non Small Cell Lung Cancer. *Int. J. Oncol.* 54 (2), 515–526. doi:10.3892/ijo.2018.4654

Yan, J., Wu, X., Yu, J., Zhu, Y., and Cang, S. (2020). Prognostic Role of Tumor Mutation Burden Combined with Immune Infiltrates in Skin Cutaneous Melanoma Based on Multi-Omics Analysis. *Front. Oncol.* 10, 570654. doi:10.3389/fonc.2020.570654

Yin, W., Jiang, X., Tan, J., Xin, Z., Zhou, Q., Zhan, C., et al. (2020). Development and Validation of a Tumor Mutation Burden-Related Immune Prognostic Model for Lower-Grade Glioma. *Front. Oncol.* 10, 1409. doi:10.3389/fonc.2020.01409

**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