ADVANCES IN AI-BASED TOOLS FOR PERSONALIZED CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT

EDITED BY: Israel Tojal Da Silva, Joel Correa Da Rosa, Liang Zhao,
Rodrigo Drummond and Jianjiong Gao

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-020-0
DOI 10.3389/978-2-83250-020-0

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

ADVANCES IN AI-BASED TOOLS FOR PERSONALIZED CANCER DIAGNOSIS, PROGNOSIS AND TREATMENT

Topic Editors:
Israel Tojal Da Silva, AC Camargo Cancer Center, Brazil
Joel Correa Da Rosa, Icahn School of Medicine at Mount Sinai, United States
Liang Zhao, University of São Paulo, Ribeirão Preto, Brazil
Rodrigo Drummond, A.C.Camargo Cancer Center, Brazil
Jianjiong Gao, Memorial Sloan Kettering Cancer Center, United States

Citation: Da Silva, I. T., Da Rosa, J. C., Zhao, L., Drummond, R., Gao, J., eds. (2022). Advances in AI-Based Tools for Personalized Cancer Diagnosis, Prognosis and Treatment. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-020-0

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/fgene.2021.757582

Pan-Cancer Analysis of the Solute Carrier Family 39 Genes in Relation to Oncogenic, Immune Infiltrating, and Therapeutic Targets

Yi-Yuan Qu[1], *Rong-Yan Guo*[2], *Meng-Ling Luo*[1] *and Quan Zhou*[1]*

[1] *Department of Gynecology and Obstetrics, the People's Hospital of China Three Gorges University/The First People's Hospital of Yichang, Yichang, China,* [2] *Emergency Services Department, HanYang Hospital Affiliated of Wuhan University of Science and Technology, Wuhan, China*

Background: Emerging pieces of evidence demonstrated that the solute carrier family 39 (SLC39A) members are critical for the oncogenic and immune infiltrating targets in multiple types of tumors. However, the precise relationship between the *SLC39A* family genes and clinical prognosis as well as the pan-cancer tumor cell infiltration has not been fully elucidated.

Methods: In this study, the pan-cancer expression profile, genetic mutation, prognostic effect, functional enrichment, immune infiltrating, and potential therapeutic targets of the SLC39A family members were investigated by analyzing multiple public databases such as the Oncomine, TIMER, GEPIA, cBioPortal, KM-plotter, PrognoScan, GeneMANIA, STRING, DAVID, TIMER 2.0, and CellMiner databases.

Results: The expression levels of most *SLC39* family genes in the tumor tissues were found to be significantly upregulated compared to the normal group. In mutation analysis, the mutation frequencies of *SLC39A4* and *SLC39A1* were found to be higher among all the members (6 and 4%, respectively). Moreover, the overall mutation frequency of the *SLC39A* family genes ranged from 0.8 to 6% pan-cancer. Also, the function of the *SLC39A* highly related genes was found to be enriched in functions such as zinc II ion transport across the membrane, steroid hormone biosynthesis, and chemical carcinogenesis. In immune infiltration analysis, the expression level of the SLC39A family genes was found to be notably related to the immune infiltration levels of six types of immune cells in specific types of tumors. In addition, the *SLC39A* family genes were significantly related to the sensitivity or resistance of 63 antitumor drugs in a variety of tumor cell lines.

Conclusion: These results indicate that the *SLC39* family genes are significant for determining cancer progression, immune infiltration, and drug sensitivity in multiple cancers. This study, therefore, provides novel insights into the pan-cancer potential targets of the *SLC39* family genes.

Keywords: ***SLC39*** **family genes, biomarker, pan-cancer, prognosis, immune infiltration, drug sensitivity**

Edited by:
Liang Zhao,
University of São Paulo Ribeirão Preto,
Brazil

Reviewed by:
Luca Zammataro,
Kiromic Biopharma Inc., Houston,
United States
Shailender Kumar Verma,
Central University of Himachal
Pradesh, India

Correspondence:
Quan Zhou
zhouquan8519@163.com

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

Received: *12 August 2021*
Accepted: *12 November 2021*
Published: *02 December 2021*

Citation:
Qu Y-Y, Guo R-Y, Luo M-L and Zhou Q (2021) Pan-Cancer Analysis of the Solute Carrier Family 39 Genes in Relation to Oncogenic, Immune Infiltrating, and Therapeutic Targets. Front. Genet. 12:757582. doi: 10.3389/fgene.2021.757582

INTRODUCTION

Cancer has gradually emerged as the leading threat to public health worldwide, as estimated by GLOBOCAN 2020, stating 19.3 million newly confirmed cancer cases and nearly 10 million cancer-related deaths (Ferlay et al., 2021; Sung et al., 2021). Indeed, efforts for cancer prevention, screening, diagnosis, and comprehensive treatment have met with tremendous success in various tumors. However, studies on the clinical outcome of most cancers need further improvisation (Chen et al., 2021). The current promising targeted therapy particularly confirms that exploring the mechanism of pan-cancer initiation, maintenance, and development will unfurl new avenues for fighting various malignant tumors (Loomans-Kropp and Umar, 2019). Therefore, identification of the hub tumor-related genes is very urgent and necessary to develop new diagnostic and prognostic biomarkers and therapeutic targets. Presently, massive high-throughput data and multiple available big data online public databases are greatly helpful for finding the tumorigenic genes and conducting pan-cancer studies in multi-omics (Loomans-Kropp and Umar, 2019; Xiao et al., 2021).

The *SLC39A* family genes encode a family of proteins belonging to the Zrt- and Irt-like protein (ZIP) transport proteins, having 14 family members (SLC39A1-14). It controls the transportation and influx of zinc, with important roles in multiple signaling pathways and physiological processes, like gene transcription, endocrine regulation, cell growth, cell differentiation, and the immune response process (Kimura and Kambe, 2016; Baltaci and Yuce, 2018). Emerging pieces of evidence indicate that the mutation or functional change in the *SLC39A* family genes leads to the development and progression of multiple malignancies, such as colorectal cancer, breast cancer, esophageal cancer, hepatocellular carcinoma, pancreas cancer, gastric cancer, prostate cancer, and lung cancer (Hoang et al., 2016; To et al., 2020; Prasad, 2012). Besides, recent multi-omics studies have confirmed certain SLC39A family genes to have differential expression and prognostic value in breast, gastric, and lung cancers, acting as potentially promising clinical markers for these cancers (Liu et al., 2020; Zhou et al., 2021; Ding et al., 2019). Some basic studies have demonstrated the targeted regulation of the *SLC39A* family genes to be capable of changing the biological characteristics of some tumor cells. For example, Jin et al. found that knockdown of SLC39A5 expression significantly inhibits the invasion, proliferation, and migration of esophageal tumor cells (Jin et al., 2015). In addition, Zhu et al. demonstrated the knockdown of SLC39A11 to attenuate the cellular proliferation of the pancreatic cancer Capan-1 with decreased activation of the ERK1/2 pathway (Zhu et al., 2021). Fan et al. have found *SLC39A4* gene knockout to inhibit the malignant behavior of the ovarian tumor cells both *in vitro* and *in vivo* (Fan et al., 2017). More importantly, the growing studies have shown that SLC transporters not only directly bring the anticancer drugs into cancer cells but also serve as a medium for the uptake of essential nutrients for the growth and survival of the tumor, thereby regulating the sensitivity and resistance of the chemotherapeutic drugs (Li and Shu, 2014). Nevertheless, the underlying mechanism and biological functions of the *SLC39A* family genes in the tumor progression and as the potential therapeutic target have not been fully elucidated.

This study systematically performed an in-depth analysis on the expression of the *SLC39A* family genes and their impact on the prognosis, to explore the relationship between the *SLC39A* family genes and pan-cancer immune cell infiltration. In addition to utilizing the multi-omics and large sample data analysis, the genetic mutation, function enrichment, and drug sensitivity of the *SLC39A* family genes were investigated across different cancer types. These analyses could provide a new direction for a promising biomarker and potential targeted therapy for treating cancer.

MATERIALS AND METHODS

Expression Profiles Analysis

Three online databases (Oncomine, TIMER, and GEPIA) were applied to investigate the differential expression profiles of the SLC39A family genes between the normal and the tumor tissues in various cancer types. The website of the Oncomine online platform is www.oncomine.org (Rhodes et al., 2007), the website of the TIMER online platform is https://cistrome.shinyapps.io/timer/ (Li et al., 2017), and the website of the GEPIA online platform is http://gepia.cancer-pku.cn/ (Tang et al., 2017). Among them, the *p*-value was set to 0.01; fold change was set to 1.5; the gene level was set to all, and data type was set to mRNA in the Oncomine database, and the relevant parameters of the TIMER and GEPIA databases were set by default.

Mutation Profiles Analysis

The cBioPortal (http://www.cbioportal.org) (Gao et al., 2013) was exploited for detecting the mutation landscape (amplification, deep deletion, and missense mutations) and general mutation count of the *SLC39A* family genes in 33 types of tumors from the TCGA database. In addition, the impact of gene mutations in the *SLC39A* family on the clinical outcomes was surveyed using the cBioPortal database.

Survival Analysis

The relationship between the expression of the *SLC39A* family gene and the overall survival (OS) and progression-free survival (PFS) in the pan-cancer patients was investigated using the pan-cancer module of the KM-plotter database (http://www.kmplot.com/) (Nagy et al., 2021). In addition, the PrognoScan database (http://dna00.bio.kyutech.ac.jp/PrognoScan/index.html) (Mizuno et al., 2009) was further utilized to confirm the relationship between the expression of the SLC39A family genes and clinical outcome in the different cohorts. Multiple types of survival parameters, including OS, PFS, relapse-free survival (RFS), disease-free survival (DFS), distant recurrence–free survival (DRFS), distant metastasis–free survival (DMFS), and disease-specific survival (DSS) were represented in the current analysis. The hazard ratio (HR), log-rank *p*-value, and 95% confidence interval were directly

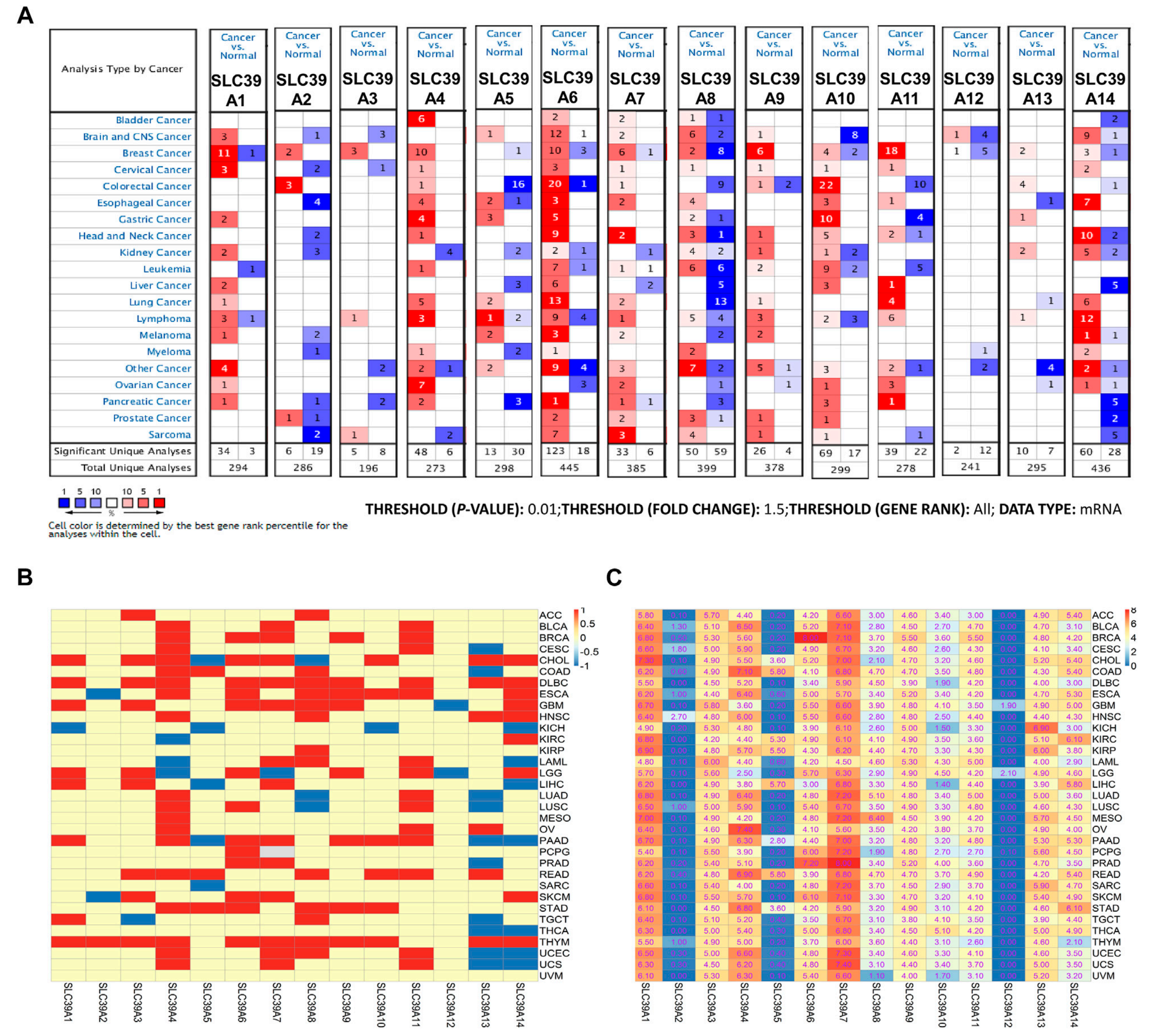


FIGURE 1 | mRNA expression profiles of the *SLC39A* family genes in pan-cancer. **(A)** Transcriptome expression profile of SLC39A family members in pan-cancer was explored in the Oncomine database. In the graph, red represents statistically significant mRNA overexpression of *SLC39A* family gene mRNA between the tumor and the corresponding normal tissue, blue represents down-expression, and the number represents the number of data sets. *p* value is set to 0.01; fold change is set to 1.5; gene level is set to all; and data type is set to mRNA in the Oncomine database. **(B)** Transcriptome expression profiles of *SLC39A* family genes in pan-cancer were explored in the GEPIA database. The red boxes represent higher *SLC39A* family gene expression in tumor tissues, while the green boxes represent the lower *SLC39A* family gene expression in tumor tissues. The inspection standard is set to *p*-value < 0.05. **(C)** Expression level of *SLC39A* family genes in 33 tumors and their normal controls in the match TCGA normal and GTEx data using GEPIA. The data in the figure represents average mRNA expression of *SLC39A* family genes in different tumors. The colors from blue to red represent the range of values in the figure.

displayed on the online platform, and the *p*-value cut-off value was set to 0.05.

Enrichment Analysis

To seek out the highly related genes of SLC39A family genes, the GeneMANIA database (http://www.genemania.org) (Warde-Farley et al., 2010) and the STRING database (https://string-db.org) (Szklarczyk et al., 2019) were exploited. Then, the DAVID database (Database for Annotation, Visualization, and Integrated Discovery, https://david.ncifcrf.gov) (Huang et al., 2009) was used to conduct the GO (gene ontology) annotation and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis of the *SLC39A* family highly related genes.

Immune Infiltration Analysis

The TIMER 2.0 (http://cistrome.shinyapps.io/timer) (Li et al., 2020) was used to evaluate the relationship between the *SLC39A* family gene expression levels and the infiltration of six common

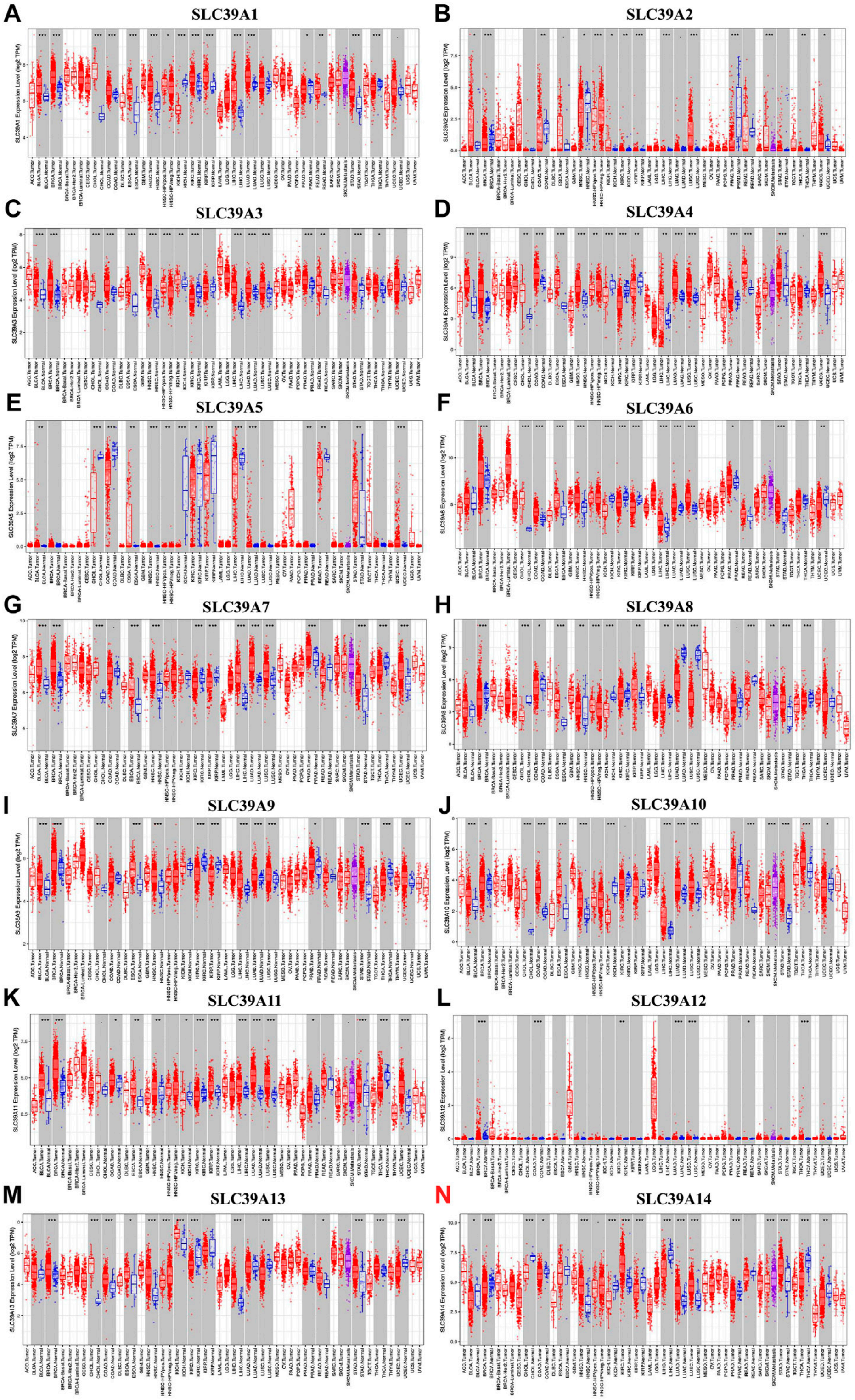


FIGURE 2 | Differential expression of the *SLC39A* family genes in pan-cancer and corresponding normal tissues. **(A)** Transcriptome expression of *SLC39A1* was explored in the TIMER database. **(B)** Transcriptome expression of *SLC39A2* was explored in the TIMER database. **(C)** Transcriptome expression of *SLC39A3* was explored in the TIMER database. **(D)** Transcriptome expression of *SLC39A4* was explored in the TIMER database. **(E)** Transcriptome expression of *SLC39A5* was explored in the TIMER database. **(F)** Transcriptome expression of *SLC39A6* was explored in the TIMER database. **(G)** Transcriptome expression of *SLC39A7* was explored in the TIMER database. **(H)** Expression of *SLC39A8* was explored in the TIMER database. **(I)** Transcriptome expression of *SLC39A 9* was explored in the (*Continued*)

FIGURE 2 | TIMER database. **(J)** Transcriptome expression of *SLC39A10* was explored in the TIMER database. **(K)** Transcriptome expression of *SLC39A11* was explored in the TIMER database. **(L)** Transcriptome expression of *SLC39A12* was explored in the TIMER database. **(M)** Transcriptome expression of *SLC39A13* was explored in the TIMER database. **(N)** Transcriptome expression of *SLC39A14* was explored in the TIMER database. (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The red box and the green box represent tumor tissue and normal control tissue, respectively. The middle line of the box represents the median and the lower and upper bounds represent the 25th and 75th percentiles, respectively.)

immune cells, including the B cells, CD4$^+$ T cells, CD8$^+$ T cells, Treg T cells, macrophages, and neutrophils.

Drug Sensitivity Analysis

The CellMiner database (https://discover.nci.nih.gov/cellminer/) (Shankavaram et al., 2009) was exploited to evaluate the relationship between the *SLC39A* family gene expression levels and the compound sensitivity or resistance through the NCI-60 analyses tools. Data processing and Pearson correlation analysis visualization used the limma and ggplot2 package, and the scatter plot showed significant correlations sorted by *p*-value from small to large, and the *p*-value cut-off value was set to 0.05.

RESULTS

The Expression Profiles of the SLC39A Family Genes in Pan-Cancer

Subsequently, the expression profiles of the *SLC39A* family genes were explored in various cancer types. The Oncomine, GEPIA, and TIMER databases were exploited to examine and verify the expression levels of the *SLC39A* family genes in the tumor tissues and the corresponding non-tumor tissues. The Oncomine database reported an increase in the mRNA expression level of other *SLC39* family genes in the tumor tissues compared to the normal control group, except for *SLC39A8* (**Figure 1A**). The median expression of the *SLC39A* family genes in the tumor tissues of all types of tumors was further compared, revealing that most of the *SLC39A* family genes show relatively high expression in the specific tumor types, such as lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), rectum adenocarcinoma (READ), and thymoma (THYM) (**Figure 1B**). In addition, the expression level of the *SLC39A* family genes was detected in 33 tumors, and their normal controls match TCGA normal and GTEx data using GEPIA. Similar to the results of the other studies, the *SLC39A* family genes have notably increased the expression in most tumors compared to the normal controls (**Figure 1C** and **Supplementary Table S1**). As shown in **Figure 2**, the TIMER2.0 database results demonstrated that the transcriptional expression levels of the *SLC39A* family genes are inconsistent between the tumor tissues and corresponding normal tissues, and most SLC39A families were over-regulated in the tumor tissues, extremely so in the bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), HNSC, kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD), and thyroid carcinoma (THCA) tumor types.

The Genetic Mutation of the SLC39A Family Genes in Pan-Cancer

The cBioPortal and TCGA database was employed to probe the mutation status of the *SLC39A* family genes in 10,967 samples in 32 studies of the pan-cancer atlas. Results showed that the mutation frequencies of *SLC39A4* and *SLC39A1* were higher than those of all the other members, 6 and 4%, respectively, and the overall mutation frequency of the *SLC39A* family genes ranged from 0.8 to 6% (**Figure 3A**). As shown in **Figure 3B**, the mutation frequency of the *SLC39A* family genes in ovarian serous cystadenocarcinoma (OV), liver hepatocellular carcinoma (LIHC), ESCA, uterine corpus endometrial carcinoma (UCEC), LUAD, skin cutaneous melanoma (SKCM), BLCA, uterine carcinosarcoma (UCS), lung squamous cell carcinoma (LUSC), STAD, and BRCA was relatively higher by more than 30%, and the other types of tumors all exhibited a very low alteration in mutation (<30%). In addition, the Kaplan–Meier plotter results demonstrated that the combined mutation of the *SLC39A* family genes has no significant effect on OS (*p*-values, 0.0664) (**Figure 3C**). However, there are statistical differences in the DSS, PFS, and DFS between the mutation group and the non-mutation group of the *SLC39A* family genes (*p*-values, 0.0308, 4.11e-5, and 7.94e-11, respectively) (**Figures 3D–F**).

The Prognostic Value of SLC39A Family Genes in Pan-Cancer

The association between the mRNA expression of the *SLC39A* family genes and the clinical outcomes in pan-cancer patients were analyzed using the KM-plotter and PrognoScan databases. As shown in **Figure 4**, the KM-plotter database revealed that the expression of the *SLC39A* family genes was significantly related to the OS and RFS in some tumor types. Among them, the high expression of most of the *SLC39A* family genes presents the risk factor for the OS of BLCA, CESC, HNSC, LIHC, LUAD, LUSC, and PAAD, as well as for the protection factors for OS of BRCA, ESCA, OV, PAAD, STAD, TGCT, and THCA. Similarly, for RFS, the high expression of most *SLC39A* family genes was significantly related to the inferior survival of BLCA, CESC, KIRP, LUAD, LUSC, PAAD, and TGCT, as well as the better prognosis of BRCA, OV, PCPG, and STAD (**Figures 4A,B** and **Supplementary Tables S2, S3**). As shown in **Figures 4C–R**, the increase in the expression of SLC39A1, SLC39A 3, SLC39A 5, SLC39A 8, SLC39A 10, SLC39A 13, and SLC39A 14 was associated with poor OS, and the upregulation of SLC39A1, SLC39A 4, SLC39A 7, SLC39A 9, and SLC39A 10 was found to lead to poor RFS in the patients with CESC.

Then, the PrognoScan platform was used to further assess and verify the prognostic value of the *SLC39A* family genes in pan-

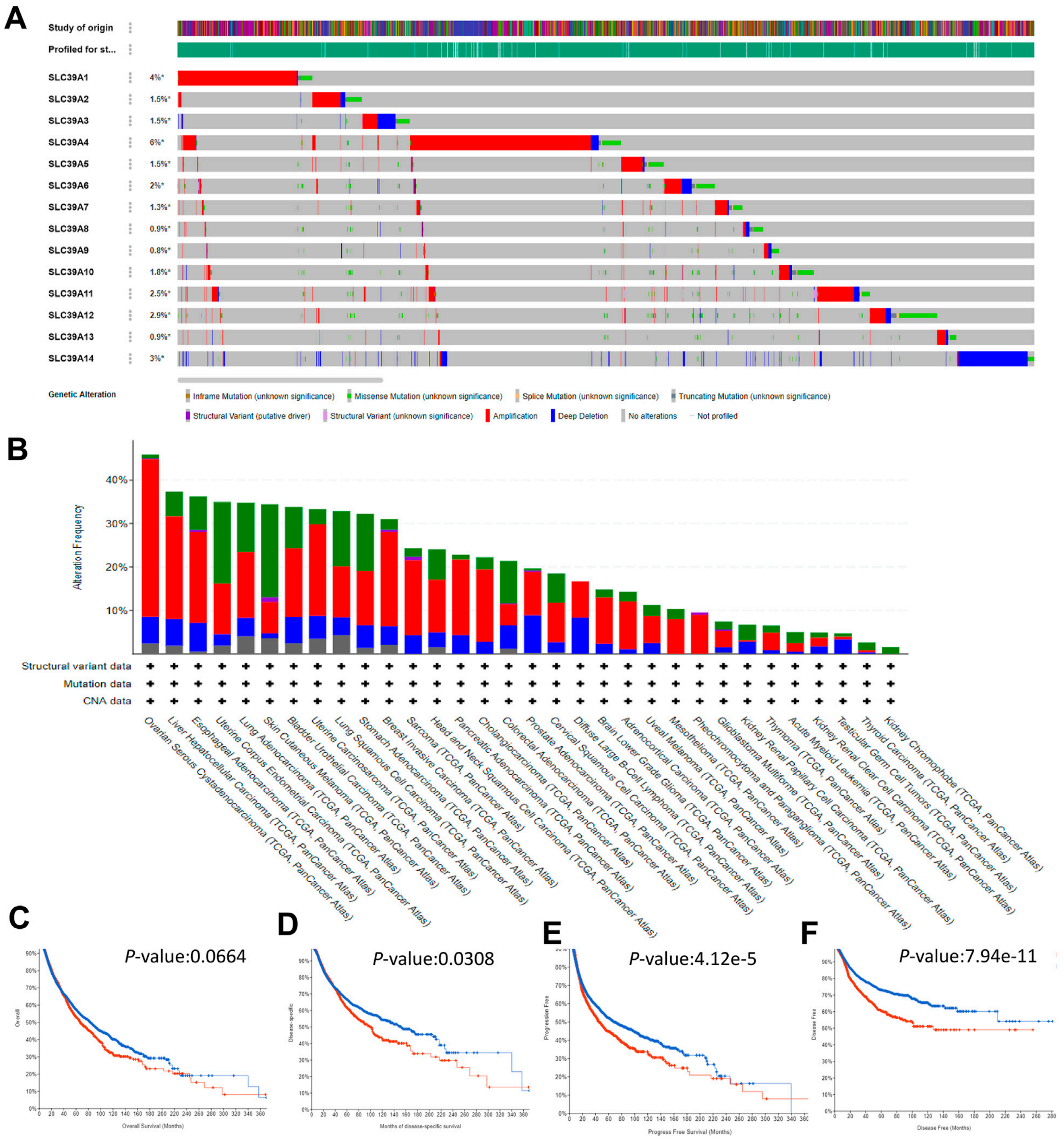


FIGURE 3 | Mutation landscape of the *SLC39A* family genes in pan-cancer derived from the cBioPortal platform. **(A)** OncoPrint summary of alteration on *SLC39A* family genes in 10,967 numbers of samples in 32 studies of the pan-cancer atlas. The mutation frequency was shown as green for mutations, red for fusions, blue for sions, and black for multiple mutations. **(B)** Rectangular graph of the general mutation counts of *SLC39A* family genes in the pan-cancer atlas. The *X*- and *Y*-axis represent the mutation frequency of *SLC39A* family genes and cancer type, respectively. It was shown as green for missense mutations, violet for fusions, deep blue for truncating, and blue for no mutations. **(C)** Kaplan–Meier chart of OS of pan-cancer with and without *SLC39A* family gene mutation. **(D)** Kaplan–Meier chart of DDS of pan-cancer with and without *SLC39A* family gene mutation. **(E)** Kaplan–Meier chart of PFS of pan-cancer with and without *SLC39A* family gene mutation. **(F)** Kaplan–Meier chart of DFS of pan-cancer with and without *SLC39A* family gene mutation. The red line and the blue line represent high and low expression, respectively. The Cox *p*-value cut-off value was set to 0.05.

cancer, based on public datasets. The results of PrognoScan indicated that the expression level of the *SLC39A* family genes was significantly related to the clinical survival of 12 types of tumors, such as colorectal cancer, breast cancer, bladder cancer, lung cancer, ovarian cancer, blood cancer, brain cancer, skin cancer, eye cancer, soft tissue cancer, prostate cancer, and head

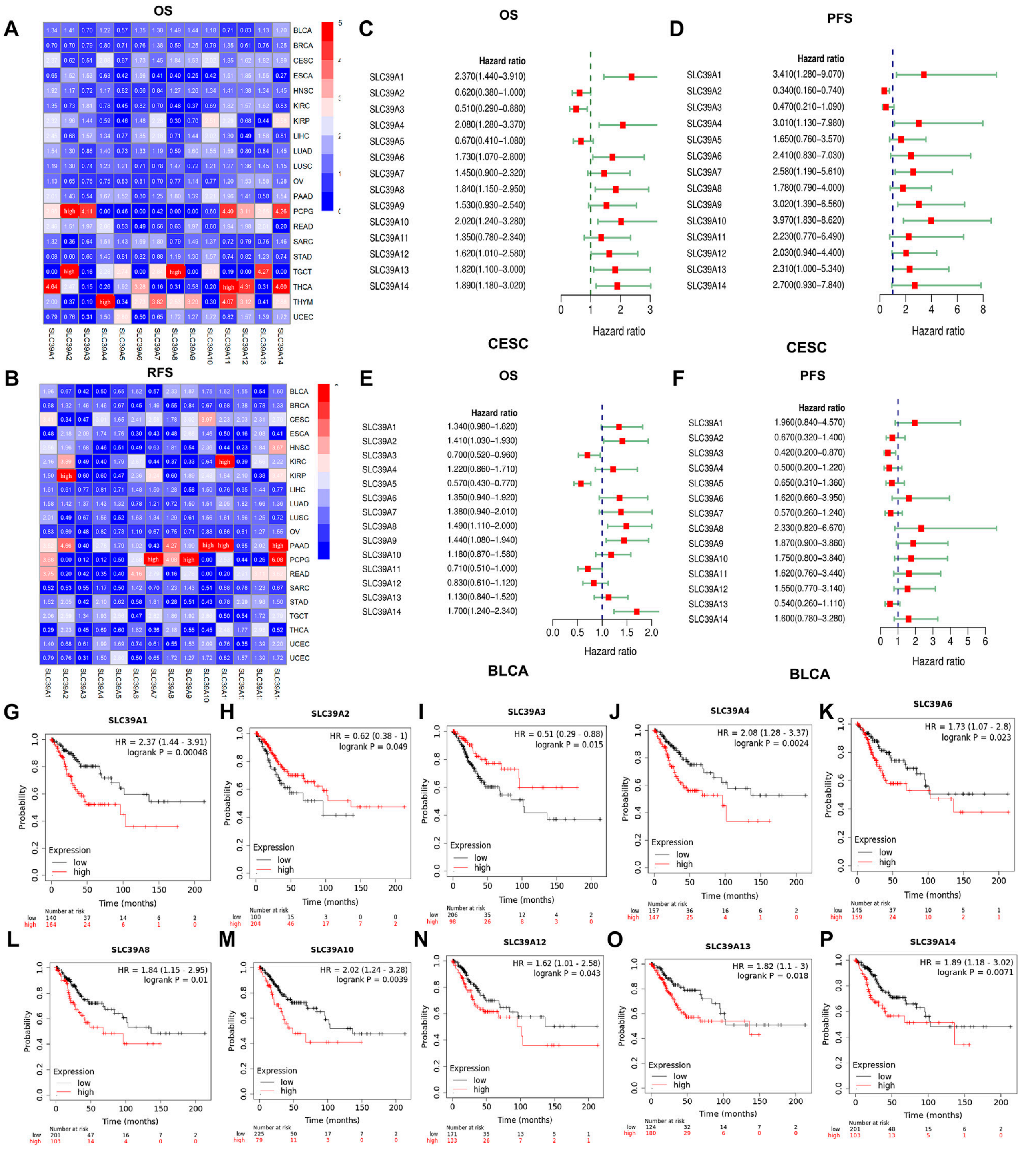


FIGURE 4 | Survival analysis of the prognostic value of the *SLC39A* family genes in pan-cancer derived from the KM-plotter dataset. **(A)** Heat map shows the hazard ratio (HR) value of the OS in pan-cancer calculated using the KM-plotter database. **(B)** Heat map shows the hazard ratio (HR) value of the RFS in pan-cancer calculated using the KM-plotter database. (The colors from blue to red represent the range of HR values in the figure. HRs over 5 were replaced with "high.") **(C)** Forest plot quantitatively synthesizes the HR and 95% confidence interval of the OS of the *SLC39A* gene family in CESC. **(D)** Forest plot quantitatively synthesizes the HR and 95% confidence interval of the PFS of the *SLC39A* gene family in CESC. **(E)** Forest plot quantitatively synthesizes the HR and 95% confidence interval of the OS of the *SLC39A* gene family in BLCA. **(F)** Forest plot quantitatively synthesizes the HR and 95% confidence interval of the PFS of the *SLC39A* gene family in BLCA. **(G)** Survival curve of *SLC39A1* on the OS in CESC. **(H)** Survival curve of *SLC39A2* on the OS in CESC. **(I)** Survival curve of *SLC39A3* on the OS in CESC. **(J)** Survival curve of *SLC39A4* on the OS in CESC. **(K)** Survival curve of *SLC39A6* on the OS in CESC. **(L)** Survival curve of *SLC39A8* on the OS in CESC. **(M)** Survival curve of *SLC39A10* on the OS in CESC. **(N)** Survival curve of *SLC39A12* on the OS in CESC. **(O)** Survival curve of *SLC39A13* on the OS in CESC. **(P)** Survival curve of *SLC39A14* on the OS in CESC. The red line and the blue line represent high and low expression, respectively. The Cox *p*-value cut-off value was set to 0.05.

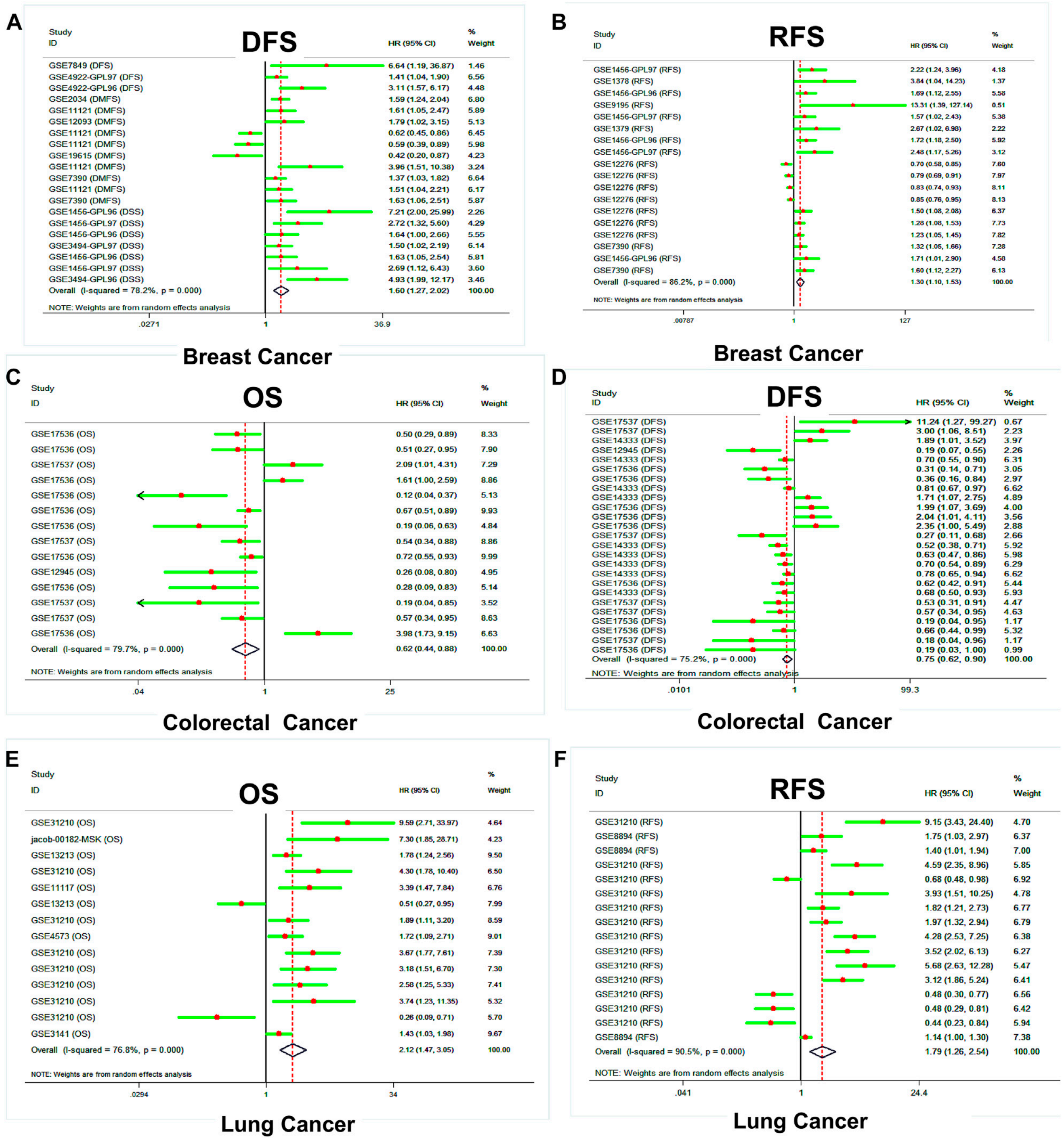


FIGURE 5 | Survival analysis of the prognostic value of the *SLC39A* family genes in pan-cancer derived from the PrognoScan dataset. **(A)** Forest plot of DDS shows the effect of *SLC39A* family gene expression on the clinical prognosis of breast cancer. **(B)** Forest plot of RFS shows the effect of *SLC39A* family gene expression on the clinical prognosis of breast cancer. **(C)** Forest plot of OS shows the effect of *SLC39A* family gene expression on the clinical prognosis of colorectal cancer. **(D)** Forest plot of DFS shows the effect of *SLC39A* family gene expression on the clinical prognosis of colorectal cancer. **(E)** Forest plot of OS shows the effect of *SLC39A* family gene expression on the clinical prognosis of lung cancer. **(F)** Forest plot of DFS shows the effect of *SLC39A* family gene expression on the clinical prognosis of lung cancer.

and neck cancer (**Supplementary Table S4**). Interestingly, most of the studies and data sets were previously focused on breast, colorectal, lung, and ovarian cancer. The results of the quantitative synthesis of related studies showed that higher expression of the *SLC39A* family genes indicated a worse survival prognosis for RFS ([HR] = 1.30, 95% confidence interval [CI] = 1.10 to 1.53) and DSS (HR = 1.60, 95% CI = 1.27 to 2.02) in breast cancer (**Figures 5A,B**). However, the

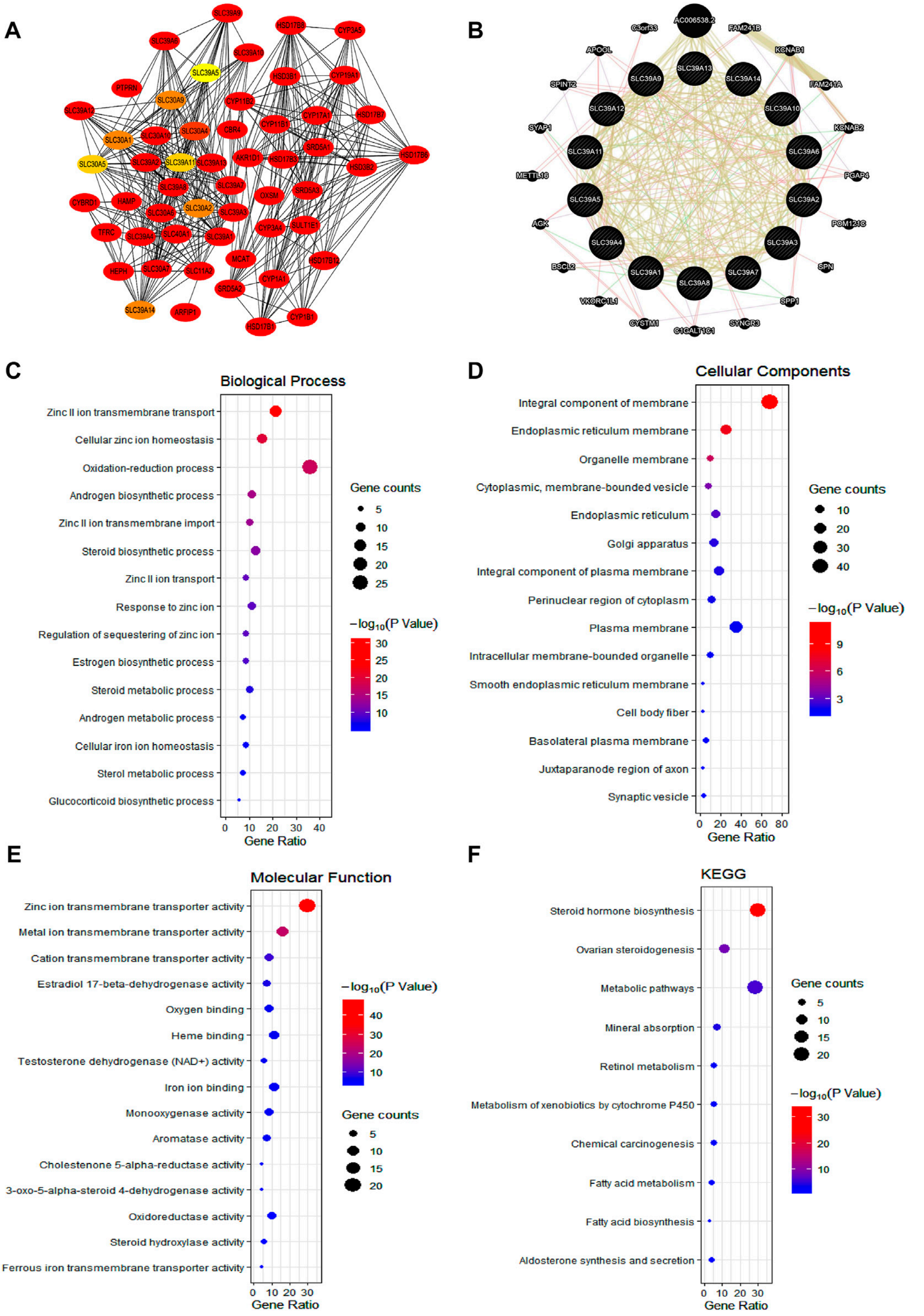


FIGURE 6 | Function enrichment of the *SLC39A* family genes in pan-cancer. **(A)** Protein–protein interaction of *SLC39A* family in the STRING dataset. **(B)** Gene–gene interaction network among *SLC39A* family members in the GeneMANIA dataset. **(C)** Bubble chart showing the BP of *SLC39A* family highly correlated genes. **(D)** Bubble chart showing the CC of *SLC39A* family highly correlated genes. **(E)** Bubble chart showing the MF of *SLC39A* family highly correlated genes. **(H)** Bubble chart showing the KEGG of SLC39A family highly correlated genes.

TABLE 1 | Top 10 GO and KEGG functional enrichment of SLC39A family highly related genes in pan-cancer derived from STRING, GENEMAIN, and DAVID datasets.

Category	GeneSet	Term description	%	*p*-value	FDR
BP	GO:0071577	Zinc II ion transmembrane transport	21.43	2.30E-31	8.15E-29
BP	GO:0006882	Cellular zinc ion homeostasis	15.71	1.13E-20	1.99E-18
BP	GO:0055114	Oxidation-reduction process	35.71	2.84E-18	3.35E-16
BP	GO:0006702	Androgen biosynthetic process	11.43	4.23E-15	3.74E-13
BP	GO:0071578	Zinc II ion transmembrane import	10.00	2.45E-14	1.74E-12
BP	GO:0006694	Steroid biosynthetic process	12.86	4.60E-13	2.72E-11
BP	GO:0006829	Zinc II ion transport	8.57	5.20E-11	2.63E-09
BP	GO:0010043	Response to zinc ion	11.43	9.89E-11	4.38E-09
BP	GO:0061088	Regulation of sequestering of zinc ion	8.57	1.17E-10	4.59E-09
BP	GO:0006703	Estrogen biosynthetic process	8.57	4.25E-10	1.50E-08
CC	GO:0016021	Integral component of membrane	68.57	7.68E-12	7.14E-10
CC	GO:0005789	Endoplasmic reticulum membrane	25.71	1.49E-08	6.93E-07
CC	GO:0031090	Organelle membrane	10.00	9.36E-07	2.90E-05
CC	GO:0016023	Cytoplasmic, membrane-bounded vesicle	8.57	1.76E-04	0.004101
CC	GO:0005783	Endoplasmic reticulum	15.71	0.001061	0.019736
CC	GO:0005794	Golgi apparatus	14.29	0.005088	0.078862
CC	GO:0005887	Integral component of plasma membrane	18.57	0.006431	0.085435
CC	GO:0048471	Perinuclear region of cytoplasm	11.43	0.008955	0.104104
CC	GO:0005886	Plasma membrane	35.71	0.014211	0.146847
CC	GO:0043231	Intracellular membrane-bounded organelle	10.00	0.01889	0.175682
MF	GO:0005385	Zinc ion transmembrane transporter activity	30.00	9.28E-48	1.13E-45
MF	GO:0046873	Metal ion transmembrane transporter activity	15.71	5.26E-24	3.21E-22
MF	GO:0008324	Cation transmembrane transporter activity	8.57	3.84E-10	1.56E-08
MF	GO:0004303	Estradiol 17-beta-dehydrogenase activity	7.14	4.67E-08	1.43E-06
MF	GO:0016712	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	7.14	2.99E-07	7.30E-06
MF	GO:0019825	Oxygen binding	8.57	1.14E-06	2.01E-05
MF	GO:0020037	Heme binding	11.43	1.15E-06	2.01E-05
MF	GO:0047035	Testosterone dehydrogenase (NAD+) activity	5.71	2.07E-06	3.15E-05
MF	GO:0005506	Iron ion binding	11.43	2.41E-06	3.27E-05
MF	GO:0016705	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	8.57	3.02E-06	3.65E-05
KEGG	hsa00140	Steroid hormone biosynthesis	30.00	5.49E-34	1.92E-32
KEGG	hsa04913	Ovarian steroidogenesis	11.43	4.26E-09	7.45E-08
KEGG	hsa01100	Metabolic pathways	28.57	1.85E-06	2.16E-05
KEGG	hsa04978	Mineral absorption	7.14	7.39E-05	6.47E-04
KEGG	hsa00830	Retinol metabolism	5.71	0.004406	0.030844
KEGG	hsa00980	Metabolism of xenobiotics by cytochrome P450	5.71	0.006615	0.038587
KEGG	hsa05204	Chemical carcinogenesis	5.71	0.008206	0.041031
KEGG	hsa01212	Fatty acid metabolism	4.29	0.025834	0.113026
KEGG	hsa00061	Fatty acid biosynthesis	2.86	0.065993	0.233413
KEGG	hsa04925	Aldosterone synthesis and secretion	4.29	0.066689	0.233413

FDR, false discovery rate; GO, gene ontology; BP, biological processes; CC, cellular component; MF, molecular function; KEGG, Kyoto Encyclopedia of Genes and Genomes.

higher expression of the *SLC39A* family genes was associated with a better prognosis for OS (HR = 0.62, 95% CI = 0.44 to 0.88) and DFS (HR = 0.75, 95% CI = 0.62 to 0.90) in colorectal cancer (**Figures 5C,D**). In addition, the upregulation of the *SLC39A* family gene expression was significantly associated with poor OS (HR = 2.12, 95% CI = 1.45 to 3.05) and RFS (HR = 1.79, 95% CI = 1.26 to 2.54) in lung cancer.

The Function Enrichment of the SLC39A Family Genes in Pan-Cancer

To investigate the potential mechanism of the *SLC39A* family genes affecting the prognosis and progression of tumors, the protein–protein interactions (PPIs) of the *SLC39A* family highly related genes were conducted by using the STRING and GeneMANIA platforms. The STRING web was used to conduct the protein–protein interaction (PPI) network analysis of the *SLC39A* family genes. As expected, 54 nodes and 410 edges were obtained in the PPI network, and the 10 top-ranked node genes were *CBR4*, *MCAT*, *CYP17A1*, *HSD17B6*, *HEPH*, *SRD5A1*, *HSD17B3*, *CYP11B2*, *CYP11B1*, and *CYP19A1* (**Figure 6A**; **Supplementary Table S5 and Supplementary Figure S1**). In addition, the GeneMANIA results revealed that a total of 34 genes (including the *SLC39A* family genes) are associated with co-expression, genetic interactions, physical interactions, and shared protein domains. Among them, relationships of co-expression were predicted between *SLC39A1* and *SLC39A13*, *SLC39A1* and *SLC39A7*, *SLC39A4* and *SLC39A14*, *SLC39A5* and *SLC39A14*, *SLC39A5* and *SLC39A4*, *SLC39A6* and *SLC39A10*, and *SLC39A8* and *SLC39A14.* Genetic interactions were predicted between *SLC39A9* and *SLC39A8*, *SLC39A9* and *SLC39A3*, *SLC39A11* and *SLC39A13*, and *SLC39A11* and *SLC39A14*. Moreover, *SLC39A1* and *SLC39A2*, *SLC39A5* and *SLC39A10*, *SLC39A5* and *SLC39A6*, and *SLC39A9* and *SLC39A2* were found to share physical interactions. Most of the *SLC39A* family genes were found to share protein domains (**Figure 6B**).

The functional enrichment of the *SLC39A* family's highly related genes was predicted by analyzing the GO annotation and KEGG pathway *via* the DAVID platform. According to the results (**Figures 6C–F** and **Table 1**), the function of the *SLC39A* highly related genes was enriched in the zinc II ion transmembrane transport, cellular zinc ion homeostasis, oxidation-reduction process, androgen biosynthetic process, and zinc II ion transmembrane import in biological processes (BPs). As for the cellular component (CC), the *SLC39A* family highly related genes were enriched in the integral components of the membrane, endoplasmic reticulum membrane, organelle membrane, cytoplasmic, membrane-bounded vesicle, and endoplasmic reticulum. Moreover, the *SLC39A* family influenced molecular functions through histone methyltransferase binding. With respect to the molecular function (MF), the *SLC39A* family of highly related genes was enriched in the zinc ion transmembrane transporter activity, metal ion transmembrane transporter activity, cation transmembrane transporter activity, estradiol 17-beta-dehydrogenase activity, and oxygen binding. Meanwhile, in the KEGG analysis, 10 pathways were significantly enriched, including the steroid hormone biosynthesis, ovarian steroidogenesis, metabolic pathways, mineral absorption, retinol metabolism, metabolism of xenobiotics by cytochrome P450, chemical carcinogenesis, fatty acid metabolism, fatty acid biosynthesis, and aldosterone synthesis and secretion.

The Immune Infiltration of the SLC39A Family of Genes in Pan-Cancer

To explore whether the *SLC39A* family genes affect the tumor immune infiltrating and microenvironment in pan-cancer, the TIMER 2.0 was used to evaluate the relationship between the *SLC39A* family gene expression levels and the infiltration of six common immune cells, including B cells, $CD4^+$ T cells, $CD8^+$ T cells, Treg T cells, macrophages, and neutrophils. The results confirmed a positive correlation between B-cell infiltration in ACC, KICH, KIRP, LIHC, PCPG, and PRAD but found a negative relation to B-cell immunity in COAD, DLBC, HNSC, KIRC, OV, SKCM, TGCT, TKYM, UCEC, and UVM (**Figure 7A**). For the CD4+ T cell, most of the *SLC39A* family members were positively related to immune infiltration in ESCA, GBM, HNSC, LIHC, and TCGT, while most of the *SLC39A* family genes were negatively correlated with the immune infiltration in BRCA and THYM (**Figure 7B**). For the CD8+ T cell, most of the *SLC39A* family genes were positively correlated with immune infiltration in the ACC, BLCA, DLBC, KICH, PRAD, and UVM, while most of the *SLC39A* family genes were negatively correlated with the immune infiltration in the HNSC, THYM, and UCEC (**Figure 7C**). Besides, a positive correlation was observed in most *SLC39A* family genes and Treg cell infiltration in TGCT and UVM, and a negative correlation was observed in the DLBC and THYM (**Figure 7D**). Particularly, a negative correlation was observed in most SLC39A family genes and macrophage cell infiltration in the BRCA, DLBC, UCEC, and UVM (**Figure 7E**). In addition, a positive correlation was observed in most of the *SLC39A* family genes and neutrophil infiltration (**Figure 7F**). It is noteworthy that the *SLC39A2*, *SLC39A3*, *SLC39A3*, and *SLC39A5* showed a significant negative correlation with the macrophages and neutrophil cell infiltration in most tumor types.

Drug Sensitivity Analysis of the SLC39A Family Genes in Pan-Cancer

To explore the potential sensitization or the effects of drug resistance of the *SLC39A* family genes on the drug response of different human cancer cell lines, a Pearson correlation analysis was performed between the mRNA expression of the *SLC39A* family genes in the NCI-60 cancer cell line and the drug activity of 263 antitumor drugs (**Figure 8**; **Table 2**; and **Supplementary Figure S2**). The results demonstrated the upregulation of *SLC39A1* expression to reduce the drug sensitivity of imisone, oxaliplatin, ifosfamide, eribulin mesylate, palbociclib, and paclitaxel but enhanced the drug sensitivity of Irofulven. An increase in the SLC39A2 expression enhanced the drug sensitivity of Isotretinoin, and the sensitivity of cladribine was found to increase by SLC39A. Various tumor cells with high expression of SLC39A4 were found to be more resistant to the okadaic acid and are more sensitive to 8-chloroadenosine and allopurinol. An increase in the SLC39A5 expression enhanced the drug sensitivity of tegafur, fluorouracil, and BML-277. Notably, the upregulation of SLC39A6 expression was found to increase the sensitivity of raloxifene and fulvestrant. High SLC39A7 expression was found to increase the drug resistance of oxaliplatin, palbociclib, dexrazoxane, entinostat, carfilzomib, epirubicin, and teniposide. However, an elevation in the *SLC39A8* gene expression was found to enhance the drug sensitivity of nelarabine, fluphenazine, chelerythrine, fenretinide, imexon, hydroxyurea, cyclophosphamide, and pipobroman. In addition, the *SLC39A10* gene expression was found to increase the drug sensitivity of gefitinib, afatinib, erlotinib, lapatinib, vandetanib, ibrutinib, and bosutinib and also increased the tolerance of cell lines to elesclomol, paclitaxel, tyrothricin, and vinorelbine. The *SLC39A12* gene expression increased the drug sensitivity of PD-98059, vemurafenib, selumetinib, hypothemycin, and dabrafenib and also increased the tolerance of the cell lines to dasatinib. The expression of SLC39A13 was found to reduce the drug sensitivity of the by-product of CUDC-305, vinorelbine, eribulin mesilate, paclitaxel, oxaliplatin, actinomycin D, nilotinib, homoharringtonine, LDK-378, vinblastine, dolastatin 10, tamoxifen, imatinib, AFP464, tanespimycin, crizotinib, palbociclib, and carfilzomib and enhance the drug sensitivity of simvastatin. At last, the expression of the *SLC39A14* gene was also found to increase the resistance of multiple drugs, including AFP464, panobinostat, cyclophosphamide, palbociclib, lificguat, and fulvestrant. On the other hand, the *SLC39A14* gene expression was found to increase the drug sensitivity of entinostat.

DISCUSSION

Zinc, as an essential trace element, participates in various physiological events, such as growth, differentiation, development, immunity, apoptosis, and other physiological

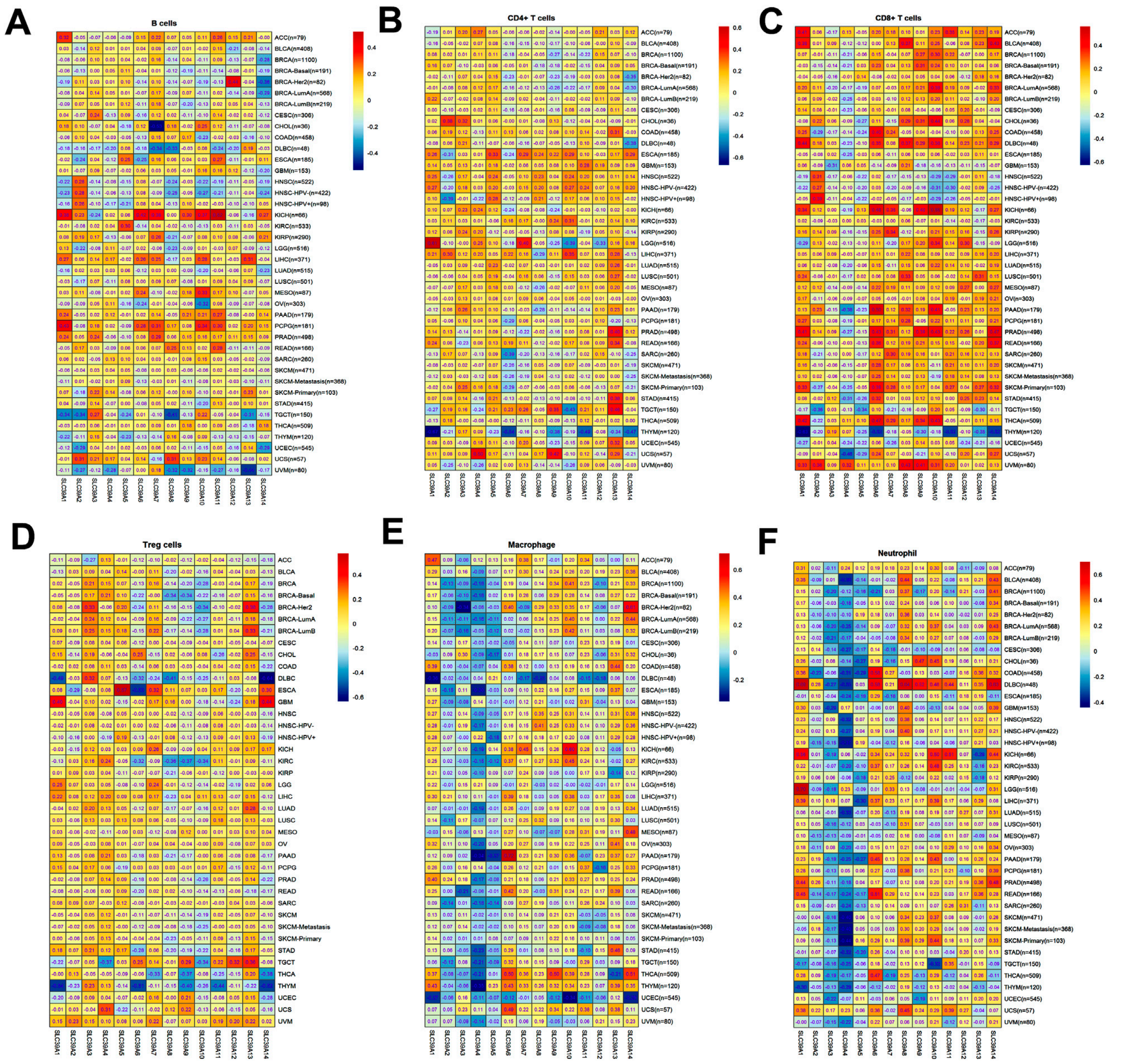


FIGURE 7 | Immune cell infiltration of the *SLC39A* family genes in pan-cancer derived from the TIMER2.0 dataset. **(A)** Correlation coefficient between *SLC39A* family gene expression and B-cell infiltration score in pan-cancer. **(B)** Correlation coefficient between *SLC39A* family gene expression and $CD4^+$ T-cell infiltration score in pan-cancer. **(C)** Correlation coefficient between SLC39A family gene expression and $CD8^+$ T-cell infiltration score in pan-cancer. **(D)** Correlation coefficient between *SLC39A* family gene expression and Treg T-cell infiltration score in pan-cancer. **(E)** Correlation coefficient between *SLC39A* family gene expression and macrophage cell infiltration score in pan-cancer. **(F)** Correlation coefficient between *SLC39A* family gene expression and neutrophil cell infiltration score in pan-cancer. The association was generated with tumor purification adjusted.

processes (Kimura and Kambe, 2016). Previous studies have reported that zinc is required for over 300 enzymes' activity and 2,000 transcription factors to work (Prasad, 2012). Thus, zinc metabolism and homeostasis regulate the normal cell functions in a complex manner. Aberrant Zn transporters have been reported to contribute to specific diseases, including endocrine diseases, neurodegenerative diseases, metabolic diseases, cardiovascular diseases, immune deficiencies, and cancers (Prasad, 2012; Kambe et al., 2014). In particular, current evidence suggests that zinc deficiency and dysregulation of the zinc metabolism are risk factors for tumorigenesis, and zinc is considered a tumor-suppressive agent and a potential tumor treatment target (Grattan and Freake, 2012; Pan et al., 2017; Zhang et al., 2021). In addition, the two groups of zinc transporters, the ZnT transporter (SLC30A) and the ZIP channel (SLC39A), exert strict control over the concentration of zinc in cells, and the SLC39A are known to operate in the influx of zinc across the cytoplasm from the extracellular environment into the cytosol

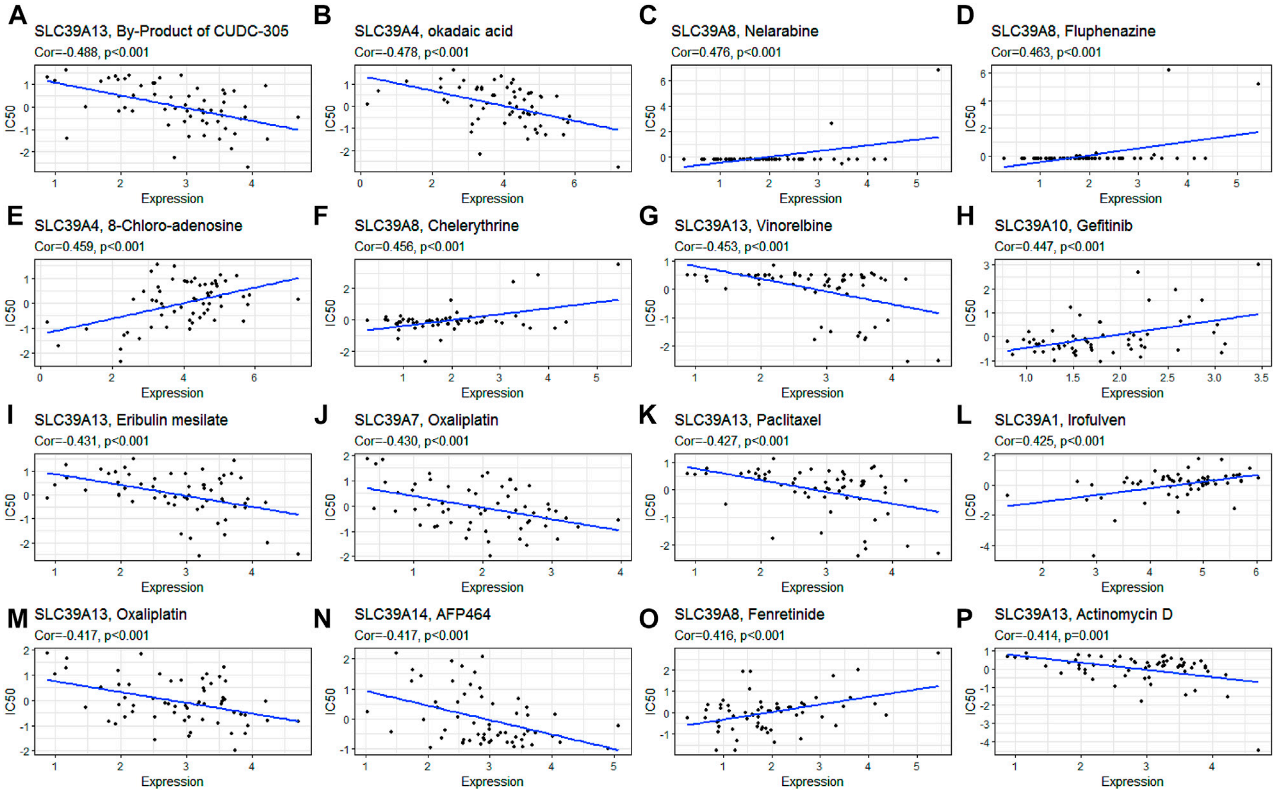


FIGURE 8 | Association of the *SLC39A* family gene expression with the drug sensitivity derived from the NCI-60 cell line data. **(A)** Scatter plot of negative correlation between *SLC39A13* expression and the sensitivity of the by-product of CUDC-305. **(B)** Scatter plot of negative correlation between *SLC39A4* expression and the sensitivity of okadaic acid. **(C)** Scatter plot of positive correlation between *SLC39A8* expression and the sensitivity of nelarabine. **(D)** Scatter plot of positive correlation between *SLC39A8* expression and the sensitivity of fluphenazine. **(E)** Scatter plot of positive correlation between *SLC39A4* expression and the sensitivity of 8-chloro-adenosine. **(F)** Scatter plot of negative correlation between *SLC39A13* expression and the sensitivity of chelerythrine. **(G)** Scatter plot of positive correlation between *SLC39A4* expression and the sensitivity of vinorelbine. **(H)** Scatter plot of positive correlation between *SLC39A10* expression and the sensitivity of gefitinib. **(I)** Scatter plot of negative correlation between *SLC39A13* expression and the sensitivity of eribulin mesilate. **(J)** Scatter plot of negative correlation between *SLC39A7* expression and the sensitivity of oxaliplatin. **(K)** Scatter plot of negative correlation between *SLC39A13* expression and the sensitivity of paclitaxel. **(L)** Scatter plot of positive correlation between *SLC39A1* expression and the sensitivity of Irofulven. **(M)** Scatter plot of negative correlation between *SLC39A13* expression and the sensitivity of oxaliplatin. **(N)** Scatter plot of negative correlation between *SLC39A14* expression and the sensitivity of AFP464. **(O)** Scatter plot of positive correlation between *SLC39A8* expression and the sensitivity of fenretinide. **(P)** Scatter plot of negative correlation between *SLC39A13* expression and the sensitivity of actinomycin D. Z-score from test by Pearson's correlation using NCI-60 cell line data.

(Hojyo and Fukada, 2016; Pan et al., 2017; Brito et al., 2020). The available data suggest that the mutation or functional change of the *SLC39A* family of genes develops various diseases such as the tumors of the digestive system, urinary system, and reproductive tract (Prasad, 2012; Hoang et al., 2016; To et al., 2020). Interestingly, the SLC39A family gene knockout animals have revealed many unique phenotypes and the possibility of the clinical targeted application and the possibility of discovery and development of the SLC39A family inhibitors as anticancer drugs and modulators regulating the sensitivity or resistance of chemotherapeutic drugs (Geng et al., 2018; Hu, 2020; Mohammadinejad et al., 2020; Cheng et al., 2021). However, the precise function of the *SLC39A* family genes in pan-cancer has not been comprehensively determined.

In the current study, compared to the normal control group, the expression levels of most of the *SLC39* family genes in the tumor tissues were found to be significantly upregulated. The high expression of SLC39A6 is a dependable marker for breast cancer (luminal A subtype), and elevated *SLC39A10* mRNA levels were evident in the cancer cell lines of the highly aggressive breast cancer (Kagara et al., 2007; Hogstrand et al., 2013). Cheng et al. (2017) have found the ESCC tissues to possess an increased mRNA expression level of *SLC39A6* compared to the non-tumor tissues. Studies by Li et al. (2007) reported that compared to the human pancreatic ductal epithelium (HPDE) cells, the expression of the *SLC39A4* mRNA was significantly increased in the human pancreatic cancer cells. In addition, a bioinformatics study also found increased expression levels of the *SLC39A* family genes with significant upregulation in the breast cancer, gastric cancer, and lung cancer tissues compared to the normal breast tissues (Ding et al., 2019; Liu et al., 2020; Zhou et al., 2021). There are not many studies on the genetic mutations of the SLC39A family in tumor tissues. Our study reported that the mutation frequencies of the *SLC39A4* and *SLC39A1* were the highest among all the

TABLE 2 | Relationship between SLC39A family gene expression and drug sensitivity based on the NCI-60 cell line.

Gene	Drug	cor	*p*-value	NSC#	PubChem SID
SLC39A13	By-product of CUDC-305	–0.488	7.51E-05	761390	-
SLC39A4	Okadaic acid	–0.478	0.000113	677083	516878
SLC39A8	Nelarabine	0.476	0.000122	755985	144074932
SLC39A8	Fluphenazine	0.463	0.000194	92339	398387
SLC39A4	8-chloro-adenosine	0.459	0.000226	354258	464177
SLC39A8	Chelerythrine	0.456	0.000254	36405	544923
SLC39A13	Vinorelbine	–0.453	0.000274	760087	144076280
SLC39A10	Gefitinib	0.447	0.000338	759856	144076186
SLC39A13	Eribulin mesilate	–0.431	0.000579	707389	529374
SLC39A7	Oxaliplatin	–0.430	0.000602	266046	569872
SLC39A13	Paclitaxel	–0.427	0.000661	758645	144075668
SLC39A1	Irofulven	0.425	0.000704	683863	520035
SLC39A13	Oxaliplatin	–0.417	0.00091	266046	569872
SLC39A14	AFP464	–0.417	0.000911	710464	530822
SLC39A8	Fenretinide	0.416	0.000953	760419	144076412
SLC39A13	Actinomycin D	–0.414	0.001015	755841	144074852
SLC39A10	Afatinib	0.413	0.00103	750691	131407778
SLC39A13	Nilotinib	–0.411	0.001111	747599	91148446
SLC39A1	Imexon	–0.411	0.001115	714597	532526
SLC39A6	Raloxifene	0.411	0.00112	747974	91148450
SLC39A14	Panobinostat	–0.409	0.001172	761190	-
SLC39A13	Homoharringtonine	–0.407	0.001261	758253	-
SLC39A13	LDK-378	–0.405	0.00134	777193	-
SLC39A13	Vinblastine	–0.403	0.001413	757384	144075282
SLC39A4	Allopurinol	0.403	0.001424	1,390	68199
SLC39A10	Erlotinib	0.397	0.001684	718781	534851
SLC39A1	Oxaliplatin	–0.394	0.001825	266046	569872
SLC39A10	Lapatinib	0.393	0.00192	745750	91147938
SLC39A8	Imexon	0.391	0.001991	714597	532526
SLC39A1	Ifosfamide	–0.390	0.002069	109724	301170
SLC39A7	Palbociclib	–0.389	0.002157	758247	-
SLC39A12	PD-98059	0.381	0.002652	679828	518213
SLC39A13	Dolastatin 10	–0.380	0.002743	376128	469333
SLC39A13	Tamoxifen	–0.378	0.002902	180973	447264
SLC39A10	Vandetanib	0.377	0.002996	760766	131408693
SLC39A12	Vemurafenib	0.375	0.003169	761431	131408691
SLC39A6	Fulvestrant	0.374	0.003201	719276	534986
SLC39A10	Elesclomol	–0.372	0.003428	174939	445356
SLC39A10	Ibrutinib	0.370	0.003626	761910	-
SLC39A10	Bosutinib	0.370	0.003646	765694	-
SLC39A12	Selumetinib	0.366	0.003978	741078	91146061
SLC39A7	Dexrazoxane	–0.366	0.004008	169780	442425
SLC39A13	Imatinib	–0.366	0.004011	743414	91146949
SLC39A5	Tegafur	0.366	0.004025	148958	430704
SLC39A13	AFP464	–0.365	0.004084	710464	530822
SLC39A13	Tanespimycin	–0.364	0.004255	330507	574817
SLC39A10	Paclitaxel	–0.361	0.004571	758645	144075668
SLC39A1	Eribulin mesilate	–0.358	0.004949	707389	529374
SLC39A13	Crizotinib	–0.358	0.005009	756645	131408690
SLC39A5	Fluorouracil	0.357	0.005048	757036	144075048
SLC39A13	Palbociclib	–0.357	0.005157	758247	-
SLC39A8	Hydroxyurea	0.355	0.00534	32065	90752
SLC39A1	Palbociclib	–0.354	0.005531	758247	-
SLC39A14	Cyclophosphamide	–0.354	0.005552	26271	87150
SLC39A10	Tyrothricin	–0.354	0.005564	757363	144075261
SLC39A5	BML-277	0.353	0.005734	741899	91146360
SLC39A7	Entinostat	–0.351	0.00593	706995	529250
SLC39A8	Cyclophosphamide	0.351	0.006	26271	87150
SLC39A10	Vinorelbine	–0.348	0.006518	760087	144076280
SLC39A13	Carfilzomib	–0.347	0.006546	758252	-
SLC39A14	Palbociclib	–0.345	0.006897	758247	-
SLC39A13	Simvastatin	0.345	0.006942	758706	144075729
SLC39A14	Lificguat	–0.344	0.007171	728165	48427734
SLC39A12	Hypothemycin	0.343	0.007328	354462	576295

(Continued on following page)

TABLE 2 | (*Continued*) Relationship between SLC39A family gene expression and drug sensitivity based on the NCI-60 cell line.

Gene	Drug	cor	*p*-value	NSC#	PubChem SID
SLC39A7	Carfilzomib	−0.342	0.007394	758252	-
SLC39A12	Dabrafenib	0.340	0.007788	764134	-
SLC39A3	Cladribine	0.339	0.008008	105014	405818
SLC39A14	Fulvestrant	−0.339	0.008077	719276	534986
SLC39A2	Isotretinoin	0.338	0.008346	122758	416403
SLC39A14	Zoledronate	0.336	0.008774	721517	536160
SLC39A1	Paclitaxel	−0.333	0.009309	758645	144075668
SLC39A7	Epirubicin	−0.333	0.009424	759195	144075933
SLC39A14	Entinostat	−0.332	0.009669	706995	529250
SLC39A7	Teniposide	−0.331	0.009681	758667	144075690
SLC39A12	Dasatinib	−0.331	0.009702	759877	144076207
SLC39A8	Pipobroman	0.331	0.009745	25154	86412

cor, correlation coefficient; NSC#, Cancer Chemotherapy National Service Center number; PubChem SID, PubChem Substance IDs.

members and were 6 and 4%, respectively. The overall mutation frequency of the *SLC39A* family genes was in the general range of 0.8–6.0% in pan-cancer. Moreover, the mutations in the *SLC39A* family genes were found to have a significant impact on the DSS, PFS, and DFS of the malignant tumor.

For prognostic analysis, the expression of SLC39A family genes was found to be significantly related to the OS and RFS in multiple types using the KM-plotter database. Most *SLC39A* family genes showed protective effects in BLCA, CESC, HNSC, BRCA, ESCA, and OV. Simultaneously, the PrognoScan results indicated that the *SLC39A* family gene expression levels were significantly correlated with the prognosis of the colorectal, breast, bladder, lung, ovarian, blood, brain, skin, eye, soft tissue, prostate, and head and neck cancers. Our findings were consistent with those of the previous study, in that high mRNA expression levels of *SLC39A6* and *SLC39A14* indicated favorable OS, but upregulated *SLC39A2-5*, *SLC39A7*, and *SLC39A12-13* were associated with poor OS in the patients with breast carcinoma (Liu et al., 2020). Higher expression of SLC39A1, 5–7, and 9 indicated better OS, FPS, and PPS, and increased SLC39A2–4, 8, and SLC39A10 expression indicated poor OS, FP, and PPS in the patients with gastric cancer (Ding et al., 2019). In addition, an increase in the SLC39A7 expression was related to better OS, while the upregulated level of SLC39A3 and SLC39A4 were associated with inferior OS in patients with LUSC (Zhou et al., 2021). Consistent with previous research, the GO function enrichment indicated the SLC39A family genes and their highly related genes to contribute to zinc transport– and homeostasis-related biological processes, such as zinc II ion transmembrane transport, cellular zinc ion homeostasis, oxidation-reduction, androgen biosynthesis, and zinc II ion transmembrane import. The KEGG analysis showed *SLC39A* family genes to be involved in the hormone regulation, metabolic pathways, mineral absorption, and chemical carcinogenesis pathways (Ding et al., 2019; Zhou et al., 2021). Therefore, the prognostic effect of the SLC39A family genes can be speculated to be closely related to zinc transfer, metabolism, and function (Guo and He, 2020).

Increasing studies have shown that zinc is involved in a variety of important functions of immune cell activation and initiation of immune response in the process of innate immunity and adaptive immunity; thus, zinc deficiency can lead to immune dysfunction (Bin et al., 2018). Given the close relationship between the SLC39A family and zinc transport, the regulatory relationship between the *SLC39A* family genes and immune infiltration deserves more attention. SLC39A6 and SLC39A10 are the first zinc transporters reported to regulate immune cell functions in mammals. Subsequent studies have confirmed the role of SLC39A8 in regulating various immune cells and playing an irreplaceable role in the process of innate immunity (Kitamura et al., 2006; Liu et al., 2013). The study of Hojyo et al. (2014) reported the SLC39A10 expression to be upregulated in pro-B lymphocytes and SLC39A10 to participate in B-cell immunity by leading the homeostasis and the function of the B cells. Our findings suggested a significant correlation between the *SLC39A* family gene expression and B-cell infiltration in broad cancer types and CD4$^+$ T cells, CD8$^+$ T cells, Treg T cells, macrophages, and neutrophils in specific tumors. These results provide new possibilities for immunotherapy to improve the prognosis by modulating the *SLC39A* family genes on the tumors or immune cells. Presently, there have been a few attempts of transformational application research using the SLC39A family to treat or alleviate diseases in animals or cell models. One study showed that in an *in vivo* xenograft model, the overexpression of the SLC39A1 leads to an increased zinc uptake, reducing the tumor growth (Golovine et al., 2008). Utilizing the characteristics of the zinc SLC39A6 transporter widely expressed in all breast cancer subtypes, Seattle Genetics has designed and constructed a new antibody–drug conjugate called SGN-LIV1A to treat metastatic breast cancer through the targeted regulation of SLC39A6 (Sun et al., 2011). In our study, the Pearson correlation analysis was performed between the mRNA expression of the *SLC39A* family genes in the NCI-60 cancer cell line and the drug activity of 263 antitumor drugs. The results showed that SLC39A family genes are significantly related to the sensitivity or resistance of 63 antitumor drugs in a variety of tumor cell lines. Among them, SLC39A13 and the by-product of the CUDC-305, SLC39A4, okadaic acid, SLC39A8, and nelarabine are the three most likely connections. Based on these results, the detection and targeted regulation of the expression of the SLC39A family gene have been found to have special potential value for the clinical selection of antitumor drugs.

Despite being the first one to perform a multidimensional and multi-omics analysis of the *SLC39A* family genes in pan-cancer, this study has some shortcomings worth considering. First, the bioinformatics analysis was carried out through multiple online big data databases, and further *in vitro* and *in vivo* experiments are required to verify the prediction results. Second, multiple databases are not completely consistent in the expression and survival prognosis of the *SLC39A* family genes in certain tumors. Large samples, different populations, and multicenter clinical studies need further clarification. Third, although we have confirmed that the *SLC39A* family gene expression was significantly related to the immune infiltration and survival outcome of a variety of tumors, the causal relationship between the immune infiltration and prognosis remains elusive. Fourth, analyzing the level of immune infiltrating cells at the tumor tissue level may be error-prone, and hence, single-cell sequencing may be required for further exploration.

CONCLUSION

This pan-cancer study performed a comprehensive and systematic investigation of the expression patterns, genetic mutation, prognostic value, function enrichment, immune infiltrating, and potential therapeutic targets of the *SLC39A* family of genes. Our results proved that most of the *SLC39* family genes' expression was significantly increased in the tumor tissues and was associated with clinical prognosis in pan-cancer. Moreover, the *SLC39A* family gene expression was significantly related to the immune cell infiltration levels of six types of immune cells and contributed to the sensitivity or resistance of the drugs in specific types of tumors. Thus, we concluded that the *SLC39A* of family genes may be crucial for tumorigenesis, the tumor microenvironment, and drug sensitivity, providing novel ideas to develop new targeted therapy for malignant tumors.

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

QZ conceptualized and designed this study. Y-YQ, R-YG, and M-LL performed the data collection and analysis. YYQ and M-LL wrote the manuscript. Y-YQ and QZ participated in manuscript revision. 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.2021.757582/full#supplementary-material

REFERENCES

Baltaci, A. K., and Yuce, K. (2018). Zinc Transporter Proteins. *Neurochem. Res.* 43 (3), 517–530. doi:10.1007/s11064-017-2454-y

Bin, B. H., Seo, J., and Kim, S. T. (2018). Function, Structure, and Transport Aspects of ZIP and ZnT Zinc Transporters in Immune Cells. *J. Immunol. Res.* 2018, 9365747. doi:10.1155/2018/9365747

Brito, S., Lee, M. G., Bin, B. H., and Lee, J. S. (2020). Zinc and its Transporters in Epigenetics. *Mol. Cell* 43 (4), 323–330. doi:10.14348/molcells.2020.0026

Chen, W., Sun, Z., and Lu, L. (2021). Targeted Engineering of Medicinal Chemistry for Cancer Therapy: Recent Advances and Perspectives. *Angew. Chem. Int. Ed.* 60 (11), 5626–5643. doi:10.1002/anie.201914511

Cheng, X., Wang, J., Liu, C., Jiang, T., Yang, N., Liu, D., et al. (2021). Zinc Transporter SLC39A13/ZIP13 Facilitates the Metastasis of Human Ovarian Cancer Cells via Activating Src/FAK Signaling Pathway. *J. Exp. Clin. Cancer Res.* 40 (1), 199. doi:10.1186/s13046-021-01999-3

Cheng, X., Wei, L., Huang, X., Zheng, J., Shao, M., Feng, T., et al. (2017). Solute Carrier Family 39 Member 6 Gene Promotes Aggressiveness of Esophageal Carcinoma Cells by Increasing Intracellular Levels of Zinc, Activating Phosphatidylinositol 3-Kinase Signaling, and Up-Regulating Genes that Regulate Metastasis. *Gastroenterology* 152 (8), 1985–1997. doi:10.1053/j.gastro.2017.02.006

Ding, B., Lou, W., Xu, L., Li, R., and Fan, W. (2019). Analysis the Prognostic Values of Solute Carrier (SLC) Family 39 Genes in Gastric Cancer. *Am. J. Transl Res.* 11 (1), 486–498.

Fan, Q., Cai, Q., Li, P., Wang, W., Wang, J., Gerry, E., et al. (2017). The Novel ZIP4 Regulation and its Role in Ovarian Cancer. *Oncotarget* 8 (52), 90090–90107. doi:10.18632/oncotarget.21435

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., et al. (2021). Cancer Statistics for the Year 2020: An Overview. *Int. J. Cancer*. doi:10.1002/ijc.33588

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 (269), pl1. doi:10.1126/scisignal.2004088

Geng, X., Liu, L., Banes-Berceli, A., Yang, Z., Kang, P., Shen, J., et al. (2018). Role of ZIP8 in Regulating Cell Morphology and NF-κB/Snail2 Signaling. *Metallomics* 10 (7), 953–964. doi:10.1039/c8mt00079d

Golovine, K., Makhov, P., Uzzo, R. G., Shaw, T., Kunkle, D., and Kolenko, V. M. (2008). Overexpression of the Zinc Uptake Transporter hZIP1 Inhibits Nuclear Factor-Kb and Reduces the Malignant Potential of Prostate Cancer Cells *In Vitro* and *In Vivo*. *Clin. Cancer Res.* 14 (17), 5376–5384. doi:10.1158/1078-0432.ccr-08-0455

Grattan, B. J., and Freake, H. C. (2012). Zinc and Cancer: Implications for LIV-1 in Breast Cancer. *Nutrients* 4 (7), 648–675. doi:10.3390/nu4070648

Guo, Y., and He, Y. (2020). Comprehensive Analysis of the Expression of SLC30A Family Genes and Prognosis in Human Gastric Cancer. *Sci. Rep.* 10 (1), 18352. doi:10.1038/s41598-020-75012-w

Hoang, B. X., Han, B., Shaw, D. G., and Nimni, M. (2016). Zinc as a Possible Preventive and Therapeutic Agent in Pancreatic, Prostate, and Breast Cancer. *Eur. J. Cancer Prev.* 25 (5), 457–461. doi:10.1097/cej.0000000000000194

Hogstrand, C., Kille, P., Ackland, M. L., Hiscox, S., and Taylor, K. M. (2013). A Mechanism for Epithelial-Mesenchymal Transition and Anoikis Resistance in Breast Cancer Triggered by Zinc Channel ZIP6 and STAT3 (Signal Transducer and Activator of Transcription 3). *Biochem. J.* 455 (2), 229–237. doi:10.1042/bj20130483

Hojyo, S., and Fukada, T. (2016). Zinc Transporters and Signaling in Physiology and Pathogenesis. *Arch. Biochem. Biophys.* 611, 43–50. doi:10.1016/j.abb.2016.06.020

Hojyo, S., Miyai, T., Fujishiro, H., Kawamura, M., Yasuda, T., Hijikata, A., et al. (2014). Zinc Transporter SLC39A10/ZIP10 Controls Humoral Immunity by Modulating B-Cell Receptor Signal Strength. *Proc. Natl. Acad. Sci.* 111 (32), 11786–11791. doi:10.1073/pnas.1323557111

Hu, J. (2020). Toward Unzipping the ZIP Metal Transporters: Structure, Evolution, and Implications on Drug Discovery against Cancer. *FEBS J.* 288, 5805–5825. doi:10.1111/febs.15658

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211

Jin, J., Li, Z., Liu, J., Wu, Y., Gao, X., and He, Y. (2015). Knockdown of Zinc Transporter ZIP5 (SLC39A5) Expression Significantly Inhibits Human Esophageal Cancer Progression. *Oncol. Rep.* 34 (3), 1431–1439. doi:10.3892/or.2015.4097

Kagara, N., Tanaka, N., Noguchi, S., and Hirano, T. (2007). Zinc and its Transporter ZIP10 Are Involved in Invasive Behavior of Breast Cancer Cells. *Cancer Sci.* 98 (5), 692–697. doi:10.1111/j.1349-7006.2007.00446.x

Kambe, T., Hashimoto, A., and Fujimoto, S. (2014). Current Understanding of ZIP and ZnT Zinc Transporters in Human Health and Diseases. *Cell. Mol. Life Sci.* 71 (17), 3281–3295. doi:10.1007/s00018-014-1617-0

Kimura, T., and Kambe, T. (2016). The Functions of Metallothionein and ZIP and ZnT Transporters: An Overview and Perspective. *Ijms* 17 (3), 336. doi:10.3390/ijms17030336

Kitamura, H., Morikawa, H., Kamon, H., Iguchi, M., Hojyo, S., Fukada, T., et al. (2006). Toll-like Receptor-Mediated Regulation of Zinc Homeostasis Influences Dendritic Cell Function. *Nat. Immunol.* 7 (9), 971–977. doi:10.1038/ni1373

Li, M., Zhang, Y., Liu, Z., Bharadwaj, U., Wang, H., Wang, X., et al. (2007). Aberrant Expression of Zinc Transporter ZIP4 (SLC39A4) Significantly Contributes to Human Pancreatic Cancer Pathogenesis and Progression. *Proc. Natl. Acad. Sci.* 104 (47), 18636–18641. doi:10.1073/pnas.0709307104

Li, Q., and Shu, Y. (2014). Role of Solute Carriers in Response to Anticancer Drugs. *Mol. Cell Therapies* 2, 15. doi:10.1186/2052-8426-2-15

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 (21), e108–e110. doi:10.1158/0008-5472.can-17-0307

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, L., Yang, J., and Wang, C. (2020). Analysis of the Prognostic Significance of Solute Carrier (SLC) Family 39 Genes in Breast Cancer. *Biosci. Rep.* 40 (8). doi:10.1042/BSR20200764

Liu, M.-J., Bao, S., Gálvez-Peralta, M., Pyle, C. J., Rudawsky, A. C., Pavlovicz, R. E., et al. (2013). ZIP8 Regulates Host Defense through Zinc-Mediated Inhibition of NF-Kb. *Cell Rep.* 3 (2), 386–400. doi:10.1016/j.celrep.2013.01.009

Loomans-Kropp, H. A., and Umar, A. (2019). Cancer Prevention and Screening: the Next Step in the Era of Precision Medicine. *npj Precision Onc* 3, 3. doi:10.1038/s41698-018-0075-9

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. Genomics* 2, 18. doi:10.1186/1755-8794-2-18

Mohammadinejad, R., Sassan, H., Pardakhty, A., Hashemabadi, M., Ashrafizadeh, M., Dehshahri, A., et al. (2020). ZEB1 and ZEB2 Gene Editing Mediated by CRISPR/Cas9 in A549 Cell Line. *Bratisl Lek Listy* 121 (1), 31–36. doi:10.4149/BLL_2020_005

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

Pan, Z., Choi, S., Ouadid-Ahidouch, H., Yang, J. M., Beattie, J. H., and Korichneva, I. (2017). Zinc Transporters and Dysregulated Channels in Cancers. *Front. Biosci.* 22, 623–643. doi:10.2741/4507

Prasad, A. S. (2012). Discovery of Human Zinc Deficiency: 50 Years Later. *J. Trace Elem. Med. Biol.* 26 (2-3), 66–69. doi:10.1016/j.jtemb.2012.04.004

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

Shankavaram, U. T., Varma, S., Kane, D., Sunshine, M., Chary, K. K., Reinhold, W. C., et al. (2009). CellMiner: a Relational Database and Query Tool for the NCI-60 Cancer Cell Lines. *BMC genomics* 10, 277. doi:10.1186/1471-2164-10-277

Sun, J., Liu, J., Pan, X., Quimby, D., Zanesi, N., Druck, T., et al. (2011). Effect of Zinc Supplementation on N-Nitrosomethylbenzylamine-Induced Forestomach Tumor Development and Progression in Tumor Suppressor-Deficient Mouse Strains. *Carcinogenesis* 32 (3), 351–358. doi:10.1093/carcin/bgq251

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

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, 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 (W1), W98–W102. doi:10.1093/nar/gkx247

To, P. K., Do, M. H., Cho, J. H., and Jung, C. (2020). Growth Modulatory Role of Zinc in Prostate Cancer and Application to Cancer Therapeutics. *Int. J. Mol. Sci.* 21 (8). doi:10.3390/ijms21082991

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function. *Nucleic Acids Res.* 38, W214–W220. doi:10.1093/nar/gkq537

Xiao, Q., Zhang, F., Xu, L., Yue, L., Kon, O. L., Zhu, Y., et al. (2021). High-throughput Proteomics and AI for Cancer Biomarker Discovery. *Adv. Drug Deliv. Rev.* 176, 113844. doi:10.1016/j.addr.2021.113844

Zhang, Y., Tian, Y., Zhang, H., Xu, B., and Chen, H. (2021). Potential Pathways of Zinc Deficiency-Promoted Tumorigenesis. *Biomed. Pharmacother.* 133, 110983. doi:10.1016/j.biopha.2020.110983

Zhou, H., Zhu, Y., Qi, H., Liang, L., Wu, H., Yuan, J., et al. (2021). Evaluation of the Prognostic Values of Solute Carrier (SLC) Family 39 Genes for Patients with Lung Adenocarcinoma. *Aging* 13 (4), 5312–5331. doi:10.18632/aging.202452

Zhu, B., Huo, R., Zhi, Q., Zhan, M., Chen, X., and Hua, Z.-C. (2021). Increased Expression of Zinc Transporter ZIP4, ZIP11, ZnT1, and ZnT6 Predicts Poor Prognosis in Pancreatic Cancer. *J. Trace Elem. Med. Biol.* 65, 126734. doi:10.1016/j.jtemb.2021.126734

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: 17 December 2021
doi: 10.3389/fgene.2021.778715

A Novel Model of Tumor-Infiltrating B Lymphocyte Specific RNA-Binding Protein-Related Genes With Potential Prognostic Value and Therapeutic Targets in Multiple Myeloma

JingJing Zhang[1,2], *Pengcheng He*[1], *Xiaoning Wang*[1], *Suhua Wei*[1], *Le Ma*[1] *and Jing Zhao*[1]*

[1]*Department of Hematology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China,* [2]*The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China*

OPEN ACCESS

Edited by:
Rodrigo Drummond,
A. C. Camargo Cancer Center, Brazil

Reviewed by:
Andrea Grioni,
Novartis Institutes for BioMedical Research, Switzerland
Abhinita S. Mohanty,
Memorial Sloan Kettering Cancer Center, United States

****Correspondence:***
Jing Zhao
vicki0922@163.com

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

Received: *17 September 2021*
Accepted: *16 November 2021*
Published: *17 December 2021*

Citation:
Zhang J, He P, Wang X, Wei S, Ma L and Zhao J (2021) A Novel Model of Tumor-Infiltrating B Lymphocyte Specific RNA-Binding Protein-Related Genes With Potential Prognostic Value and Therapeutic Targets in Multiple Myeloma.
Front. Genet. 12:778715.
doi: 10.3389/fgene.2021.778715

Background: RNA-binding proteins (RBPs) act as important regulators in the progression of tumors. However, their role in the tumorigenesis and prognostic assessment in multiple myeloma (MM), a B-cell hematological cancer, remains elusive. Thus, the current study was designed to explore a novel prognostic B-cell-specific RBP signature and the underlying molecular mechanisms.

Methods: Data used in the current study were obtained from the Gene Expression Omnibus (GEO) database. Significantly upregulated RBPs in B cells were defined as B cell-specific RBPs. The biological functions of B-cell-specific RBPs were analyzed by the cluster Profiler package. Univariate and multivariate regressions were performed to identify robust prognostic B-cell specific RBP signatures, followed by the construction of the risk classification model. Gene set enrichment analysis (GSEA)-identified pathways were enriched in stratified groups. The microenvironment of the low- and high-risk groups was analyzed by single-sample GSEA (ssGSEA). Moreover, the correlations among the risk score and differentially expressed immune checkpoints or differentially distributed immune cells were calculated. The drug sensitivity of the low- and high-risk groups was assessed via Genomics of Drug Sensitivity in Cancer by the pRRophetic algorithm. In addition, we utilized a GEO dataset involving patients with MM receiving bortezomib therapy to estimate the treatment response between different groups.

Results: A total of 56 B-cell-specific RBPs were identified, which were mainly enriched in ribonucleoprotein complex biogenesis and the ribosome pathway. ADAR, FASTKD1 and SNRPD3 were identified as prognostic B-cell specific RBP signatures in MM. The risk model was constructed based on ADAR, FASTKD1 and SNRPD3. Receiver operating characteristic (ROC) curves revealed the good predictive capacity of the risk model. A nomogram based on the risk score and other independent prognostic factors exhibited excellent performance in predicting the overall survival of MM patients. GSEA showed enrichment of the Notch signaling pathway and mRNA cis-splicing via spliceosomes in the high-risk group. Moreover, we found that the infiltration of diverse immune cell subtypes and the expression of CD274, CD276, CTLA4 and VTCN1 were significantly different between the two groups. In addition, the IC50 values of 11 drugs were higher in the low-risk group. Patients in the low-risk group exhibited a higher complete response rate to bortezomib therapy.

Conclusion: Our study identified novel prognostic B-cell-specific RBP biomarkers in MM and constructed a unique risk model for predicting MM outcomes. Moreover, we explored the immune-related mechanisms of B cell-specific RBPs in regulating MM. Our findings could pave the way for developing novel therapeutic strategies to improve the prognosis of MM patients.

Keywords: multiple myeloma, tumor-infiltrating B lymphocyte, RNA-binding protein, prognostic signature, immune-related signature, immunotherapy

1 INTRODUCTION

Multiple myeloma (MM) is a B-cell hematological malignancy. The proliferation of plasma cells further induces end organ dysfunctions, including anemia, hypercalcemia, bone lesions and renal failure (Palumbo and Anderson, 2011). The incidence rate of MM has rapidly increased by 126% globally over the past 2decades (Cowan et al., 2018). The rapidly increasing incidence rate has underscored the urgent need for treatment improvement. Although the overall survival of multiple myeloma has been rapidly improved by the widespread application of stem cell transplantation and novel drugs represented by proteasome inhibitors and immunomodulatory drugs (Attal et al., 2017; Facon et al., 2019; Mikkilineni and Kochenderfer, 2021), MM remains incurable. The highly heterogeneous clinical outcomes of MM patients depend on the tumor burden, tumor cell characteristics, and especially genetic abnormalities. Currently, a risk classification model based on more detailed genetic and molecular information was created by the International Multiple Myeloma Working Group in 2015 (Palumbo et al., 2015). This staging system is widely used in clinical practice. Approximately 75% of patients who present without cytogenetic abnormalities are considered as low risk. These patients present heterogenous clinical outcomes (Binder et al., 2017). There remains a group of patients who are divided into low-risk groups characterized by therapy resistance, rapid refractory periods and short overall survival. Meanwhile, existing classification model fail to identify some of patients with 1q21 amplification and del 17p for very poor outcome. However, no attempts have been made to further sub-stratify such amount of patients (Walker et al., 2018). In light of the limitations of the current staging system, it is necessary to identify novel biomarkers and establish a prognostic model based on cytogenetic characterization to distinguish good prognosis from poor prognosis patients, thereby improving patients' final prognosis.

The highly heterogeneous outcome of MM is mainly ascribed to the complex genomic landscape, including chromosomal gains or losses, structural variations, and cancer driver gene mutations (Manier et al., 2017). These genomic instabilities contribute to the clonal expansion of disease. As a result of the rapid development of high-throughput sequencing, posttranscript regulation (PTGR) has gained attention throughout the whole process of tumors (Gerstberger et al., 2014). RNA-binding proteins (RBPs) play key roles in posttranscript regulation by affecting gene expression and cellular metabolism (Yan et al., 2021). Studies have found that RBPs are functionally associated with tumor progression in different types of cancers, including multiple myeloma (Konishi et al., 2021; Wang et al., 2021).

The crucial role of the complex bone marrow microenvironment in MM progression and therapeutic response has been well established. The interactive relationship between tumor cells and the bone marrow environment is critical in promoting chromosomal instability in MM(Neuse et al., 2020). Single-cell RNA-sequencing datasets revealed in-depth interactions of stomal cells, myeloma cells and immune cells within the bone marrow microenvironment. These analyses found bone marrow mesenchymal stromal cells, accompanied by immune cells and aberrant genes involved in immune modulation and tumorigenesis (de Jong et al., 2021).

It has been gradually recognized that the success of chemotherapy and immunotherapy relies on the anticancer immune response (Fridman et al., 2017). The correlation between tumor-infiltrating lymphocytes and the clinical outcomes of cancers has been investigated (Fridman et al., 2012). The prognostic value of infiltrating T lymphocytes has been widely accepted. In contrast to T cells, the effects of infiltrating B cells in tumorigenesis and treatment are far from clear.

Growing evidence has indicated that Tumor-infiltrating B (TIL-B) cells contribute to the prognostic effect of tumors by inducing $CD4^+$ T cells and $CD8^+$ T cells, which help to regulate tumor invasion and metastasis (Wouters and Nelson, 2018).

Multiple myeloma is a plasmocytic disease. The core biological process of MM is genetic dysfunction throughout the multistep progression of B cell development (Pawlyn and Morgan, 2017). In the present study, we investigated the TIL-B-related RBP signature in MM. Furthermore, we propose a B-cell-specific RBP prognostic model of MM for the first time by combining immune, RBP and clinical characteristics. This model enables us to predict the clinical prognosis and therapeutic response of MM patients.

2 MATERIALS AND METHODS

2.1 Patient and Tumor Cell Line Data Preparation

Transcriptional data of MM patients were downloaded from GSE24080, GSE4204 and GSE39754. GSE24080, including 559 newly diagnosed patients with MM(Mitchell et al., 2016), was used as the training set. GSE4204, including 538 newly diagnosed MM patients (Driscoll et al., 2010), was used as the validation set. These samples were analyzed on platforms GPL570, Affymetrix Human Genome U133 plus 2.0 array. GSE39754, including gene expression profiling of 170 newly diagnosed MM patients receiving bortezomib therapy (Chauhan et al., 2012), was used to compare the treatment response between different groups. A total of 1,542 RBPs were obtained from a previous study (Gerstberger et al., 2014). Expression data of RBPs in different cell types were downloaded from GSE42058 (4 samples of CD11c + cells), GSE49910 (4 samples of B cells, four samples of neutrophils, 24 samples of T cells, six samples of monocytes, eight samples of erythroblasts and a sample of bone marrow and progenitors), GSE51540 (9 samples of T cells), GSE59237 (10 samples of dendritic cells), GSE6863 (3 samples of dendritic cells), GSE8059 (3 samples of NK cells), GSE13906 (2 samples of gamma-delta T cells and two samples of lymphocytes), GSE23371 (12 samples of dendritic cells), GSE25320 (11 samples of mast cells), GSE27291 (12 samples of T cells), GSE27838 (16 samples of NK cells), GSE28490 (10 samples of monocytes, five samples of B cells, 10 samples of T cells, five samples of NK cells, four samples of eosinophils, five samples of mDCs, three samples of neutrophils and five samples of pDCs), GSE28726 (10 samples of NKT cells, eight samples of CD1d-aGC + Va24- T cells and eight samples of CD4 T cells), and GSE37750 (8 samples of plasmacytoid dendritic cells) and GSE39889 (16 samples of neutrophils). Each dataset was normalized, and all subsequent analyses were performed on normalized datasets.

2.2 Identification and Functional Analysis of Robust Prognostic B-Cell Specific RBP Signatures in MM

The Limma package (Ritchie et al., 2015) was used to screen differentially expressed RBPs among B cells and other cell types by following model: design < - model.matrix (~group+0). Genes with FDR-corrected *p*-values below 0.01 were considered differentially expressed genes. Significantly upregulated RBPs in B cells were defined as B cell-specific RBPs. Gene ontology (GO) (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) enrichment analyses of B-cell-specific RBPs were applied by clusterProfiler in the R package (Yu et al., 2012; Wu et al., 2021). K-M analysis was performed to screen B-cell-specific RBPs associated with survival ($p < 0.05$). Then, univariate and multivariate Cox regressions were performed to further obtain a robust prognostic B-cell-specific RBP signature in MM ($p < 0.05$).

2.3 Construction of the Risk Model and Nomogram

The calculation formula for the risk score was defined as follows:

ExpGene1*Coef1 + ExpGene2*Coef2+ ExpGene3*Coef3.where Coef indicates the regression coefficients of genes, and Exp is the normalized expression value of each prognostic B cell-specific RBP signature. According to the median value of the risk score, MM patients in the training set were grouped according to the value of the risk score. K-M analysis was performed to identify the overall survival of all risk groups. ROC curves were plotted to evaluate the effectiveness of the risk model using the "survivalROC" routine in the R package. The risk model was tested in the validation set. Thereafter, Cox regression analyses were performed to identify independent prognostic factors for MM patients. The risk predictive model was plotted as a nomogram based on independent prognostic factors. The performance of the nomogram was evaluated by calibration and decision curves.

2.4 Immune Microenvironment of MM Patients in High- and Low-Risk Groups

Twenty-nine immune-related gene sets were used to perform ssGSEA (Subramanian et al., 2005) to calculate the enrichment infiltration of immune cells, pathways or functions in the MM samples. The 29 gene sets represented all types of subtypes of immune cells, potential functions, and related pathways described in a previous study (He Y. et al., 2018). Moreover, the correlations between the risk score and differentially enriched immune cells, pathways or functions and the correlations between the prognostic B-cell specific RBP signature and differentially enriched immune cells, pathways or functions were calculated. At the same time, the expression of immune checkpoints, including CD274 (also named PD-L1), CD276, CTLA4, PDCD1 and VTCNA, was compared between different groups. Additionally, the correlations between the risk score and differentially expressed immune checkpoints were calculated. Correlations were evaluated using Pearson tests.

2.5 Observation of Chemotherapeutic Response in Different Risk Groups

The IC50 values of 20 common chemotherapeutic drugs in the low- and high-risk groups were calculated by the pRRophetic algorithm via the GDSC database (Yang et al., 2013; Geeleher et al., 2014), while the percentages of complete response (CR), very good partial response (VGPR), no response, progression-free (NR) or progression and no response (Prog) were calculated to evaluate the treatment response to bortezomib therapy in the low- and high-risk groups.

2.6 Statistical Analysis

All of the data were analyzed by R software (version 4.0.0). Comparisons between low- and high-risk groups were calculated using Wilcoxon's test.

3 RESULTS

3.1 Identification and Functional Analysis of 56B-cell-specific RBPs

By comparing the expression of RBPs among B cells and other cell types, we found a total of 56 significantly upregulated RBPs ($p < 0.01$) and defined them as B cell-specific RBPs. Heatmap displaying differential gene expression in B cells and other cells (**Figure 1A**). GO enrichment analysis indicated that these B cell-specific RBPs were enriched in ribosome-related and RNA

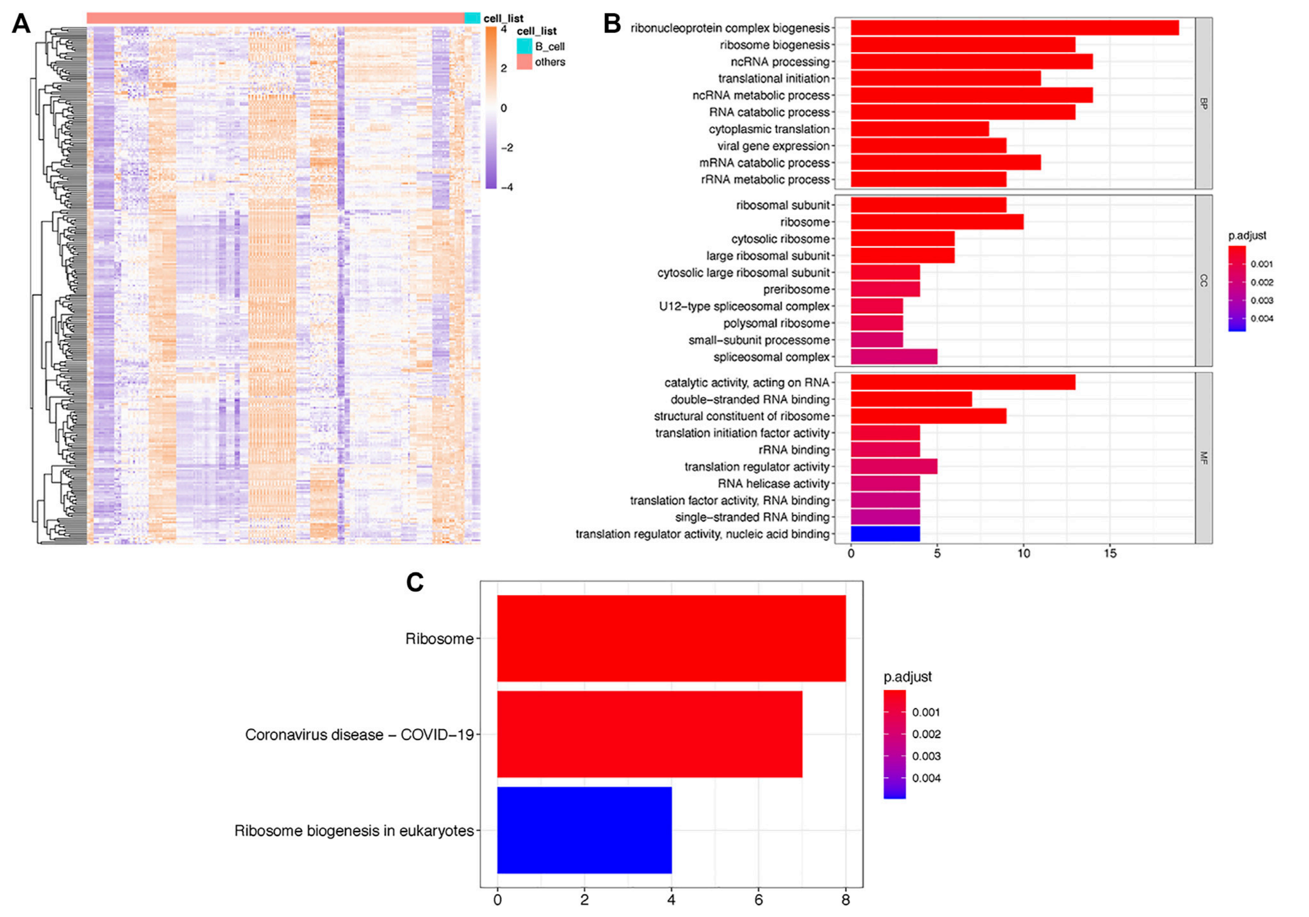


FIGURE 1 | Identification of TILB-RBP related mRNAs. **(A)**. Heatmap of differential RPB gene expression in B cells and other immune cells. **(B)**. GO enrichment analysis of TILB-RBP-related mRNAs. **(C)**. KEGG pathway of TILB-RBPs-related mRNAs.

metabolism- and catabolism-associated BP, CC and MF, including ribonucleoprotein complex biogenesis, ribosome biogenesis, mRNA catabolic process, ribosomal subunit and catalytic activity, and acted on RNA. The top 10 BP, CC and MF are shown in **Figure 1B**. Similar to the GO results, we found that these B cell-specific RBPs were significantly enriched in KEGG pathways of ribosome and ribosome biogenesis in eukaryotes (**Figure 1C**).

3.2 Identification of ADAR, FASTKD1 and SNRPD3 as Prognostic Signatures in MM

3.2.1 Purified Immune Cell Data

Thereafter, we investigated the prognostic value of these B cell-specific RBPs. First, according to the expression of each RBP, we divided MM patients into low- and high-RBP expression groups. K-M analysis revealed that patients in the groups with lower expression of FASKD1, SNRPD3, DDX21, MRPL3, ADAR, CPSF3, DROSHA, and CAPRIN2 and higher expression of SART1 had better survival (**Figure 2**), suggesting that FASKD1, SNRPD3, DDX21, MRPL3, ADAR, CPSF3, DROSHA, CAPRIN2 and SART1 might play important roles in the prognosis of MM patients. Next, univariate Cox regression analysis showed that FASKD1, SNPPD3, DDX21, MRPL3, ADAR, CPSF3 and DROSHA were closely related to the outcomes of patients (p < 0.05, **Table 1**). To further obtain a robust prognostic signature, we performed multivariate Cox regression algorithm analysis and found that ADAR, FASKD1 and SNRPD3 were significantly correlated with prognosis (p < 0.05, **Table 2**), and all of them acted as risk factors (HR > 1, **Figure 3**). Thus, ADAR, FASKD1 and SNRPD3 were identified as prognostic B cell-specific RBP signatures in MM and were used for subsequent construction of the risk model.

3.2.2 Construction of the Risk Score Model and Nomogram Based on the Prognostic B-Cell Specific RBP Signature in MM

The risk scores of each patient were calculated according to the expression levels and coefficients of ADAR, FASKD1 and SNRPD3. Patients in the training set were divided into high- and low-risk groups according to the median risk score (**Figure 4A**). The distribution of all patients' survival status in the training set is shown in **Figure 4B**. Patients in the low-risk group had a survival advantage over patients in the high-risk

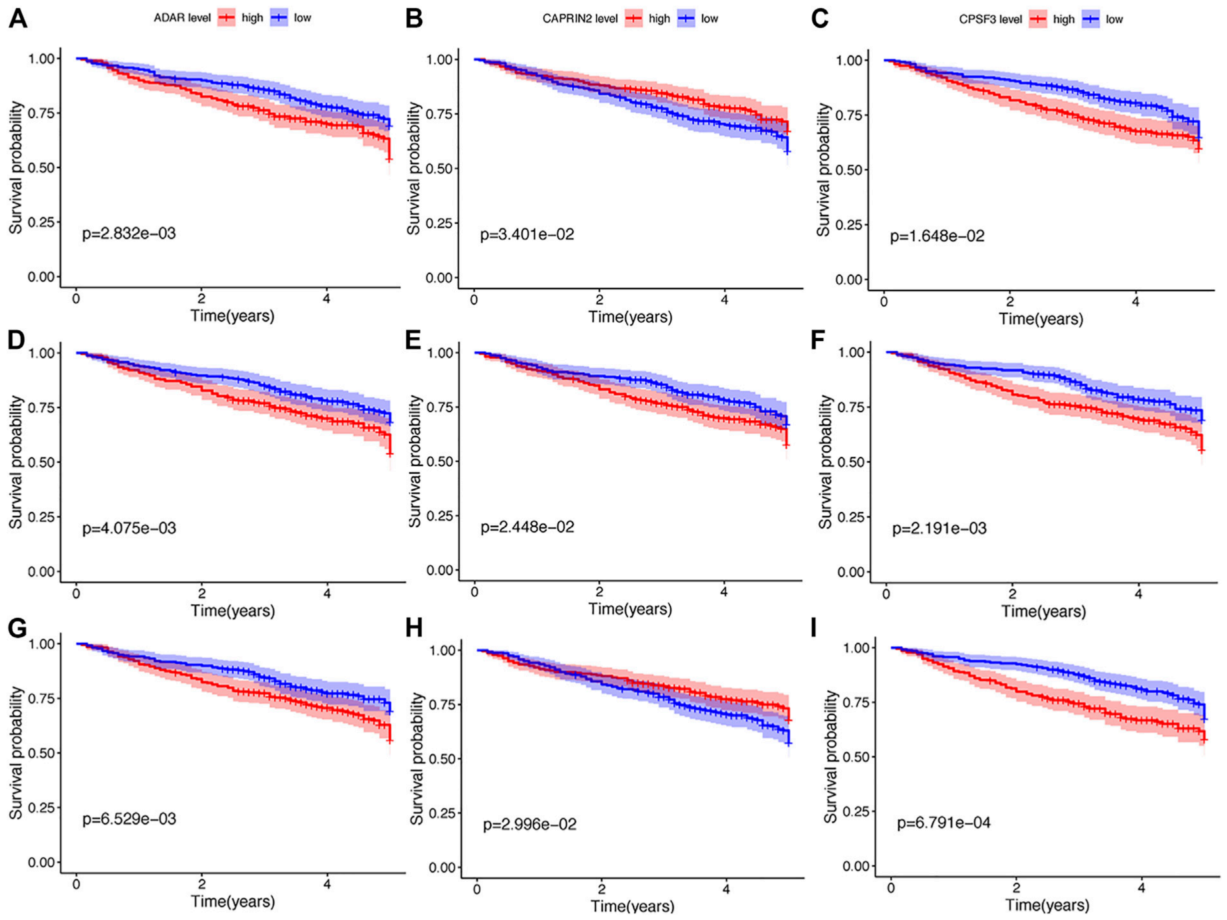


FIGURE 2 | K-M analysis of nine differential genes regarding survival. **(A–G)**. Kaplan-Meier survival curves of Multiple myeloma with different expression levels of ADAR, CAPRIN2, CPSF3, DDX21, FASKD1, DROSHA, MRPL3, SART1, and SNRPD3.

TABLE 1 | Univariate Cox regression analysis results of differential RBPs.

Gene	Hazard ratios	CL95	*p*-value
ADAR	1.76	1.28–2.41	0.000
CPSF3	1.83	1.29–2.60	0.001
DDX21	1.58	1.17–2.11	0.002
DROSHA	1.64	1.12–2.39	0.010
FASTKD1	1.74	1.32–2.31	0.000
MRPL3	1.73	1.18–2.54	0.001
SNRPD3	1.48	1.19–1.84	0.000
CAPRIN2	0.82	0.58–1.16	0.262
SART1	0.79	0.58–1.06	0.118

group (**Figure 4C**). The areas under the ROC curves for 1-,3–5 years survival were 0.648, 0.642 and 0.626, respectively, suggesting good performance of the risk model in the training set (**Figure 4D**). The risk model was further tested in the validation set, and similar results were obtained (**Figures 5A–D**).

Next, we conducted univariate and multivariate analyses to detect independent prognostic factors. The univariate results showed that age, B2M, BMPC and risk score were significantly associated with the overall survival of MM patients (**Figure 6A**). Age, B2M, BMPC and risk score were then included in the multivariate analysis, and we found that the risk score was remarkably related to prognosis (**Figure 6B**), indicating that the risk score was an independent prognostic factor for poor prognosis in MM.

Thereafter, we constructed a nomogram with a C-index of 0.667 to predict the 1-, 3–5 years survival of MM patients, combined with independent prognostic factors (age, B2M and risk score) obtained by the above multivariate analysis (**Figure 7A**). The slopes of the calibration curves for 1-, 3–5 years survival were close to 1 (**Figure 7B**), indicating the high accuracy of the nomogram. In addition, the decision curves, which displayed the clinical utility of each model, indicated that the nomogram had better survival prediction performance than the risk model (**Figure 7C**).

3.3 Functional Analysis of Prognostic B-cell-specific RBP Genes by GSEA

To better understand the underlying mechanisms of the prognostic B cell-specific RBP signature in regulating MM, we first analyzed the functions of genes by GSEA. We found that the Notch signaling pathway, prespliceosome, mRNA cis-splicing via spliceosome and

TABLE 2 | Multivariate Cox regression algorithm analysis of ADAR, FASKD1 and SNRPD3 gene expression with prognosis.

Gene	Coef	HR	HR.95L	HR.95H	*p*-value
ADAR	0.40216328	1.49505543	1.07431169	2.08057937	0.01707585
FASTKD1	0.44163922	1.55525453	1.17418704	2.05999263	0.00207,193
SNRPD3	0.25941536	1.29617207	1.03397688	1.62485454	0.02446633

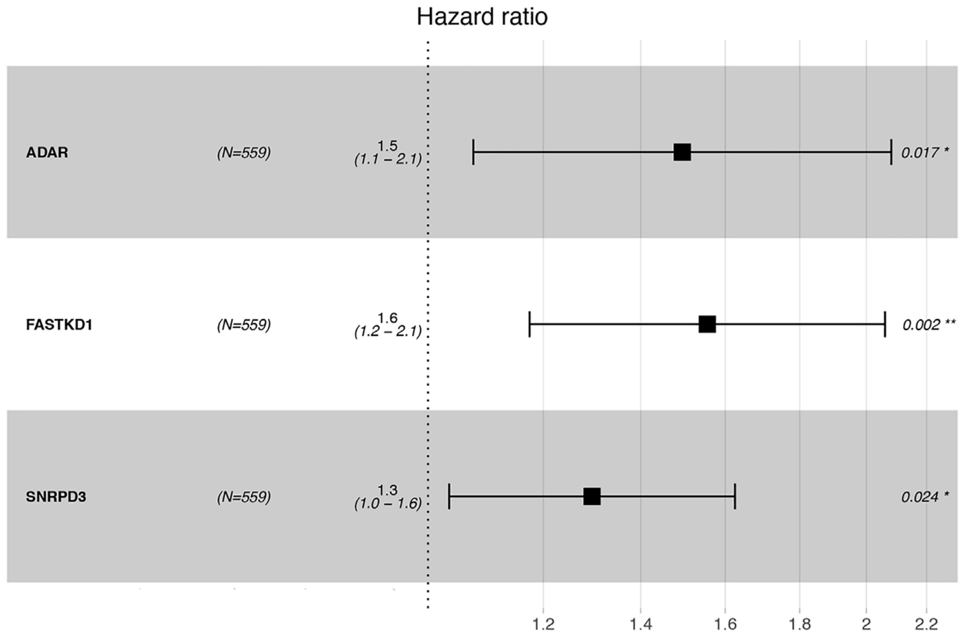


FIGURE 3 | Forest plot visualizing the HRs of univariate Cox analysis of the TILB-RBPs and prognosis.

U5 snRNP were notably enriched in the high-risk group ($p < 0.01$), while olfactory receptor activity, sensory perception of smell, response to amphetamine, establishment of pigment granule localization, regulation of renal system process, pigment granule localization, olfactory transduction, mating, and odorant binding were enriched in the low-risk group ($p < 0.01$, **Figure 8**).

3.4 Immune Microenvironment of Low- and High-Risk Groups

Next, we performed ssGSEA to detect the enriched distribution of different immune cells, pathways or functions. We analyzed the expression of immune checkpoints further to evaluate the immune microenvironment differences between the two groups. The ssGSEA results showed that the enrichment level of Tfhs was lower in the low-risk group, and other immune cells, including B cells, CD8$^+$ T cells, T cell coinhibition, T cell costimulation, Th1 cells and type II IFN response, DCs, iDCs, APC costimulation, CCR, checkpoint, HLA, inflammation-promoting, macrophages, mast cells, MHC class Ⅰ, and neutrophils, were more highly enriched in the low-risk group (**Figure 9A**). All of the enrichment levels of the immune cells, except Tfh and MHC class I, were negatively correlated with the expression of ADAR, FASKD1 and SNRPD3 ($p < 0.05$, **Figure 9B**). We also found that the risk score was negatively correlated with enrichment levels of the immune cells ($p < 0.05$, **Figure 9C**). Consistent with the ssGSEA results, we observed that the expression of immune checkpoints, including CD274, CD276, CTLA4 and VTCN1, was remarkably higher in the low-risk group ($p < 0.01$, **Figure 9D**), while the risk score was negatively correlated with the expression of CD274, CD276, CTLA4 and VTCN1 ($p < 0.01$, **Figure 9E**). All of these results suggested that the prognostic B-cell-specific RBP signature influenced the immune microenvironment of MM patients, and a higher risk score could indicate lower antitumor immunity in MM patients.

3.5 Validation of the Prognostic Value of TBIL-RBPs in the Chemotherapeutic Response of MM

Given the different immune microenvironments between the low- and high-risk groups, we hypothesized that the response to drugs might be different between the two groups. The IC50 of A.443,654, A.770,041, ABT.888, AG.014699, AICAR, AKT. inhibitors VIII, ATRA, AUY922, axitinib, AZ628 and AZD7762 were significantly higher in the low-risk group

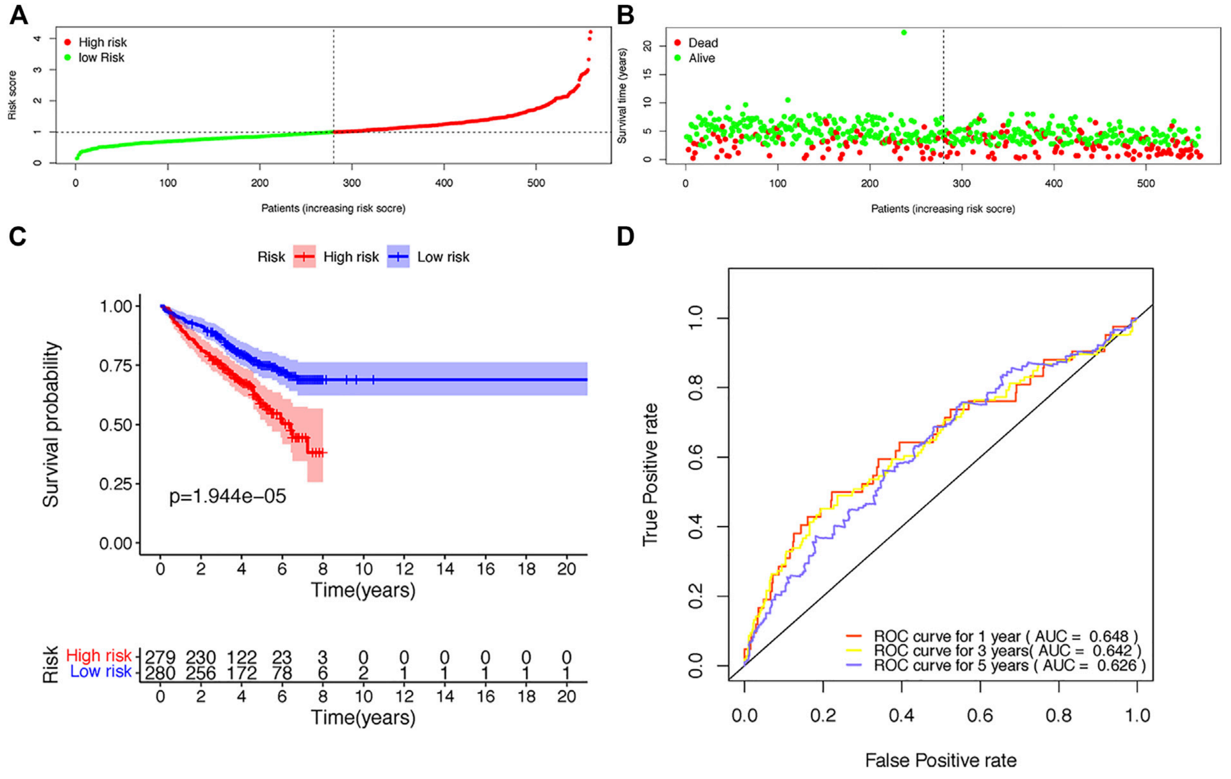


FIGURE 4 | Construction of the risk score model based on the prognostic B-cell specific RBP signature in MM. **(A)**. Patient distribution by different risk scores in the training set. **(B)**. Survival status of all patients in the training set. **(C)**. Kaplan-Meier survival curves of patients in the high-risk and low-risk groups. **(D)**. ROC curve analysis according to the 1–5 years survival of the area under the ROC curve value in the training set.

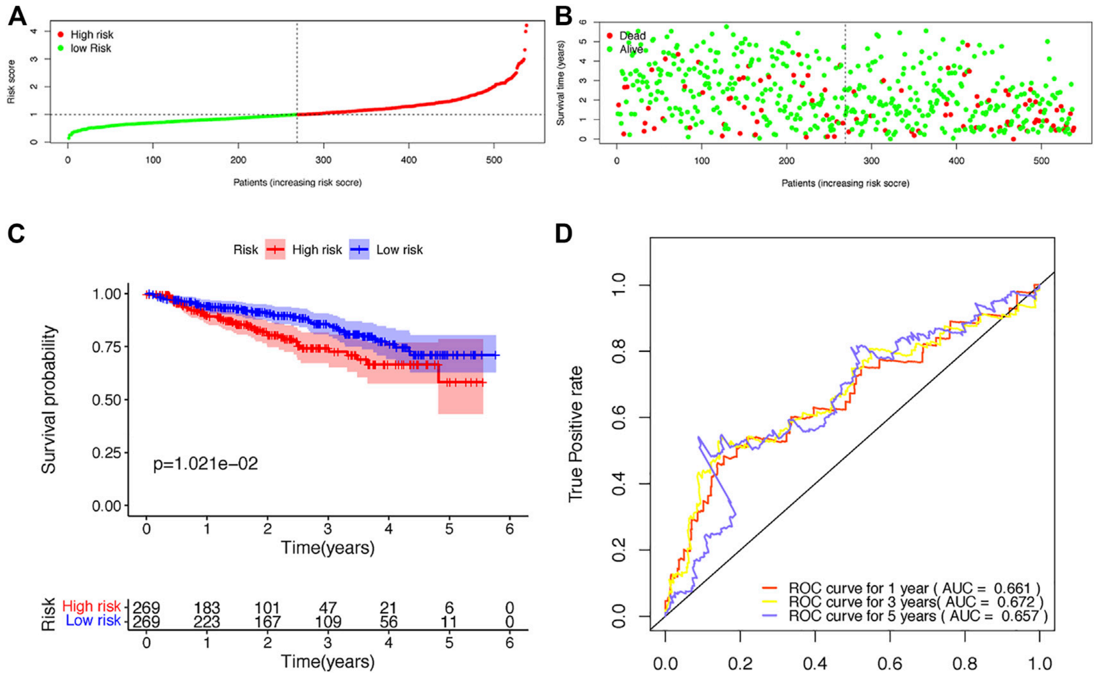


FIGURE 5 | Validation of the risk score model based on the prognostic B-cell specific RBP signature in MM. **(A)**. Patient distribution by different risk scores in the validation set. **(B)**. Survival status of all patients in the validation set. **(C)**. Kaplan-Meier survival curves of patients in the high-risk and low-risk groups. **(D)**. ROC curve analysis according to the 1–5 years survival of the area under the ROC curve value in the validation set.

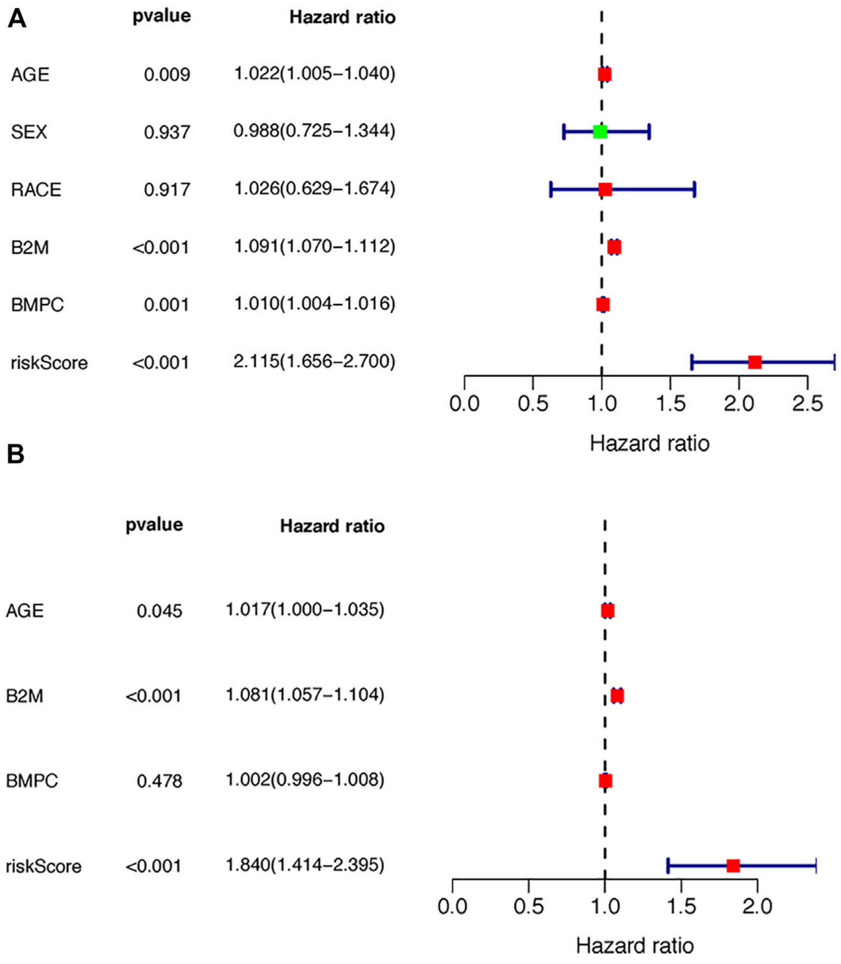


FIGURE 6 | Independence of the TILB-RBPs. A. Forest plot visualizing the HRs of univariate Cox analysis of the TILBlncSig and clinicopathological factors in **(A)** the TCGA discovery dataset **(B)** the TCGA testing dataset; and **(C)** the GSE31684 dataset.

(**Figure 10A**), indicating that patients in the low-risk group were more sensitive to these drugs. In addition, we compared the treatment response to bortezomib therapy in different risk classification groups. We found that a larger proportion of patients (36.6%) in the low-risk group had CR to bortezomib therapy than that (27.7%) in the high-risk group (**Figure 10B**), suggesting that the TBIL-RBPs might be a potential biomarker of bortezomib treatment response for MM patients.

4 DISCUSSION

Multiple myeloma is a B cell hematological malignancy with insidious onset. Once diagnosed, most patients suffer from multiorgan dysfunction. The incidence rate of MM has increased rapidly over the last decade. Substantial strides have been made in the treatment of MM. However, for some reasons, including a lack of early detection and complex cytogenetic abnormalities, the majority of MM patients continue to relapse, and a minority of MM patients even suffer from early relapse and resistance to chemotherapy and immunotherapy, gaining little benefit from advances in therapy (Kumar et al., 2017). The application of genomic technologies has led to a better understanding of the underlying biology of MM(Lohr et al., 2014). At the same time, it is widely accepted that dysregulated immunological processes in the tumor microenvironment are closely related to the progression of MM(Nakamura et al., 2020; Botta et al., 2021). Thus, we concentrated on the cytogenetic heterogeneity of MM and the correlation between tumor immune cell infiltration and tumor cells. Using RNA sequencing data and clinical data in GEO, we constructed a novel prognostic model based on B cell-specific RBP-associated genes, which are of remarkable importance in the early diagnosis, prognosis prediction and therapeutic evaluation. Subsequently, we verified the predictive value of the model in validation datasets. Furthermore, a nomogram with high accuracy for predicting the overall survival of MM patients was constructed based on the TBIL-RBPs and other independent prognostic factors, as evidenced by calibration and decision curves.

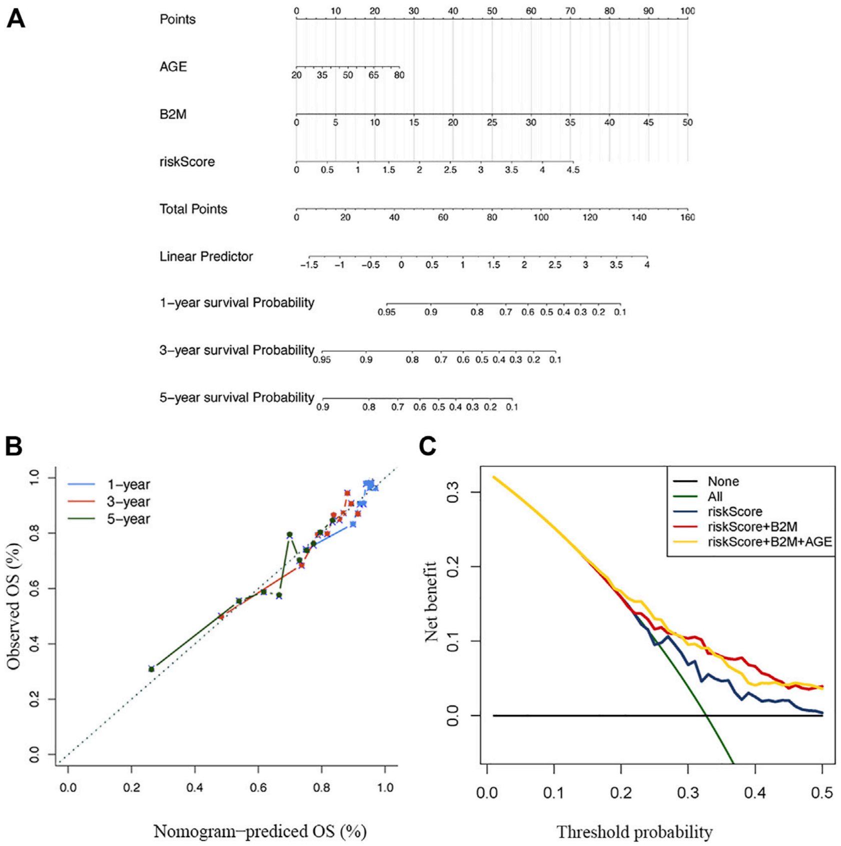


FIGURE 7 | Construction and verification of the nomogram **(A)** A nomogram combining clinical signatures and prognostic factors to predict the 1–5 years survival rate of MM patients **(B)** The 5 years calibration chart verifies the predictive ability of the nomogram **(C)** The 5 years decision curve analysis of the clinical benefit rate.

In the present study, we first conducted a comparison analysis among different immune cell lines. Fifty-six highly specifically expressed TILB-RBPs were preferentially observed in B cell lines compared with other immune cell lines. Furthermore, functional enrichment analysis revealed that these B cell-specific RBPs were closely related to the immune response, ribosome biogenesis and RNA metabolism. RBPs play key roles in posttranscriptional regulation via genetic changes, epigenetic alterations, and noncoding RNA mediation, which are essential in the malignant transformation of cancers (Bitaraf et al., 2021). RBPs are also essential in tumorigenesis in hematological malignancies. Insulin-like growth factor 2 mRNA binding proteins (IGF2BPs) are described as major regulators of stem cells. IGF2BP1 and IGF2BP3 are overexpressed in translocation-ETV6/RUNX1-positive B-ALL (Elcheva et al., 2020). Acute myelocytic leukemia patients with high expression of IGF2BP2 had worse overall survival (He X. et al., 2018). Musashi-2 protein (MSI2) is overexpressed in acute myeloid leukemia (AML) cell lines, and high expression of MSI2 promotes proliferation and inhibits apoptosis of AML cells. High expression of MSI2 in AML patients correlates with poorer survival in patients, thereby defining MSI2 as a prognostic biomarker for therapy in AML (Kharas et al., 2010). There have also been several reports of specific low-occurrence mutations in RPL5 and RPL10 and overexpression of RPS9 in MM that were closely related to tumorigenesis and clinical outcomes (Dabbah et al., 2021). These studies were in accordance with our findings that RBPs could be potential prognostic biomarkers.

To further define the role of the TILB-RBPs in the clinical outcomes of MM, the relationship between TILB-RBPs and overall survival was assessed. We identified 3B cell-specific RBP genes -- ADAR, FASTKD1 and SNRPD3—which were significantly correlated with the outcomes of MM patients. ADAR-mediated A-to-I editing is a key form of posttranscriptional regulation in human physiology (Vesely and Jantsch, 2021). ADAR1 is the most abundant and active RNA editing enzyme in MM and is recognized as an oncogenic central driver of cancer cell proliferation (Teoh et al., 2018). ADAR1 promotes malignant regeneration of MM by mediating the recoding of the self-renewal agonist GLI1, which activates the Hh pathway and promotes the production of cancer stem cells (Lazzari et al., 2017). FASTK family proteins have been verified to be linked to mitochondrial diseases by regulating mitochondrial RNA homeostasis (Boehm et al., 2017). Some studies have confirmed that FASTKD1 is related to the occurrence of tumors. For example, FASTKD1 was associated with poor prognosis of ALL in children and adults (Wang et al., 2015). FASTKD1 could also be used as a biomarker of primary

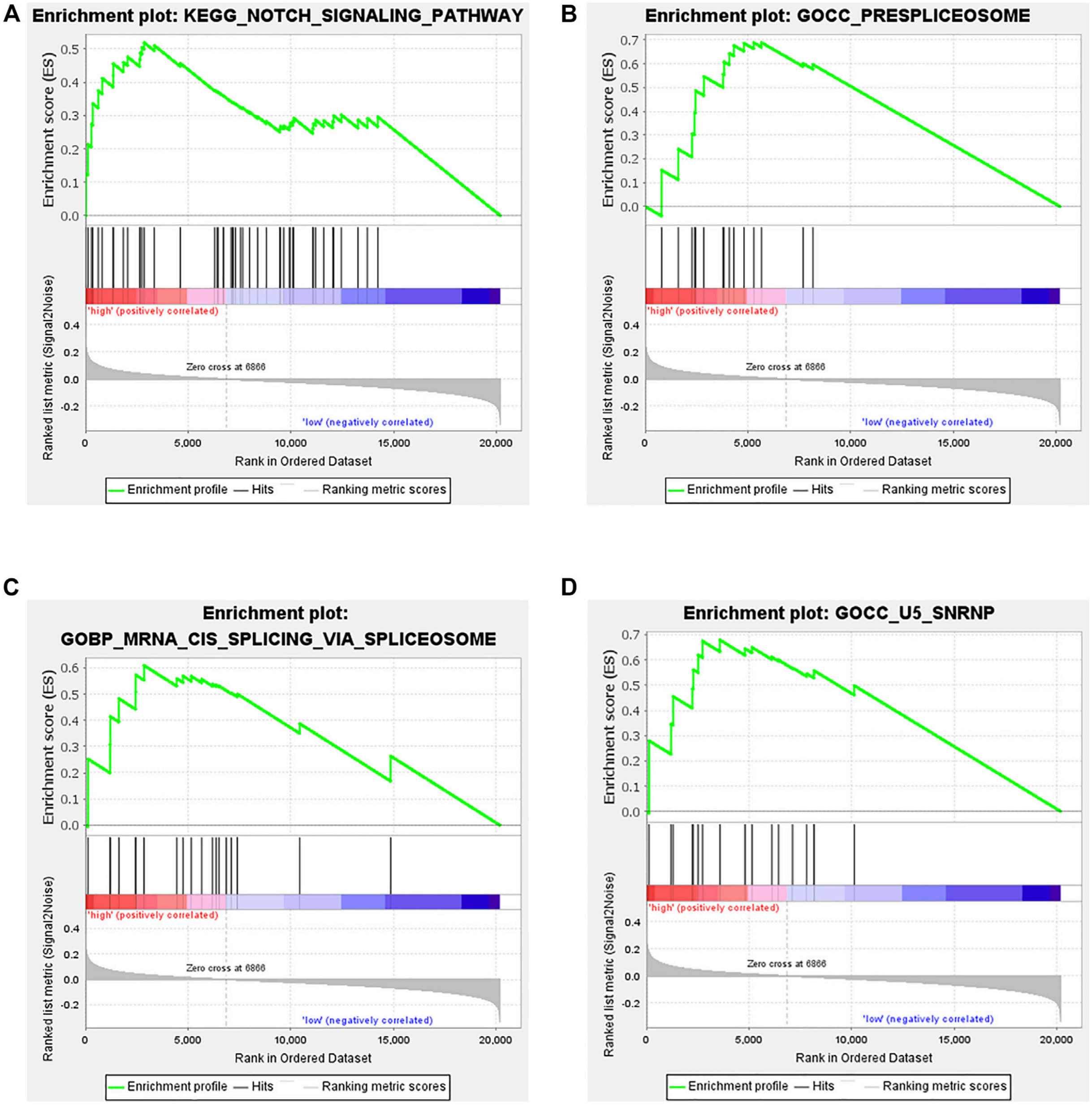


FIGURE 8 | Functional analysis of genes in the low- and high-risk groups by GSEA. **(A)**. Notch signaling pathway, **(B)**. prespliceosome, **(C)**. mRNA cis-splicing via spliceosome, **(D)**. U5 snRNP.

endometrial tumors (Colas et al., 2011). SNRPD3, also called SMD3, binds to small nuclear RNA to affect the formation of small nuclear ribonucleoprotein particles (Camasses et al., 1998). Studies have revealed that silencing of SNRPD3 causes overexpression of p53 levels, thereby modulating CDKN1A expression and further influencing the cell cycle arrest and cell death of NSCLC cells (Siebring-van Olst et al., 2017). In addition, a study also found that SNRPD3 might be a novel breast cancer-related biomarker (Zhang et al., 2015). In our study, we found for the first time that FASTKD1 and SNRPD3 are related to the prognosis of multiple myeloma, and the specific function and mechanism of these genes in tumorigenesis in multiple myeloma require further study. At the same time, we calculated the risk score and constructed a predictive model based on these three genes. The results of the ROC curve analysis showed that the model has good predictive effects. In addition, univariate and multivariate regression analyses indicated that the risk score was an independent prognostic factor. At the same time, a nomogram was constructed for predicting the survival of patients with multiple myeloma at 1, 3 and 5 years. The C index and correction curve of the nomogram showed that the

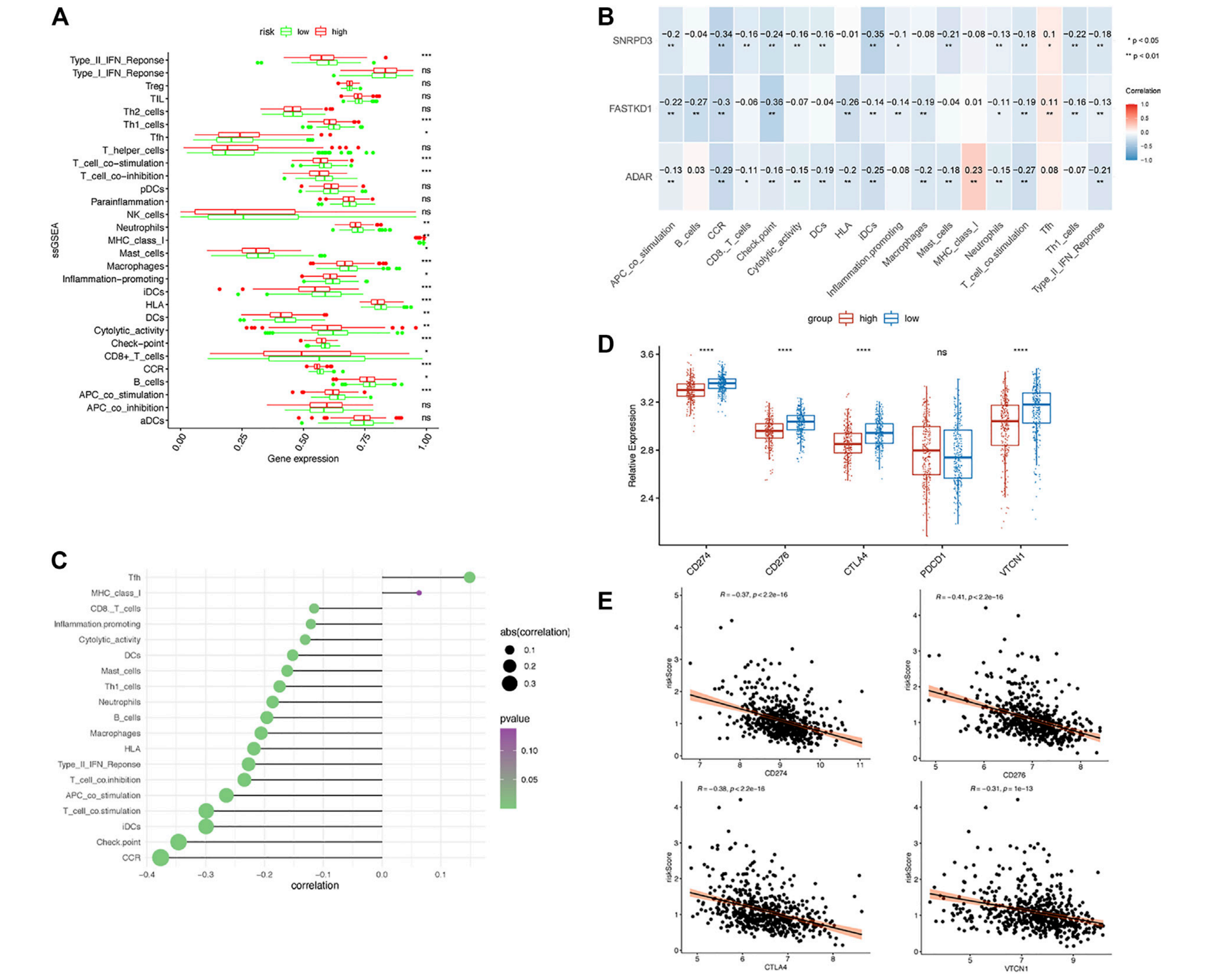


FIGURE 9 | Immune microenvironment of the low- and high-risk groups. **(A)**. Boxplots of the immune cell infiltration cluster in the high- (red)- and low- (green)-risk groups stratified by the TILB RBP prognostic model. **(B)**. Correlation between the expression of ADAR, SNRPD3 and FASTKD1 and the immune cell infiltration cluster. **(C)**. Correlation between the risk score and immune cell infiltration cluster. **(D)**. The differential expression of immune checkpoints, including CD274, CD276, CTLA4 and VTCN1, in the low-risk and high-risk groups. **(E)**. Association between the risk score and the expression of CD274, CD276, CTLA4 and VTCN1.

prediction model has high prediction accuracy for 1, 3 and 5 years and has clinical value.

To further clarify the role of the TILB-RBPs in stratifying survival, the association of the TILB-RBPs and survival in MM was assessed. Patients were grouped based on the risk score. First, GSEA functional enrichment analysis was performed for all genes in different groups. The results revealed that the Notch signaling pathway and biological processes and cellular components related to RNA splicing were significantly enriched in the high-risk groups. The Notch pathway is crucial to cell cycle regulation. Accumulating evidence has shown that the Notch pathway deregulates MM in tumorigenesis and drug resistance, especially in proteasome inhibitor resistance (Colombo et al., 2013). Deregulation of Notch signaling in MM occurs throughout the pathogenesis of plasma cells (Saltarella et al., 2019). Notch receptors and their ligands affect not only MM cells but also bone marrow stroma to further regulate the adhesive behavior of MM (Nefedova et al., 2004). In addition, the Notch pathway plays a vital role in immune regulation by stimulating the proliferation of T regulatory cells and upregulating TGF-β receptor II to suppress antitumor T-cell responses (Hue et al., 2012). An increasing number of studies have shown that the RNA spliceosome pathway is a major factor in cancer progression. A study revealed that aberrant RNA splicing patterns were relevant to worse survival outcomes of MM patients, which could be used for the risk stratification of patients (Bauer et al., 2021). Moreover, a study showed that inhibition of the spliceosome could synergize with proteasome inhibitors to potentiate antitumor effects. This unreported mechanism of the spliceosome suggests that spliceosome targeting could serve as a potential therapeutic

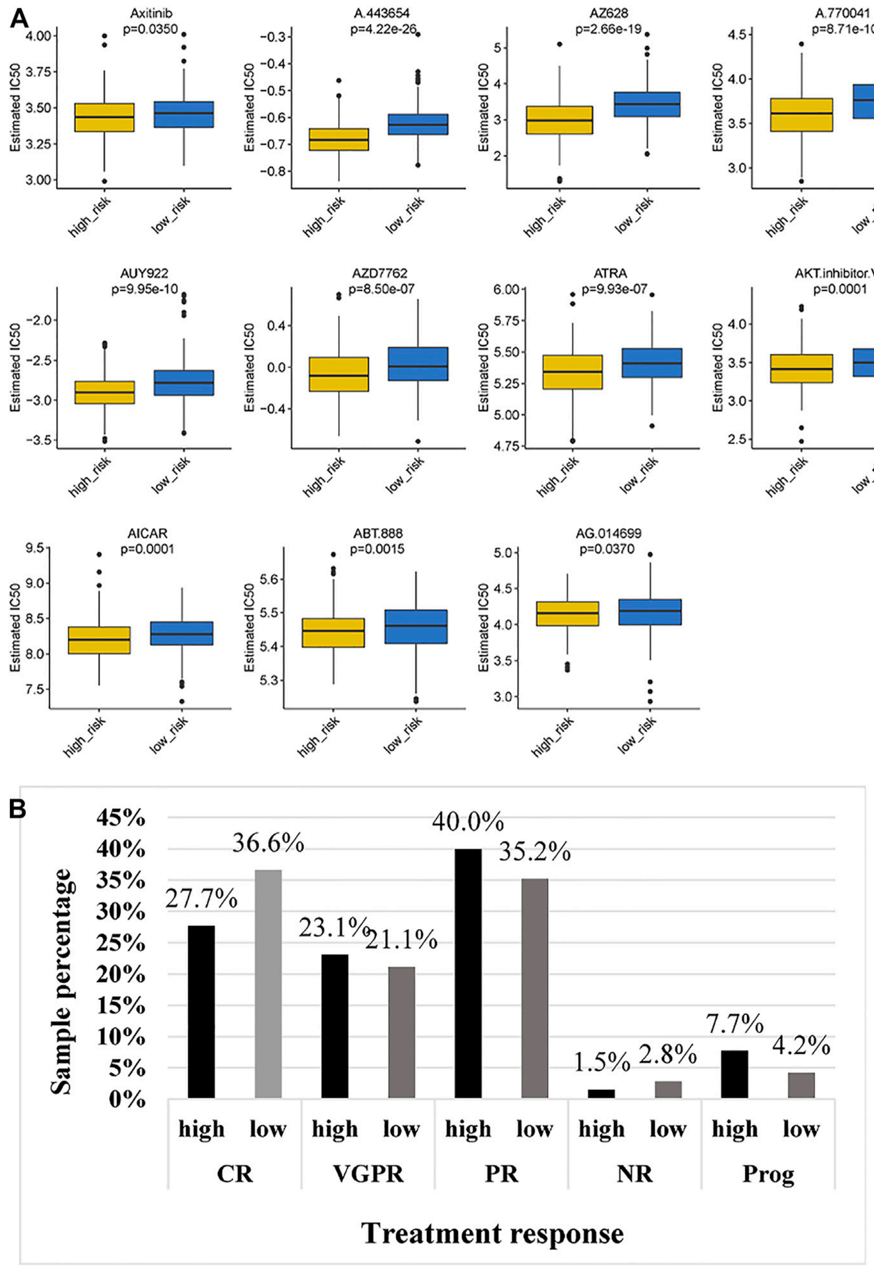


FIGURE 10 | Chemotherapeutic response of MM patients in the low- and high-risk groups. **(A)**. Comparison of IC50 of chemotherapeutic drugs between the high-risk and low-risk groups. **(B)**. Bortezomib treatment response of MM patients in the high-risk and low-risk groups.

target in myeloma (Huang et al., 2020). The above results are in accordance with our findings that prognostic characteristic genes could affect the prognosis of patients with multiple myeloma by regulating the splicing of precursor mRNA, activation of the Notch pathway and RNase L and ribosomal nucleoprotein synthesis.

Subsequently, we also compared the immune microenvironment in different groups. We found that there were significant differences in immune cell infiltration, immune-related functions, immune-related pathways and the expression of immune checkpoint genes between the two groups. Several single-cell transcriptional studies have revealed that transcriptional programs are associated with aggressive myeloma progression and immune evasion (Ryu et al., 2020; Liu et al., 2021). According to the above findings, we present the hypotheses that the prognostic characteristic genes are highly associated with different immune microenvironments in the two groups. Subsequently, we conducted a correlation analysis of TBIL-RBPs and immune cell infiltration. We found that the expression of ADAR, FASTKD1 and SNRPD3 was negatively correlated with the infiltration, functions and pathways of immune cells. The risk score was also negatively correlated with the expression of immune checkpoints, indicating that ADAR, FASTKD1 and SNRPD3 might interact with the immune microenvironment of multiple myeloma. TBIL-RBPs might further influence the immune response of MM patients, response to treatment, and prognosis.

We finally analyzed the Genomics of Drug Sensitivity in Cancer (GDSC) dataset to further validate the prognostic effect of the risk score. The GDSC is a large dataset including cell viability and response to drugs (Yang et al., 2013). We found that the IC50 of 11 drugs in the low-risk group was significantly higher than that in the high-risk group, indicating that patients in the low-risk group might have greater sensitivity to these 11 drugs. Strikingly, the high-risk group presented less sensitivity to bortezomib treatment. These results, together with previous observations, supported the risk score based on TILB-RBPs and demonstrated good accuracy for prognostic assessment. The TILB-RBPs were shown to have prognostic value not only for chemotherapy but also for immunotherapy. Nonetheless, there are limitations of our current study. First the prognostic model still needs to be further validated in other independent large sample cohorts to ensure the reliability of the model before clinical use. In addition, more functional experiments *in vivo* and vitro are still needed to further reveal the possible mechanisms for TILB-RBPs.

5 CONCLUSION

In conclusion, in this study, we identified 3 B lymphocyte-specific RBPs significantly related to the overall survival of MM patients and further established a risk model based on these genes. The good predictive value of the model was verified in the validation set. Application of the TBIL-RBPs to immunotherapy datasets revealed that the risk model can assess not only chemotherapy but also immunotherapy response. To the best of our knowledge, our study is the first to investigate B lymphocyte specific RBPs in MM, emphasizing the impact of TILB-RBPs on clinical outcomes and treatment response. The results of this study could provide a basis for individualized precision therapy in the future. The three prognostic genes—ADAR, FASTKD1 and SNRPD3 -- could be potential new prognostic and therapeutic biomarkers of MM.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE24080, GSE4204, and GSE39754.

AUTHOR CONTRIBUTIONS

JJZ contributed to Data curation, writing—original draft preparation. PH and XW contributed to analysis and validation of data SW and LM contributed to processed figures and tables. JZ contributed to design of the study and writing—reviewing. All authors have read and approved the final manuscript.

FUNDING

The present study was funded by grants from the Fundamental Research Funds for the Central Universities (grant no. xjj2018jchz15).

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Attal, M., Lauwers-Cances, V., Hulin, C., Leleu, X., Caillot, D., Escoffre, M., et al. (2017). Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma. *N. Engl. J. Med.* 376 (14), 1311–1320. doi:10.1056/NEJMoa1611750

Binder, M., Rajkumar, S. V., Ketterling, R. P., Dispenzieri, A., Lacy, M. Q., Gertz, M. A., et al. (2017). Substratification of Patients with Newly Diagnosed Standard-Risk Multiple Myeloma. *Blood* 130 (Suppl. 1), 1789. doi:10.1111/bjh.15800

Bitaraf, A., Razmara, E., Bakhshinejad, B., Yousefi, H., Vatanmakanian, M., Garshasbi, M., et al. (2021). The Oncogenic and Tumor Suppressive Roles of RNA-binding Proteins in Human Cancers. *J. Cel Physiol* 236 (9), 6200–6224. doi:10.1002/jcp.30311

Boehm, E., Zaganelli, S., Maundrell, K., Jourdain, A. A., Thore, S., and Martinou, J.-C. (2017). FASTKD1 and FASTKD4 Have Opposite Effects on Expression of Specific Mitochondrial RNAs, Depending upon Their Endonuclease-like RAP Domain. *Nucleic Acids Res.* 45 (10), 6135–6146. doi:10.1093/nar/gkx164

Botta, C., Mendicino, F., Martino, E. A., Vigna, E., Ronchetti, D., Correale, P., et al. (2021). Mechanisms of Immune Evasion in Multiple Myeloma: Open Questions and Therapeutic Opportunities. *Cancers* 13 (13), 3213. doi:10.3390/cancers13133213

Bauer, M. A., Cody Ashby, C., Wardell, C., Boyle, E. M., Maria Ortiz, M., Erin Flynt, E., et al. (2021). Differential RNA Splicing as a Potentially Important Driver Mechanism in Multiple Myeloma. *haematol* 106 (3), 736–745. doi:10.3324/haematol.2019.235424

Camasses, A., Bragado-Nilsson, E., Martin, R., Se´raphin, B., and Bordonne´, R. (1998). Interactions within the Yeast Sm Core Complex: from Proteins to Amino Acids. *Mol. Cel Biol* 18 (4), 1956–1966. doi:10.1128/mcb.18.4.1956

Chauhan, D., Tian, Z., Nicholson, B., Kumar, K. G. S., Zhou, B., Carrasco, R., et al. (2012). A Small Molecule Inhibitor of Ubiquitin-specific Protease-7 Induces Apoptosis in Multiple Myeloma Cells and Overcomes Bortezomib Resistance. *Cancer Cell* 22 (3), 345–358. doi:10.1016/j.ccr.2012.08.007

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 (10), 2435–2444. doi:10.1002/ijc.25901

Colombo, M., Mirandola, L., Platonova, N., Apicella, L., Basile, A., Figueroa, A. J., et al. (2013). Notch-directed Microenvironment Reprogramming in Myeloma: a Single Path to Multiple Outcomes. *Leukemia* 27 (5), 1009–1018. doi:10.1038/leu.2013.6

Cowan, A. J., Allen, C., Barac, A., Basaleem, H., Bensenor, I., Curado, M. P., et al. (2018). Global Burden of Multiple Myeloma. *JAMA Oncol.* 4 (9), 1221–1227. doi:10.1001/jamaoncol.2018.2128

Dabbah, M., Lishner, M., Jarchowsky-Dolberg, O., Tartakover-Matalon, S., Brin, Y. S., Pasmanik-Chor, M., et al. (2021). Ribosomal Proteins as Distinct "passengers" of Microvesicles: New Semantics in Myeloma and Mesenchymal Stem Cells' Communication. *Translational Res.* 236, 117–132. doi:10.1016/j.trsl.2021.04.002

de Jong, M. M. E., Kellermayer, Z., Papazian, N., Tahri, S., Hofste Op Bruinink, D., Hoogenboezem, R., et al. (2021). The Multiple Myeloma Microenvironment Is

Defined by an Inflammatory Stromal Cell Landscape. *Nat. Immunol.* 22 (6), 769–780. doi:10.1038/s41590-021-00931-3

Driscoll, J. J., Pelluru, D., Lefkimmiatis, K., Fulciniti, M., Prabhala, R. H., Greipp, P. R., et al. (2010). The Sumoylation Pathway Is Dysregulated in Multiple Myeloma and Is Associated with Adverse Patient Outcome. *Blood* 115 (14), 2827–2834. doi:10.1182/blood-2009-03-211045

Elcheva, I. A., Wood, T., Chiarolanzio, K., Chim, B., Wong, M., Singh, V., et al. (2020). RNA-binding Protein IGF2BP1 Maintains Leukemia Stem Cell Properties by Regulating HOXB4, MYB, and ALDH1A1. *Leukemia* 34 (5), 1354–1363. doi:10.1038/s41375-019-0656-9

Facon, T., Kumar, S., Plesner, T., Orlowski, R. Z., Moreau, P., Bahlis, N., et al. (2019). Daratumumab Plus Lenalidomide and Dexamethasone for Untreated Myeloma. *N. Engl. J. Med.* 380 (22), 2104–2115. doi:10.1056/NEJMoa1817249

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The Immune Contexture in Human Tumours: Impact on Clinical Outcome. *Nat. Rev. Cancer* 12 (4), 298–306. doi:10.1038/nrc3245

Fridman, W. H., Zitvogel, L., Sautès–Fridman, C., and Kroemer, G. (2017). The Immune Contexture in Cancer Prognosis and Treatment. *Nat. Rev. Clin. Oncol.* 14 (12), 717–734. doi:10.1038/nrclinonc.2017.101

Geeleher, P., Cox, N., and Huang, R. S. (2014). pRRophetic: an R Package for Prediction of Clinical Chemotherapeutic Response from Tumor Gene Expression Levels. *PLoS One* 9 (9), e107468. doi:10.1371/journal.pone.0107468

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A Census of Human RNA-Binding Proteins. *Nat. Rev. Genet.* 15 (12), 829–845. doi:10.1038/nrg3813

He, X., Li, W., Liang, X., Zhu, X., Zhang, L., Huang, Y., et al. (2018a). IGF2BP2 Overexpression Indicates Poor Survival in Patients with Acute Myelocytic Leukemia. *Cell Physiol Biochem* 51 (4), 1945–1956. doi:10.1159/000495719

He, Y., Jiang, Z., Chen, C., and Wang, X. (2018b). Classification of Triple-Negative Breast Cancers Based on Immunogenomic Profiling. *J. Exp. Clin. Cancer Res.* 37 (1), 327. doi:10.1186/s13046-018-1002-1

Huang, H. H., Ferguson, I. D., Thornton, A. M., Bastola, P., Lam, C., Lin, Y.-H. T., et al. (2020). Proteasome Inhibitor-Induced Modulation Reveals the Spliceosome as a Specific Therapeutic Vulnerability in Multiple Myeloma. *Nat. Commun.* 11 (1), 1931. doi:10.1038/s41467-020-15521-4

Hue, S., Kared, H., Mehwish, Y., Mouhamad, S., Balbo, M., and Levy, Y. (2012). Notch Activation on Effector T Cells Increases Their Sensitivity to Treg Cell-Mediated Suppression through Upregulation of TGF-Brii Expression. *Eur. J. Immunol.* 42 (7), 1796–1803. doi:10.1002/eji.201142330

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kharas, M. G., Lengner, C. J., Al-Shahrour, F., Bullinger, L., Ball, B., Zaidi, S., et al. (2010). Musashi-2 Regulates normal Hematopoiesis and Promotes Aggressive Myeloid Leukemia. *Nat. Med.* 16 (8), 903–908. doi:10.1038/nm.2187

Konishi, H., Kashima, S., Goto, T., Ando, K., Sakatani, A., Tanaka, H., et al. (2021). The Identification of RNA-Binding Proteins Functionally Associated with Tumor Progression in Gastrointestinal Cancer. *Cancers* 13 (13), 3165. doi:10.3390/cancers13133165

Kumar, S. K., Rajkumar, V., Kyle, R. A., van Duin, M., Sonneveld, P., Mateos, M.-V., et al. (2017). Multiple Myeloma. *Nat. Rev. Dis. Primers* 3, 17046. doi:10.1038/nrdp.2017.46

Lazzari, E., Mondala, P. K., Santos, N. D., Miller, A. C., Pineda, G., Jiang, Q., et al. (2017). Alu-dependent RNA Editing of GLI1 Promotes Malignant Regeneration in Multiple Myeloma. *Nat. Commun.* 8 (1), 1922. doi:10.1038/s41467-017-01890-w

Liu, R., Gao, Q., Foltz, S. M., Fowles, J. S., Yao, L., Wang, J. T., et al. (2021). Co-evolution of Tumor and Immune Cells during Progression of Multiple Myeloma. *Nat. Commun.* 12 (1), 2559. doi:10.1038/s41467-021-22804-x

Lohr, J. G., Stojanov, P., Carter, S. L., Cruz-Gordillo, P., Lawrence, M. S., Auclair, D., et al. (2014). Widespread Genetic Heterogeneity in Multiple Myeloma: Implications for Targeted Therapy. *Cancer Cell* 25 (1), 91–101. doi:10.1016/j.ccr.2013.12.015

Manier, S., Salem, K. Z., Park, J., Landau, D. A., Getz, G., and Ghobrial, I. M. (2017). Genomic Complexity of Multiple Myeloma and its Clinical Implications. *Nat. Rev. Clin. Oncol.* 14 (2), 100–113. doi:10.1038/nrclinonc.2016.122

Mikkilineni, L., and Kochenderfer, J. N. (2021). CAR T Cell Therapies for Patients with Multiple Myeloma. *Nat. Rev. Clin. Oncol.* 18 (2), 71–84. doi:10.1038/s41571-020-0427-6

Mitchell, J. S., Li, N., Weinhold, N., Försti, A., Ali, M., van Duin, M., et al. (2016). Genome-wide Association Study Identifies Multiple Susceptibility Loci for Multiple Myeloma. *Nat. Commun.* 7, 12050. doi:10.1038/ncomms12050

Nakamura, K., Smyth, M. J., and Martinet, L. (2020). Cancer Immunoediting and Immune Dysregulation in Multiple Myeloma. *Blood* 136 (24), 2731–2740. doi:10.1182/blood.2020006540

Nefedova, Y., Cheng, P., Alsina, M., Dalton, W. S., and Gabrilovich, D. I. (2004). Involvement of Notch-1 Signaling in Bone Marrow Stroma-Mediated De Novo Drug Resistance of Myeloma and Other Malignant Lymphoid Cell Lines. *Blood* 103 (9), 3503–3510. doi:10.1182/blood-2003-07-2340

Neuse, C. J., Lomas, O. C., Schliemann, C., Shen, Y. J., Manier, S., Bustoros, M., et al. (2020). Genome Instability in Multiple Myeloma. *Leukemia* 34 (11), 2887–2897. doi:10.1038/s41375-020-0921-y

Palumbo, A., and Anderson, K. (2011). Multiple Myeloma. *N. Engl. J. Med.* 364 (11), 1046–1060. doi:10.1056/NEJMra1011442

Palumbo, A., Avet-Loiseau, H., Oliva, S., Lokhorst, H. M., Goldschmidt, H., Rosinol, L., et al. (2015). Revised International Staging System for Multiple Myeloma: A Report from International Myeloma Working Group. *Jco* 33 (26), 2863–2869. doi:10.1200/jco.2015.61.2267

Pawlyn, C., and Morgan, G. J. (2017). Evolutionary Biology of High-Risk Multiple Myeloma. *Nat. Rev. Cancer* 17 (9), 543–556. doi:10.1038/nrc.2017.63

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Ryu, D., Kim, S. J., Hong, Y., Jo, A., Kim, N., Kim, H.-J., et al. (2020). Alterations in the Transcriptional Programs of Myeloma Cells and the Microenvironment during Extramedullary Progression Affect Proliferation and Immune Evasion. *Clin. Cancer Res.* 26 (4), 935–944. doi:10.1158/1078-0432.Ccr-19-0694

Saltarella, I., Frassanito, M. A., Lamanuzzi, A., Brevi, A., Leone, P., Desantis, V., et al. (2019). Homotypic and Heterotypic Activation of the Notch Pathway in Multiple Myeloma-Enhanced Angiogenesis: A Novel Therapeutic Target? *Neoplasia* 21 (1), 93–105. doi:10.1016/j.neo.2018.10.011

Siebring-van Olst, E., Blijlevens, M., de Menezes, R. X., van der Meulen-Muileman, I. H., Smit, E. F., and van Beusechem, V. W. (2017). A Genome-wide siRNA Screen for Regulators of Tumor Suppressor P53 Activity in Human Non-small Cell Lung Cancer Cells Identifies Components of the RNA Splicing Machinery as Targets for Anticancer Treatment. *Mol. Oncol.* 11 (5), 534–551. doi:10.1002/1878-0261.12052

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U S A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Teoh, P. J., An, O., Chung, T.-H., Chooi, J. Y., Toh, S. H. M., Fan, S., et al. (2018). Aberrant Hyperediting of the Myeloma Transcriptome by ADAR1 Confers Oncogenicity and Is a Marker of Poor Prognosis. *Blood* 132 (12), 1304–1317. doi:10.1182/blood-2018-02-832576

Vesely, C., and Jantsch, M. F. (2021). An I for an A: Dynamic Regulation of Adenosine Deamination-Mediated RNA Editing. *Genes* 12 (7), 1026. doi:10.3390/genes12071026

Walker, B. A., Mavrommatis, K., Wardell, C. P., Ashby, T. C., Bauer, M., Rosenthal, A., et al. (2018). A High-Risk, Double-Hit, Group of Newly Diagnosed Myeloma Identified by Genomic Analysis. *Leukemia* 33 (1), 159–170. doi:10.1038/s41375-018-0196-8

Wang, J., Mi, J.-Q., Debernardi, A., Vitte, A.-L., Emadali, A., Meyer, J. A., et al. (2015). A Six Gene Expression Signature Defines Aggressive Subtypes and Predicts Outcome in Childhood and Adult Acute Lymphoblastic Leukemia. *Oncotarget* 6 (18), 16527–16542. doi:10.18632/oncotarget.4113

Wang, W., Xu, S.-w., Zhu, X.-y., Guo, Q.-y., Zhu, M., Mao, X.-l., et al. (2021). Identification and Validation of a Novel RNA-Binding Protein-Related Gene-Based Prognostic Model for Multiple Myeloma. *Front. Genet.* 12, 665173. doi:10.3389/fgene.2021.665173

Wouters, M. C. A., and Nelson, B. H. (2018). Prognostic Significance of Tumor-Infiltrating B Cells and Plasma Cells in Human Cancer. *Clin. Cancer Res.* 24 (24), 6125–6135. doi:10.1158/1078-0432.Ccr-18-1481

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *The Innovation* 2 (3), 100141. doi:10.1016/j.xinn.2021.100141

Yan, S., Zhao, D., Wang, C., Wang, H., Guan, X., Gao, Y., et al. (2021). Characterization of RNA-Binding Proteins in the Cell Nucleus and Cytoplasm. *Analytica Chim. Acta* 1168, 338609. doi:10.1016/j.aca.2021.338609

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* 41 (Database issue), D955–D961. doi:10.1093/nar/gks1111

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, Y., Zhang, J., Liu, Z., Liu, Y., and Tuo, S. (2015). A Network-Based Approach to Identify Disease-Associated Gene Modules through Integrating DNA Methylation and Gene Expression. *Biochem. Biophysical Res. Commun.* 465 (3), 437–442. doi:10.1016/j.bbrc.2015.08.033

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: 17 January 2022
doi: 10.3389/fgene.2021.793494

Prediction of Two Molecular Subtypes of Gastric Cancer Based on Immune Signature

Dan Wu[1], *Mengyao Feng*[1], *Hongru Shen*[1], *Xilin Shen*[1], *Jiani Hu*[1], *Jilei Liu*[1], *Yichen Yang*[1], *Yang Li*[1], *Meng Yang*[1], *Wei Wang*[2], *Qiang Zhang*[3], *Fangfang Song*[2], *Ben Liu*[2], *Kexin Chen*[2]* *and Xiangchun Li*[1]*

[1]*National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin Cancer Institute, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China,* [2]*Department of Epidemiology and Biostatistics, Key Laboratory of Molecular Cancer Epidemiology of Tianjin, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China,* [3]*Department of Maxillofacial and Otorhinolaryngology Oncology, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China*

OPEN ACCESS

Edited by:
Israel Tojal da Silva,
AC Camargo Cancer Center, Brazil

Reviewed by:
Ehsan Nazemalhosseini-Mojarad,
Shahid Beheshti University of Medical Sciences, Iran
Abhinita S. Mohanty,
Memorial Sloan Kettering Cancer Center, United States

Correspondence:
Xiangchun Li
lixiangchun2014@foxmail.com
Kexin Chen
chenkexin@tmu.edu.cn

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

Received: *12 October 2021*
Accepted: *16 December 2021*
Published: *17 January 2022*

Citation:
Wu D, Feng M, Shen H, Shen X, Hu J, Liu J, Yang Y, Li Y, Yang M, Wang W, Zhang Q, Song F, Liu B, Chen K and Li X (2022) Prediction of Two Molecular Subtypes of Gastric Cancer Based on Immune Signature. Front. Genet. 12:793494. doi: 10.3389/fgene.2021.793494

Gastric cancer is the fifth most common type of human cancer and the third leading cause of cancer-related death. The purpose of this study is to investigate the immune infiltration signatures of gastric cancer and their relation to prognosis. We identified two distinct subtypes of gastric cancer (C1/C2) characterized by different immune infiltration signatures. C1 is featured by immune resting, epithelial–mesenchymal transition, and angiogenesis pathways, while C2 is featured by enrichment of the MYC target, oxidative phosphorylation, and E2F target pathways. The C2 subtype has a better prognosis than the C1 subtype (HR = 0.61, 95% CI: 0.44–0.85; log-rank test, $p = 0.0029$). The association of C1/C2 with prognosis remained statistically significant (HR = 0.62, 95% CI: 0.44–0.87; $p = 0.006$) after controlling for age, gender, and stage. The prognosis prediction of C1/C2 was verified in four independent cohorts (including an internal cohort). In summary, our study is helpful for better understanding of the association between immune infiltration and the prognosis of gastric cancer.

Keywords: gastric cancer, immune signature, molecular subtypes, prognosis, computational biology

INTRODUCTION

Gastric cancer (GC) ranks the fifth most commonly diagnosed cancer type globally and the third leading cause for cancer-related death, which was attributed to its diagnosis usually made at an advanced stage. Although gastric cancer incidence has declined in most countries over the past century, the aging population may contribute to increased diagnosis of gastric cancer (Smyth et al., 2020).

Gastric cancer is a highly heterogeneous disease characterized by histopathologic and epidemiologic features based on molecular and phenotypic levels (Van Cutsem et al., 2016). Next-generation sequencing has showed new insights into the heterogeneity of gastric cancer, and subtyping systems have been proposed (Cancer Genome Atlas Research, 2014; Cristescu et al., 2015; Sohn et al., 2017; Kim et al., 2019; Zhang et al., 2020). The Lauren classification system categorizes gastric cancer into the intestinal and diffuse subtypes (Lauren, 1965), while the WHO system divides gastric cancer into four subtypes (papillary, tubular, mucinous, and poorly cohesive) (Hu et al., 2012). Apart from the aforementioned classification subtypes, researchers from TCGA proposed four subtypes for gastric cancer: Epstein–Barr virus (EBV) positive, microsatellite unstable (MSI), genomically stable (GS), and chromosomal instability (CIN). The EBV subtype has the best prognosis among these four subtypes. Patients with the CIN subtype experienced the greatest benefit from adjuvant chemotherapy (Sohn et al., 2017). The Asia Cancer Research Group (ACRG) proposed four molecular subtypes for GC based on microsatellite instability, epithelial–mesenchymal transition, and TP53 mutation: MSI, MSS/EMT, MSS/TP53$^+$, and MSS/TP53$^-$ (Cristescu et al., 2015). The EBV subtype and MSI subtype were reported to potentially benefit from immunotherapy (Pang et al., 2009; Le et al., 2017; Amatatsu et al., 2018; Sundar et al., 2018). However, few gastric cancer subtypes development is based on immune signature and can be used to predict the prognosis of gastric cancer patients.

Immune processes play critical roles in carcinogenesis and progression of solid tumors, and they also affect the treatment and prognosis of patients. Researchers are confused with the association between the immune environment and the prognosis, and much attention has been paid to the tumor immune environment (Cully, 2018; Lim et al., 2018; Kubota et al., 2021). All tumors are potentially immunogenic, and the new knowledge about the interactions between tumor cells, immune cells, and tumor microenvironment allowed for reversal of possible immune resistance (Refolo et al., 2020; Ceresoli and Pasello, 2021). The immune response is a complex multistep process that finely regulates the balance between the recognition of non-self and the prevention of autoimmunity. Cancer cells can use these pathways to suppress tumor immunity as a major mechanism of immune resistance. The recent molecular classifications of gastric cancer by The Cancer Genome Atlas (TCGA) and by the Asian Cancer Research Group (ACRG) networks, together with the identification of multiple biomarkers, open new perspectives for stratification of patients who might benefit from a long-term immune checkpoint therapy (Newman et al., 2015; Jiang et al., 2018; Thorsson et al., 2019; Ceresoli and Pasello, 2021).

The purposes of this study are to characterize different potential molecular classification systems operative in gastric cancer and to identify previously unreported significant immune environments and independent prognosis factors for patients with gastric cancer. We collected 1386 samples from five datasets and applied molecular subtyping on each dataset. We achieved two distinct molecular subtypes of gastric cancer (i.e., C1/C2). The C2 subtype has a better prognosis and more activated immune microenvironment than the C1 subtype.

METHODS

Data Collection and Processing

In total, our study included five gastric cancer datasets: TCGA dataset, three datasets from Gene Expression Omnibus (i.e., GSE62254, GSE15459, and GSE84437), and one internal dataset. TCGA dataset was used as discovery set, whereas the other four datasets as validation sets. The expression matrix of each dataset was normalized individually. The 3151 immune-related genes were collected from previous studies (**Supplementary Table S1**) (Thorsson et al., 2019). We applied removeBatchEffect from the R *limma* package (version 3.34.9) to remove the batch effect while combining these five datasets.

Consensus Molecular Subtyping

We conducted survival analysis in TCGA dataset. Genes significant in survival analysis from TCGA dataset were reversed in follow-up molecular subtyping. We applied consensus non-negative matrix factorization (CNMF) clustering for finding molecular patterns from high-dimensional biological datasets. It is combined with a quantitative evaluation of the robustness of the number of clusters. The CNMF method is included in *CancerSubtypes* (version 1.18.0), an R package for clustering cancer subtypes (Brunet et al., 2004; Xu et al., 2017).

Survival Analysis and Multivariable Cox Regression Analysis

The association of C1/C2 with overall survival was estimated using Kaplan–Meier plots and log-rank tests. Multivariable Cox regression analysis was used to evaluate independent prognostic factors associated with overall survival, including age, gender, and tumor stage. A p value of less than 0.05 was considered statistically significant.

Statistical Analysis

Differentially expressed genes (DEGs) between C1 and C2 were evaluated by *edgeR* package (version 3.28.1). The *clusterProfiler* package (version 3.14.3) was used for pathway enrichment.

The CIBERSORT method was used to characterize the composition of 22 kinds of immune cells from RNA expression (Newman et al., 2015). The t-test was used for comparing the CIBERSORT score, and a p value of less than 0.05 was considered statistically significant.

The mutation of significantly mutated genes (SMGs) in C1/C2 was compared in this part (Li et al., 2016). The chi-square test was used for comparing the proportion of some somatic mutations in between C1 and C2. A p value of less than 0.05 was considered statistically significant.

Tumor mutational burden (TMB) was calculated as the number of mutation events per sample. The t-test was used for comparing the CIBERSORT score, and a p value of less than 0.05 was considered statistically significant. All statistical analyses were done by R software (version 3.4.3).

RESULTS

Patients and Clinical Information

The study flowchart was shown in **Figure 1**. In total, we collected 1386 gastric cancer samples from five datasets. These five datasets include 371 samples from TCGA cohort, 300 from GSE62254, 192 from GSE15459, 433 from GSE84437, and an internal dataset of 90 samples (Tianjin

TABLE 1 | Baseline characteristics.

	TCGA[a] (*n* = 371)	GSE62254 (*n* = 300)	GSE15459 (*n* = 192)	GSE84437 (*n* = 433)	Tianjin (*n* = 90)
Gender					
Male	238 (64%)	199 (67%)	125 (65%)	296 (68%)	59 (32%)
Female	133 (36%)	101 (33%)	67 (35%)	137 (32%)	34 (68%)
Age	68 (36–90)	64 (24–86)	67 (23–92)	62 (27–86)	58 (33–87)
Age ≤60 years male	78 (21%)	72 (24%)	31 (16%)	130 (30%)	28 (31%)
Age >60 years male	156 (42%)	127 (42%)	94 (49%)	166 (38%)	31 (35%)
Age ≤60 years female	31 (8%)	45 (15%)	28 (15%)	64 (15%)	21 (23%)
Age >60 years female	100 (27%)	56 (19%)	39 (20%)	73 (17%)	10 (11%)
Stage					
1	50 (14%)	30 (10%)	31 (16%)	21 (5%)	0
2	111 (30%)	97 (32%)	29 (15%)	138 (32%)	24 (27%)
3	149 (40%)	96 (32%)	72 (38%)	274 (63%)	22 (24%)
4	38 (10%)	77 (26%)	60 (31%)	0	44 (49%)
Missing	23 (6%)	0	0	0	0
Lauren type					
Intestinal	NA	150 (50%)	75 (38%)	NA	NA
Diffuse	NA	142 (47%)	99 (52)	NA	NA
Mixed	NA	8 (3%)	18 (10%)	NA	NA
Survival time (years)	1.21 (0–5)	4.82 (0.08–8.81)	1.58 (0–13.15)	5.75 (0–13.42)	2.62 (0–13.81)

[a] *In TCGA cohort, six samples (2%) have no age information.*

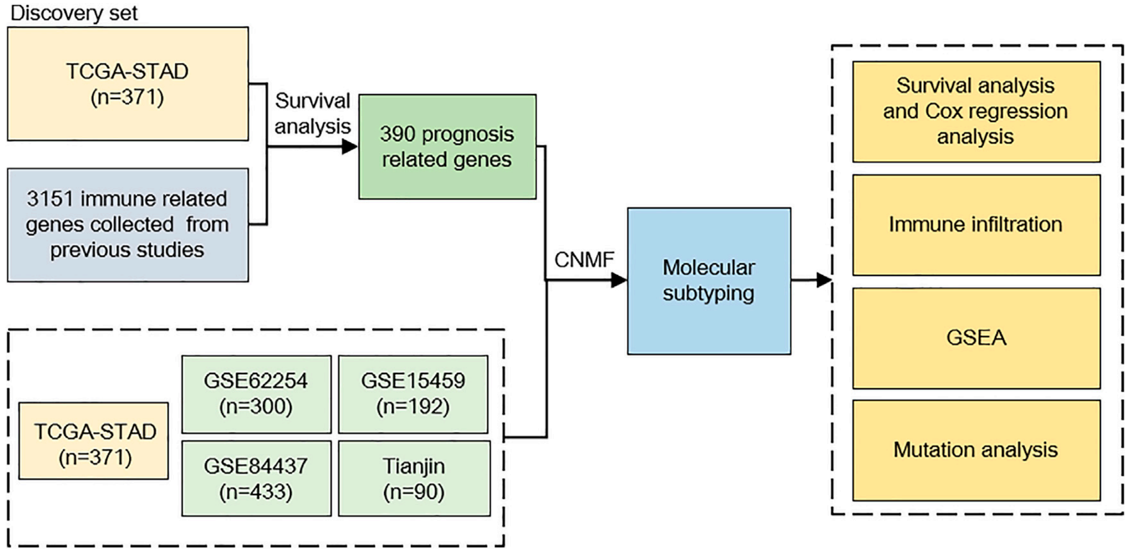


FIGURE 1 | Study flowchart. We incorporated 1386 gastric cancer (GC) samples in this study. TCGA cohort was used as a discovery set; after evaluating prognosis-related genes, 390 genes were selected to predict GC subtypes. The CNMF model was used to develop GC subtypes on five cohorts; after molecular subtyping, we did survival analysis, immune infiltration analysis, GSEA analysis, and mutation analysis to describe different characteristics between our subtypes. CNMF, consensus non-negative matrix factorization; GSEA, gene set enrichment analysis.

cohort). Patients' stages ranged from stage 1 to stage 4 (stage 1: 9.5%; stage 2: 28.8%; stage 3: 44.2%; stage 4:15.8%), and all these five cohorts have clinical information including overall survival (OS), vital stage, age, gender, and tumor stage. In addition to this, GSE62254 cohort and GSE15459 cohort also has Lauren type information. In the GSE62254 cohort, 150 (50%) patients are intestinal type, 142 (47%) patients are diffuse type, and 8 (3%) patients are mixed type. In the GSE15459 cohort, 75 (38%) patients are intestinal type, 99 (52%) patients are diffuse type, and 18 (9%) patients are mixed type. The basic information of these cohorts is reported in **Table 1**.

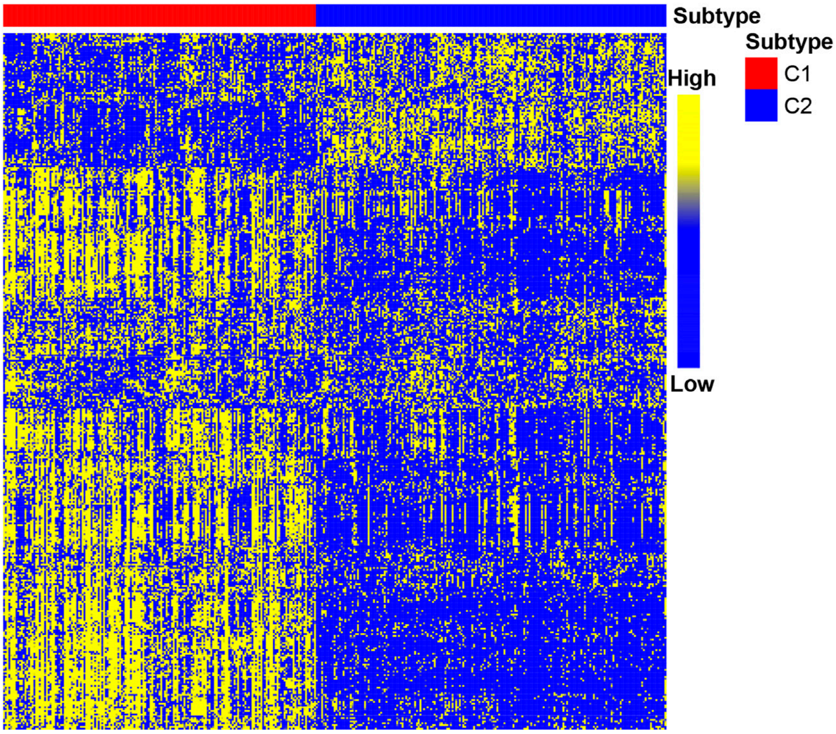


FIGURE 2 | Heat map of C1/C2 in TCGA cohort. TCGA cohort was used as a discovery set, and 390 immune-related genes were selected for the development of subtypes. Data are presented in a matrix format, in which row represents an individual gene and each column represents a sample. The color in the cells reflects relatively high (yellow) and low (blue) expression levels.

Classification of Gastric Cancer Molecular Based on Immune Genes

After determining the correlation with prognosis, 390 genes were chosen to be used to define subtype (**Supplementary Table S2**). TCGA cohort was used as a discovery set, and four other cohorts (GSE62254 cohort; GSE84437 cohort; GSE15459 cohort; Tianjin cohort) were used as the validation sets. The clustering metrics showed that the optimal cluster number is 2 among these five datasets (**Supplementary Figure S1**). These two clusters were regarded as two subtypes of gastric cancer, which are called C1 and C2. The heat map of C1 and C2 is shown in **Figure 2**, which illustrated that the immune gene expression of C1/C2 was significantly different in TCGA cohort. The C1/C2 subtype was compared with other GC microenvironment signatures. The results of Fisher's exact test illustrate that C1/C2 has association with TME signature and immune landscape subtypes in TCGA cohort, and C1/C2 has association with ACRG subtypes in GSE62254 cohort (p < 0.05) (Cristescu et al., 2015; Thorsson et al., 2019; Zeng et al., 2019) (**Supplementary Figure S2**).

C1/C2 Predict the Survival of Gastric Cancer

The prognosis of C2 is significantly better than that of C1 in TCGA (HR = 0.61, 95% CI: 0.44–0.85, log-rank test: p = 0.0029), GSE62254 (HR = 0.64, 95% CI: 0.46–0.88, log-rank test, p = 0.0055), GSE84437 (HR = 0.70, 95% CI: 0.53–0.92, log-rank test, p = 0.0094),Tianjin (HR = 0.47, 95% CI: 0.25–0.86, log-rank test: p = 0.012), and combined cohorts (HR = 0.67, 95% CI: 0.57–0.78, log-rank test: p < 0.0001) (**Figures 3A,B,D–F**). The prognosis with C1 and C2 in the GSE15459 cohort has the same trend as other cohorts (GSE15459 cohort, HR = 0.78, 95% CI: 0.52–1.17, log-rank test, p = 0.24) (**Figure 3C**). In the GSE62254 cohort, C2 has lower recurrence rate (38%) than C1 (56%) (**Figure 4**). The chi-square test results showed a significant difference in the recurrence rate of C1/C2 (p = 0.0048).

In addition, multivariable Cox regression in discovery set and validation set demonstrates that together with age, gender, and stage, C1 and C2 are still independent prognostic factors (TCGA cohort, HR = 0.62, 95% CI: 0.44–0.87, log-rank test: p = 0.006; GSE62254 cohort, HR = 0.66, 95% CI: 0.47–0.97, log-rank test, p = 0.019; GSE15459 cohort, HR = 0.69, 95% CI: 0.45–1.05, log-rank test, p = 0.08; GSE84437 cohort, HR = 0.73, 95% CI: 0.56–0.97, log-rank test, p = 0.029; Tianjin cohort, HR = 0.42, 95% CI: 0.22–0.80, log-rank test: p = 0.008) (**Figures 5A–E**). In the combined cohort, the C2 subtype remained a better prognostic factor than the C1 subtype (**Figure 5F**).

In the combined cohort, the proportion of C1/C2 in four stages was calculated. The proportion of C1 is relatively higher in advanced gastric cancer (stage 3: 42.3%; stage 4: 39.2%) (**Figure 6B**). The proportion of C1/C2 in the Lauren type in the GSE62254 and GSE15459 cohorts was calculated, and C1 accounted for 60.3%, 21.2, and 25.8%, respectively, in the diffuse

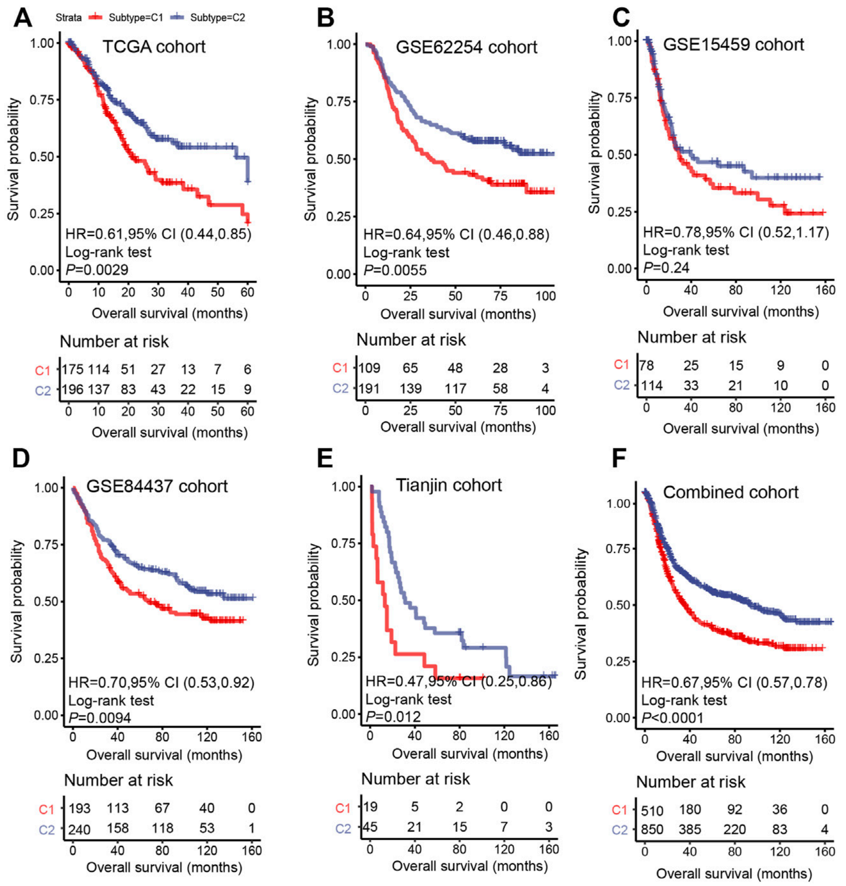


FIGURE 3 | Survival analysis for C1/C2 subtype in **(A)** TCGA cohort, **(B)** GSE62254 cohort, **(C)** GSE15459 cohort, **(D)** GSE84437 cohort, **(E)** Tianjin cohort and **(F)** Combined cohort in C1/C2. Kaplan–Meier plots of overall survival (OS) among patients stratified by C1/C2. Hazard ratio (HR) was calculated by Cox regression analysis. A *p* value was obtained using the log-rank test.

type, intestinal type, and mixed type. It shows that a higher proportion of C1 was found in the diffuse type, which is the most malignant type (**Figure 6C**).

C1 and C2 successfully stratify patients by survival in several gastric cancer cohorts. It is also an independent prognosis factor. The results show the reproducibility and clinical significance of C1/C2.

Biological Characteristics of C1/C2

The results of CIBERSORT demonstrated the immune infiltration of C1/C2 in the combined cohort. Most of the immune cells have significant differences between C1 and C2. In C2, such as T-cell CD4 memory activated, NK cells activated, mast cells activated, and dendritic cells activated, the composition of all these four kinds of activated immune cells was significantly higher than C1 ($p < 0.05$). In contrast, in C1, such as B cells were naive, T-cell memory resting, dendritic cells resting, and mast cells resting, and these kinds of immune resting cells were significantly higher than C2 ($p < 0.05$) (**Figure 6D**).

To further investigate the potential biological behavior of the molecular subtype, the DEGs were used for pathway enrichment in the combined cohort (**Supplementary Table S3**). Finally, 20 cancer-related pathways were enriched (**Figure 6A**). Genes highly expressed in C1 were enriched in "Epithelial Mesenchymal Transition," "Angiogenesis," and "UV Response." Genes highly expressed in C2 were enriched in "MYC Target," "Oxidative Phosphorylation," and "E2F Target."

The driver gene mutation between C1 and C2 was compared in TCGA cohort. It was observed that C2 had significantly more mutation events than C1 in *APC* (p = 0.0024), *NBEA* (p = 0.0026), *PIC3CA* (p = 0.0114), *XIRP2* (p = 0.0131), RNF43 (p = 0.0211), *SMAD4* (p = 0.0369), *TP53* (p = 0.0398), *KRAS* (p = 0.043), and *BNCA* (p = 0.0459), while C1 has significantly more mutation events than C2 in *BNC2* (p = 0.0459), *CDH1* (p = 0.0488), and *CTNNB1* (p = 0.05). The results showed differences in driver genes mutation of C1 and C2 (**Supplementary Table S1**). The tumor mutation burden (TMB) was also calculated, where C2 has higher TMB than C1 ($p < 0.05$) (**Supplementary Figure S3**).

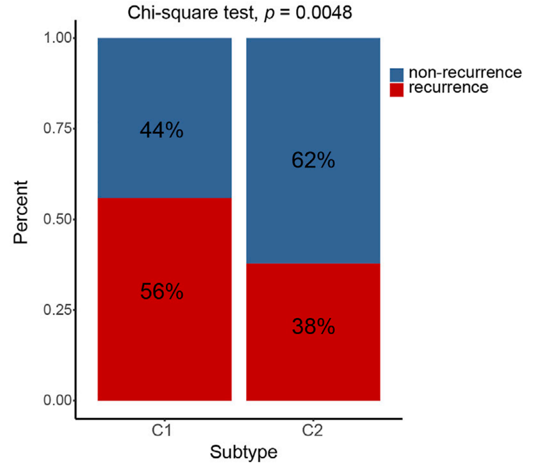
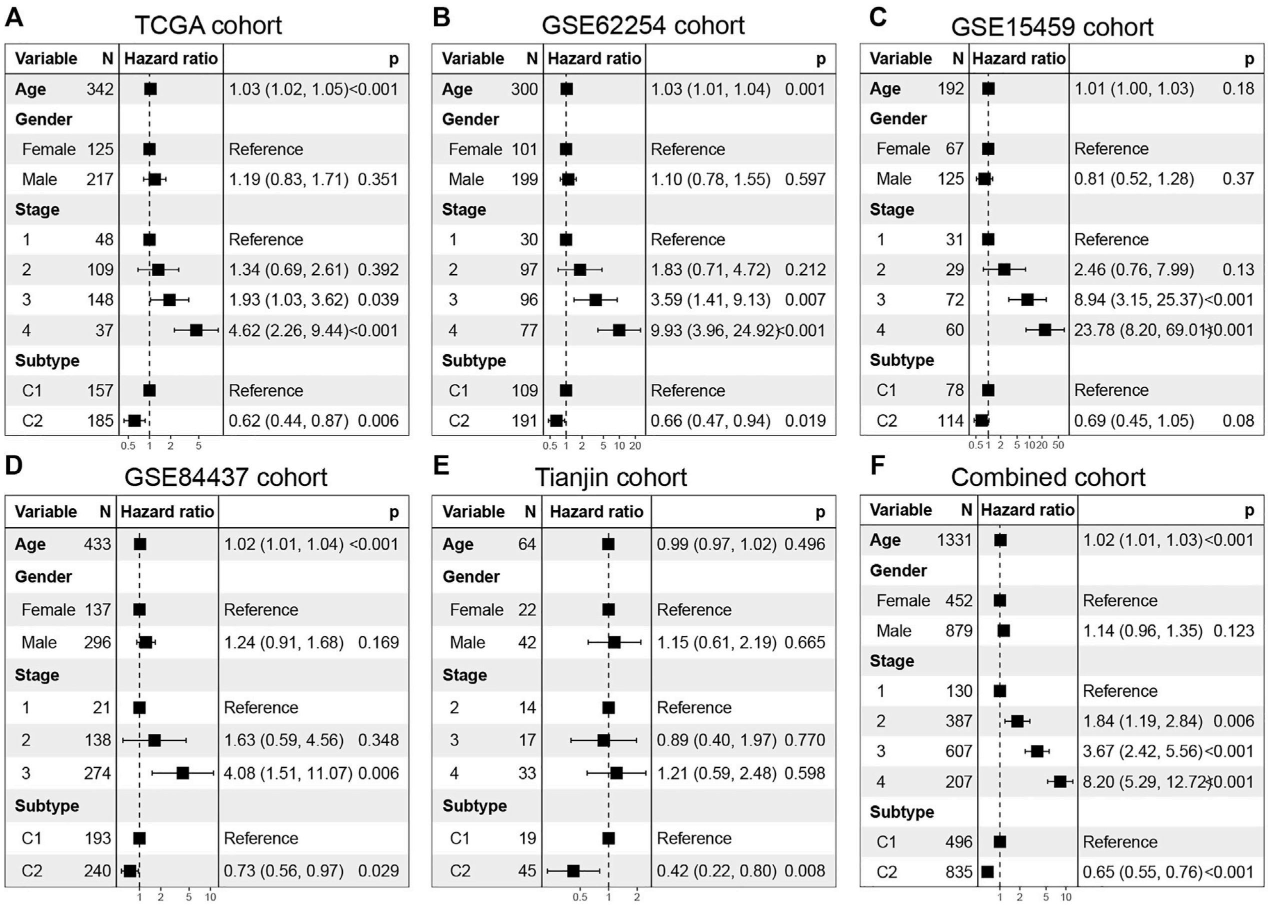


FIGURE 4 | C1/C2 has different recurrence rates in GSE62254 cohort. In GSE62254 cohort, C2 has lower recurrence rate (38%) than C1 (56%). The chi-square test results showed a significant difference in the recurrence rate of C1/C2 (*p* = 0.0048).

DISCUSSION

The clinical significance of the molecular subtype has been demonstrated in many kinds of cancers. However, few researchers have used immune signatures to predict gastric cancer subtypes.

A major clinically relevant finding in this study is based on signature of 390 immune-related genes; we classify gastric cancer into two prognostically distinct subgroups, namely, C1 and C2. The prognostic significance of C1/C2 was independent of age, gender, and stage. The C1 subtype is featured by "Epithelial Mesenchymal Transition (EMT)" and "Angiogenesis," which had poorer overall survival, whereas the C2 subtype is characterized by "MYC Target," "Oxidative Phosphorylation," and "E2F Target," with better overall survival than those in C1. Notably, previous research studies reported that EMT was shown to strongly enhance cancer cell motility and dissemination; it plays an important role in cancer metastasis (Brabletz, 2012; Brabletz et al., 2018). Angiogenesis is essential for the late stages of carcinogenesis, allowing the tumor to grow beyond 1–2 mm in diameter; it is associated with the malignancy of tumor (Sharma et al., 2001; Albini et al., 2012). Such processes may cause poor survival in C1.



TCGA cohort

Variable	N	Hazard ratio		p
Age	342		1.03 (1.02, 1.05)	<0.001
Gender				
Female	125		Reference	
Male	217		1.19 (0.83, 1.71)	0.351
Stage				
1	48		Reference	
2	109		1.34 (0.69, 2.61)	0.392
3	148		1.93 (1.03, 3.62)	0.039
4	37		4.62 (2.26, 9.44)	<0.001
Subtype				
C1	157		Reference	
C2	185		0.62 (0.44, 0.87)	0.006

GSE62254 cohort

Variable	N	Hazard ratio		p
Age	300		1.03 (1.01, 1.04)	0.001
Gender				
Female	101		Reference	
Male	199		1.10 (0.78, 1.55)	0.597
Stage				
1	30		Reference	
2	97		1.83 (0.71, 4.72)	0.212
3	96		3.59 (1.41, 9.13)	0.007
4	77		9.93 (3.96, 24.92)	<0.001
Subtype				
C1	109		Reference	
C2	191		0.66 (0.47, 0.94)	0.019

GSE15459 cohort

Variable	N	Hazard ratio		p
Age	192		1.01 (1.00, 1.03)	0.18
Gender				
Female	67		Reference	
Male	125		0.81 (0.52, 1.28)	0.37
Stage				
1	31		Reference	
2	29		2.46 (0.76, 7.99)	0.13
3	72		8.94 (3.15, 25.37)	<0.001
4	60		23.78 (8.20, 69.01)	<0.001
Subtype				
C1	78		Reference	
C2	114		0.69 (0.45, 1.05)	0.08

GSE84437 cohort

Variable	N	Hazard ratio		p
Age	433		1.02 (1.01, 1.04)	<0.001
Gender				
Female	137		Reference	
Male	296		1.24 (0.91, 1.68)	0.169
Stage				
1	21		Reference	
2	138		1.63 (0.59, 4.56)	0.348
3	274		4.08 (1.51, 11.07)	0.006
Subtype				
C1	193		Reference	
C2	240		0.73 (0.56, 0.97)	0.029

Tianjin cohort

Variable	N	Hazard ratio		p
Age	64		0.99 (0.97, 1.02)	0.496
Gender				
Female	22		Reference	
Male	42		1.15 (0.61, 2.19)	0.665
Stage				
2	14		Reference	
3	17		0.89 (0.40, 1.97)	0.770
4	33		1.21 (0.59, 2.48)	0.598
Subtype				
C1	19		Reference	
C2	45		0.42 (0.22, 0.80)	0.008

Combined cohort

Variable	N	Hazard ratio		p
Age	1331		1.02 (1.01, 1.03)	<0.001
Gender				
Female	452		Reference	
Male	879		1.14 (0.96, 1.35)	0.123
Stage				
1	130		Reference	
2	387		1.84 (1.19, 2.84)	0.006
3	607		3.67 (2.42, 5.56)	<0.001
4	207		8.20 (5.29, 12.72)	<0.001
Subtype				
C1	496		Reference	
C2	835		0.65 (0.55, 0.76)	<0.001

FIGURE 5 | Multivarate Cox regression analysis of C1/C2 subtype in **(A)** TCGA cohort, **(B)** GSE62254 cohort, **(C)** GSE15459 cohort, **(D)** GSE84437 cohort, **(E)** Tianjin cohort and **(F)** Combined cohort. Multivariable Cox regression analysis was used to evaluate independent prognostic factors associated with overall survival, including age, gender, and tumor stage. A *p* value of less than 0.05 was considered statistically significant.

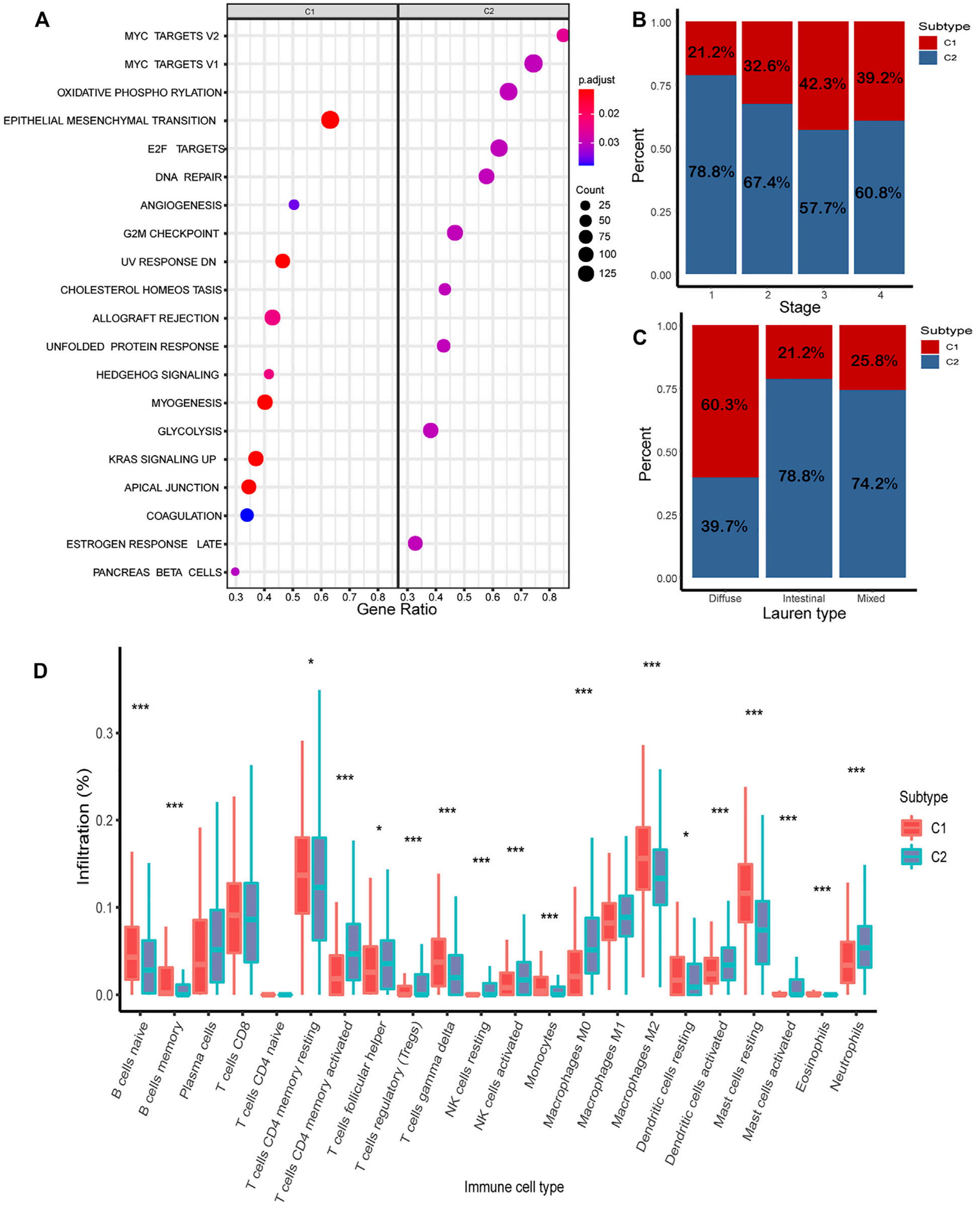


FIGURE 6 | Analysis of biological characteristics of C1/C2. **(A)** Highly expressed genes in C1/C2 were enriched in 20 cancer-related pathways. **(B)** Proportion of C1/C2 in tumor stages 1–4. **(C)** Proportion of C1/C2 in Lauren type. **(D)** Result of CIBERSORT in C1/C2.

The results of CIBERSORT suggest that C1 and C2 have very different immune environments. In C2, such as CD4 memory activated, NK cells activated, mast cells activated, and dendritic cells activated, and the composition of all these four kinds of activated immune cells are significantly higher than C1. In contrast, in C1, such as B cells were naive, T cells memory resting, dendritic cells (DC) resting, and mast cells resting, and these kinds of immune resting cells significantly higher than C2. The microenvironment between C1 and C2 shows a marked difference. Previous research studies reported that memory CD4 T cells could make effector cytokines early in response and they could enhance B-cell and CD8 T-cell responses, which enhance immune response (MacLeod et al., 2009). NK cells are important immune cells; they could swiftly kill multiple adjacent cells which show surface markers associated with oncogenic transformation; and they could also magnify immune responses (Shimasaki et al., 2020). Mast cells are evolutionarily ancient cells, and they finely modulate not only immune responses but also the mechanism of several inflammatory disorders, including cancer, autoimmunity, and infection (Frossi et al., 2017). DCs are a diverse group of specialized antigen-presenting cells; they play key roles in the initiation and regulation of innate and adaptive immune responses (Wculek et al., 2020). The activation of these immune cells in C2 may indicate higher immune activity, which leads better prognosis. The resting of these immune cells may cause poor immune activity in C1, which leads to worse prognosis in C1.

The proportion of C1/C2 in four tumor stages demonstrates C1 has higher proportion in advanced gastric cancer, while C2 has higher proportion in early stages. The proportion of C1/C2 in the Lauren type shows that C1 has the highest proportion in the diffuse type, while C2 has higher proportion in the intestinal type. This indicates that C1 has some characteristics of malignant gastric cancer.

The somatic mutation event of SMGs shows significant differences between C1 and C2, and C2 has higher TMB than C1. Previous research reported that TMB can be used as an indicator to predict the response to immunotherapy, and patients with high TMB were observed to have better clinical outcomes (Gibney et al., 2016; Gandara et al., 2018; Mandal et al., 2019). It also reported that high TMB is associated with a better prognosis in gastric cancer (Cai et al., 2020; Wang et al., 2021). The differences in mutation characteristics may lead to different clinical outcomes of C1/C2, and it also could offer some new insights into immunotherapy in gastric cancer.

In total, in this research, we predict C1 and C2, two subtypes of gastric cancer. Much evidence has shown that there are many different biological characteristics between C1 and C2. It makes two subtypes that could predict prognosis in gastric cancer patients. However, this research still has some limitation. First, the sample size is not large enough; therefore, research may not cover all types of gastric cancer. Second, due to the lack of clinical data, the subtypes in this research could only be used to predict the survival of gastric cancer patients but could not predict their response to chemoradiotherapy and immunotherapy. If more gastric cancer chemotherapy and immunotherapy data could be combined, C1/C2 could be given more clinical significance and immune characteristics could provide more insights into gastric cancer treatment.

Our research has developed two molecular subtypes of gastric cancer, and we have analyzed their immune signature and biological function. These findings may offer some new knowledge of molecular mechanisms for study on treatment of gastric 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

XL and KC designed and supervised the study; DW and XL performed data collection, analysis, and wrote the manuscript; DW, HS, XS, MF, JH, JL, YY, YL, and MY collected the data; and XL, KC, and DW revised the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (no. 31801117 to XL, no. 31900471 to MY, no. 82172894 to KC, no. 82073028 to BL, and no. 82073287 to QZ), Tianjin Municipal Health Commission Foundation (grant no. RC20027 to YL), and IRT_14R40 from the Program for Changjiang Scholars and Innovative Research Team in University in China (KC).

ACKNOWLEDGMENTS

We are grateful for researchers for their generosity to made their data publicly available.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.793494/full#supplementary-material

Supplementary Figure S1 | The CNMF results of five cohorts and combined cohort. The ordinate "Cophenetic" means cophenetic correlation coefficient. We select the rank value where the magnitude of the cophenetic correlation coefficient begins to fall. The ordinate "Dispersion" between 0 and 1 measures the class assignments' robustness concerning random initial conditions, closing to 0 or 1 means the model's stability. In five cohorts and combined cohort, 2 is the best number of clusters.

Supplementary Figure S2 | Association between molecular subtype C1/C2 and tumor microenvironment signature. **(A)** Association between C1/C2 and TME signature in TCGA cohort. **(B)** Association between C1/C2 and immune landscape subtype in TCGA cohort. **(C)** Association between C1/C2 and ACRG subtype in GSE62254 cohort

Supplementary Figure S3 | TMB of C1/C2 in TCGA cohort. TMB of TCGA cohort was calculated. In this part, TMB equal to mutation events occur in each sample. TMB of C2 significantly higher than C1 (p <0.05).

Supplementary Table S1 | Mutation prevalence of driver genes in C1 versus C2 in TCGA cohort.

REFERENCES

Albini, A., Tosetti, F., Li, V. W., Noonan, D. M., and Li, W. W. (2012). Cancer Prevention by Targeting Angiogenesis. *Nat. Rev. Clin. Oncol.* 9 (9), 498–509. doi:10.1038/nrclinonc.2012.120

Amatatsu, M., Arigami, T., Uenosono, Y., Yanagita, S., Uchikado, Y., Kijima, Y., et al. (2018). Programmed Death-Ligand 1 Is a Promising Blood Marker for Predicting Tumor Progression and Prognosis in Patients with Gastric Cancer. *Cancer Sci.* 109 (3), 814–820. doi:10.1111/cas.13508

Brabletz, T. (2012). EMT and MET in Metastasis: where Are the Cancer Stem Cells? *Cancer Cell* 22 (6), 699–701. doi:10.1016/j.ccr.2012.11.009

Brabletz, T., Kalluri, R., Nieto, M. A., and Weinberg, R. A. (2018). EMT in Cancer. *Nat. Rev. Cancer* 18 (2), 128–134. doi:10.1038/nrc.2017.118

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *Proc. Natl. Acad. Sci.* 101 (12), 4164–4169. doi:10.1073/pnas.0308531101

Cai, L., Li, L., Ren, D., Song, X., Mao, B., Han, B., et al. (2020). Prognostic Impact of Gene Copy Number Instability and Tumor Mutation burden in Patients with Resectable Gastric Cancer. *Cancer Commun.* 40 (1), 63–66. doi:10.1002/cac2.12007

Cancer Genome Atlas Research, N. (2014). Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* 513 (7517), 202–209. doi:10.1038/nature13480

Ceresoli, G. L., and Pasello, G. (2021). Immune Checkpoint Inhibitors in Mesothelioma: a Turning point. *The Lancet* 397 (10272), 348–349. doi:10.1016/S0140-6736(21)00147-1

Cristescu, R., Lee, J., Nebozhyn, M., Kim, K.-M., Ting, J. C., Wong, S. S., et al. (2015). Molecular Analysis of Gastric Cancer Identifies Subtypes Associated with Distinct Clinical Outcomes. *Nat. Med.* 21 (5), 449–456. doi:10.1038/nm.3850

Cully, M. (2018). Fibroblast Subtype Provides Niche for Cancer Stem Cells. *Nat. Rev. Cancer* 18 (3), 136. doi:10.1038/nrc.2018.18

Frossi, B., Mion, F., Tripodo, C., Colombo, M. P., and Pucillo, C. E. (2017). Rheostatic Functions of Mast Cells in the Control of Innate and Adaptive Immune Responses. *Trends Immunol.* 38 (9), 648–656. doi:10.1016/j.it.2017.04.001

Gandara, D. R., Paul, S. M., Kowanetz, M., Schleifman, E., Zou, W., Li, Y., et al. (2018). Blood-based Tumor Mutational burden as a Predictor of Clinical Benefit in Non-small-cell Lung Cancer Patients Treated with Atezolizumab. *Nat. Med.* 24 (9), 1441–1448. doi:10.1038/s41591-018-0134-3

Gibney, G. T., Weiner, L. M., and Atkins, M. B. (2016). Predictive Biomarkers for Checkpoint Inhibitor-Based Immunotherapy. *Lancet Oncol.* 17 (12), e542–e551. doi:10.1016/S1470-2045(16)30406-5

Hu, B., El Hajj, N., Sittler, S., Lammert, N., Barnes, R., and Meloni-Ehrig, A. (2012). Gastric Cancer: Classification, Histology and Application of Molecular Pathology. *J. Gastrointest. Oncol.* 3 (3), 251–261. doi:10.3978/j.issn.2078-6891.2012.021

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T Cell Dysfunction and Exclusion Predict Cancer Immunotherapy Response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Kim, J. Y., Kim, W. G., Kwon, C. H., and Park, D. Y. (2019). Differences in Immune Contextures Among Different Molecular Subtypes of Gastric Cancer and Their Prognostic Impact. *Gastric Cancer* 22 (6), 1164–1175. doi:10.1007/s10120-019-00974-4

Kubota, S. I., Takahashi, K., Mano, T., Matsumoto, K., Katsumata, T., Shi, S., et al. (2021). Whole-organ Analysis of TGF-β-Mediated Remodelling of the Tumour Microenvironment by Tissue Clearing. *Commun. Biol.* 4 (1), 294. doi:10.1038/s42003-021-01786-y

Laurén, P. (1965). The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-type Carcinoma. *Acta Pathol. Microbiol. Scand.* 64, 31–49. doi:10.1111/apm.1965.64.1.31

Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., et al. (2017). Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade. *Science* 357 (6349), 409–413. doi:10.1126/science.aan6733

Li, X., Wu, W. K. K., Xing, R., Wong, S. H., Liu, Y., Fang, X., et al. (2016). Distinct Subtypes of Gastric Cancer Defined by Molecular Characterization Include Novel Mutational Signatures with Prognostic Capability. *Cancer Res.* 76 (7), 1724–1732. doi:10.1158/0008-5472.CAN-15-2443

Lim, B., Woodward, W. A., Wang, X., Reuben, J. M., and Ueno, N. T. (2018). Author Correction: Inflammatory Breast Cancer Biology: the Tumour Microenvironment Is Key. *Nat. Rev. Cancer* 18 (8), 526. doi:10.1038/s41568-018-0022-7

MacLeod, M. K. L., Clambey, E. T., Kappler, J. W., and Marrack, P. (2009). CD4 Memory T Cells: what Are They and what Can They Do? *Semin. Immunol.* 21 (2), 53–61. doi:10.1016/j.smim.2009.02.006

Mandal, R., Samstein, R. M., Lee, K.-W., Havel, J. J., Wang, H., Krishna, C., et al. (2019). Genetic Diversity of Tumors with Mismatch Repair Deficiency Influences Anti-PD-1 Immunotherapy Response. *Science* 364 (6439), 485–491. doi:10.1126/science.aau0447

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Pang, M.-F., Lin, K.-W., and Peh, S.-C. (2009). The Signaling Pathways of Epstein-Barr Virus-Encoded Latent Membrane Protein 2A (LMP2A) in Latency and Cancer. *Cell Mol Biol Lett* 14 (2), 222–247. doi:10.2478/s11658-008-0045-2

Refolo, M. G., Lotesoriere, C., Messa, C., Caruso, M. G., and D'Alessandro, R. (2020). Integrated Immune Gene Expression Signature and Molecular Classification in Gastric Cancer: New Insights. *J. Leukoc. Biol.* 108 (2), 633–646. doi:10.1002/JLB.4MR0120-221R

Sharma, R. A., Harris, A. L., Dalgleish, A. G., Steward, W. P., and O'Byrne, K. J. (2001). Angiogenesis as a Biomarker and Target in Cancer Chemoprevention. *Lancet Oncol.* 2 (12), 726–732. doi:10.1016/S1470-2045(01)00586-1

Shimasaki, N., Jain, A., and Campana, D. (2020). NK Cells for Cancer Immunotherapy. *Nat. Rev. Drug Discov.* 19 (3), 200–218. doi:10.1038/s41573-019-0052-1

Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C., and Lordick, F. (2020). Gastric Cancer. *The Lancet* 396 (10251), 635–648. doi:10.1016/S0140-6736(20)31288-5

Sohn, B. H., Hwang, J.-E., Jang, H.-J., Lee, H.-S., Oh, S. C., Shim, J.-J., et al. (2017). Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by the Cancer Genome Atlas Project. *Clin. Cancer Res.* 23, 4441–4449. doi:10.1158/1078-0432.CCR-16-2211

Sundar, R., Qamra, A., Tan, A. L. K., Zhang, S., Ng, C. C. Y., Teh, B. T., et al. (2018). Transcriptional Analysis of Immune Genes in Epstein-Barr Virus-Associated Gastric Cancer and Association with Clinical Outcomes. *Gastric Cancer* 21 (6), 1064–1070. doi:10.1007/s10120-018-0851-9

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2019). The Immune Landscape of Cancer. *Immunity* 51 (2), 411–412. doi:10.1016/j.immuni.2019.08.004

Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K., and Prenen, H. (2016). Gastric Cancer. *The Lancet* 388 (10060), 2654–2664. doi:10.1016/S0140-6736(16)30354-3

Wang, D., Wang, N., Li, X., Chen, X., Shen, B., Zhu, D., et al. (2021). Tumor Mutation burden as a Biomarker in Resected Gastric Cancer via its Association with Immune Infiltration and Hypoxia. *Gastric Cancer* 24 (4), 823–834. doi:10.1007/s10120-021-01175-8

Wculek, S. K., Cueto, F. J., Mujal, A. M., Melero, I., Krummel, M. F., and Sancho, D. (2020). Dendritic Cells in Cancer Immunology and Immunotherapy. *Nat. Rev. Immunol.* 20 (1), 7–24. doi:10.1038/s41577-019-0210-z

Xu, T., Le, T. D., Liu, L., Su, N., Wang, R., Sun, B., et al. (2017). CancerSubtypes: an R/Bioconductor Package for Molecular Cancer Subtype Identification, Validation and Visualization. *Bioinformatics* 33 (19), 3131–3133. doi:10.1093/bioinformatics/btx378

Zeng, D., Li, M., Zhou, R., Zhang, J., Sun, H., Shi, M., et al. (2019). Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures. *Cancer Immunol. Res.* 7 (5), 737–750. doi:10.1158/2326-6066.CIR-18-0436

Zhang, B., Wu, Q., Li, B., Wang, D., Wang, L., and Zhou, Y. L. (2020). m6A Regulator-Mediated Methylation Modification Patterns and Tumor Microenvironment Infiltration Characterization in Gastric Cancer. *Mol. Cancer* 19 (1), 53. doi:10.1186/s12943-020-01170-0

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 February 2022
doi: 10.3389/fgene.2022.806842

MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis

Xiao Li[1†], Jie Ma[1†], Ling Leng[2], Mingfei Han[1], Mansheng Li[1], Fuchu He[1]* and Yunping Zhu[1]*

[1]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing, China, [2]Stem Cell and Regenerative Medicine Lab, Department of Medical Science Research Center, State Key Laboratory of Complex Severe and Rare Diseases, Translational Medicine Center, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

In light of the rapid accumulation of large-scale omics datasets, numerous studies have attempted to characterize the molecular and clinical features of cancers from a multi-omics perspective. However, there are great challenges in integrating multi-omics using machine learning methods for cancer subtype classification. In this study, MoGCN, a multi-omics integration model based on graph convolutional network (GCN) was developed for cancer subtype classification and analysis. Genomics, transcriptomics and proteomics datasets for 511 breast invasive carcinoma (BRCA) samples were downloaded from the Cancer Genome Atlas (TCGA). The autoencoder (AE) and the similarity network fusion (SNF) methods were used to reduce dimensionality and construct the patient similarity network (PSN), respectively. Then the vector features and the PSN were input into the GCN for training and testing. Feature extraction and network visualization were used for further biological knowledge discovery and subtype classification. In the analysis of multi-dimensional omics data of the BRCA samples in TCGA, MoGCN achieved the highest accuracy in cancer subtype classification compared with several popular algorithms. Moreover, MoGCN can extract the most significant features of each omics layer and provide candidate functional molecules for further analysis of their biological effects. And network visualization showed that MoGCN could make clinically intuitive diagnosis. The generality of MoGCN was proven on the TCGA pan-kidney cancer datasets. MoGCN and datasets are public available at https://github.com/Lifoof/MoGCN. Our study shows that MoGCN performs well for heterogeneous data integration and the interpretability of classification results, which confers great potential for applications in biomarker identification and clinical diagnosis.

Keywords: multi-omics integration, graph convolutional network, autoencoder, similarity network fusion, cancer subtype classification

Edited by:
Joel Correa Da Rosa,
Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Bolin Chen,
Northwestern Polytechnical University, China
Jiazhou Chen,
South China University of Technology, China

Correspondence:
Yunping Zhu
zhuyunping@ncpsb.org.cn
Fuchu He
hefc@bmi.ac.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: *01 November 2021*
Accepted: *14 January 2022*
Published: *02 February 2022*

Citation:
Li X, Ma J, Leng L, Han M, Li M, He F and Zhu Y (2022) MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. Front. Genet. 13:806842. doi: 10.3389/fgene.2022.806842

1 INTRODUCTION

Owing to the recent rapid developments in high-throughput sequencing technology, multi-omics research has strongly promoted the development of precision medicine. However, the application of precision medicine for the prevention, diagnosis, and treatment of tumors is far from satisfactory (Lu and Zhan, 2018). Multi-omics approaches are novel frameworks that can integrate multiple omics datasets generated from the same patients (Heo et al., 2021); thus, an increasing number of studies have tried to characterize the molecular and clinical features of cancers from a multi-omics perspective (Sun and Hu, 2016).

Integrated multi-omics approaches can be divided into two types: the integration of Euclidean structure data or the integration of non-Euclidean structure data (Eicher et al., 2020). The first approach uses the expression matrix as the input, and then trains machine learning models for clustering and classification. For example, Chaudhary et al. were the first to use a deep autoencoder (AE) (Hinton and Salakhutdinov, 2006) model to predict the survival of patients with hepatocellular carcinoma (Chaudhary et al., 2018); Chen et al. designed a deep-learning framework, DeepType, that performs a joint model of supervised classification, unsupervised clustering, and dimensionality reduction to learn cancer-relevant data representation (Chen et al., 2020). These methods can handle large-scale datasets, but require substantial effort to interpret how specific features contribute to the predicted results. On the other hand, the non-Euclidean data integration approach trains models using the network topology data. These methods can identify cancer subtypes by fusing the similarities derived from various omics data, such as similarity network fusion (SNF) (Wang et al., 2014), GrassmannCluster (Ding et al., 2019), and high-order path elucidated similarity (HOPES) (Xu et al., 2019). These network-based processes are clinically intuitive, but existing studies have focused on the unsupervised integration of multi-omics datasets.

Meanwhile, classification of tumor subtypes plays a leading role in the treatment and prognosis of cancers. This is a multi-class classification task and has always presented a challenge in the field of integrating multi-omics using machine learning. There is an urgent need for a multi-class network classification model to handle cancer subtype classification and biomarker identification. Graph convolutional network (GCN) (Kipf and Welling, 2017) is a recently developed approach to incorporate graph structures into a deep learning framework. It classifies unlabeled nodes using information from the topology of the network as well as the feature vectors of the nodes. The network structure makes GCN naturally interpretable. Several studies have been reported to use this model to predict the complex genome-disease association (Xuan et al., 2019) and drug-disease associations (Liu et al., 2020; Yu et al., 2021).

Herein, we developed MoGCN, a multi-omics integration model based on graph convolutional network, for cancer subtype analysis. This study creatively proposes developing a network diagnosis model based on the pipeline of "integrating multi-omics data first and then performing classification". Specifically, we utilized AE to integrate multi-omics expression data and SNF to integrate a typical network topology data patient similarity network (PSN) (Pai and Bader, 2018), to construct a comprehensive view of cancer patients. Then, we used GCN to combine the AE and SNF results and construct the final model for cancer subtype classification. By applying MoGCN on the breast invasive carcinoma (BRCA) data in The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/), we demonstrated that MoGCN could achieve the best performance in cancer subtype classification among the current algorithms. Similarly, MoGCN achieved good results on the TCGA pan-kidney cancer validation dataset. The case study for breast cancer also shows that our method has great potential for heterogeneous data integration, marker identification, and clinical diagnosis.

2 MATERIALS AND METHODS

MoGCN uses multi-omics expression datasets from patients as inputs, including but not limited to genomics, transcriptomics, and proteomics datasets. First, we applied the autoencoder model to extract patient expression features (expression matrix), and applied the similarity network fusion model to construct a patient similarity network. Then, we used the GCN model to integrate these two types of heterogeneous features and to train the cancer subtype classification model. By integrating network and vector characteristics, MoGCN was able to achieve good classification performance, and effectively addressed the issue of deep learning interpretability in clinical applications. MoGCN is a command-line tool that allows users to integrate multi-omics datasets for cancer subtyping classifications efficiently. The overall workflow is shown in **Figure 1**.

2.1 Data Preparation

BRCA datasets were downloaded from the UCSC Xena browser (https://xenabrowser.net/) and the Cancer Proteome Atlas (TCPA) portal (https://tcpaportal.org/tcpa/) and processed. Copy number variation (CNV) data at the genomic level, RNA-seq data at the transcriptomic level, reverse phase protein array data (RPPA) at the proteomic level, and clinical data were all available. The breast tumors were classified into four subtypes (Cancer Genome Atlas Network, 2012): Basal-like, typically with no expres-sion of hormone receptors or ERBB2; Her2-enriched, overexpressing the oncogene ERBB2; and Luminal A and B, generally estrogen receptor (ER)-positive tumors expressing epithelial markers (Luminal B shows a higher Ki67 index and worse prognosis than Luminal A); these were similar to results generated by the established and widely used PAM50 assay. Common samples were collected from each omics level; therefore, data from a total of 511 patients in the BRCA dataset were obtained. The details of the dataset are shown in **Table 1**.

A 10-fold cross validation method was applied to all algorithms implemented in this study. The dataset of 511 samples was first divided randomly into 10 subsets. We successively selected one subset to become a testing dataset, while the others were used as a training dataset. Therefore, 10 combinations of the training dataset and testing dataset were obtained. In each run, we used the training dataset to train the model and the testing dataset to test the model's performance; the average result of the 10 runs was taken as the final result of the model.

2.2 Multi-Modal Autoencoder

The autoencoder consists of two modules simultaneously: an encoder and a decoder. The encoder (*f*) maps the original domain *X* to a new domain named latent space *Z* with dimension *L*. Then, the decoder (*g*) maps *Z* back to the original input space *X*. The encoder and decoder are defined as $z = f(x)$ and $\tilde{x} = g(z)$. By

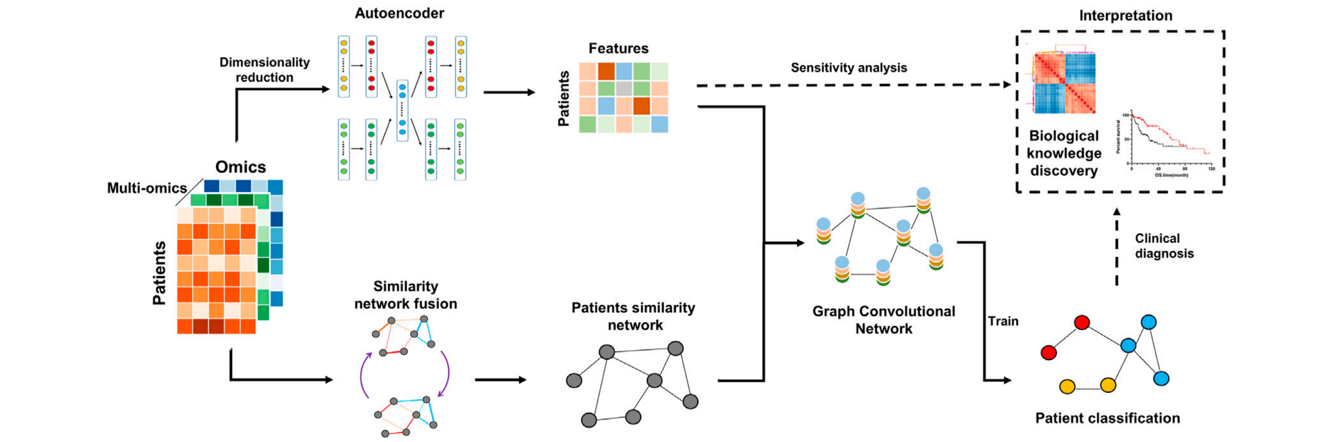


FIGURE 1 | MoGCN workflow schematic. The input is the multi-omics data. First, the AE and SNF methods are used to reduce dimensionality and to construct the patient similarity network, respectively. Next, the vector features and adjacency matrix are fed into the GCN for training. Feature extraction and network visualization can be used for further biological knowledge discovery.

TABLE 1 | Summary of the BRCA dataset.

Number of samples		Number of features	
Basal-like	112	CNV	19,273
Her2-enriched	53	mRNA	19,580
Luminal A	248	RPPA	223
Luminal B	98	—	—
Total	511	Total	39,076

minimizing the reconstruction loss, the model captures the significant features of the data. The loss function to minimize is formalized as: $E = argmin_{f,g}\ [Loss\ (x,\ g\ (f\ (x)))]$.

As the input data are characterized by multi-omics data types and represented by multiple matrices X_1, X_2, X_n, corresponding to the genome, transcriptome, proteome matrices, and so on, the autoencoder must have multiple inputs and outputs. A multi-modal autoencoder architecture is proposed. As shown in **Figure 1**, the model consists of multiple encoders and decoders, which share the same latent layer. The loss function to minimize is formalized as:

$$E = argmin_{f,g}\left(\alpha Loss_1\left(x1, g1\left(f1\left(x1\right)\right)\right) + \ldots + \beta Loss_2\left(x1, g1\left(f1\left(x1\right)\right)\right)\right) \tag{1}$$

Where $\alpha, \ldots, \beta$ are the weights (prior knowledge) of each data type, and $\alpha + \ldots + \beta = 1$.

2.3 Similarity Network Fusion

The SNF algorithm integrates different types of omics data, creating a network for each data type, and ultimately establishing a comprehensive view of the disease or biological process. SNF is able to compute and fuse PSNs for each data type, which enable the exploitation of complementary information from multi-omics data types and outperforms other single data analysis methods. Specifically, the algorithm computes patient-patient similarity matrices for each data type and constructs patient-patient similarity networks. Then, network fusion is performed to enhance strong connections and remove weak connections. Finally, a fused patient similarity network is established.

Based on the assumption that there were n samples (such as patients) and m different data types, for the v_{th} ($v = 1, 2, \ldots, m$) data type, a scaled exponential similarity matrix was calculated:

$$W\left(i, j\right) = exp\left(-\frac{\rho^2\left(x_i, x_j\right)}{\mu\varepsilon_{i,j}}\right) \tag{2}$$

$\rho\left(i, j\right)$ represents the Euclidean distance between the patient x_i and x_j $W\left(i, j\right)$ represents the $n \times n$ similarity matrix between patient x_i and x_j μ is a hyperparameter that can be empirically set, and ε is used to eliminate the scaling problem. Then, the similarity matrix $P^{(v)}$ of all patients and K-nearest similarity matrix $S^{(v)}$ can be defined as

$$P\left(i, j\right) = \begin{cases} \dfrac{W\left(i, j\right)}{2\sum_{k \neq i} W\ \left(i, k\right)}, & j \neq i \\ \dfrac{1}{2}, & j = i \end{cases},$$
$$S\left(i, j\right) = \begin{cases} \dfrac{W\left(i, j\right)}{\sum_{k \in N_i} W\left(i, k\right)}, & j \in N_i \\ 0, & otherwise \end{cases} \tag{3}$$

Then, for the case in which there was two types of data, the process was as follows:

a Calculate $P^{(1)}, P^{(2)}, S^{(1)}, S^{(2)}$. Let $P^{(1)}_{t=0} = P^{(1)}$ and $P^{(2)}_{t=0} = P^{(2)}$ represent the initial two status matrices at $t = 0$.
b Iteratively update the similarity matrix.

$$P^{(1)}_{t+1} = S^{(1)} \times P^{(2)}_t \times \left(S^{(1)}\right)^T, P^{(2)}_{t+1} = S^{(2)} \times P^{(1)}_t \times \left(S^{(2)}\right)^T \tag{4}$$

c After t steps, the overall state matrix can be calculated by:

$$P^{(t)} = \frac{P_t^{(1)} + P_t^{(2)}}{2} \tag{5}$$

For the generalization of $m > 2$, the update process is:

$$P^{(v)} = S^{(v)} \times \left(\frac{\sum_{k \neq v} P^{(k)}}{m-1} \right) \times \left(S^{(v)} \right)^T, \; v = 1, 2, \ldots, m \tag{6}$$

2.4 Graph Convolutional Network

GCN analysis requires two inputs: the structure of the graph and the features of each node. In this manual, one input is the multi-omics feature matrix $X \in R^{n \times d}$, where n is the number of nodes and d is the number of features. Another input is the PSN, which can be represented by the form of an adjacency matrix $A \in R^{n \times n}$. The GCN is built by stacking multiple convolutional layers. Specifically, each layer is defined as:

$$H^{(1+1)} = \sigma\left(LH^{(I)}W^{(I)}\right) \tag{7}$$

Where $L = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ or $L = \tilde{D}^{-1}\tilde{A}$ denotes the normalized graph laplacian; $\widetilde{A} = A + I$ denotes the adjacency matrix with added self-connections; $\tilde{D}$ is the degree matrix of $\tilde{A}$; W is the weight matrix learned from training; σ denotes the nonlinear activation function, generally the *ReLU* activation function; and $H^{(l)}$ is the input of each layer, and notably, $H^{(0)} = X$.

2.5 Interpretability of MoGCN

Machine learning has great potential for improving products, processes, and research. However, computers usually do not explain their predictions, which is a barrier to the adoption of machine learning. In this study, the interpretability of MoGCN is reflected in both AE feature extraction and PSN visualization. In the autoencoder model, we used sensitivity analysis (Garson, 1991) for feature extraction: 1) multiplying the standard deviation of each input node by its connection weights in the network; 2) extracting top features every 10 epochs; and 3) merging and sorting the extracted features. The weights analysis method allows feature extraction during the training process, but consumes relatively little extra time. Meanwhile, the visualization of the PSN also provides an intuitive explanation for the clinical diagnosis of patient subtyping.

Sensitivity analysis is a valuable method used to describe the importance of input variables in neural networks quantitatively. The importance of a node can be determined by the variance of this feature (also known as variable sensitivity) and the weighted connections that the node contributes to the network (also known as weight sensitivity). Therefore, the importance score of a feature x_i can be defined as:

$$S_i = \sigma_i \times \sum_{j=1}^{L} \left| W_{ij} \right| \tag{8}$$

Where σ_i represents the standard deviation of x_i, L is the number of nodes in the next layer, and W is the connection weight of the input nodes to the output nodes.

In order to obtain stable characteristics of AE during training, the process for each omics layer is as follows:

a Calculate *Si* and extract top *N* features every 10 epochs to get feature sets G_1, G_2, . . ., G_m.
b After training, merge G_1, G_2, . . ., G_m and obtain the stable set of essential genes.

The case study on breast cancer demonstrates the promising potential of MoGCN in biological knowledge mining.

2.6 Mainstream Feature Extraction Methods and Classification Methods

We compared AE with the following unsupervised feature extraction algorithms: principal component analysis (PCA), factor analysis (FA), independent component analysis (ICA), and singular value decomposition (SVD). These methods were implemented by calling the built-in functions in the Python scikit-learn library (https://scikit-learn.org/stable/).

We compared GCN with the following state-of-the-art methods: decision tree (DT), K-nearest neighbors (KNN), Gaussian naïve Bayes (GNB), random forests (RF), support vector machine (SVM), deep neural network (DNN) with four layers, GrassmannCluster and HOPES. GrassmannCluster and HOPES were implemented using Matlab. Moreover, other methods were also implemented by calling the built-in functions in the Python scikit-learn library (https://scikit-learn.org/stable/).

2.7 Evaluation Index of Model Performance

In the classification tasks, the prediction results of a model have four basic indicators: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The accuracy represents the proportion of all samples judged correctly by the classifier, and is defined as:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

The *F1* score is a measure of classification tasks. It is often used as the final evaluation index in most machine learning competitions. It is the harmonic average of the precision rate and the recall rate, which has a maximum of 1 and a minimum of 0. It is defined as:

$$F1\,score = 2 \times \frac{precision \times recall}{precision + recall} \tag{10}$$

Where $precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN}$.

In addition, all results were subjected to 10-fold cross validation.

2.8 Functional Enrichment Analysis

Biological Process (BP) annotation, Molecular Function (MF) annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses for selected genes were conducted using David (https://david.ncifcrf.gov/). Gene set

TABLE 2 | The accuracy of different dimensionality reduction algorithms.

	PCA	FA	ICA	SVD	AE
mRNA	0.8318 ± 0.0427	0.7808 ± 0.0351	0.6889 ± 0.0301	0.8278 ± 0.0380	0.8357 ± 0.0396
CNV	0.6008 ± 0.0417	0.5949 ± 0.0263	0.5030 ± 0.0339	0.6047 ± 0.0488	0.5695 ± 0.0497
RPPA	0.7730 ± 0.0199	0.7495 ± 0.0411	0.5440 ± 0.0294	0.7847 ± 0.0474	0.8082 ± 0.0438
mRNA + CNV + RPPA	0.8258 ± 0.0459	0.7044 ± 0.0440	0.6283 ± 0.0441	0.8337 ± 0.0402	0.8787 ± 0.0477

**10-fold cross validation (mean ± standard deviation).*

TABLE 3 | The F1 score of different dimensionality reduction algorithms.

	PCA	FA	ICA	SVD	AE
mRNA	0.8086 ± 0.0534	0.7499 ± 0.0450	0.6226 ± 0.0324	0.8129 ± 0.0428	0.8144 ± 0.0520
CNV	0.5578 ± 0.0504	0.5493 ± 0.0231	0.4295 ± 0.0410	0.5461 ± 0.0413	0.5209 ± 0.0548
RPPA	0.7313 ± 0.0388	0.7098 ± 0.0547	0.4430 ± 0.0438	0.7498 ± 0.0573	0.7935 ± 0.0489
mRNA + CNV + RPPA	0.8080 ± 0.0490	0.6541 ± 0.0489	0.5670 ± 0.0542	0.8172 ± 0.0493	0.8722 ± 0.0529

**10-fold cross validation (mean ± standard deviation).*

variation analysis (GSVA) (Hanzelmann et al., 2013) was performed on the MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/) "c2. cp.reactome.v7.4. symbols.gmt" gene set using the "GSVA" package in R software. p-value of <0.05 was considered statistically significant.

2.9 Kaplan-Meier Survival Analysis

We used the validation breast cancer cohort (n = 1880) from Kaplan–Meier (KM) plotter (https://kmplot.com/analysis/) to validate the prognostic value of genes. 10-years overall survival analysis was performed.

3 RESULTS

3.1 Multi-Omics Integration Using AE Can Improve Classification Performance

Multi-omics data sets are inherently high-dimensional, and their processing may be computationally intensive. Dimensionality reduction is a general strategy to reduce computational burden. Moreover, multi-omics data are highly heterogeneous and the relationship between different data types (also named as layers in data form) is not linear. The extraction of important features from the various layers remains a huge challenge. Here, we used random forest as a benchmark classifier to compare the performance of different dimensionality reduction algorithms (**Tables 2**, **3**). The results in the rows show that AE performs best in most cases. More importantly, the results in the columns show that after the integration of different omics features, the performance of AE-based classification improved, whereas that of other methods slightly decreased or remained unchanged. The potential reasons for this were: 1) a large amount of noise in multi-omics data, so the classification relative information density is low, which interferes with the traditional algorithms; 2) traditional algorithms are linear methods and cannot uncover potential nonlinear relationships within complex biological data.

3.2 Integration of PSN for Greater Performance Improvement

After integration of multi-omics data using the AE, we applied the SNF model to construct the patient similarity network (PSN) (**Figure 1**). Then, we used GCN to integrate the expression data and PSNs to establish a complete pipeline for multi-omics biological data. We compared MoGCN with DT, KNN, GNB, RF, SVM, DNN, GrassmannCluster, and HOPES. Considering that GrassmannCluster and HOPES are algorithms used to construct PSNs and cannot be directly used for classification, we used the GCN to integrate the GrassmannCluster or HOPES algorithm for classification separately. The results showed that the MoGCN method was able to achieve state-of-the-art classification results (**Figure 2**). The standard deviation of MoGCN was the smallest among all compared methods, indicating that integration of the vector features and the PSN can improve the overall prediction stability. In addition, we found that using the features extracted by AE can help other classification algorithms improve their classification performance further (**Supplementary Table S1**). We also implemented the ablation experiments to prove that a combination of AE and SNF with GCN (MoGCN) can achieve better prediction performance. As AE and SNF are both unsupervised algorithms, the classifier method GCN is needed for subtype prediction. As shown in **Figure 2**, MoGCN performed better than AE + GCN and SNF + GCN in accuracy and F1 score.

3.3 The Interpretability of MoGCN From AE Feature Extraction and PSN Visualization

3.3.1 AE Captured Cancer Gene Mutation Patterns at the CNV Level

We trained the AE for 100 epochs to converge, extracted top 100 genes with the highest scores every 10 epochs, and finally obtained 183 genes. The BP enrichment analysis of the top-

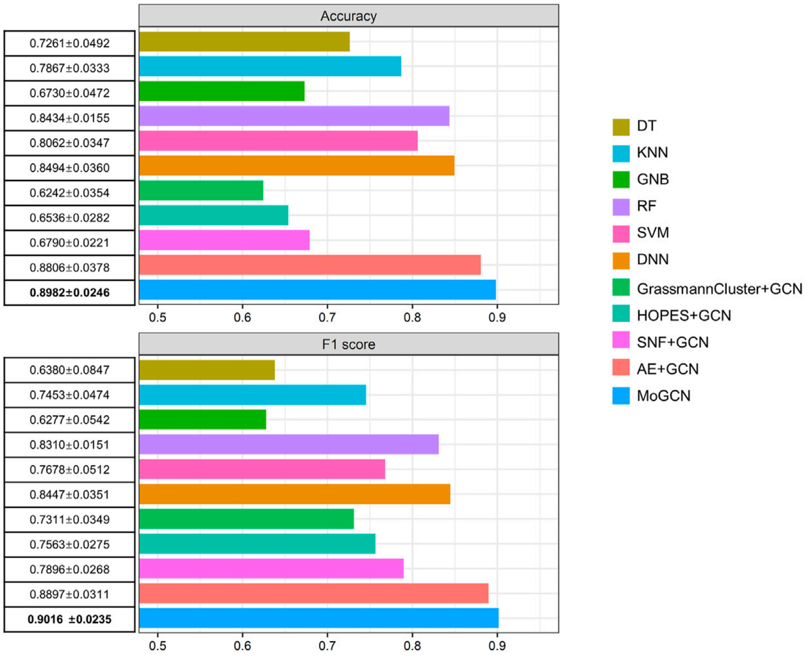


FIGURE 2 | Performance comparison of different algorithms. 10-fold cross validation (mean ± standard deviation).

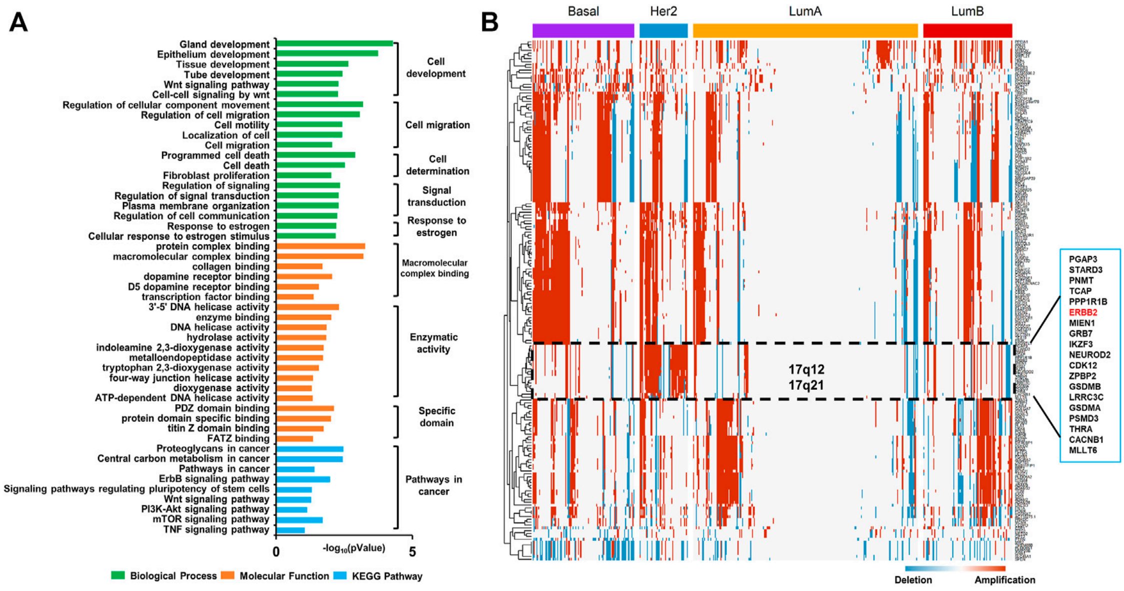


FIGURE 3 | Copy number variation characteristics of breast cancer. **(A)** Biological Process (BP), Molecular Function (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations for the top-scoring genes using David (p < 0.05). **(B)** Hierarchical clustering heat map of the mutation distribution of the top-scoring genes selected by AE.

scoring genes using David showed that their biological function focused on cell development, cell migration, cell death, signal transduction, and response to estrogen (**Figure 3A**). The KEGG annotation showed that these genes are significantly enriched in the Wnt, ErbB, PI3K-AKT-mTOR, and tumor necrosis factor (TNF) signaling pathways. The Wnt signaling pathway is highly

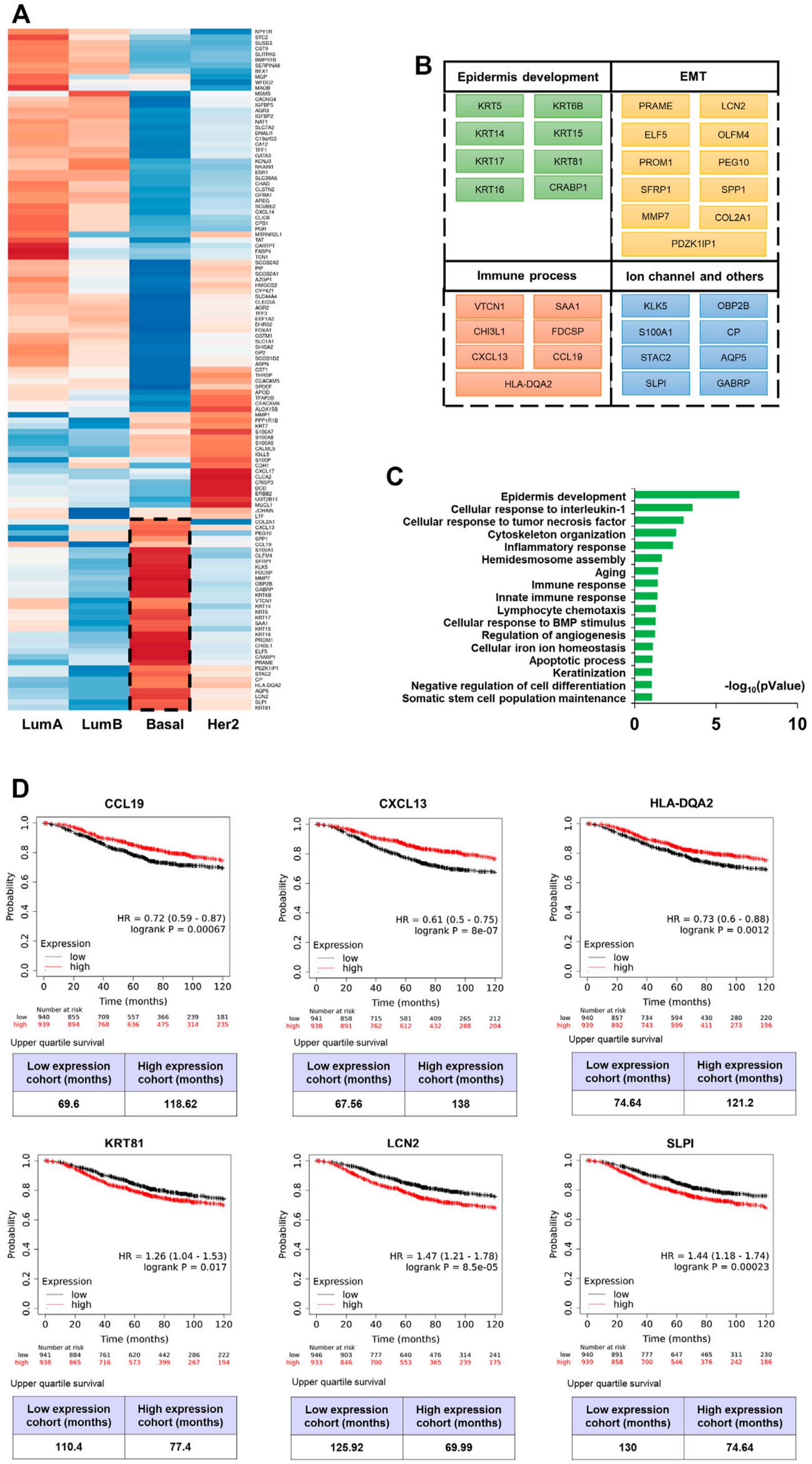


FIGURE 4 | mRNA molecular characteristics of breast cancer. **(A)** Hierarchical clustering heat map of the top-scoring genes selected by AE. **(B) (C)** List of genes which are high expressed in basal-like breast cancer (BLBC) subtype and biological process (BP) annotation of these genes using David ($p < 0.05$). **(D)** 10-years overall survival analysis (logrank $p < 0.05$) of CCL19, CXCL13, HLA-DQA2, KRT81, LCN2 and SLPI.

conserved and it plays a key role in cancer progression. Mutations in the PI3K-AKT-mTOR signaling pathway are the key drivers of tumorigenesis and are related to the resistance of endocrine therapy in breast cancer. The TNF family is a group of cytokines that can cause cell apoptosis, and their expression is strongly associated with the development of various cancers. These results indicated that AE has captured genes with significant mutation patterns in BRCA.

Furthermore, we performed hierarchical clustering analysis of the selected mutation genes in all samples (**Figure 3B**). We found the local co-amplification of ERBB2 in the Her2-enriched subtype on 17q12-21. The role of ERBB2 as an important predictor of patient outcome and response to various therapies in breast cancer has been clearly established. It is well known that amplification of the 17q12-q21 region is the most common mechanism for ERBB2 activation in breast cancer and that it leads to the simultaneous activation of several other genes. These co-amplified and co-activated genes may have an impact on disease progression and the clinical behavior of ERBB2-positive tumors and thus represent important targets of research (Kauraniemi and Kallioniemi, 2006).

3.3.2 AE Captured the EMT and Epidermal Development Characteristics of Basal-Like Subtype on Transcriptome Level

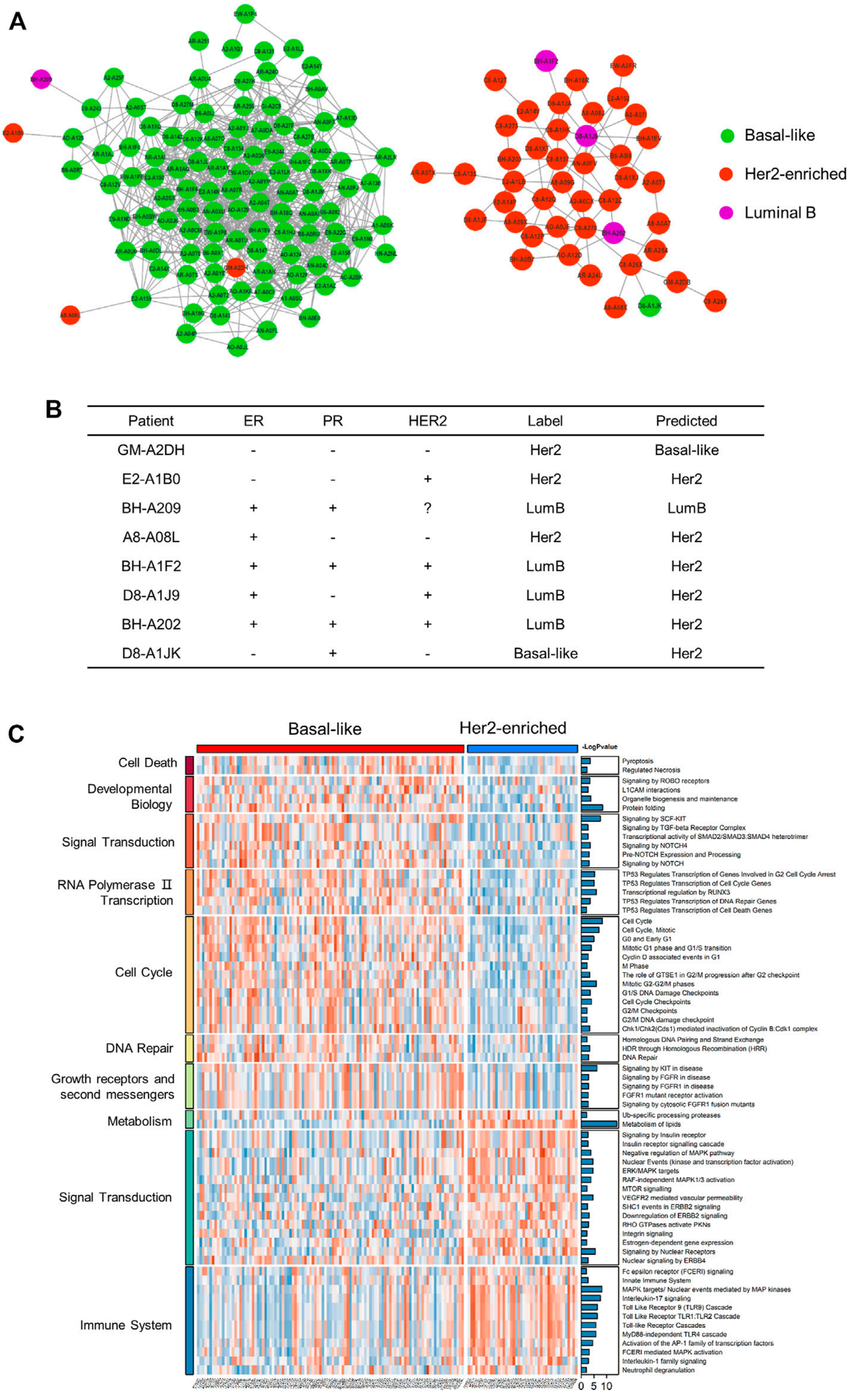
Similar as CNV data, AE selected a total of 121 candidate genes at the transcriptome level after training for 100 epochs. A cohort of 1880 patients from the KM plotter was used to validate the prognostic value of these genes. We found 70 genes that were significantly associated with 10-years overall survival (logrank $p < 0.05$), suggesting that they were potential biomarkers for BRCA prognosis (**Supplementary Table S2**).

The expression heatmap (**Figure 4A**) of the 121 genes was presented according to the four known subtypes (Luminal A and B, Her2-enriched, and basal-like). The BLBC patients were associated with aggressive behavior and poor prognosis, and do not typically express hormone receptors or HER-2 (the "triple-negative" phenotype). Therefore, patients with basal-like cancers are unlikely to benefit from the currently available targeted systemic therapy (Rakha et al., 2008). We focused on those genes and found that this subtype was related to epidermal development and the epithelial-to-mesenchymal transition (EMT) (**Figures 4B,C**). Specifically, KRT5, KRT6B, KRT14, and KRT17 are all well-described BLBC markers. KRT81 is one of the main hair proteins that is expressed in the hair cortex. However, it was reported that KRT81 is expressed in clinical specimens from patients with breast cancer (Nanashima et al., 2017). KM survival analysis showed that KRT81 is associated with poor prognosis (**Figure 4D**). Our results are consistent with previous studies that BLBC expresses basal cytokeratin and other markers of healthy breast myoepithelial cells. The EMT has been associated with various tumor functions, including tumor initiation, malignant progression, tumor stemness, tumor cell migration, intravasation to the blood, metastasis, and resistance to therapy. Matrix metalloproteinase (MMPs) are considered as target genes of the EMT pathway and MMP expression is a late event of the EMT (Han et al., 2018). PRAME plays a tumor-promoting role in triple-negative breast cancer by increasing cancer cell motility through EMT-gene reprogramming (Al-Khadairi et al., 2019). ELF5 is a suppressor of EMT and metastasis through the transcriptional repression of Snail2 in breast cancer (Chakrabarti et al., 2012). LCN2 modulates the degradation, allosteric events, and enzymatic activity of matrix metalloprotease-9 (Santiago-Sanchez et al., 2020). And we found LCN2 is an unfavourable prognostic factor (**Figure 4D**). SLPI were overexpressed preferentially in human patients that had lung-metastatic relapse (Zhang et al., 2002), its poor prognosis (**Figure 4D**) suggests that it may be widely related to the drivers of human cancer metastasis progression. Additionally, we found some immune factors with good prognosis, CCL19, CXCL13 and HLA-DQA2 (**Figure 4D**). In conclusion, these characteristics of the basal-like subtype were supported by the association between basal cytokeratins and poor outcome.

3.3.3 Network Visualization and Pathway Analysis at the Proteome Level

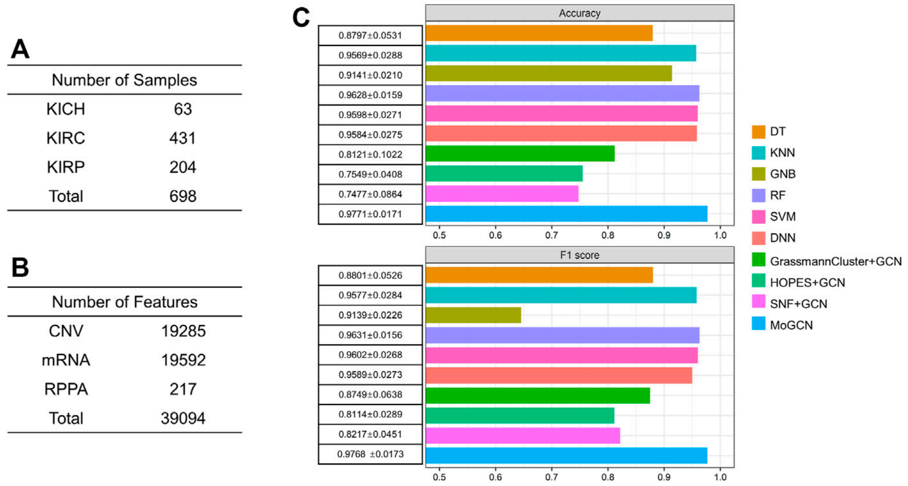
After training the model, we reclassified the subtypes of all patients. We visualized the patient network using Cytoscape (https://cytoscape.org/) and identified the two largest subgraphs with high similarity and strong connections. These were dominated by patients with the basal-like subtype and with the Her2-enriched subtype (**Figure 5A**). We compared the classification results with the immunohistochemistry results (**Figure 5B**). In the basal-like subgroup, there were four abnormal patients (**Figure 5A**). Specifically, the status of GM-A2DH is ER-negative, PR-negative, HER2-negative, and it is located in the center of the basal-like subgraph. Compared with the original label Her2-enriched, it is more reasonable for MoGCN to classify GM-A2DH as basal-like subtype. Although E2-A1B0 (ER$^-$, PR$^-$, HER2$^+$), BH-A209 (ER$^+$, PR$^+$, HER2$^-$), and A8-A08L (ER$^+$, PR$^-$, HER2$^-$) were connected to basal-like patients in the subgraph, their prediction results were consistent with the original labels. We suggested that this was the result of a combination of two features: 1) these patients were located at the edge of the basal-like subgraph, and 2) the multi-omics feature extracted by AE complemented the decision-making of the network. In the Her2-enriched subgroup, there were also four abnormal patients (**Figure 5A**), which were all predicted by MoGCN predicted as the Her2-enriched subtype. BH-A1F2, D8-A1J9, and BH-A202 were HER2$^+$, indicating that they could benefit from HER2-targeting therapy. D8-A1JK (ER$^-$, PR$^+$, HER2$^-$) did not meet the classification criteria of Her2-enriched and basal-like subtypes. Considering that it is in the Her2-enriched subgroup, MoGCN diagnosed it as Her2-enriched. These results suggested that by integrating the network structure and multi-omics features, MoGCN was able to make clinically interpretable decisions.

Considering the significant enrichment of the two subgraphs of the basal-like subtype and the Her2-enriched subtype, we performed GSVA analysis on the RPPA data of



Patient	ER	PR	HER2	Label	Predicted
GM-A2DH	-	-	-	Her2	Basal-like
E2-A1B0	-	-	+	Her2	Her2
BH-A209	+	+	?	LumB	LumB
A8-A08L	+	-	-	Her2	Her2
BH-A1F2	+	+	+	LumB	Her2
D8-A1J9	+	-	+	LumB	Her2
BH-A202	+	+	+	LumB	Her2
D8-A1JK	-	+	-	Basal-like	Her2

FIGURE 5 | Analysis of the results for the proteome and patient similarity network. **(A)** Visualization of basal-like and Her2-enriched subgroups using Cytoscape. **(B)** The IHC, original label, and MoGCN -predicted label of patients. "–", IHC-negative; "+", IHC-positive; "?", missing data. **(C)** GSVA of basal-like subgroup and Her2-enriched subgroup (p < 0.05).



Number of Samples	
KICH	63
KIRC	431
KIRP	204
Total	698

Number of Features	
CNV	19285
mRNA	19592
RPPA	217
Total	39094

FIGURE 6 | Performance of MoGCN on KIPAN dataset. **(A)**, **(B)** Summary of the KIPAN dataset. **(C)** Performance comparison of different algorithms. 10-fold cross validation (mean ± standard deviation).

these samples (**Supplementary Table S3**). The results showed the statistically different pathways in two subgroups. The basal-like subgroup was more enriched in intense cell cycle activity, DNA damage repair, and the fibroblast growth factor receptors (FGFR) pathways (**Figure 5C**). The basal-like cell lines express an autocrine FGF2 signaling loop that may also be targetable by monoclonal antibodies (Sharpe et al., 2011), suggesting that patients harboring those tumors may be candidates for FGFR-based targeted therapies. The Her2-enriched subgroup overexpressed ErbB signaling, insulin receptor signaling, and MTOR signaling pathways, which was consistent with the genome level changes. Therefore, combination therapy targeting HER2 can effectively improve patient survival.

3.4 Validation of MoGCN on the TCGA Pan-Kidney Cancer Dataset

To verify the generality of MoGCN, we applied this analysis model to the TCGA pan-kidney cancer (KIPAN) dataset, which consisted of three main subtypes: kidney chromophobe (KICH), kidney clear cell carcinoma (KIRC), and kidney papillary cell carcinoma (KIRP) (**Figure 6A**). The CNV, mRNA, and RPPA data for the 698 patients were obtained (**Figure 6B**). In the subtype analysis, the accuracy and F1 score of the MoGCN model reached 97.71 and 97.68% and outperformed all other compared methods (**Figure 6C**). These results showed that MoGCN has potential applicability for a wide range of multi-omics data mining.

4 DISCUSSION

Cancer has been widely regarded as a highly heterogeneous disease, and the early diagnosis and prognostic of a cancer type have become the focus of cancer research. The ultimate goal of biology is to achieve systems biology understanding, that is, the integration, interpretation and insight of multi-omics. In the era of big data, efficient data mining of massive biomedical data is an important challenge for bioinformatics research.

We developed MoGCN, a network-based multi-omics integration pipeline for cancer subtype classification. Our study focused on the issues of feature reduction and the interpretation of prediction results. Notably, AE improved performance after integrating multi-omics features, and it also achieved the most optimal performance, which implied that it has the ability to capture the complex nonlinear relationships between multi-omics data. Whereas other mainstream algorithms slightly decreased or remained unchanged. Moreover, by using GCN to integrate the omics features and the PSN, the classification performance of our method was further improved, and displayed the highest accuracy (0.8982) and F1 score (0.9016) compared with the current mainstream cancer subtype prediction algorithms.

MoGCN is interpretative in terms of feature extraction and clinically intuitive diagnosis. Once the model has been trained, MoGCN was able to extract the most signification features of each omics layer for downstream biological knowledge discovery. The mutated genes at genome level were significantly enriched in functions or signaling pathways for cancer development, such as epidermal development, cell migration, Wnt signaling, ErbB signaling, and mTOR signaling. In addition, the genes highly expressed in the basal-like subtype with the worst clinical prognosis were characterized by enrichment in epidermal development and the epithelial-mesenchymal transition. Finally, through the visualization of the PSN, we found that the topological network and omics data features were complementary and could provide intuitive information for clinical diagnosis. The generality of MoGCN was proven on the TCGA pan-kidney cancer dataset. These case studies show that MoGCN performs well for heterogeneous data integration and the interpretability of classification results, which confers great potential for applications in biomarker identification and clinical diagnosis.

5 CONCLUSION

In conclusion, we developed an interpretable deep learning multi-omics integration model, for cancer subtype analysis. The captured features could reveal the molecular characteristics of cancer subtypes and the patient similarity network could provide intuitive information for clinical diagnosis. This study provided a novel method of the multi-omics integration. And the graph-based approach could provide new possibilities to the precision medicine.

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

XL gathered all the data and performed the data analysis. JM designed the study and drafted the manuscript. LL, MH, and ML contributed to results analysis and discussions. FH and YZ supervised the study, revised the manuscript and gave the final approval of the version to be published.

FUNDING

This work was supported by the National Key Research Program of China (2017YFA0505002, 2021YFA1301603, and 2015AA020108).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.806842/full#supplementary-material

REFERENCES

Al-Khadairi, G., Naik, A., Thomas, R., Al-Sulaiti, B., Rizly, S., and Decock, J. (2019). PRAME Promotes Epithelial-To-Mesenchymal Transition in Triple Negative Breast Cancer. *J. Transl Med.* 17 (1), 9. doi:10.1186/s12967-018-1757-3

Cancer Genome Atlas Network. (2012). Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* 490 (7418), 61–70. doi:10.1038/nature11412

Chakrabarti, R., Hwang, J., Blanco, M. A., Wei, Y., Lukačišin, M., Romano, R.-A., et al. (2012). Elf5 Inhibits the Epithelial-Mesenchymal Transition in Mammary Gland Development and Breast Cancer Metastasis by Transcriptionally Repressing Snail2. *Nat. Cel Biol* 14 (11), 1212–1222. doi:10.1038/ncb2607

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* 24 (6), 1248–1259. doi:10.1158/1078-0432.CCR-17-0853

Chen, R., Yang, L., Goodison, S., and Sun, Y. (2020). Deep-learning Approach to Identifying Cancer Subtypes Using High-Dimensional Genomic Data. *Bioinformatics* 36 (5), 1476–1483. doi:10.1093/bioinformatics/btz769

Ding, H., Sharpnack, M., Wang, C., Huang, K., and Machiraju, R. (2019). Integrative Cancer Patient Stratification via Subspace Merging. *Bioinformatics* 35 (10), 1653–1659. doi:10.1093/bioinformatics/bty866

Eicher, T., Kinnebrew, G., Patt, A., Spencer, K., Ying, K., Ma, Q., et al. (2020). Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources. *Metabolites* 10 (5), 202. doi:10.3390/metabo10050202

Garson, G. D. (1991). Interpreting Neural Network Connection Weights. *Artif. Intelligence Expert* 6, 46–51.

Han, B., Zhou, B., Qu, Y., Gao, B., Xu, Y., Chung, S., et al. (2018). FOXC1-induced Non-canonical WNT5A-MMP7 Signaling Regulates Invasiveness in Triple-Negative Breast Cancer. *Oncogene* 37 (10), 1399–1408. doi:10.1038/s41388-017-0021-2

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14 (7), 7. doi:10.1186/1471-2105-14-7

Heo, Y. J., Hwa, C., Lee, G.-H., Park, J.-M., and An, J.-Y. (2021). Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. *Mol. Cell* 44 (7), 433–443. doi:10.14348/molcells.2021.0042

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science* 313 (5786), 504–507. doi:10.1126/science.1127647

Kauraniemi, P., and Kallioniemi, A. (2006). Activation of Multiple Cancer-Associated Genes at the ERBB2 Amplicon in Breast Cancer. *Endocr. Relat. Cancer* 13 (1), 39–49. doi:10.1677/erc.1.01147

Kipf, N. T., and Welling, M. (2017). Semi-supervised Classification with Graph Convolutional Networks, International Conference on Learning Representations(ICLR), 22 Feb 2017 (Toulon, France: Palais des Congrès Neptune).

Liu, Q., Hu, Z., Jiang, R., and Zhou, M. (2020). DeepCDR: a Hybrid Graph Convolutional Network for Predicting Cancer Drug Response. *Bioinformatics* 36 (Suppl. l_2), i911–i918. doi:10.1093/bioinformatics/btaa822

Lu, M., and Zhan, X. (2018). The Crucial Role of Multiomic Approach in Cancer Research and Clinically Relevant Outcomes. *EPMA J.* 9 (1), 77–102. doi:10.1007/s13167-018-0128-8

Nanashima, N., Horie, K., Yamada, T., Shimizu, T., and Tsuchida, S. (2017). Hair Keratin KRT81 Is Expressed in normal and Breast Cancer Cells and Contributes to Their Invasiveness. *Oncol. Rep.* 37 (5), 2964–2970. doi:10.3892/or.2017.5564

Pai, S., and Bader, G. D. (2018). Patient Similarity Networks for Precision Medicine. *J. Mol. Biol.* 430 (18 Pt A), 2924–2938. doi:10.1016/j.jmb.2018.05.037

Rakha, E. A., Reis-Filho, J. S., and Ellis, I. O. (2008). Basal-like Breast Cancer: a Critical Review. *Jco* 26 (15), 2568–2581. doi:10.1200/JCO.2007.13.1748

Santiago-Sánchez, G. S., Pita-Grisanti, V., Quiñones-Díaz, B., Gumpper, K., Cruz-Monserrate, Z., and Vivas-Mejía, P. E. (2020). Biological Functions and Therapeutic Potential of Lipocalin 2 in Cancer. *Ijms* 21 (12), 4365. doi:10.3390/ijms21124365

Sharpe, R., Pearson, A., Herrera-Abreu, M. T., Johnson, D., Mackay, A., Welti, J. C., et al. (2011). FGFR Signaling Promotes the Growth of Triple-Negative and Basal-like Breast Cancer Cell Lines BothIn VitroandIn Vivo. *Clin. Cancer Res.* 17 (16), 5275–5286. doi:10.1158/1078-0432.CCR-10-2727

Sun, Y. V., and Hu, Y.-J. (2016). Integrative Analysis of Multi-Omics Data for Discovery and Functional Studies of Complex Human Diseases. *Adv. Genet.* 93, 147–190. doi:10.1016/bs.adgen.2015.11.004

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* 11 (3), 333–337. doi:10.1038/nmeth.2810

Xu, A., Chen, J., Peng, H., Han, G., and Cai, H. (2019). Simultaneous Interrogation of Cancer Omics to Identify Subtypes with Significant Clinical Differences. *Front. Genet.* 10, 236. doi:10.3389/fgene.2019.00236

Xuan, P., Pan, S., Zhang, T., Liu, Y., and Sun, H. (2019). Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations. *Cells* 8 (9), 1012. doi:10.3390/cells8091012

Yu, Z., Huang, F., Zhao, X., Xiao, W., and Zhang, W. (2021). Predicting Drug-Disease Associations through Layer Attention Graph Convolutional Network. *Brief Bioinform* 22 (4). doi:10.1093/bib/bbaa243

Zhang, D., Simmen, R. C. M., Michel, F. J., Zhao, G., Vale-Cruz, D., and Simmen, F. A. (2002). Secretory Leukocyte Protease Inhibitor Mediates Proliferation of Human Endometrial Epithelial Cells by Positive and Negative Regulation of Growth-Associated Genes. *J. Biol. Chem.* 277 (33), 29999–30009. doi:10.1074/jbc.M203503200

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

SYSTEMATIC REVIEW
published: 04 February 2022
doi: 10.3389/fgene.2022.829283

Association of Interleukin-10 Polymorphism (rs1800896, rs1800871, and rs1800872) With Breast Cancer Risk: An Updated Meta-Analysis Based on Different Ethnic Groups

Lijun Li[1], *Wei Xiong*[2], *Donghua Li*[1] *and Jiangang Cao*[3]*

[1]The Second Affiliated Hospital, Department of Pharmacy, Hengyang Medical School, University of South China, Hengyang, China, [2]The Second Affiliated Hospital, Department of Breast and Thyroid Surgical, Hengyang Medical School, University of South China, Hengyang, China, [3]The Affiliated Nanhua Hospital, Clinical Research Institute, Hengyang Medical School, University of South China, Hengyang, China

OPEN ACCESS

Edited by:
Rodrigo Drummond,
A.C.Camargo Cancer Center, Brazil

Reviewed by:
Seyed Reza Mohebbi,
Shahid Beheshti University of Medical Sciences, Iran
Hamidreza Mahboobi,
Tehran University of Medical Sciences, Iran

Correspondence:
Jiangang Cao
814709270@qq.com

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *05 December 2021*
Accepted: *21 January 2022*
Published: *04 February 2022*

Citation:
Li L, Xiong W, Li D and Cao J (2022) Association of Interleukin-10 Polymorphism (rs1800896, rs1800871, and rs1800872) With Breast Cancer Risk: An Updated Meta-Analysis Based on Different Ethnic Groups.
Front. Genet. 13:829283.
doi: 10.3389/fgene.2022.829283

Background: The interleukin10 (IL-10) gene polymorphisms have been indicated to be associated with breast cancer (BC) risk, but the findings are still controversial. To derive a more precise evaluation, we performed a comprehensive meta-analysis.

Methods: A systematic literature search was conducted using PubMed, Embase, CNKI, China biomedical (CBM), and Google Scholar to 29 March 2020. Revman5.3 and Stata 12.0 software analyzed the data, and the strength of the association was identified using the odds ratio (OR) and the corresponding 95% confidence interval (CI).

Results: A total of 23 studies (7,250 cancer cases and 7,675 case-free controls) were included in this meta-analysis. The results show that IL-10 gene polymorphisms were significantly correlated with BC risk based on subgroup analysis by ethnicity. The IL-10 rs1800896 polymorphism was significantly associated with the risk of BC in Asians (G vs. A: OR = 0.78, 95% CI = 0.65–0.95, $p = 0.01$; GG vs. AA: OR = 0.51, 95% CI = 0.31–0.84, $p = 0.007$; GA vs. AA: OR = 0.6, 95% CI = 0.44–0.81, $p = 0.0009$; GG + GA vs. AA: OR = 0.6, 95% CI = 0.45–0.81, $p = 0.0007$); Moreover, an increased BC risk in Asians were also associated with the IL-10 rs1800872 polymorphism (AA vs CC: OR = 0.74, 95% CI = 0.55–0.99, $p = 0.04$; A vs C: OR = 0.85, 95% CI = 0.74–0.98, $p = 0.03$). In addition, The IL-10 rs1800871 (CT vs. TT: OR = 1.8, 95% CI = 1.03–3.13, $p = 0.04$) and rs1800872 polymorphism (A vs C: OR = 0.65, 95% CI 0.43–0.98, $p = 0.04$) were associated with BC risk in Caucasians.

Conclusion: Collectively, this meta-analysis demonstrated that IL-10 rs1800896 and rs1800872 (AA vs. CC; A vs. C) polymorphisms significantly increased the risk of BC in Asians, while the rs1800871 and rs1800872 (A vs. C) were associated with the risk of BC in Caucasians. Therefore, this may provide new ideas for predicting and diagnosing BC susceptibility through the detection of IL-10 gene polymorphism.

Systematic Review Registration: [https://www.crd.york.ac.uk/ PROSPERO], identifier [CRD42021266635].

Keywords: interleukin-10, gene polymorphism, breast cancer, meta-analysis, species variation

INTRODUCTION

Breast cancer (BC) is the leading cause of female cancer-related death worldwide and is one of the most common cancer forms (Anastasiadi et al., 2017). BC incidence varies widely, ranging from 27/100,000^2 (Central-East Asia and Africa) to 85–94/100,000^2 (Australia, North America, and Western Europe). And the incidence of BC in France is the highest in Europe (Sancho-Garnier and Colonna, 2019). In Asian countries, the incidence rate of BC has also been increasing rapidly (Mubarik et al., 2020; Oblak et al., 2020). The pathogenesis of BC is multifactorial. Hereditary BC accounts for only 5–10% of all BC cases and germline mutations, with the two significant BC susceptibility genes, *BRCA1* and *BRCA2* is responsible for approximately 2–3% of all cases (Kwong et al., 2016). Besides gene tests for identifying high-risk BRCA1 or BRCA2 mutations carriers (Ha et al., 2017), the ability to predict BC development is not well established yet. Although genetic, environmental, and lifestyle factors are associated with BC occurrence, the biological mechanism that causes BC remains unclear.

Inflammation plays a significant role in BC development and is an important part of the BC microenvironment (Mohamed et al., 2018). Interleukin-10 (IL-10) is an important anti-inflammatory and immunomodulatory cytokine in the human immune response. IL-10 is located on chromosome 1 (1q31-1q32), composed of five exons and four introns (Roh et al., 2002). Single nucleotide polymorphism (SNP) is the most common genetic variation. In the SNP database (http://www.ncbi.nlm.nih.gov/snp), three promoter SNPs of IL-10, rs1800896 (-1082A/G), rs1800871 (-819T/C), and rs1800872 (-592A/C) were extensively investigated in many diseases. Because they might affect IL-10 gene transcription and translation, resulting in abnormal cell proliferation and cancer development (Howell and Rose-Zerilli, 2007). The possible mechanism is that IL-10 is activated by the Janus kinase (JAK)/signal transducer and activator of transcription (STAT) signaling pathways through its receptor IL-10 R1 which binds to STAT3. Then STAT3 is translocated into the nucleus, where it binds to STAT-binding elements in the promoters of proliferation-related genes. It has been reported that IL-10 gene polymorphism plays an important role in the occurrence and development of cancers such as BC, gastric cancer, lung cancer (Bhattacharjee et al., 2016; Chen et al., 2019; Zhao et al., 2019). And some studies reported the high IL-10 expression levels in the BC paraffin section and its expression is correlated with worse outcomes in patients with malignant tumors (Li et al., 2014; Zhao et al., 2015).

In recent years, several studies have reported the relationship between IL-10 polymorphisms and BC susceptibility. A study found that: rs1800896 (-1082A/G) polymorphism was correlated with cancer staging and associated with the progression of BC at AA genotype (Abedinzadeh et al., 2018). In the research based on Caucasians, it was found that there was a significant association between the IL10-1082 G/G genotype and the increased risk of BC (Zhu et al., 2020). Another study found that the rs1800871 (-819T/C) polymorphism increased the risk of BC in Han Chinese women (Li et al., 2020). And a study shows the rare allele of rs1800872 (-592A/C) polymorphism may be a potential prognostic indicator of disease-free survival in BC patients (Gerger et al., 2010). These suggest that IL-10 gene polymorphism may affect the risk of human BC (Setrerrahmane and Xu, 2017). However, these results are inconsistent. Moreover, IL-10 polymorphism and BC susceptibility studies are constantly updated, and adjustments vary between included studies based on race, age, lifestyle, and other covariates (Patricia Gallegos-Arreola et al., 2019). Considering the critical role of IL-10 in the development of BC, we conducted this systematic review. And compared with previous meta-analyses, we comprehensively included the latest relevant studies to evaluate the association of IL-10 rs1800896, rs1800871, and rs1800872 polymorphisms with the risk of BC in different ethnic groups. It will provide theoretical evidence for the genetic mechanism of BC.

METHODS

This meta-analysis was conducted according to the PRISMA reporting criteria (Moher et al., 2009).

Search Strategy

Research articles on the relationship between IL-10 gene polymorphisms and BC risk were searched in different databases, including PubMed, Web of Knowledge, Embase, CNKI, CBM, and Google Scholar up to 29 March 2020. And we retrieved with the keywords: ("breast cancer" or "breast tumor" or "breast neoplasm" or "malignant breast tumor" or "breast carcinoma") and ("Interleukin-10" or "IL-10") and ("polymorphism" or "SNP" or "single nucleotide polymorphism" or "variation" or "mutation").

Inclusion and Exclusion Criteria

Inclusion criteria: (1) Clinical BC patients were selected as the case group and healthy people as the control group; (2) Case-control or cohort studies about associations between IL-10 gene polymorphism and BC in humans; (3) Full manuscript in English or Chinese is retrievable; (4) Reporting the number of cases and controls for each genotype and detailed genotyping data, or knowing the odds ratio (OR) helped to calculate the 95% confidence interval (CI).

Exclusion criteria: (1) Abstracts, reviews; (2) Studies on the relationship between IL-10 gene polymorphism and prognosis of BC; (3) studies on the apparent imbalance of baseline between the case group and control group; (4) The cases and control sources were not provided; (5) Repeatedly published literature. If multiple studies from the same case series were available, the one including the most individuals were used in the analysis.

Data Extraction

Two researchers selected the literature according to the inclusion and exclusion criteria, extracted the data, and cross-checked them independently into a standard data collection form. If there were any disputes, we would reach an agreement by discussion or by a third party and strive to reach a consensus on each project. Data were collected from each article included: the first author, year of publication, study location, type of study, ethnicity (classified as Asian, Caucasian, or mixed descent), total number of cases and

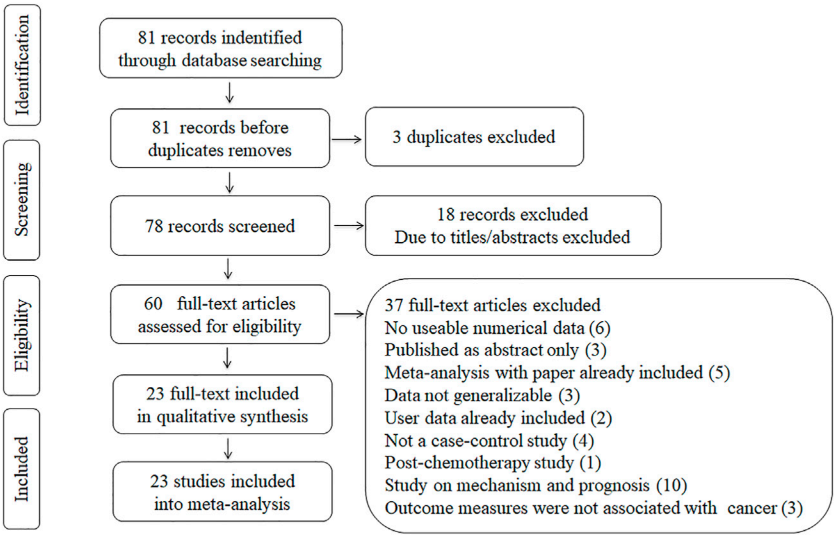

Identification
Screening
Eligibility
Included
81 records indentified through database searching
81 records before duplicates removes
3 duplicates excluded
78 records screened
18 records excluded Due to titles/abstracts excluded
60 full-text articles assessed for eligibility
37 full-text articles excluded
No useable numerical data (6)
Published as abstract only (3)
Meta-analysis with paper already included (5)
Data not generalizable (3)
User data already included (2)
Not a case-control study (4)
Post-chemotherapy study (1)
Study on mechanism and prognosis (10)
Outcome measures were not associated with cancer (3)
23 full-text included in qualitative synthesis
23 studies included into meta-analysis

FIGURE 1 | Flowchart of study selection for the present study.

controls, genotype frequency, genotype detection method, and the source of authority.

Sensitivity Analysis

Sensitivity analysis was performed to assess the stability of the results. The Funnel plot, Begg's test, and Egger's test were used to evaluate publication bias. RevMan5.3 and Stata 12.0 software was used for the above statistical analysis.

Statistical Analysis

The correlation between IL-10 gene polymorphisms and BC risk was evaluated by OR and 95% CI as the effect size. 95% CI without one and $P(\mathrm{OR}) < 0.05$ was considered statistically significant. The Z-test determines the significance of the OR value. The effects of heterogeneity were quantified by I^2 and $P(\mathrm{H})$ values. In addition, the I^2 value is used to quantify the degree of heterogeneity ($I^2 < 25\%$: low/no heterogeneity; $25 < I^2 < 75\%$: moderate heterogeneity; $I^2 > 75\%$: extreme high heterogeneity). The fixed-effect model is adopted when the $I^2 < 25\%$; otherwise, the random effect model is adopted. We further carried out subgroup analyses by ethnicity to get ethnic-specific results.

RESULTS

Search Results

We had a total of 78 articles after removing three duplicated pieces. After the layer-by-layer screening, a total of 23 articles finally met the criteria for inclusion in this meta-analysis. Eligible papers were published between 2004 and 2019. This meta-analysis updated three 2019 case-control studies compared to previous meta-analyses (Dai et al., 2014; Abedinzadeh et al., 2018; Moghimi et al., 2018). A flow diagram schematizing the inclusion and exclusion process of identified articles with the inclusion criteria is presented in **Figure 1**.

Data Extraction and Quality Assessment

The 23 eligible articles had a total sample size of 14,925 participants, including 7,250 BC patients and 7,675 healthy controls (Smith et al., 2004; Abdolrahim-Zadeh et al., 2005; Guzowski et al., 2005; Langsenlehner et al., 2005; Balasubramanian et al., 2006; Onay et al., 2006; Scola et al., 2006; Gonullu et al., 2007; Pharoah et al., 2007; Kong et al., 2010; He et al., 2012; Pooja et al., 2012; Meijiang, 2014; Wang et al., 2014; Vinod et al., 2015; AlSuhaibani et al., 2016; Atoum, 2016; Maruthi et al., 2017; Sabet et al., 2017; Tian et al., 2017; Azher et al., 2019; Fanyu et al., 2019; Patricia Gallegos-Arreola et al., 2019). The samples were involved in three IL-10 polymorphism sites: rs1800896, rs1800871, and rs1800872. There were 17 studies on rs1800896 (3,308 cases and 3,425 controls), twelve studies on rs1800871 (2,530 cases and 2,698 controls), and 13 studies on rs1800872 (4,702 cases and 4,818 controls). Ten studies were based on Caucasians, six were based on Asians, and the remaining seven studies were mixed-race in the 23 criteria studies. Of these studies, eighteen were hospital-based, and five were population-based. The Newcastle-Ottawa Scale (NOS) was used to assess the quality of the included articles (Stang, 2010). And NOS scores ranged from zero to nine. We considered the study's methodological quality good if the score was ≥ seven. Two authors independently completed our data extraction and quality evaluation. **Table 1** and **Table 2** show the basic characteristics of the included literature, the distribution of polymorphisms at the studied gene sites, allele frequency, and the quality assessment of the included studies.

Meta-Analysis Results

The association between IL-10 gene polymorphisms (rs1800896, rs1800871, and rs1800872) and BC is shown in **Table 3** and **Figures 2–5**. Squares and horizontal lines correspond to study-specific OR and 95% CI. The area of a square reflects the weight (inversely proportional to the variance). The diamond represents the sum of OR and 95% CI.

TABLE 1 | Characteristics of the studies included in the meta-analysis.

First author	Year	Country	Ethnicity	Genotyping method	SOC	Case/control	Study design	SNP No.	NOS score
Gallegos-Arreola	2019	Mexican	Mixed	PCR-RFLP	HB	368/320	CC	3	8
Al-Ankoshy	2019	Iraq	Caucasian	PCR–SSP	HB	70/70	CC	1	8
Zeng	2019	China	Asian	PCR-RFLP	HB	208/215	CC	3	8
Sabet	2017	Egypt	Caucasian	PCR-RFLP	HB	105/50	CC	1,2,3	7
Tian	2017	China	Asian	Mass ARRAY	PB	312/312	CC	1,2,3	7
Maruthi	2017	India	Mixed	PCR-RFLP	HB	285/285	CC	1	7
Atoum	2016	Jordan	Mixed	PCR-RFLP	HB	202/210	CC	1,2,3	7
Alsuhaibani	2016	Egypt	Caucasian	PCR-RFLP	HB	80/80	CC	1	7
Vinod	2015	India	Mixed	ASPCR	HB	125/160	CC	1	8
Li	2014	China	Asian	PCR–SSP	PB	128/128	CC	1,2	7
Wang	2014	China	Asian	PCR-RFLP	HB	474/501	CC	2	8
Pooja	2012	India	Mixed	PCR-RFLP	PB	200/200	CC	1,2,3	7
He	2012	China	Asian	MALDI-TOF MS	HB	347/500	CC	2	7
Kong2010	2010	China	Asian	PCR-RFLP	HB	315/322	CC	1,2,3	7
Pharoah	2007	European	Caucasian	TaqMan	PB	2045/2218	CC	3	8
Gonullu	2007	Turkey	Caucasian	Mass ARRAY	HB	38/24	CC	1,2,3	7
Scola	2006	Italy	Caucasian	PCR-RFLP	HB	84/106	CC	1,2,3	7
Onay	2006	Canada	Mixed	TaqMan	PB	398/372	CC	1	8
Balasu bramanian	2006	United Kingdom	Caucasian	TaqMan	HB	497/498	CC	1	7
Guzowski	2005	America	Mixed	DHPLC	HB	50/25	CC	1,2,3	7
Langsenlehner	2005	Australia	Caucasian	TaqMan	PB	500/496	CC	3	8
Abdolrahim	2005	Iran	Caucasian	PCR-RFLP	HB	275/320	CC	1,2,3	8
Smith	2004	United Kingdom	Caucasian	ARMS-PCR	HB	144/263	CC	1	8

SOC, source of controls; HB: hospital-based; PB, population-based; CC, case–control; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; DHPLC, denaturing highperformance liquid chromatography; EPIC, European Prospective Investigation of Cancer (a prospective study of diet and cancer being carried out in nine European countries); ASPCR, allele-specific PCR; SNP, single-nucleotide polymorphisms; SNP No. 1, - 1082A > G (rs1800896); 2: - 819T > C (rs1800871); 3, - 592A > C (rs1800872); NOS, Newcastle-ottawa scale.

Correlation Between rs1800896 Polymorphism and Breast Cancer

A total of 17 studies were conducted on the association between IL-10 rs1800896 polymorphism and BC risk, with a total sample size of 6,733 cases, including 3,308 patients and 3,425 healthy controls. Overall population heterogeneity test I^2 was 82%. The random-effect model results showed that the comparison results of the five gene models showed no statistical significance between rs1800896 polymorphism and BC (**Table 3**). Subgroups by ethnicity showed that under the four genetic models (allele G vs. A: OR = 0.78, 95% CI = 0.65–0.95, p = 0.01; homozygous GG vs. AA: OR = 0.51, 95% CI = 0.31–0.84, p = 0.007; heterozygous GA vs. AA: OR = 0.6, 95% CI = 0.44–0.81, p = 0.0009; dominant GG + GA vs. AA: OR = 0.6, 95% CI = 0.45–0.81, p = 0.0007) (**Figures 2A–D**), rs1800896 polymorphism was significantly associated with BC risk in Asians. This result suggests that ethnicity is likely to be the source of heterogeneity, and the rs1800896 polymorphism is significantly associated with the BC risk in Asians.

Correlation Between rs1800871 Polymorphism and Breast Cancer

A total of 12 studies with 2,530 patients and 2,698 controls evaluated the strength of the association between the IL-10 rs1800871 polymorphism and BC. There was no association between BC risk and rs1800871 polymorphism in any genetic model in the overall population. However, when stratified by ethnicity, the rs1800871 polymorphism was associated with BC risk in the heterozygous model in Caucasians (CT vs. TT: OR = 1.8, 95% CI = 1.03–3.13, p = 0.04) (**Figure 3**). This result indicates that Caucasians with the rs1800871 heterozygous model are more likely to develop BC than individuals with other genotypes.

Correlation Between rs1800872 Polymorphism and Breast Cancer

Thirteen studies (4,702 cases and 4,818 controls) assessed the strength of the association between IL-10 rs1800872 polymorphism and BC susceptibility. As shown in **Table 3**, the five gene model comparison results showed that the association between IL-10 rs1800872 polymorphism and BC in the overall population was not statistically significant. However, after stratification by ethnicity, the homozygous model of rs1800872 polymorphism was associated with BC risk in Asians (AA vs. CC: OR = 0.74, 95% CI = 0.55–0.99, p = 0.04) (**Figure 4A**). Allele model of rs1800872 polymorphism was associated with the risk of BC in Asians (A vs. C: OR = 0.85, 95% CI = 0.74–0.98, p = 0.03) (**Figure 4B**) and Caucasians (A vs. C: OR = 0.65, 95% CI = 0.43–0.98, p = 0.04) (**Figure 4C**).

Publication Bias

Funnel plot, Begg's test, and Egger's test were used to evaluate the publication bias (Stata12.0). As shown in **Figure 5**, the funnel plot was essentially symmetrical, and the p values of Begg's test and Egger's test are all greater than 0.05. It was

TABLE 2 | IL-10 polymorphisms genotype distribution and allele frequency in cases and controls.

First author	Case	Control	Cases					Control					MAF
			Genotypes			Alleles		Genotypes			Alleles		
-1082A > G rs1800896			AA	AC	CC	A	C	AA	AC	CC	A	C	
Atoum (2016)	202	210	157	29	16	343	61	151	42	17	344	76	0.181
AlSuhaibani et al. (2016)	80	80	16	47	17	79	81	14	50	16	78	82	0.512
Vinod et al. (2015)	125	160	76	31	18	183	67	67	78	15	212	108	0.337
Pooja et al. (2012)	200	200	132	60	8	324	76	145	50	5	34	60	0.638
Kong et al. (2010)	315	322	285	29	1	599	31	285	35	2	605	39	0.061
Gonullu et al. (2007)	38	24	13	22	3	48	28	16	7	1	39	9	0.187
Guzowski et al. (2005)	50	25	10	28	12	48	52	9	12	4	30	20	0.400
Sabet et al. (2017)	105	50	15	41	49	71	139	27	21	2	75	25	0.250
Tian et al. (2017)	312	312	51	132	129	234	390	27	154	131	208	416	0.666
Abdolrahim-Zadeh et al. (2005)	275	320	119	116	40	171	373	146	125	49	417	223	0.348
Scola et al. (2006)	84	106	28	40	16	96	72	40	45	21	125	87	0.410
Onay et al. (2006)	398	372	90	205	103	385	411	107	194	71	408	336	0.451
Balasubramanian et al. (2006)	497	498	121	253	123	499	495	117	260	121	494	502	0.504
Maruthi et al. (2017)	285	285	80	146	59	262	308	89	159	37	234	336	0.589
Azher et al. (2019)	70	70	36	10	24	97	43	16	17	37	44	96	0.690
Smith et al. (2004)	144	263	32	58	39	136	122	46	120	57	250	276	0.524
Li et al. (2014)	128	128	96	30	2	222	34	80	44	4	204	52	0.203
-819T > C (rs1800871)			TT	TC	CC	T	C	TT	TC	CC	T	C	
Atoum (2016)	202	210	88	47	67	223	181	93	41	76	227	193	0.459
Wang et al. (2014)	474	501	90	198	186	378	570	48	219	234	315	687	0.685
Pooja et al. (2012)	200	200	54	92	54	200	200	65	78	57	208	192	0.480
Kong et al. (2010)	315	322	119	135	61	273	257	134	131	57	399	245	0.380
Gonullu et al. (2007)	38	24	5	17	16	27	49	4	10	10	18	30	0.625
Guzowski et al. (2005)	50	25	3	19	28	25	75	1	10	14	12	38	0.760
Sabet et al. (2017)	105	50	16	47	42	79	131	26	22	2	74	26	0.260
Tian et al. (2017)	312	312	124	141	47	389	235	144	128	40	416	208	0.333
Abdolrahim-Zadeh et al. (2005)	275	320	129	120	26	375	172	166	122	32	454	186	0.290
He et al. (2012)	347	500	177	141	29	495	199	229	223	44	681	311	0.313
Scola et al. (2006)	84	106	5	30	49	40	128	12	35	59	59	177	0.721
Li et al. (2014)	128	128	105	22	1	232	23	96	28	4	220	36	0.203
-592C > A (rs1800872)			AA	AC	CC	A	C	AA	AC	CC	A	C	
Patricia Gallegos-Arreola et al. (2019)	368	320	42	154	172	238	498	11	100	209	122	518	0.190
Zeng et al. (2019)	208	215	10	88	110	108	308	22	95	98	139	291	0.323
Atoum (2016)	202	210	76	84	42	236	168	79	91	40	249	171	0.593
Pooja et al. (2012)	200	200	45	67	88	157	243	38	84	78	160	240	0.400
Kong et al. (2010)	315	322	119	135	61	373	257	134	131	57	399	245	0.620
Gonullu et al. (2007)	38	24	5	17	16	27	49	4	10	10	18	30	0.375
Guzowski et al. (2005)	50	25	3	17	30	23	77	3	10	12	16	34	0.320
Langsenlehner et al. (2005)	500	496	21	210	269	252	748	36	199	261	271	721	0.273
Sabet et al. (2017)	105	50	4	36	65	42	166	31	16	6	78	28	0.736
Tian et al. (2017)	312	312	131	130	51	392	232	141	127	44	409	215	0.655
Abdolrahim-Zadeh et al. (2005)	275	320	27	100	148	154	396	29	132	159	190	450	0.297
Scola et al. (2006)	84	106	5	30	49	40	128	12	35	59	59	153	0.278
Pharoah et al. (2007)	2045	218	116	679	1,251	367	3,181	116	764	1,338	996	3,440	0.225

MAFs: minor allele frequencies.

indicated that there was almost no obvious publication bias at the three loci.

DISCUSSION

IL-10, known initially as cytokine synthesis inhibitory factor (CSIF), is a potent anti-inflammatory cytokine. IL-10 can stimulate the expression of carboxypeptidase B2 (CPB2) in inflammatory BC cells. Thus it increases the cancer cells' aggressiveness (Mohamed et al., 2018). Moreover, IL-10 is involved in the abnormal proliferation of breast ducts and lobules and stimulates mitotic activity, leading to increased cancer risk (Kong et al., 2010; Moghimi et al., 2018). IL-10 can also induce tumor progression by inhibiting many cytokines such as IL-1a, IL-1b, IL-6, IL-8, IL-12, and IL-18. And IL-10 gene silencing down-regulates the expression of phosphoinositide 3-kinase (PI3K)/protein kinase B (AKT) and B cell lymphoma 2 (Bcl2) and increases the expression levels of BCL2 binding component 3(BBC3), Bax, and caspase3 (Alotaibi et al., 2018). Studies on the mechanism of IL-10 promoting BC have shown that the production of IL-10 may represent a new escape mechanism for BC cells to escape the destruction of the immune system. It might be closely related to the fact that

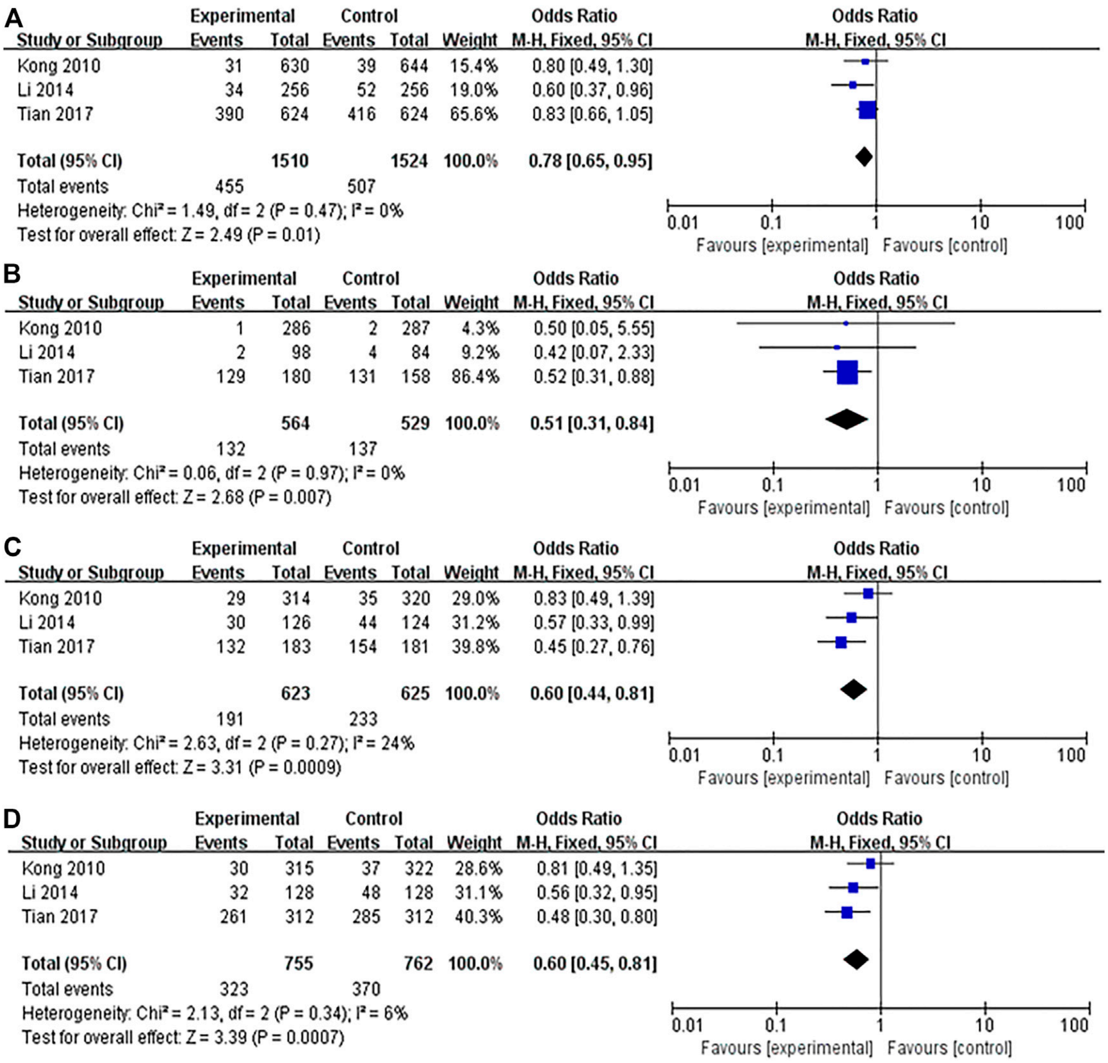
TABLE 3 | Results of the association of IL-10 polymorphisms with BC risk.

Subgroup	Genetic model	Type of model	Heterogeneity		Odds Ratio			
			I^2 (%)	PH	OR	95% CI	Z test	POR
rs1800896								
Overall	G vs. A	Random	82	<0.00001	1.06	0.87–1.28	0.56	0.57
	GG vs. AA	Random	71	<0.00001	1.14	0.81–1.59	0.73	0.46
	GA vs. AA	Random	73	<0.00001	0.91	0.71–1.17	0.74	0.46
	GG + GA vs. AA	Random	79	<0.00001	0.99	0.76–1.28	0.11	0.92
	GG vs. GA + AA	Random	53	0.006	1.18	0.94–1.48	1.45	0.15
Ethnicity								
Asian	G vs. A	Fixed	0	0.47	0.78	0.65–0.95	2.49	**0.01**
	GG vs. AA	Fixed	0	0.97	0.51	0.31–0.84	2.68	**0.007**
	GA vs. AA	Fixed	24	0.27	0.6	0.44–0.81	3.31	**0.0009**
	GG + GA vs. AA	Fixed	6	0.34	0.6	0.45–0.81	3.39	**0.0007**
	GG vs. GA + AA	Fixed	0	0.66	0.94	0.69–1.28	0.39	0.7
Caucasian	G vs. A	Random	0.0089	<0.00001	1.19	0.83–1.72	0.94	0.35
	GG vs. AA	Random	0.008	0.00001	1.21	0.67–2.16	0.63	0.53
	GA vs. AA	Random	72	0.0008	1.09	0.73–1.62	0.43	0.67
	GG + GA vs. AA	Random	84	<0.00001	1.19	0.74–1.91	0.7	0.48
	GG vs. GA + AA	Random	70	0.002	1.14	0.75–1.72	0.61	0.55
rs1800871								
Overall	C vs. T	Random	84	<0.00001	1.11	088–1.39	0.87	0.39
	CC vs. TT	Random	75	<0.00001	1.12	0.75–1.66	0.55	0.58
	CT vs. TT	Random	67	0.0004	1.11	0.85–1.44	0.77	0.44
	CC + CT vs. TT	Random	78	<0.00001	1.12	0.84–1.50	0.77	0.44
	CC vs. CT + TT	Random	50	0.03	1	0.80–1.25	0.02	0.98
Ethnicity								
Asian	C vs. T	Random	89	<0.00001	0.94	0.68–1.31	0.35	0.72
	CC vs. TT	Random	79	0.0008	0.81	0.48–1.40	0.74	0.46
	CT vs. TT	Random	76	0.002	0.86	0.61–1.21	0.88	0.38
	CC + CT vs. TT	Random	82	0.0001	0.84	0.58–1.22	0.93	0.35
	CC vs. CT + TT	Random	38	0.17	0.93	0.72–1.20	0.59	0.56
Caucasian	C vs. T	Random	88	<0.0001	1.57	0.81–3.05	1.33	0.18
	CC vs. TT	Random	84	0.0004	1.11	0.85–1.44	0.77	0.44
	CT vs. TT	Random	45	0.14	1.8	1.03–3.13	2.07	**0.04**
	CC + CT vs. TT	Random	79	0.002	2.13	0.89–5.13	1.69	0.09
	CC vs. CT + TT	Random	78	0.003	1.64	0.69–3.90	1.12	0.26
rs1800872								
Overall	A vs. C	Random	89	<0.00001	0.82	0.66–1.03	1.7	0.09
	AA vs. CC	Random	83	<0.00001	0.71	0.46–1.08	1.6	0.11
	AC vs. CC	Random	59	0.004	0.93	0.78–1.11	0.8	0.43
	AA + AC vs. CC	Random	79	0.00001	0.86	0.68–1.08	1.31	0.19
	AA vs. AC + CC	Random	60	0.004	0.95	0.75–1.20	0.44	0.66
Ethnicity								
Asian	A vs. C	Fixed	0	0.54	0.85	0.74–0.98	2.22	**0.03**
	AA vs. CC	Fixed	23	0.27	0.74	0.55–0.99	2.04	**0.04**
	AC vs. CC	Fixed	0	0.88	0.88	0.69–1.13	0.96	0.34
	AA + AC vs. CC	Fixed	0	0.81	0.82	0.65–1.04	1.65	0.1
	AA vs. AC + CC	Fixed	25	0.26	0.82	0.66–1.01	1.84	0.07
Caucasian	A vs. C	Random	93	<0.00001	0.65	0.43–0.98	2.07	**0.04**
	AA vs. CC	Random	89	<0.00001	0.43	0.19–1.00	1.96	0.05
	AC vs. CC	Random	48	0.09	0.89	0.72–1.11	1.02	0.31
	AA + AC vs. CC	Random	82	<0.00001	0.74	0.52–1.05	1.69	0.09
	AA vs. AC + CC	Random	36	0.18	0.87	0.62–1.21	0.84	0.4

OR, odds ratio; PH, p *value of Heterogeneity; CI, confidence intervals; POR,* p *value of odds ratio.* p *value, significant at <0.05. Bold numbers denote statistical significance.*

polymorphic variations in the promoter sequences of the IL-10 gene might influence the gene expression and consequently play a specific role in susceptibility and the clinical course of BC. The IL-10 promoter region polymorphisms affected IL-10 gene transcription and translation, resulting in abnormal cell proliferation and cancer development (Moghimi et al., 2018; Sheikhpour et al., 2018).

Studies have shown that the three most common single nucleotide polymorphisms (SNPs) play an important role in regulating IL-10 activity. They are located at the transcriptional starting point of rs1800896 (-1082A/G), rs1800871 (-819T/C), and rs1800872 (-592A/C). And they encode high (GCC), medium (ACC), and low (ATA) expression of IL-10, respectively (Westendorp et al.,



A

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Kong 2010	31	630	39	644	15.4%	0.80 [0.49, 1.30]
Li 2014	34	256	52	256	19.0%	0.60 [0.37, 0.96]
Tian 2017	390	624	416	624	65.6%	0.83 [0.66, 1.05]
Total (95% CI)		**1510**		**1524**	**100.0%**	**0.78 [0.65, 0.95]**
Total events	455		507			

Heterogeneity: Chi² = 1.49, df = 2 (P = 0.47); I² = 0%
Test for overall effect: Z = 2.49 (P = 0.01)

B

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Kong 2010	1	286	2	287	4.3%	0.50 [0.05, 5.55]
Li 2014	2	98	4	84	9.2%	0.42 [0.07, 2.33]
Tian 2017	129	180	131	158	86.4%	0.52 [0.31, 0.88]
Total (95% CI)		**564**		**529**	**100.0%**	**0.51 [0.31, 0.84]**
Total events	132		137			

Heterogeneity: Chi² = 0.06, df = 2 (P = 0.97); I² = 0%
Test for overall effect: Z = 2.68 (P = 0.007)

C

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Kong 2010	29	314	35	320	29.0%	0.83 [0.49, 1.39]
Li 2014	30	126	44	124	31.2%	0.57 [0.33, 0.99]
Tian 2017	132	183	154	181	39.8%	0.45 [0.27, 0.76]
Total (95% CI)		**623**		**625**	**100.0%**	**0.60 [0.44, 0.81]**
Total events	191		233			

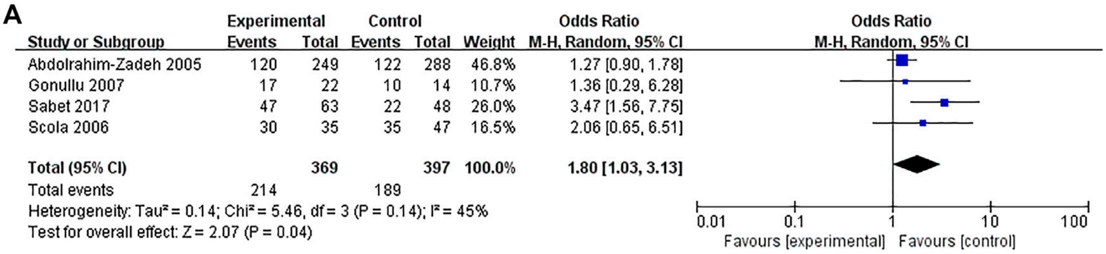
Heterogeneity: Chi² = 2.63, df = 2 (P = 0.27); I² = 24%
Test for overall effect: Z = 3.31 (P = 0.0009)

D

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Kong 2010	30	315	37	322	28.6%	0.81 [0.49, 1.35]
Li 2014	32	128	48	128	31.1%	0.56 [0.32, 0.95]
Tian 2017	261	312	285	312	40.3%	0.48 [0.30, 0.80]
Total (95% CI)		**755**		**762**	**100.0%**	**0.60 [0.45, 0.81]**
Total events	323		370			

Heterogeneity: Chi² = 2.13, df = 2 (P = 0.34); I² = 6%
Test for overall effect: Z = 3.39 (P = 0.0007)

FIGURE 2 | Forest plots showed a significant association of IL-10 rs1800896 polymorphism and breast cancer risk in Asians. **(A)** (allele model: G vs. A); **(B)** (homozygous model: GG vs. AA); **(C)** (heterozygous model: GA vs. AA); **(D)** (dominant model: GG + GA vs. AA). The squares and horizontal lines correspond to the study-specific odds ratio (OR) and 95% confidence interval (CI). The area of the squares reflects the weight (inverse of the variance). The diamond represents the summary OR and 95% CI.

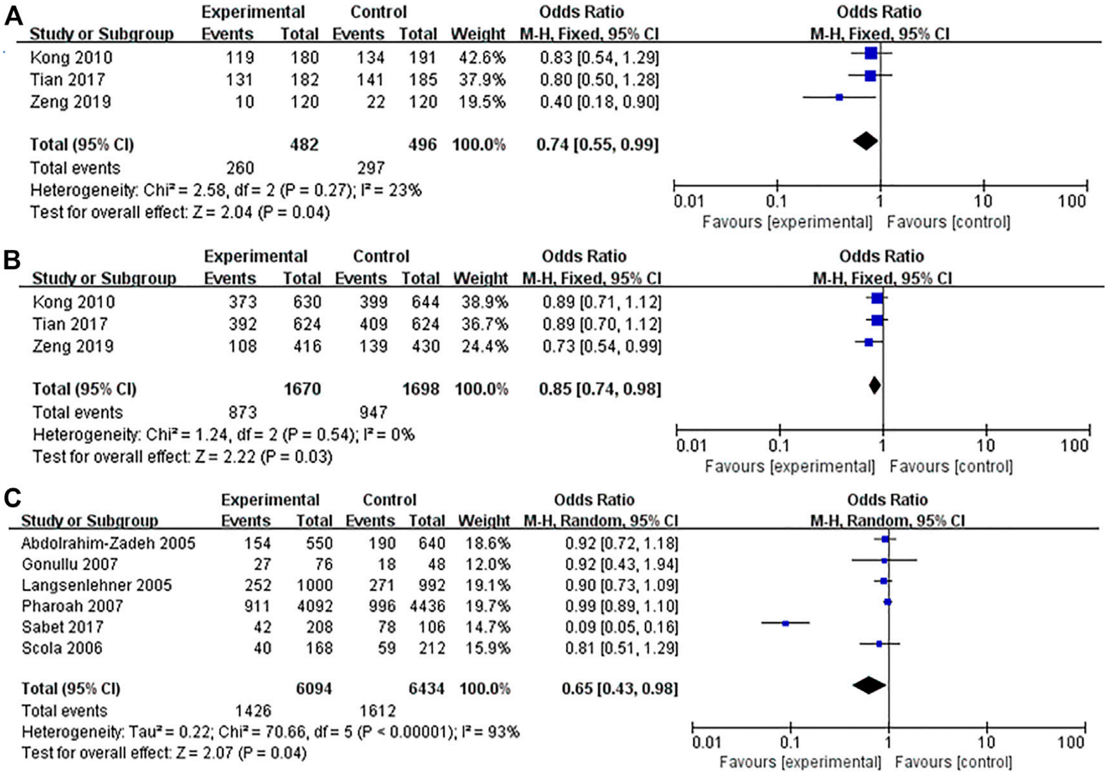


A

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Random, 95% CI
Abdolrahim-Zadeh 2005	120	249	122	288	46.8%	1.27 [0.90, 1.78]
Gonullu 2007	17	22	10	14	10.7%	1.36 [0.29, 6.28]
Sabet 2017	47	63	22	48	26.0%	3.47 [1.56, 7.75]
Scola 2006	30	35	35	47	16.5%	2.06 [0.65, 6.51]
Total (95% CI)		**369**		**397**	**100.0%**	**1.80 [1.03, 3.13]**
Total events	214		189			

Heterogeneity: Tau² = 0.14; Chi² = 5.46, df = 3 (P = 0.14); I² = 45%
Test for overall effect: Z = 2.07 (P = 0.04)

FIGURE 3 | Forest plots showed a significant association of IL-10 rs1800871 polymorphism and breast cancer risk in the Caucasians (heterozygous model: CT vs. TT). The squares and horizontal lines correspond to the study-specific odds ratio (OR) and 95% confidence interval (CI). The area of the squares reflects the weight (inverse of the variance). The diamond represents the summary OR and 95% CI.



A

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Kong 2010	119	180	134	191	42.6%	0.83 [0.54, 1.29]
Tian 2017	131	182	141	185	37.9%	0.80 [0.50, 1.28]
Zeng 2019	10	120	22	120	19.5%	0.40 [0.18, 0.90]
Total (95% CI)		**482**		**496**	**100.0%**	**0.74 [0.55, 0.99]**
Total events	260		297			

Heterogeneity: Chi² = 2.58, df = 2 (P = 0.27); I² = 23%
Test for overall effect: Z = 2.04 (P = 0.04)

B

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Kong 2010	373	630	399	644	38.9%	0.89 [0.71, 1.12]
Tian 2017	392	624	409	624	36.7%	0.89 [0.70, 1.12]
Zeng 2019	108	416	139	430	24.4%	0.73 [0.54, 0.99]
Total (95% CI)		**1670**		**1698**	**100.0%**	**0.85 [0.74, 0.98]**
Total events	873		947			

Heterogeneity: Chi² = 1.24, df = 2 (P = 0.54); I² = 0%
Test for overall effect: Z = 2.22 (P = 0.03)

C

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Random, 95% CI
Abdolrahim-Zadeh 2005	154	550	190	640	18.6%	0.92 [0.72, 1.18]
Gonullu 2007	27	76	18	48	12.0%	0.92 [0.43, 1.94]
Langsenlehner 2005	252	1000	271	992	19.1%	0.90 [0.73, 1.09]
Pharoah 2007	911	4092	996	4436	19.7%	0.99 [0.89, 1.10]
Sabet 2017	42	208	78	106	14.7%	0.09 [0.05, 0.16]
Scola 2006	40	168	59	212	15.9%	0.81 [0.51, 1.29]
Total (95% CI)		**6094**		**6434**	**100.0%**	**0.65 [0.43, 0.98]**
Total events	1426		1612			

Heterogeneity: Tau² = 0.22; Chi² = 70.66, df = 5 (P < 0.00001); I² = 93%
Test for overall effect: Z = 2.07 (P = 0.04)

FIGURE 4 | Forest plots showed a significant association of IL-10 rs1800872 polymorphism and breast cancer risk in Asians and Caucasians. **(A)** IL-10 rs1800872 polymorphism in Asians (homozygous model: AA vs. CC); **(B)** IL-10 rs1800872 polymorphism in Asians (allele model: A vs. C); **(C)** IL-10 rs1800872 polymorphism in Caucasians (allele model: A vs. C). The squares and horizontal lines correspond to the study-specific odds ratio (OR) and 95% confidence interval (CI). The area of the squares reflects the weight (inverse of the variance). The diamond represents the summary OR and 95% CI.

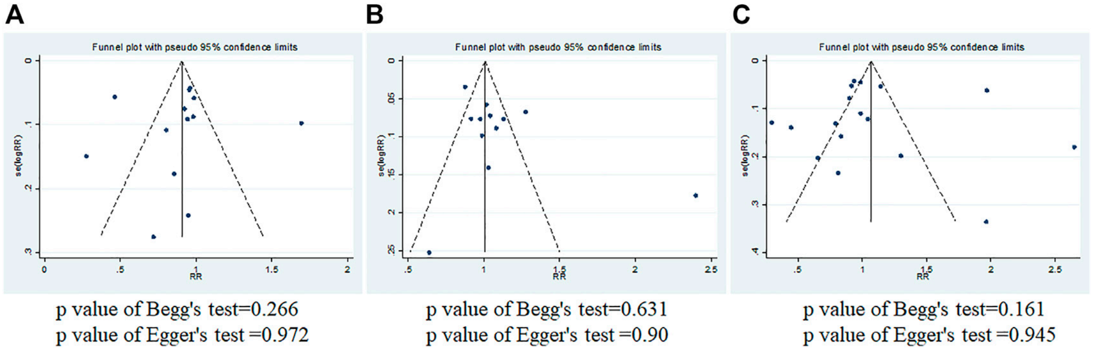


FIGURE 5 | Begg's and Egger's funnel plots of IL-10 gene polymorphism and breast cancer risk for publication bias test. **(A)** rs1800896; **(B)** rs1800871; **(C)** rs1800872.

1997; Zupin et al., 2014; Hofmann et al., 2018). Several other polymorphic loci of IL-10 (rs1800890, rs6703630, and rs6693899) are also controversial, but few relevant studies are present. Many studies have reported the relationship between race and IL-10 gene polymorphism and BC risk in recent years. For example, the IL-10 rs1800872 polymorphism was associated with BC susceptibility in the Mexican population (Patricia Gallegos-Arreola et al., 2019). Also, the mutant allele and genotypes of IL-10 rs1800896 were significantly associated with Indian postmenopausal BC (Pooja et al., 2012). Since the IL-10 gene polymorphisms were associated with the risk of BC, we hypothesized that race is the key to the association between IL-10 gene polymorphisms and BC. This meta-analysis conducted the most comprehensive analysis of the relationship between three IL-10 polymorphisms (rs1800896, rs1800871, and rs1800872) and the BC risk of different races. In a subgroup analysis by ethnicity (Asian and Caucasian/mixed race), the three IL-10 polymorphisms (rs1800896, rs1800871, and rs1800872) were significantly associated with BC. It showed that rs1800896 (allele G vs. A: OR = 0.78, 95% CI = 0.65–0.95, p = 0.01; homozygous GG vs. AA: OR = 0.51, 95% CI = 0.31–0.84, p = 0.007; heterozygous GA vs. AA: OR = 0.6, 95% CI = 0.44–0.81, p = 0.0009; dominant GG + GA vs. AA: OR = 0.6, 95% CI = 0.45–0.81, p = 0.0007) were significantly correlated with BC risk in Asians. The rs1800871 heterozygote model (CT vs. TT: OR = 1.8, 95% CI = 1.03–3.13, p = 0.04) was associated with BC risk in Caucasians. The rs1800872 homozygous model (AA vs CC: OR = 0.74, 95% CI = 0.55–0.99, p = 0.04) was associated with BC risk in Asians, and the allelic model (A vs. C: OR = 0.85, 95% CI = 0.74–0.98, p = 0.03) was associated with BC risk in Asians and Caucasians (A vs C: OR = 0.65, 95% = CI 0.43–0.98, p = 0.04). The above results indicate that the ethnic subgroup of IL-10 gene polymorphisms is the key factor affecting the susceptibility to BC. It is consistent with the results of previous studies: the relationship between IL-10 gene polymorphism and BC risk is strongly associated with ethnicity (Patricia Gallegos-Arreola et al., 2019).

Previously, three researchers (Dai et al., 2014; Abedinzadeh et al., 2018; Moghimi et al., 2018) have analyzed the correlation between IL-10 gene polymorphisms and BC risk, but their analysis is not comprehensive enough. Because there are few studies included and the ethnic division is not accurate enough in their articles. In addition, *Xu* and *Wang* did a meta-analysis on the relationship between various interleukins and BC. Still, their correlations between IL-10 gene polymorphisms and BC risk were inconsistent with ours (Xu and Wang, 2020). It may be related to the different criteria for inclusion and exclusion and quality assessment of the article. Because the quality, quantity, and new studies included in the meta-analysis will directly affect the credibility and stability of the results, we used a broad search strategy to capture all relevant information. This meta-analysis conducted a more comprehensive analysis of the relationship between three IL-10 polymorphisms (rs1800896, rs1800871, and rs1800872) and BC risk by including 23 studies (published between 2004 and 2019) and ruling out the researches with low quality. Moreover, this meta-analysis showed no significant publication bias, and the heterogeneity of the subgroups was small. Sensitivity analysis results were also stable. Therefore, the conclusion of the association between the three IL-10 gene polymorphisms (rs1800896, rs1800871, and rs1800872) and BC in this meta-analysis was reliable and had certain clinical guidance values.

However, this meta-analysis has several limitations that should be acknowledged. Firstly, due to the limited research on the interaction between these three polymorphic sites and their interaction with the environment, it is impossible to estimate the impact of gene-gene and gene-environment interaction on the study results. Secondly, we found that heterogeneity existed in the meta-analysis as indicated by the I^2 values. Despite using a random-effects model in some studies, the heterogeneity remained. It is predictable because other factors that affect BC should be considered, such as staging and grading of tumors, age, genetic background, environment, and lifestyle. However, due to the lack of some qualified original data, we cannot calculate the impact of these factors on BC. Moreover, in the future we need to consider more factors influencing BC, such as age, menopausal state, environment, and lifestyle factors, to further validate gene-gene and gene-environment interactions on IL-10 polymorphisms and BC risk.

CONCLUSION

In summary, this meta-analysis provides a new idea for clinical, genetic, and epidemiological studies of BC. Our results show that alleles, homozygotes, and dominant genotypes of IL-10 rs1800896 are significantly associated with the risk of BC in Asians. The homozygous and allele patterns of rs1800872 increase the risk of BC in Asians, while the heterozygous pattern of rs1800871 and the allele pattern of rs1800872 increase the risk of BC in Caucasians. IL-10 gene polymorphisms may be a key regulator of BC susceptibility. Different ethnic groups can predict BC susceptibility by detecting other IL-10 polymorphisms locus. However, the etiology of BC is complex, so we strongly recommend further genetic association studies to explore the effects of gene-gene interactions on disease susceptibility. Large-scale multicenter studies can be conducted in the future to verify further the results of gene-gene and gene-environment interactions on IL-10 gene polymorphisms and BC risk in different environments.

AUTHOR CONTRIBUTIONS

LL designed and managed the study. WX interpreted the data and drafted the manuscript. LL and JC performed data analysis and wrote the manuscript. DL conceived the idea and supervised the study. All authors read and approved the final manuscript.

FUNDING

The present study was supported by the University-level Foundation of the University of South China (nk2020205).

ACKNOWLEDGMENTS

The also authors would like to thank reviewers for their careful reading and valuable suggestions.

REFERENCES

Abdolrahim-Zadeh, H., Hakkakian, N., Asadollahi, R., Gharesifard, B., Sarvari, J., Kamali, E., et al. (2005). Interleukin-10 Promoter Polymorphisms and Breast Cancer Risk in Iranian Women. *Iranian J. Immunol.* 2. 158–165.

Abedinzadeh, M., Neamatzadeh, H., Jafari, M., Forat-Yazdi, M., Nasiri, R., Farahnak, S., et al. (2018). Association of Interleukin-10 -1082A>G (Rs1800896) Polymorphism with Predisposition to Breast Cancer: a Meta-Analysis Based on 17 Case-Control Studies. *Rev. Assoc. Med. Bras.* 64, 756–764. doi:10.1590/1806-9282.64.08.756

Alotaibi, M. R., Hassan, Z. K., Al-Rejaie, S. S., Alshammari, M. A., Almutairi, M. M., Alhoshani, A. R., et al. (2018). Characterization of Apoptosis in a Breast

Cancer Cell Line after IL-10 Silencing. *Asian Pac. J. Cancer Prev.* 19, 777–783. doi:10.22034/APJCP.2018.19.3.777

Alsuhaibani, E. S., Kizilbash, N. A., Malik, S., Dasti, J. I., Al Beladi, F., and El-Morshedi, N. (2016). Polymorphisms in Promoter Regions of IL-6 and IL-10 Genes in Breast Cancer: a Case-Control Study. *Genet. Mol. Res.* 15, gmr.150173. doi:10.4238/gmr.15017360

Anastasiadi, Z., Lianos, G. D., Ignatiadou, E., Harissis, H. V., and Mitsis, M. (2017). Breast Cancer in Young Women: an Overview. *Updates Surg.* 69, 313–317. doi:10.1007/s13304-017-0424-1

Atoum, M. (2016). ACC Interleukin-10 Gene Promoter Haplotype as a Breast Cancer Risk Factor Predictor Among Jordanian Females. *Onco Targets Ther.* 9, 3353–3357. doi:10.2147/OTT.S101628

Azher, A., Al-Ankoshy, M., Hamid, A., and Dawood Alatbee, A. (2019). The Impact of IL-10 Gene Polymorphism on Progressive Breast Cancer. *J. Pharm. Sci. Res.* 11 (1), 93–97.

Balasubramanian, S., Azmy, I., Higham, S., Wilson, A., Cross, S., Cox, A., et al. (2006). Interleukin Gene Polymorphisms and Breast Cancer: a Case Control Study and Systematic Literature Review. *BMC Cancer* 6, 188. doi:10.1186/1471-2407-6-188

Bhattacharjee, H. K., Bansal, V. K., Nepal, B., Srivastava, S., Dinda, A. K., and Misra, M. C. (2016). Is Interleukin 10 (IL10) Expression in Breast Cancer a Marker of Poor Prognosis? *Indian J. Surg. Oncol.* 7, 320–325. doi:10.1007/s13193-016-0512-6

Chen, L., Shi, Y., Zhu, X., Guo, W., Zhang, M., Che, Y., et al. (2019). IL-10 Secreted by Cancer-associated Macrophages Regulates Proliferation and Invasion in Gastric Cancer Cells via c-Met/STAT3 Signaling. *Oncol. Rep.* 42, 595–604. doi:10.3892/or.2019.7206

Dai, Z.-J., Wang, X.-J., Zhao, Y., Ma, X.-B., Kang, H.-F., Min, W.-L., et al. (2014). Effects of Interleukin-10 Polymorphisms (Rs1800896, Rs1800871, and Rs1800872) on Breast Cancer Risk: Evidence from an Updated Meta-Analysis. *Genet. Test. Mol. Biomarkers* 18, 439–445. doi:10.1089/gtmb.2014.0012

Fanyu, Z., Wenhui, L., Shan, Z., Wei, T., Xiaofen, Z., and Qiujin, Z. (2019). Association of Interleukin-10 Gene Polymorphism with Susceptibility of Breast Cancer in Chinese Women of Guangxi Province. *Guo ji mian yi xue za zhi* 042, 464–467. doi:10.1016/j.jclinepi.2009.06.005

Gerger, A., Renner, W., Langsenlehner, T., Hofmann, G., Knechtel, G., Szkandera, J., et al. (2010). Association of Interleukin-10 Gene Variation with Breast Cancer Prognosis. *Breast Cancer Res. Treat.* 119, 701–705. doi:10.1007/s10549-009-0417-y

Gonullu, G., Basturk, B., Evrensel, T., Oral, B., Gozkaman, A., and Manavoglu, O. (2007). Association of Breast Cancer and Cytokine Gene Polymorphism in Turkish Women. *Saudi Med. J.* 28, 1728–1733.

Guzowski, D., Chandrasekaran, A., Gawel, C., Palma, J., Koenig, J., Wang, X. P., et al. (2005). Analysis of Single Nucleotide Polymorphisms in the Promoter Region of Interleukin-10 by Denaturing High-Performance Liquid Chromatography. *J. Biomol. Tech.* 16, 154–166.

Ha, S. M., Chae, E. Y., Cha, J. H., Kim, H. H., Shin, H. J., and Choi, W. J. (2017). Association of BRCA Mutation Types, Imaging Features, and Pathologic Findings in Patients with Breast Cancer with BRCA1 and BRCA2 Mutations. *Am. J. Roentgenol.* 209, 920–928. doi:10.2214/AJR.16.16957

He, J.-R., Chen, L.-J., Su, Y., Cen, Y.-L., Tang, L.-Y., Yu, D.-D., et al. (2012). Joint Effects of Epstein-Barr Virus and Polymorphisms in Interleukin-10 and Interferon-γ on Breast Cancer Risk. *J. Infect. Dis.* 205, 64–71. doi:10.1093/infdis/jir710

Hofmann, S. R., Laass, M. W., Fehrs, A., Schuppan, D., Zevallos, V. F., Salminger, D., et al. (2018). IL10 Promoter Haplotypes May Contribute to Altered Cytokine Expression and Systemic Inflammation in Celiac Disease. *Clin. Immunol.* 190, 15–21. doi:10.1016/j.clim.2018.02.010

Howell, W. M., and Rose-Zerilli, M. J. (2007). Cytokine Gene Polymorphisms, Cancer Susceptibility, and Prognosis. *J. Nutr.* 137, 194S–199S. doi:10.1093/jn/137.1.194S

Kong, F., Liu, J., Liu, Y., Song, B., Wang, H., and Liu, W. (2010). Association of Interleukin-10 Gene Polymorphisms with Breast Cancer in a Chinese Population. *J. Exp. Clin. Cancer Res.* 29, 72. doi:10.1186/1756-9966-29-72

Kwong, A., Shin, V. Y., Ho, J. C. W., Kang, E., Nakamura, S., Teo, S.-H., et al. (2016). Comprehensive Spectrum ofBRCA1andBRCA2deleterious Mutations in Breast Cancer in Asian Countries. *J. Med. Genet.* 53, 15–23. doi:10.1136/jmedgenet-2015-103132

Langsenlehner, U., Krippl, P., Renner, W., Yazdani-Biuki, B., Eder, T., Köppel, H., et al. (2005). Interleukin-10 Promoter Polymorphism Is Associated with Decreased Breast Cancer Risk. *Breast Cancer Res. Treat.* 90, 113–115. doi:10.1007/s10549-004-3607-7

Li, M., Yue, C., Zuo, X., Jin, G., Wang, G., Guo, H., et al. (2020). The Effect of Interleukin 10 Polymorphisms on Breast Cancer Susceptibility in Han Women in Shaanxi Province. *PLoS One* 15, e0232174. doi:10.1371/journal.pone.0232174

Li, Y., Yu, H., Jiao, S., and Yang, J. (2014). Prognostic Value of IL-10 Expression in Tumor Tissues of Breast Cancer Patients. *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi* 30, 517–520.

Maruthi, G., Ramachander, V., Komaravalli, P. L., Sureka, T., and Jahan, P. (2017). Association of Il10 Gene Polymorphisms (Rs 1800896, Rs1800872) in Breast Cancer Patients. *Int. J. Med. Sci. Clin. Invention* 4, 3050–3061. doi:10.18535/ijmsci/v4i6.19

Meijiang, L. (2014). The Relationship between Interleukin Gene Polymorphism and Females Genetic Susceptibility to Breast Cancer in Western Guangxi. *J. Youjiang Med. Univ. Nationalities* 36, 339–343. doi:10.3969/j.issn.1001-5817.2014.03.007

Moghimi, M., Ahrar, H., Karimi-Zarchi, M., Aghili, K., Salari, M., Zare-Shehneh, M., et al. (2018). Association of IL-10 Rs1800871 and Rs1800872 Polymorphisms with Breast Cancer Risk: A Systematic Review and Meta-Analysis. *Asian Pac. J. Cancer Prev.* 19, 3353–3359. doi:10.31557/apjcp.2018.19.12.3353

Mohamed, H. T., El-Husseiny, N., El-Ghonaimy, E. A., Ibrahim, S. A., Bazzi, Z. A., Cavallo-Medved, D., et al. (2018). IL-10 Correlates with the Expression of Carboxypeptidase B2 and Lymphovascular Invasion in Inflammatory Breast Cancer: The Potential Role of Tumor Infiltrated Macrophages. *Curr. Probl. Cancer* 42, 215–230. doi:10.1016/j.currproblcancer.2018.01.009

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA Statement. *J. Clin. Epidemiol.* 62, 1006–1012. doi:10.1016/j.jclinepi.2009.06.005

Mubarik, S., Wang, F., Fawad, M., Wang, Y., Ahmad, I., and Yu, C. (2020). Trends and Projections in Breast Cancer Mortality Among Four Asian Countries (1990-2017): Evidence from Five Stochastic Mortality Models. *Sci. Rep.* 10, 5480. doi:10.1038/s41598-020-62393-1

Oblak, T., Zadnik, V., Krajc, M., Lokar, K., and Zgajnar, J. (2020). Breast Cancer Risk Based on Adapted IBIS Prediction Model in Slovenian Women Aged 40-49 Years - Could it Be Better? *Radiol. Oncol.* 54, 335–340. doi:10.2478/raon-2020-0040

Onay, V. Ü., Briollais, L., Knight, J. A., Shi, E., Wang, Y., Wells, S., et al. (2006). SNP-SNP Interactions in Breast Cancer Susceptibility. *BMC Cancer* 6, 114. doi:10.1186/1471-2407-6-114

Patricia Gallegos-Arreola, M., Zúñiga González, G. M., Figuera, L. E., Puebla Pérez, A. M., and Delgado Saucedo, J. I. (2019). Association of the IL-10 Gene Rs1800872 (-592 C>A) Polymorphism with Breast Cancer in a Mexican Population. *J. BUON* 24, 2369–2376.

Pharoah, P. D. P., Tyrer, J., Dunning, A. M., Easton, D. F., Ponder, B. A. J., and Investigators, S. (2007). Association between Common Variation in 120 Candidate Genes and Breast Cancer Risk. *Plos Genet.* 3, e42. doi:10.1371/journal.pgen.0030042

Pooja, S., Chaudhary, P., Nayak, L. V., Rajender, S., Saini, K. S., Deol, D., et al. (2012). Polymorphic Variations in IL-1β, IL-6 and IL-10 Genes, Their Circulating Serum Levels and Breast Cancer Risk in Indian Women. *Cytokine* 60, 122–128. doi:10.1016/j.cyto.2012.06.241

Roh, J. W., Kim, M. H., Seo, S. S., Kim, S. H., Kim, J. W., Park, N. H., et al. (2002). Interleukin-10 Promoter Polymorphisms and Cervical Cancer Risk in Korean Women. *Cancer Lett.* 184, 57–63. doi:10.1016/s0304-3835(02)00193-3

Sabet, S., El-Sayed, S. K., Mohamed, H. T., El-Shinawi, M., and Mohamed, M. M. (2017). Inflammatory Breast Cancer: High Incidence of GCC Haplotypes (−1082A/G, −819T/C, and −592A/C) in the Interleukin-10 Gene Promoter Correlates with Over-expression of Interleukin-10 in Patients' Carcinoma Tissues. *Tumour Biol.* 39, 101042831771339. doi:10.1177/1010428317713393

Sancho-Garnier, H., and Colonna, M. (2019). Épidémiologie des cancers du sein. *La Presse Médicale* 48, 1076–1084. doi:10.1016/j.lpm.2019.09.022

Scola, L., Vaglica, M., Crivello, A., Palmeri, L., Forte, G. I., Macaluso, M. C., et al. (2006). Cytokine Gene Polymorphisms and Breast Cancer Susceptibility. *Ann. N.Y Acad. Sci.* 1089, 104–109. doi:10.1196/annals.1386.017

Setrerrahmane, S., and Xu, H. (2017). Tumor-related Interleukins: Old Validated Targets for New Anti-cancer Drug Development. *Mol. Cancer* 16, 153. doi:10.1186/s12943-017-0721-9

Sheikhpour, E., Noorbakhsh, P., Foroughi, E., Farahnak, S., Nasiri, R., and Neamatzadeh, H. (2018). A Survey on the Role of Interleukin-10 in Breast Cancer: a Narrative. *Rep. Biochem. Mol. Biol.* 7, 30–37.

Smith, K. C., Bateman, A. C., Fussell, H. M., and Howell, W. M. (2004). Cytokine Gene Polymorphisms and Breast Cancer Susceptibility and Prognosis. *Eur. J. Immunogenet.* 31, 167–173. doi:10.1111/j.1365-2370.2004.00462.x

Stang, A. (2010). Critical Evaluation of the Newcastle-Ottawa Scale for the Assessment of the Quality of Nonrandomized Studies in Meta-Analyses. *Eur. J. Epidemiol.* 25, 603–605. doi:10.1007/s10654-010-9491-z

Tian, K., Zhang, R., and Wang, X. (2017). Association of Interleukin-10 Polymorphisms and Haplotypes with the Risk of Breast Cancer in Northern China. *Int. J. Clin. Exp. Pathol.* 10, 6989–6996.

Vinod, C., Jyothy, A., Vijay Kumar, M., Raghu Raman, R., Nallari, P., and Venkateshwari, A. (2015). A Common SNP of IL-10 (-1082A/G) Is Associated with Increased Risk of Premenopausal Breast Cancer in South Indian Women. *Iran J. Cancer Preven* 8, e3434. doi:10.17795/ijcp-3434

Wang, Z., Liu, Q.-L., Sun, W., Yang, C.-J., Tang, L., Zhang, X., et al. (2014). Genetic Polymorphisms in Inflammatory Response Genes and Their Associations with Breast Cancer Risk. *Croat. Med. J.* 55, 638–646. doi:10.3325/cmj.2014.55.638

Westendorp, R. G., Langermans, J. A., Huizinga, T. W., Elouali, A. H., Verweij, C. L., Boomsma, D. I., et al. (1997). Genetic Influence on Cytokine Production and Fatal Meningococcal Disease. *The Lancet* 349, 170–173. doi:10.1016/s0140-6736(96)06413-6

Xu, G., and Wang, F. (2020). Associations of Polymorphisms in Interleukins with Susceptibility to Breast Cancer: Evidence from a Meta-Analysis. *Cytokine* 130, 154988. doi:10.1016/j.cyto.2020.154988

Zeng, F., Liu, W., Zhang, S., Tang, W., Zhao, X., and Zhang, Q. (2019). Correlation Analysis of IL-10 Gene Polymorphism and Breast Cancer Susceptibility in Women from Guangxi, China. *Guo Ji Mian Yi Xue Za Zhi* 42, 464–467. doi:10.3760/cma.j.issn.1673-4394.2019.05.004

Zhao, S., Wu, D., Wu, P., Wang, Z., and Huang, J. (2015). Serum IL-10 Predicts Worse Outcome in Cancer Patients: A Meta-Analysis. *PLoS One* 10, e0139598. doi:10.1371/journal.pone.0139598

Zhao, Y., Chen, S., Shen, F., Long, D., Yu, T., Wu, M., et al. (2019). *In Vitro* neutralization of Autocrine IL-10 Affects Op18/stathmin Signaling in Non-Small Cell Lung Cancer Cells. *Oncol. Rep.* 41, 501–511. doi:10.3892/or.2018.6795

Zhu, Z., Liu, J.-B., Liu, X., and Qian, L. (2020). Association of Interleukin 10 Rs1800896 Polymorphism with Susceptibility to Breast Cancer: a Meta-Analysis. *J. Int. Med. Res.* 48, 030006052090486. doi:10.1177/0300060520904863

Zupin, L., Polesello, V., Catamo, E., Crovella, S., and Segat, L. (2014). Interleukin-10 Gene Promoter Polymorphisms in Celiac Patients from north-eastern Italy. *Hum. Immunol.* 75, 656–661. doi:10.1016/j.humimm.2014.04.011

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/fgene.2022.768971

An Effective Hypoxia-Related Long Non-Coding RNA Assessment Model for Prognosis of Lung Adenocarcinoma

Yuanshuai Li[1,2] *and Xiaofang Sun*[1,2]*

[1]*Department of Obstetrics and Gynecology, Key Laboratory for Major Obstetric Diseases of Guangdong Province, The Third Affiliated Hospital of Guangzhou Medical University, Guangzhou, China,* [2]*Key Laboratory of Reproduction and Genetics of Guangdong Higher Education Institutes, Guangzhou, China*

Background: Lung adenocarcinoma (LUAD) represents one of the highest incidence rates worldwide. Hypoxia is a significant biomarker associated with poor prognosis of LUAD. However, there are no definitive markers of hypoxia-related long non-coding RNAs (lncRNAs) in LUAD.

Methods: From The Cancer Genome Atlas (TCGA) and the Molecular Signatures Database (MSigDB), we acquired the expression of hypoxia-related lncRNAs and corresponding clinical information of LUAD patients. The hypoxia-related prognostic model was constructed by univariable COX regression analysis, least absolute shrinkage and selection operator (LASSO), and multivariable Cox regression analysis. To assess the performance of the model, the Kaplan–Meier (KM) survival and receiver operating characteristic (ROC) curve analyses were performed.

Results: We found seven lncRNAs, AC022613.1, AC026355.1, GSEC, LINC00941, NKILA, HSPC324, and MYO16-AS1, as biomarkers of the potential hypoxia-related prognostic signature. In the low-risk group, patients had a better overall survival (OS). In addition, the results of ROC analysis indicated that the risk score predicted LUAD prognosis exactly. Furthermore, combining the expression of lncRNAs with clinical features, two predictive nomograms were constructed, which could accurately predict OS and had high clinical application value.

Conclusion: In summary, the seven-lncRNA prognostic signature related to hypoxia might be useful in predicting clinical outcomes and provided new molecular targets for the research of LUAD patients.

Keywords: lung adenocarcinoma, hypoxia-related prognostic signature, immune infiltrates, long non-coding RNAs, nomogram

OPEN ACCESS

Edited by:
Jianjiong Gao,
Memorial Sloan Kettering Cancer Center, United States

Reviewed by:
Pinaki Saha,
KU Leuven, Belgium
Bin Liu,
Beijing Institute of Technology, China

Correspondence:
Xiaofang Sun
xiaofangsun@gzhmu.edu.cn

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

Received: *01 September 2021*
Accepted: *03 February 2022*
Published: *16 March 2022*

Citation:
Li Y and Sun X (2022) An Effective Hypoxia-Related Long Non-Coding RNA Assessment Model for Prognosis of Lung Adenocarcinoma. Front. Genet. 13:768971. doi: 10.3389/fgene.2022.768971

Abbreviations: C-index, c concordance index; DELs, differentially expressed lncRNAs; GO, Gene Ontology; GSEA, Gene Set Enrichment Analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; KM, Kaplan–Meier analysis; LUAD, lung adenocarcinoma; LASSO, least absolute shrinkage and selection operator; LncRNAs, long non-coding RNAs; MsigDB, The Molecular Signatures Database; OS, overall survival time; PCA, principal component analysis; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas database; WGCNA, weighted gene co-expression network analysis.

INTRODUCTION

LUAD is the most common pathological subtype in lung cancer, accounting for 40% of all lung cancer incidences (Lemjabbar-Alaoui et al., 2015; Shi et al., 2016). The mean 5-year survival rate of patients was only 15% (Odintsov et al., 2021). The major reason for the high mortality of LUAD is that LUAD is diagnosed at an advanced stage in most patients. Currently, the diagnosis of LUAD is primarily based on symptoms, which included the size and location of the tumor, the location of the tumor in the lymph nodes, and where the cancer has spread (Lemjabbar-Alaoui et al., 2015; Rami-Porta et al., 2017; Carter et al., 2018). Although many potential biomarkers for early detection of LUAD have been studied, such as autophagy-related survival model, immune-related survival model, and ferroptosis-related gene signature, there is still lack of clinically used biomarkers due to lack of sensitivity and validity of these biomarkers on the development of LUAD (Hirsch et al., 2017; Chen et al., 2021; Shi et al., 2021; Wu et al., 2021; Jiang et al., 2021; Li et al., 2021).

As a non–protein-coding RNA, lncRNA has approximately 200 nucleotides in length, and it has attracted much attention because of its ability to regulate gene expression in epigenetic, transcriptional, and posttranscriptional dimensions (Chang et al., 2016; Fang and Fullwood, 2016; Li et al., 2016; Bhan et al., 2017; Peng et al., 2017). LncRNAs significantly affected the development of tumors (Chang et al., 2016; Choudhry et al., 2016). Recently, lncRNA-related prognostic models have been extensively developed for many cancers, including gastric cancer, lung cancer, pancreatic cancer, breast cancer, and colorectal cancer (Guo et al., 2020; Wang et al., 2020; Zhang H et al., 2021).

Hypoxia is one of the main characteristics of the tumor microenvironment (TME) and usually associated with poor prognosis. According to the study, many lncRNAs play a regulatory role in the hypoxia of tumors, such as participating in the regulation of tumor growth, vascular formation, invasion, and metastasis. Under hypoxic conditions, the lncRNA HABON could promote growth and proliferation of hepatocarcinoma cells (Ma C et al., 2021; Ma T et al., 2021). In the hypoxic environment of gastric cancer, the expression of the lncRNA LINC00460 is upregulated and promotes tumor invasiveness (Chen et al., 2020). The hypoxia-regulated lncRNA H19 and PDK1 (pyruvate dehydrogenase kinase 1) expression exhibits strong correlations in primary breast carcinomas, and they promote reprogramming of cancer stem cells (Peng et al., 2018). However, there are no exact prognostic markers related to hypoxia-related lncRNAs in LUAD.

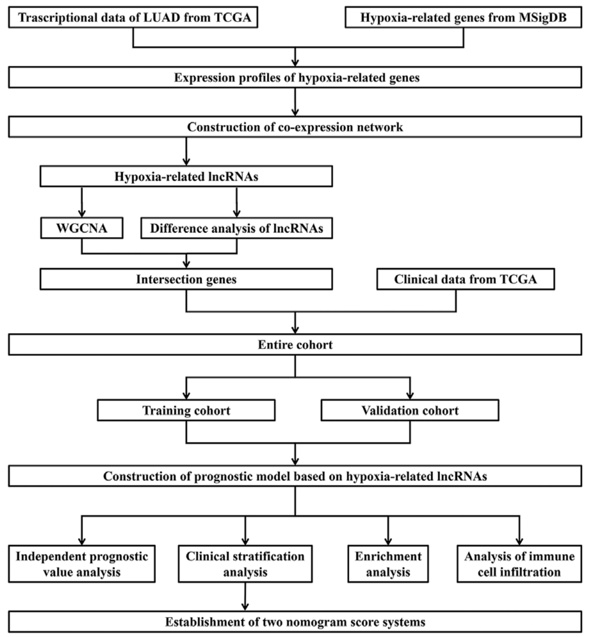

Trascriptional data of LUAD from TCGA
Hypoxia-related genes from MSigDB
Expression profiles of hypoxia-related genes
Construction of co-expression network
Hypoxia-related lncRNAs
WGCNA
Difference analysis of lncRNAs
Intersection genes
Clinical data from TCGA
Entire cohort
Training cohort
Validation cohort
Construction of prognostic model based on hypoxia-related lncRNAs
Independent prognostic value analysis
Clinical stratification analysis
Enrichment analysis
Analysis of immune cell infiltration
Establishment of two nomogram score systems

FIGURE 1 | Flow chart of data acquisition and analysis.

In this study, we identified seven hypoxia-related lncRNAs strongly associated with OS. Meanwhile, a risk signature was constructed. By using the other cohort, the accuracy and reliability of this model were validated. Moreover, we found that the signature was independent of clinical features. In conclusion, we successfully established a risk model associated with hypoxia. Moreover, it may be used for clinical treatment and diagnosis.

MATERIALS AND METHODS

Ethics Statement

The RNA-sequencing and clinical data of LUAD were downloaded from TCGA database (https://cancergenome.nih.gov/). Our study was based on the open resource data that were free for researching and

TABLE 1 | Basic clinical information of 499 LUAD patients from TCGA.

Variables	LUAD patients (N = 499)
Gender	
Female	270 (54%)
Male	229 (46%)
Age	
≤65 years	240 (48%)
>65 years	259 (52%)
Pathologic T Stage	
T1	170 (34%)
T2	263 (53%)
T3	45 (9%)
T4	18 (4%)
Unknown	3 (0%)
Pathologic M Stage	
M0	331 (66%)
M1	24 (5%)
Unknown	144 (29%)
Pathologic N Stage	
N0	322 (65%)
N1	95 (19%)
N2	69 (14%)
N3	2 (0%)
Unknown	11 (2%)
Pathologic Stage	
Stage I	266 (53%)
Stage II	120 (24%)
Stage III	80 (16%)
Stage IV	25 (5%)
Unknown	8 (2%)

Note: T, tumor size; M, distant metastasis; N, lymph node.

publishing relevant articles with no ethical issues and other conflicts of interest. The process of this study is presented in **Figure 1**.

Data Acquirement of TCGA

The mRNA expression data were derived from 535 LUAD patients and 59 healthy controls. Meanwhile, corresponding clinico-pathological data, including gender; age; pathologic T, M, and N stage; tumor clinical stage; and overall survival (OS) time were also obtained from TCGA database. Twenty-three out of 522 patients were excluded due to lack of information of OS (or the OS time was zero); therefore, lncRNA expression of 499 patients and their clinico-pathologic data were used for analysis. We presented the basic clinical information of these patients in **Table 1**.

Hypoxia-Related LnRNA Extraction

We performed Gene Set Enrichment Analysis (GSEA) to research two datasets associated with hypoxia (HARRIS_HYPOXIA, WINTER_HYPOXIA_METAGENE) from the MsigDB database. Then, the genes in all samples were analyzed in the abovementioned gene sets. There were 239 hypoxia-related genes found from the statistically significant gene set. Finally, by Pearson's correlation analysis, we identified hypoxia-related lncRNAs with the criteria of |correlation coefficient| > 0.3 and p-value < 0.001 and then constructed an mRNA–lncRNA coexpression network connected with hypoxia.

Identification of Hypoxia-Related LncRNAs

The Wilcoxon test was utilized to screen the differential expressed lncRNAs (DELs) between tumor and normal samples. The genes with p-value < 0.05 and $|\log_2$ fold-change (FC)$|$ > 1 were defined as DELs. Then, we utilized the R package "WGCNA" to construct a scale-free coexpression network for the all hypoxia-related lncRNAs by setting the soft threshold power value to 4. Finally, we selected two models highly correlated with cancer samples for followed analysis.

Construction and Validation of the Hypoxia-Related LncRNA Prognostic Signature

We first took the intersection of 601 DELs and 617 lncRNAs in the two modules (blue: 394, brown: 223). By using univariate Cox analysis, survival-related lncRNAs associated with hypoxia were identified. Then, LASSO regression analysis was used to further screen genes. At this time, we randomly divided the LUAD samples into two cohorts, training and validation cohorts. Finally, the lncRNAs were selected to construct a multivariate Cox regression model, and the risk score was calculated. Based on the median score of the training cohort, the patients were divided into high- and low-risk subgroups. Between the two groups, KM analysis was used to compare the survival time, and the ROC curve was used to evaluate the predictive power of the signature. In this way, the prognostic signature was constructed. In addition, in the other cohort, we performed the same procedure to evaluate the correctness of it.

Functional Enrichment Analysis

We utilized GSEA v4.1 (http://www.gsea-msigdb.org/gsea/index.jsp) to perform GSEA between the low- and high-risk groups. After that, we carried out the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of the differentially expressed mRNAs (DEGs) between the two groups, using "limma" and "clusterProfiler" packages.

Correlation Analysis of the Tumor Microenvironment in 33 the Cancer Genome Atlas Pan-Cancers

We then downloaded 33 cancer types from the UCSC Xena database (https://xenabrowser.net/datapages/), and they are adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), esophageal carcinoma (ESCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), testicular germ cell tumors (TGCT), thyroid carcinoma (THCA), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), cholangiocarcinoma (CHOL), stomach adenocarcinoma (STAD), and uveal melanoma (UVM). Meanwhile, six types of immune infiltration, namely, C1 (wound healing), C2 (INF-r dominant), C3 (inflammatory), C4 (lymphocyte depleted), C5 (immunologically quiet), and C6 (TGF-β dominant) were also downloaded from it. Spearman's analysis was used to calculate the association of the lncRNAs with immune subtypes, tumor mutation burden (TMB), and stemness score (RNAss and DNAss).

Statistical Analysis

All statistical analyses were accomplished with R software (version 4.0.5). The DEGs were identified by the Wilcoxon test. We used Pearson's correlation analysis to calculate the correlation between lncRNAs and mRNAs associated with hypoxia. Meanwhile, we applied univariate and multivariate Cox regression analyses to evaluate the correlation between the risk score and clinical features. P-value < 0.05 was regarded as a significant outcome.

RESULTS

Identification of Hypoxia-Related LncRNAs in LUAD

Through GSEA, we found that the WINTER_HYPOXIA_METAGEN gene set was significantly

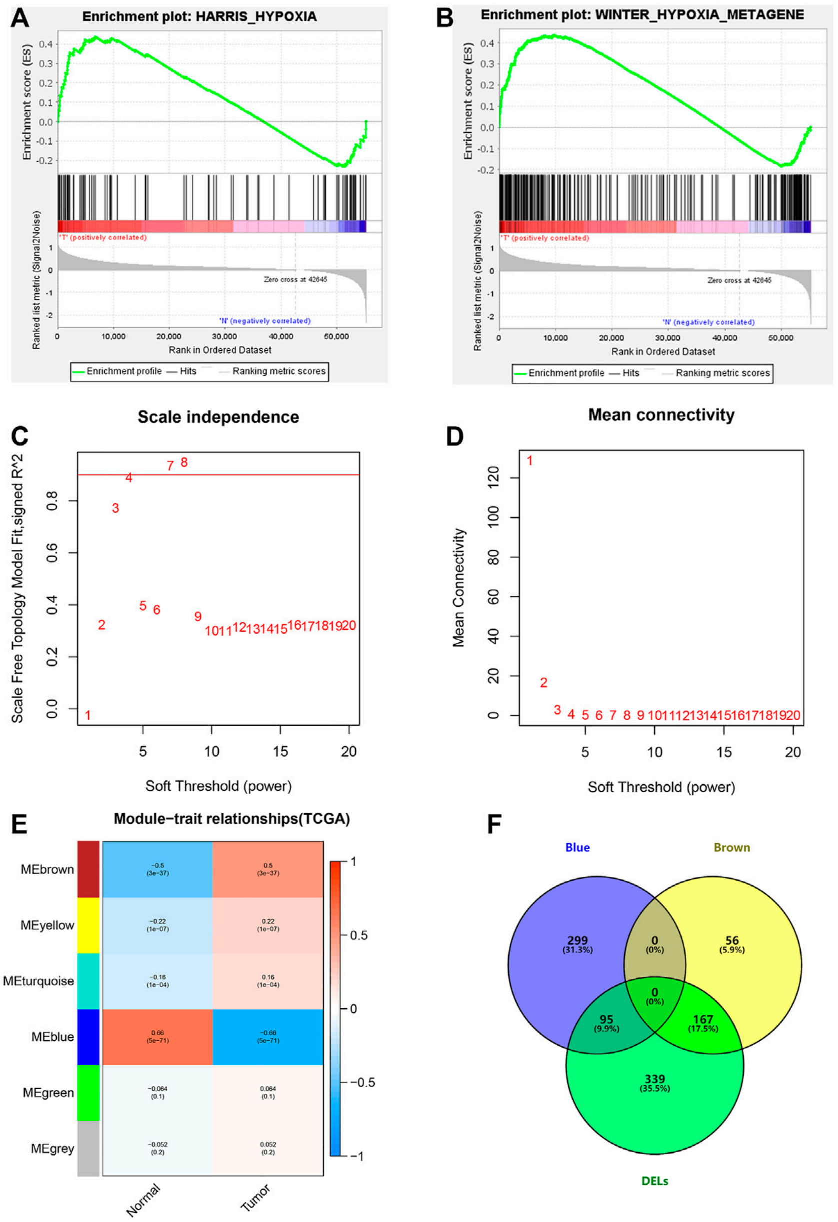


FIGURE 2 | Hypoxia-related lncRNA extraction. **(A,B)** GSEA analysis showing the most enriched hypoxia-related pathways. **(C,D)** Determination of the most suitable power value for scale-free coexpression network. **(E)** Brown and blue modules were the most correlated with the tumor state. **(F)** Identification of common genes between DELs and the brown and the blue modules by overlapping them.

enriched [(FDR = 0.040), (**Supplementary Table S1**)], while the HARRIS_HYPOXIA gene set was not [(FDR = 0.164), (**Figures 2A,B**)]. Next, 1,629 hypoxia-related lncRNAs were screened according to the significant gene set (**Supplementary Table S2**).

We further studied hypoxia-related lncRNAs by the means of the Wilcoxon test and the weighted gene co-expression network (WGCNA) analysis, and we used the 601 differentially expressed hypoxia-related lncRNAs between normal and tumor samples to intersect with the 617 lncRNAs from the two modules of WGCNA analysis (**Figures 2C–F**).

Construction of Hypoxia-Related LncRNA Prognostic Signature for LUAD

After taking the intersection, we got 262 hypoxia-related lncRNAs (**Supplementary Table S3**). We then used these lncRNAs to construct the prognostic signature. First, 20 lncRNAs were screened by univariate Cox analysis [(p-value < 0.05), (**Supplementary Table S4**)]. Then, by using LASSO analysis, 20 variables were reduced to 11 potential predictors (**Figures 3A,B**). Finally, seven lncRNAs were identified by the multivariate Cox regression analysis in the training cohort.

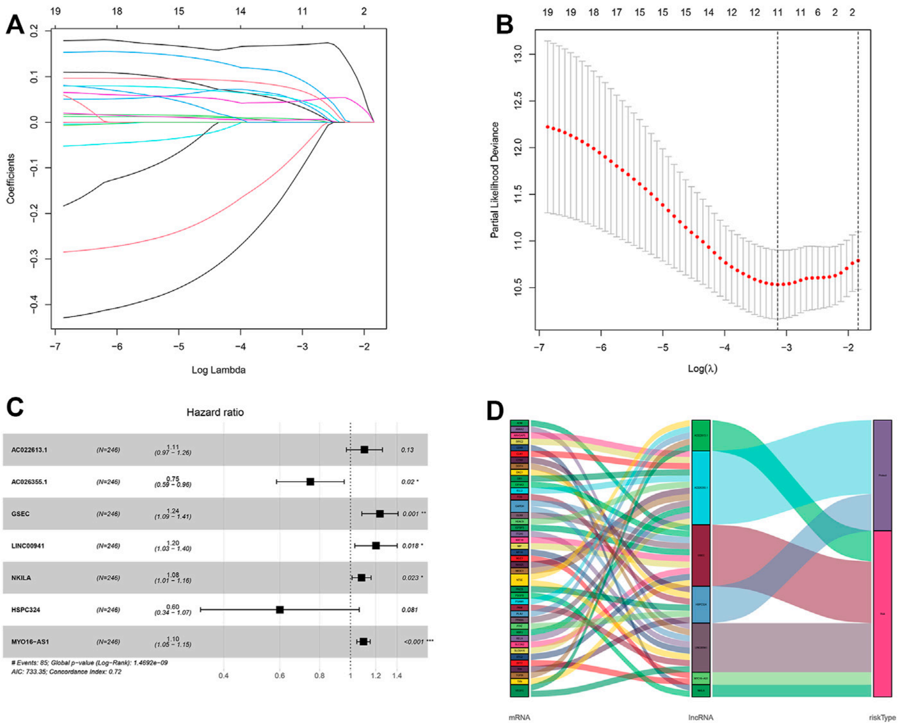


FIGURE 3 | Identification and construction of a hypoxia-related lncRNA signature in the training cohort. **(A,B)** Robust lncRNAs were screened by LASSO analysis. **(C)** Forest plot of the multivariate Cox regression model. **(D)** Sankey diagram shows the lncRNA–mRNA interaction about hypoxia. * represents p < 0.05, ** represents p < 0.01, and *** represents p < 0.001.

TABLE 2 | Information of seven hypoxia-related lncRNAs associated with OS in patients with LUAD.

lncRNA symbol	Cox (β)	HR
AC022613.1	0.100829689	1.106088246
AC026355.1	−0.29010788	0.748182849
GSEC	0.213042416	1.237437137
LINC00941	0.184433915	1.202537509
NKILA	0.080054987	1.083346635
HSPC324	−0.510440682	0.600231009
MYO16-AS1	0.094137118	1.098710388

LncRNAs, AC022613.1, AC026355.1, GSEC, LINC00941, NKILA, HSPC324, and MYO16-AS1, were used to calculate the risk score (**Figures 3C–E**). Prognostic risk genes correlated with hypoxia were constructed (**Table 2**).

The unique risk score of patients was calculated through multivariate Cox analysis and the expression level. Risk score = expression of AC010980.2 × 0.490289217 + expression of AC026355.1 × −0.417505811 + expression of AL606489.1 × 0.293303854 + expression of ITGB1-DT × 0.289653582 + expression of AL034397.3 × −0.277395699 + expression of LINC01116 × 0.192509771 + expression of LINC01150 × −0.506323639. According to the median of the risk score, there were 123 patients in the high- and low-risk groups, respectively. Additionally, in the other cohort, the number was 130 and 114 patients using the same middle score.

By applying ROC curve analysis, in the training cohort, the area under the curve (AUC) was 0.740 at year 3 and 0.736 at year 5. KM analysis also showed that the model could be a valid prognostic indicator for patients (**Figure 4A**). Likewise, in the other cohort, the AUC values were 0.600 and 0.634, respectively (**Figure 4B**). From the result of KM analysis, we got the same trend with the training cohort (**Figures 4C,D**). Meanwhile, we found that in both cohorts, patients in the low-risk group had more survival time than those in the high-risk group (**Figures 4E,F**). In addition, in the training cohort, the result of the C-index was 0.715 and 0.643 in the other cohort.

Independent Prognostic Value of the Signature

Among these lncRNAs, two lncRNAs in the training cohort (AC026355.1 and HSPC324) were upregulated, while MYO16–AS1, GSEC, NKILA, AC022613.1 and LINC00941 were downregulated in the low-risk group. In addition, similar results were obtained in the validation cohort (**Figures 5A,B**). Moreover, in both cohorts, we used univariate and multivariate Cox regression analyses to assess whether the risk score could serve as an independent prognostic factor. The risk score was an independent factor revealed by the univariate Cox regression, and the HR of it was 1.442. In the multivariate analysis, the risk score

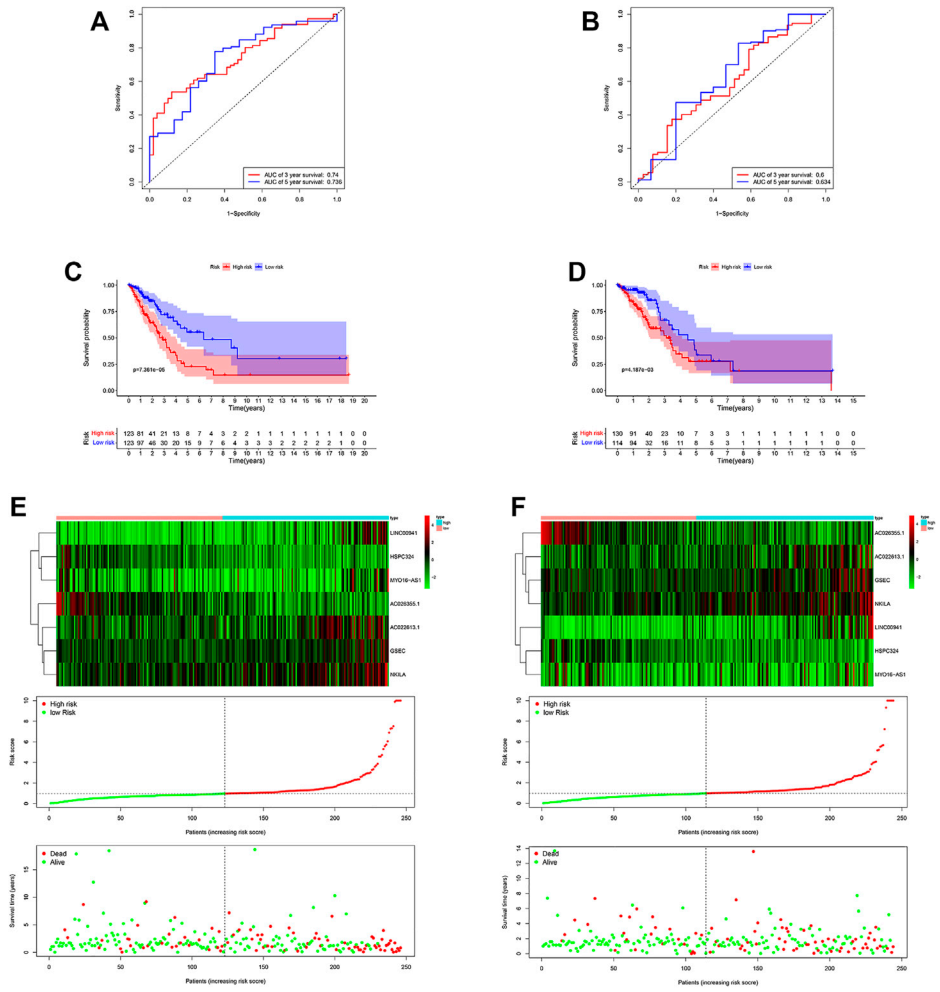


FIGURE 4 | Risk score of the hypoxia-related lncRNA signature for survival prediction in the two cohorts. The ROC analysis showed that the signature was stable **(A)** in the training and **(B)** in the validation cohorts. Between low- and high-risk groups, KM analysis was used for comparison of the OS **(C)** in the training and **(D)** validation cohorts. Distributions of risk score, survival status, and expression of lncRNAs **(E)** in the training, and **(F)** validation cohorts.

also remained an independent prognostic indicator [p < 0.001, HR = 1.391, 95% CI: 1.258–1.540 (**Figures 5C,D**)] in the training cohort. In the validation cohort, the risk score was an independent prognostic indicator, too (**Figures 5E,F**).

In addition, to study the applicability of this model, we also conducted validations in different clinical subgroups. The patients were sorted by age (the 65 years or younger group and the more than 65 years group), gender (the male group and the female group), T stage of the tumor (the T1–T2 group and the T3–T4 group), M stage of the tumor (the M0 group and the M1 group), N stage of the tumor (the N0 group and the N1–N3 group), and tumor stage (the stage I–II group and the stage III–IV group). The results showed that the survival rate of high-risk patients with different age, gender, M0, N0, T1–T2, and stage I–II group was significantly different from that of low-risk patients (p-value < 0.05) in the training cohort (**Figures 6A–H**). In the validation cohort, we obtained the similar result (**Supplementary Figure S1**).

We then constructed two nomograms that integrated the risk score of the seven-lncRNA models and clinico-pathological features to predict survival probability of patients. Based on these, we predicted the patient's 1-, 3-, 5-, and 10-year survival probabilities (**Figures 6I,J**). In both the nomograms, the higher the total points calculated, the worse the prognosis. Meanwhile, the calibration plot for the prediction of 3-year and 5-year survival also indicated the consistency between observation and prediction in both the cohorts (**Figures 6K–N**).

Functional Analyses Based on the Risk Model

We used GSEA software to perform KEGG analysis for exploring which pathways were enriched. The results identified that in the high-risk group, processes such as cell cycle, DNA replication, and mismatch repair were enriched by using the training cohort (**Figures 7A,B**).

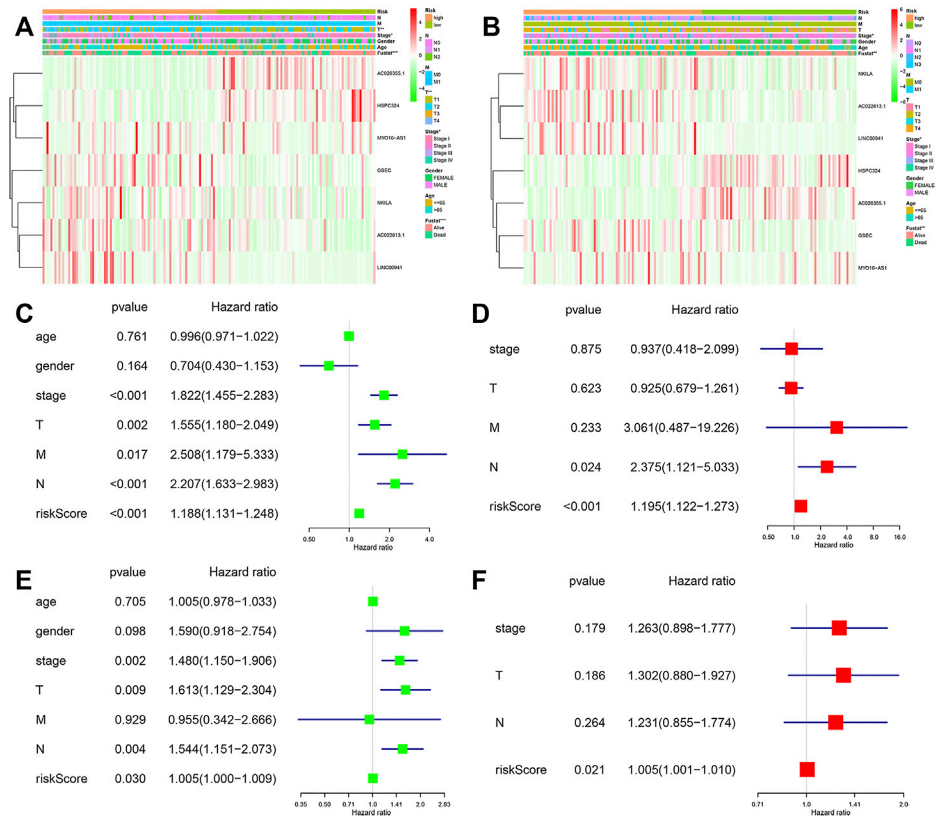


FIGURE 5 | Forest plots of the univariate and multivariate Cox analysis in LUAD patients. Heatmap and clinico-pathologic features of the seven hypoxia-related lncRNAs **(A)** in the training, and **(B)** validation cohorts. Univariate and multivariate analyses for OS **(C,D)** in the training and **(E,F)** validation cohorts.

We further screened 660 different genes (adjust p-value < 0.05 and |log2FC (fold-change)| > 1) between the high- and low-risk groups and carried out enrichment analysis based on these in the training cohort. GO enrichment analysis indicated that many nuclear biological processes or molecular functions were significantly enriched (**Figures 7C,D**). Next, KEGG pathway analysis indicated that the biological processes related to cell proliferation were enriched, based on the upregulated genes, while downregulated genes were highly correlated with the immune process (**Figures 7E,F** and **Supplementary Figures S2, S3**). It was further verified that hypoxia is related to cell proliferation and immunity.

Comparison of Immune Cell Infiltration Between Subgroups

We calculated the proportion of 22 immune cells of all samples in the training and validation cohorts by using the CIBERSORT algorithm (**Figures 8A,C**). The violin plot showed that patients in the high-risk group had a higher proportion of activated memory CD4+ T cells, resting NK cells, M0 macrophages, and activated mast cells and a lower proportion of regulatory T cells, resting mast cells, and resting dendritic cells than those in the low-risk group in the training cohort (**Figure 8B**). In addition, in the validation cohort, CD8+ T cells, activated memory CD4+ T cells, resting NK cells, M0 macrophages, and activated mast cells were higher in the high-risk group than in the other group, whereas monocytes, M2 macrophages, resting mast cells, and resting dendritic cells were lower (**Figure 8D**). This result indicated that immune-related activities were associated with hypoxia.

In addition, based on the prognostic signature, there was an observably different distribution between high- and low-risk groups through the principal component analysis, which indicated that there was a difference in the hypoxia phenotype of the model (**Figures 9A,B**).

Association of Model LncRNAs With the Tumor Microenvironment and Immune Infiltration in Pan-Cancer

From the abovementioned results, the seven-model lncRNAs played an important role in LUAD. We then downloaded 33 cancer types to understand the function of the lncRNAs and selected AC022613.1, GSEC, LINC00941, and NKILA for further study. We found that AC022613.1, GSEC, LINC00941, and NKILA were mainly upregulated in tumor samples compared with normal samples (**Figures 9D–G**). In addition, the expression of these lncRNAs varied in different tumors (**Figure 9C**). The expressions of LINC00941 and NKILA were highly expressed in CHOL samples than those in the other tumors. In LUAD tissues,

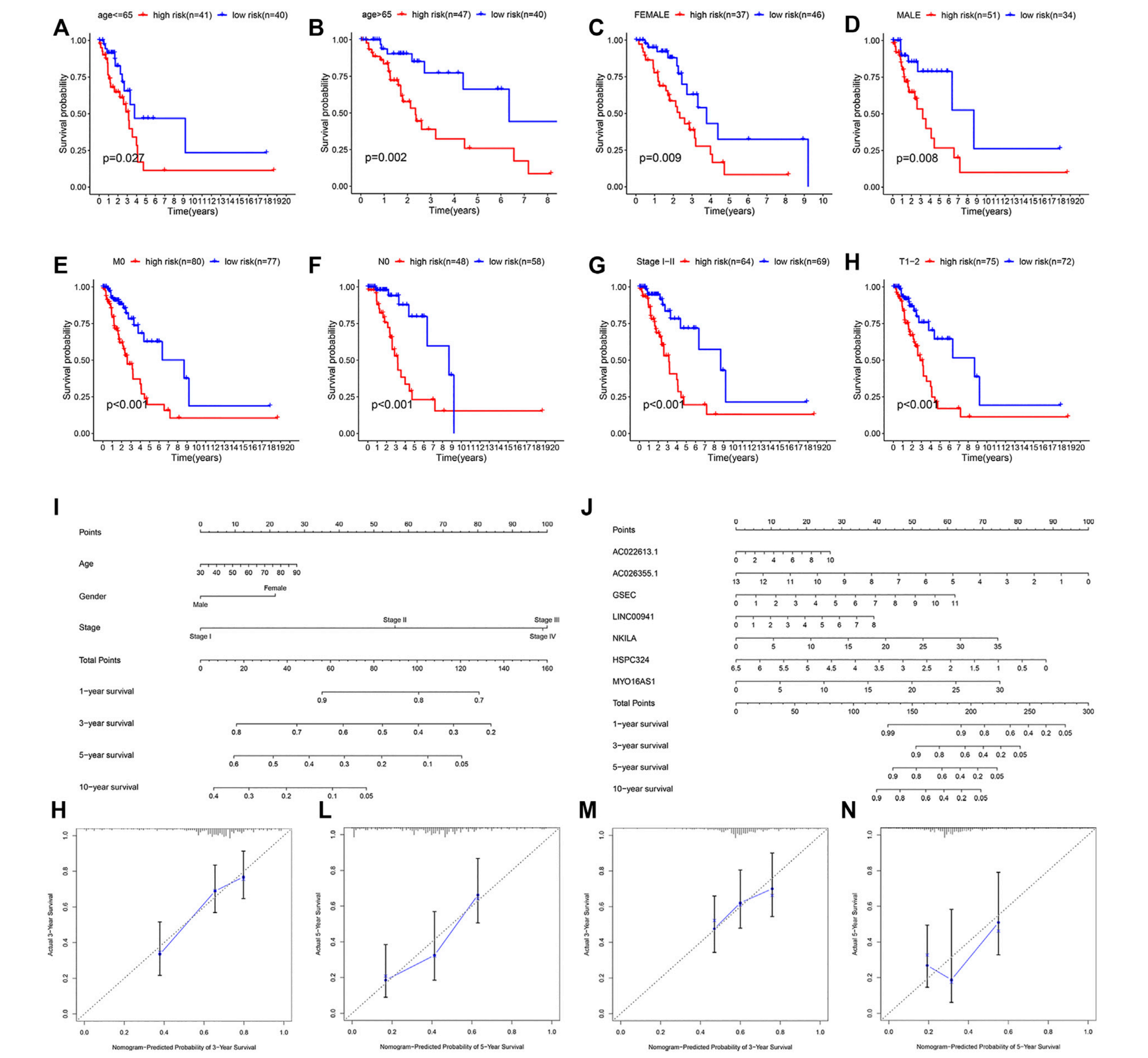


FIGURE 6 | Stratification analysis of various clinico-pathological factors, nomograms used to predict the OS prognosis, and calibration plots used to predict prognosis in these patients with LUAD. KM curves of OS in the subgroups of **(A,B)** both the age groups, **(C,D)** both gender groups, **(E)** M0 group, **(F)** T0 group, **(G)** clinical stage I–II group, and **(H)** T1–2 group. **(I,J)** Nomograms to predict the 1-, 3-, 5-, and 10-year OS. Calibration plots for 3-year survival and 5-year survival **(K,L)** in the training and **(M,N)** validation cohorts.

the expressions of the four lncRNAs were less than zero (**Figure 9H**). In addition, we found that LINC00941 and NKILA may have similar functions (**Figure 9I**).

By plotting KM curves for 32 cancer types, we found that only AC022613.1 significantly affected the survival of patients in many cancer types (**Figures 10A–F**). Furthermore, we used univariate Cox analysis to investigate the relationship between the expression of AC022613.1 and patient survival. The result showed that the relationship was different in different tumors (**Figure 10G**). We then explored their role in the six immune subtypes, stromal, ESTIMATE, tumor purity, immune score, tumor stem cells, and TMB. In LUAD patients, we found that AC022613.1 was strongly connected with the tumor stage and GSEC, LINC00941, and NKILA were significantly connected with immune subtypes (**Figures 10H,I**). Based on the ESTIMATE analysis, we researched the connection between the four lncRNAs and tumor microenvironment in LUAD. The results showed that GSEC had a negative association with stromal, ESTIMATE, and immune score, while LINC00941 had the opposite result (**Figure 10J**). Meanwhile, the results indicated

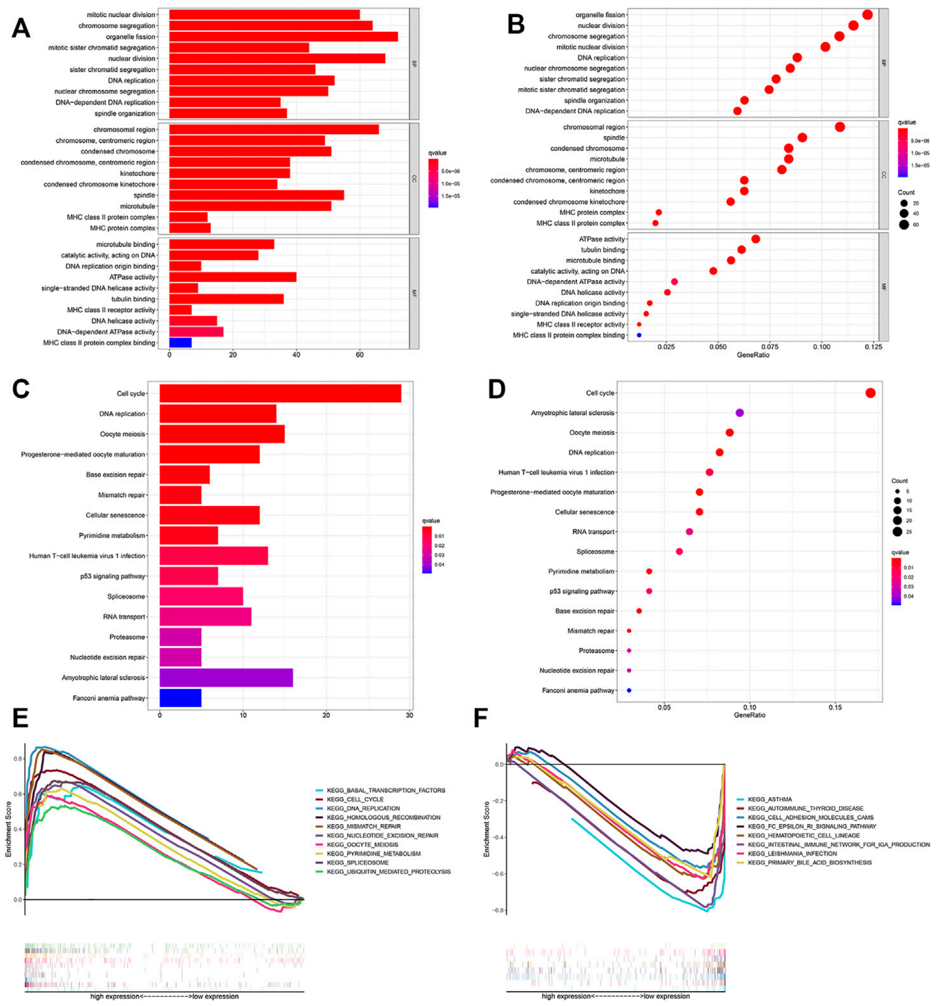


FIGURE 7 | Functional analysis. Bubble and barplot graph for **(A,B)** KEGG pathways and **(C,D)** GO enrichment based on the DEGs in the training cohort. **(E,F)** KEGG pathways of GSEA analysis.

that in pan-cancer, the four lncRNAs were strongly associated with immune subtypes (**Figure 9J**). Moreover, we found that they were mainly positively connected with stromal, ESTIMATE, and immune score, while there was a negative correlation between these and tumor purity, RNAss, and DNAss in pan-cancer (**Figures 10K–N**). Furthermore, we used the radar plots to distribute their association between the four lncRNAs and TMB. Distinctly, we found that NKILA and GSEC had strongly correlation in LUAD patients (**Figures 10O–R**).

DISCUSSION

Overall, in this study, we obtained 239 hypoxia-related genes. According to the expression levels of the 239 genes, we identified 1,629 hypoxia-related lncRNAs using Pearson's correlation analysis. |Correlation coefficient| > 0.3 and p-value < 0.001 were our selection criteria. In addition, there were many useful tools that could help extract hypoxia-related lncRNAs, such as BioSeq-BLM (Li et al., 2021), BioSeq-Analysis 2.0 (Liu et al., 2019), and starBase v2.0 (Li et al., 2014). However, they were mainly used for residue-level analysis and sequence-level analysis. In this study, according to the expression levels of the transcriptome, we used Pearson's correlation analysis to identify lncRNAs closely associated with hypoxia. The correlation between the expression levels of mRNAs and lncRNAs was fully considered. Meanwhile, this method was widely used in the computational genomics field of tumors, such as hepatocellular carcinoma (Zhou et al., 2021), bladder cancer (Ma et al., 2021), soft tissue sarcomas (Zhang J et al., 2021), and breast cancer (Zhang L et al., 2021). There were 601 DELs associated with hypoxia between normal and tumor samples. Of them, 530 differentially expressed hypoxia-related lncRNAs were upregulated in the tumor samples, and 71 lncRNAs were downregulated. In addition, by performing WGCNA, we obtained 617 hypoxia-related lncRNAs that were associated with tumor samples. By taking the intersection of the 601 DELs related to hypoxia and 617 hypoxia-related

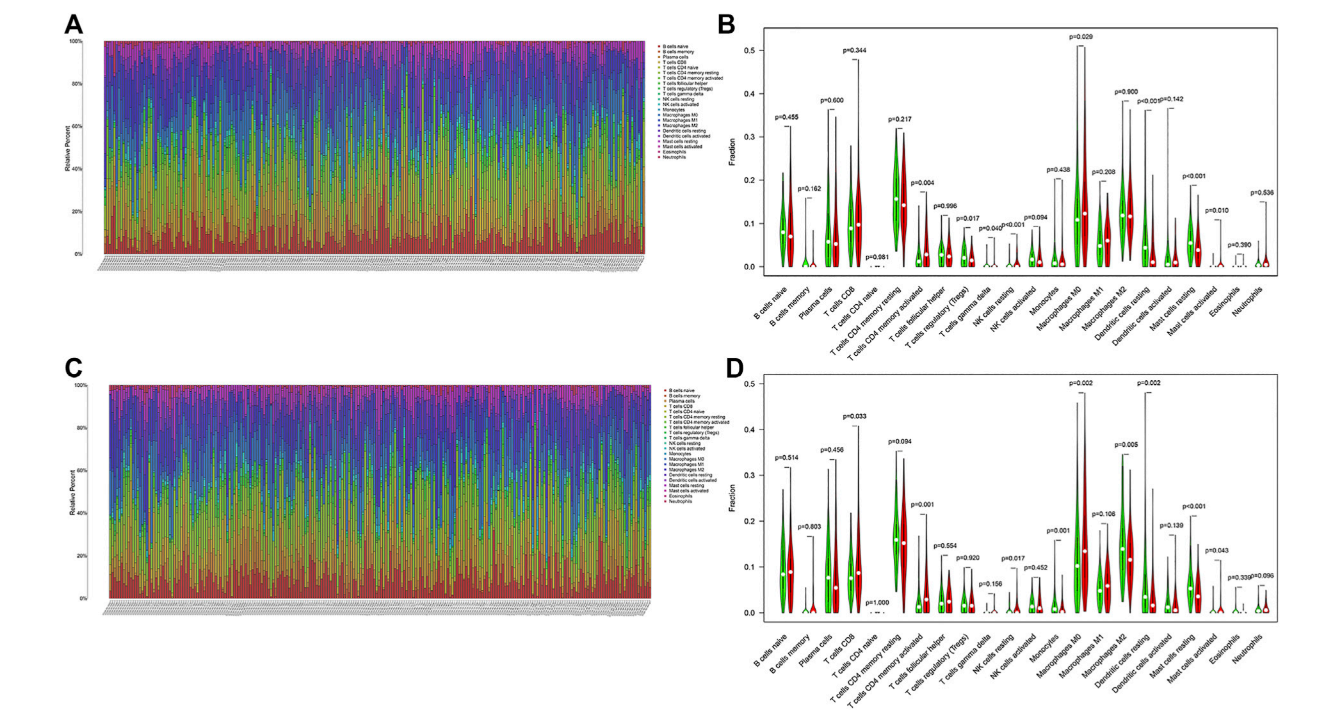


FIGURE 8 | Landscape of immune cell infiltration in LUAD. Immune landscape of the patients with LUAD **(A)** in the training and **(C)** validation cohorts. Relationships between the risk score and immune cell infiltration **(B)** in the training and **(D)** validation cohorts. Red and green represent the high- and low-risk groups, respectively.

lncRNAs, we got 262 hypoxia-related lncRNAs. Finally, we used univariate Cox analysis, LASSO analysis, and multivariate Cox analysis to generate the hypoxia-related lncRNA signature. We identified seven lncRNAs associated with hypoxia as potential prognostic biomarkers. KM analysis indicated that in the high-risk group, the OS of patients was shorter than that of patients in the low-risk group. Meanwhile, the seven-lncRNA signature was highly sensitive in the prediction of OS time of LUAD patients by taking the ROC analysis, and the results were further verified in the validation cohort. Finally, we constructed two nomograms to calculate a score representing the OS of LUAD patients.

In the signature, there were seven different lncRNAs in total. These lncRNAs were AC026355.1, AC022613.1, GSEC, LINC00941, NKILA, HSPC324, and MYO16-AS1. Among these hypoxia-related lncRNAs, according to reports, AC026355.1 is connected with the development of multiple tumors. It had important prognostic significance in both immune- and autophagy-related models (Li et al., 2020; You et al., 2021; Jiang et al., 2021). LINC00941 is one of the immune-related prognostic models comprising 7 lncRNAs in LUAD (Jin et al., 2020; Li et al., 2021). GSEC has only been described in osteosarcoma cells. In osteosarcoma cells, the overexpression of GSEC can enhance the proliferation and migration of tumor cells (Liu et al., 2020). LINC00941 usually was the risk factor and connected with worse survival (Chang et al., 2021; Fang et al., 2021; Wang et al., 2021). Wang Jie et al. found that in pancreatic cancer, LINC00941 was overexpressed and patients yielded worse prognosis (Wang et al., 2021; Chang et al., 2021). However, this lncRNA has not been reported in LUAD. NKILA is a tumor suppressor that affects the proliferation and metastasis of cancer cells by regulating the STAT3 pathway (Ashrafizadeh et al., 2021). LncRNA HSPC324 plays a crucial role in tumorigenesis of LUAD (Jafarzadeh et al., 2020). MYO16-AS1 was an oncogenic lncRNA in bladder cancer (Jafarzadeh et al., 2020). It has rarely been reported in LUAD. In conclusion, these lncRNAs all play a significant role in the occurrence and development of tumors.

A growing body of evidence suggests that the prognosis of tumor patients is connected with the level of immune invasion of the tumor, and the state of immune invasion is a key determinant of tumor development in the tumor microenvironment (Tao et al., 2021). Hypoxia of tumor tissue plays a vital role in promoting tumor immunosuppression and immunotherapy resistance. In this state, there are often abundant tumor-associated macrophages and Tregs, which inhibit the function of CD8+T cells and CD4+T cells (Dehghani et al., 2012; Sanchez-Martinez et al., 2018). Hypoxia inhibits the activity of effector T cells and NK cells, leading to decreased immune function. In our study CD8+ T cells; resting NK cells; M0, M1, and M2 macrophages; resting dendritic cells; and resting mast cells were found to be differentially infiltrated in LUAD and normal tissues, which is closely related to the development of tumors. This finding supported that the hypoxia-related lncRNA signatures reflected immune infiltration to some extent, providing meaningful information for immunotherapy (Kumar and Gabrilovich, 2014; Labiano et al., 2015; Aponte-Lopez and Munoz-Cruz, 2020).

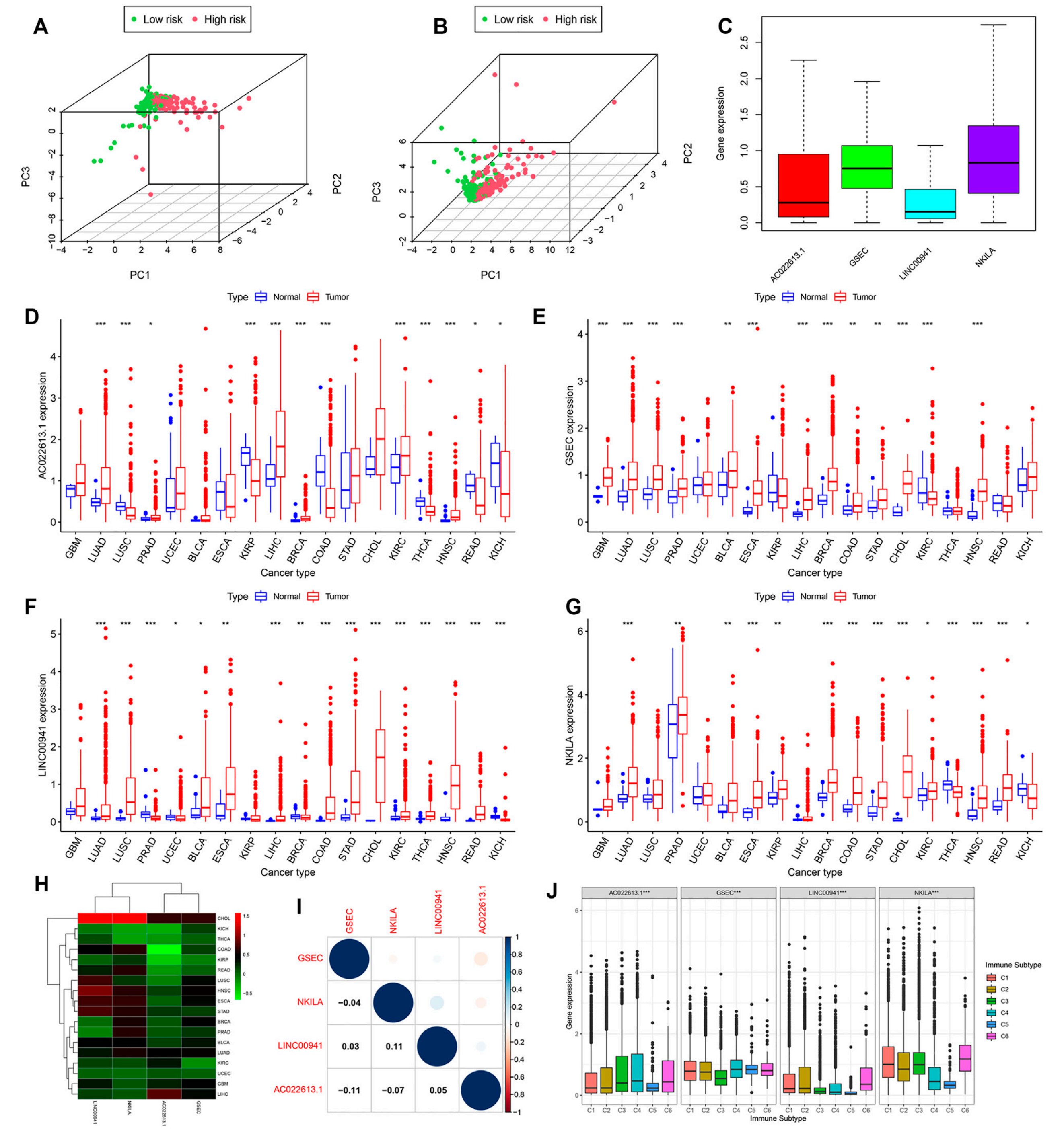


FIGURE 9 | Expression of lncRNAs in pan-cancer which had more than five normal samples. Principal component analysis based on the lncRNAs of the model **(A)** in the training and **(B)** validation cohorts. **(C)** Expression of lncRNAs of the model is shown by the boxplot in pan-cancer. **(D–G)** In 18 cancer types, the difference of the expression of lncRNAs between tumor and normal samples is shown. **(H)** Heatmap showing the difference between normal and tumor samples of the expression of lncRNAs. **(I)** Relationship calculated by Spearman's correlation analysis of the lncRNAs in pan-cancer. **(J)** Association tested by ANOVA in all cancers of four lncRNAs with six immune subtypes.

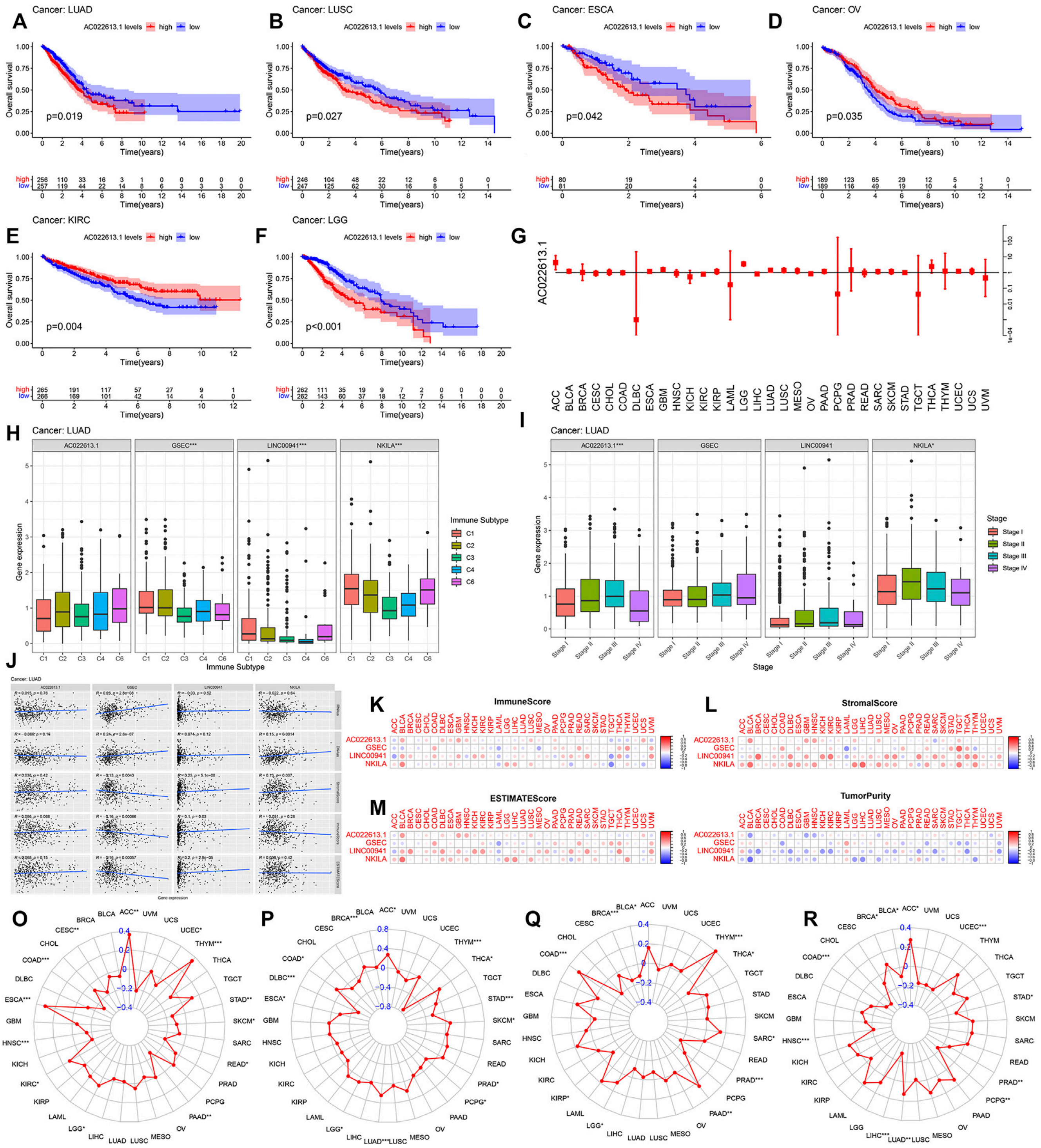


FIGURE 10 | Correlation analysis between the four lncRNAs and patient prognosis and TMB in all cancer types. **(A–F)** Connection of the four lncRNAs with the prognosis of patients in pan-cancer. **(G)** Forest plot showing the hazard ratio of AC022613.1 across all cancer types. The correlation between the four lncRNAs and **(H)** immune subtypes and **(I)** clinical stage of LUAD. **(J)** Association of four lncRNAs with tumor stem cell scores and stromal, immune, and ESTIMATE score by Spearman's correlation analysis. **(K–N)** The correlation relationship between four lncRNAs and stromal, immune, ESTIMATE, and tumor purity score. In 33 TCGA cancer types, the radar graph showing the association of the expression of **(O)** AC022613.1, **(P)** GSEC, **(Q)** LINC00941, and **(R)** NKILA with TMB. $p < 0.05$, $p < 0.01$, and $p < 0.0001$ were represented by *, **, and ***, respectively.

CONCLUSION

In summary, our study demonstrated that hypoxia is connected with the development of LUAD. Meanwhile, the two predictive nomograms were established for predicting the prognosis of LUAD patients. We anticipated that the study will provide an important basis for studies on the correlation between hypoxia-related genes and LUAD.

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 below: Publicly available datasets were analyzed in this study. These data can be found here: The Cancer Genome Atlas (https://portal.gdc.cancer.gov/), the Molecular Signatures Database (https://immport.niaid.nih.gov), and CIBERSORT (https://cibersort.stanford.edu/).

AUTHOR CONTRIBUTIONS

YL: investigation, data curation, investigation, data curation, methodology, and writing—original draft. XS: conceptualization, writing—review and editing, and project administration. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (32070582, 31872800) and Guangzhou City Science and Technology Key Topics Project (201904020025).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.768971/full#supplementary-material

REFERENCES

Aponte-López, A., and Muñoz-Cruz, S. (2020). Mast Cells in the Tumor Microenvironment. *Adv. Exp. Med. Biol.* 1273, 159–173. doi:10.1007/978-3-030-49270-0_9

Ashrafizadeh, M., Gholami, M. H., Mirzaei, S., Zabolian, A., Haddadi, A., Farahani, M. V., et al. (2021). Dual Relationship between Long Non-coding RNAs and STAT3 Signaling in Different Cancers: New Insight to Proliferation and Metastasis. *Life Sci.* 270, 119006. doi:10.1016/j.lfs.2020.119006

Bhan, A., Soleimani, M., and Mandal, S. S. (2017). Long Noncoding RNA and Cancer: A New Paradigm. *Cancer Res.* 77 (15), 3965–3981. doi:10.1158/0008-5472.CAN-16-2634

Carter, B. W., Lichtenberger, J. P., Benveniste, M. K., de Groot, P. M., Wu, C. C., Erasmus, J. J., et al. (2018). Revisions to the TNM Staging of Lung Cancer: Rationale, Significance, and Clinical Application. *Radiographics* 38 (2), 374–391. doi:10.1148/rg.2018170081

Chang, L., Zhou, D., and Luo, S. (2021). Novel lncRNA LINC00941 Promotes Proliferation and Invasion of Colon Cancer through Activation of MYC. *Ott* Vol. 14, 1173–1186. doi:10.2147/OTT.S293519

Chang, Y.-N., Zhang, K., Hu, Z.-M., Qi, H.-X., Shi, Z.-M., Han, X.-H., et al. (2016). Hypoxia-regulated lncRNAs in Cancer. *Gene* 575 (1), 1–8. doi:10.1016/j.gene.2015.08.049

Chen, D., Wang, Y., Zhang, X., Ding, Q., Wang, X., Xue, Y., et al. (2021). Characterization of Tumor Microenvironment in Lung Adenocarcinoma Identifies Immune Signatures to Predict Clinical Outcomes and Therapeutic Responses. *Front. Oncol.* 11, 581030. doi:10.3389/fonc.2021.581030

Chen, Q., Hu, L., and Chen, K. (2020). Construction of a Nomogram Based on a Hypoxia-Related lncRNA Signature to Improve the Prediction of Gastric Cancer Prognosis. *Front. Genet.* 11, 570325. doi:10.3389/fgene.2020.570325

Choudhry, H., Harris, A. L., and McIntyre, A. (2016). The Tumour Hypoxia Induced Non-coding Transcriptome. *Mol. Aspects Med.* 47-48, 35–53. doi:10.1016/j.mam.2016.01.003

Dehghani, M., Sharifpour, S., Amirghofran, Z., and Zare, H. R. (2012). Prognostic Significance of T Cell Subsets in Peripheral Blood of B Cell Non-hodgkin's Lymphoma Patients. *Med. Oncol.* 29 (4), 2364–2371. doi:10.1007/s12032-012-0176-1

Fang, L., Wang, S. H., Cui, Y. G., and Huang, L. (2021). LINC00941 Promotes Proliferation and Metastasis of Pancreatic Adenocarcinoma by Competitively Binding miR-873-3p and Thus Upregulates ATXN2. *Eur. Rev. Med. Pharmacol. Sci.* 25 (4), 1861–1868. doi:10.26355/eurrev_202102_25081

Fang, Y., and Fullwood, M. J. (2016). Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics & Bioinformatics* 14 (1), 42–54. doi:10.1016/j.gpb.2015.09.006

Guo, Y., Yang, P. T., Wang, Z. W., Xu, K., Kou, W. H., and Luo, H. (2020). Identification of Three Autophagy-Related Long Non-coding RNAs as a Novel Head and Neck Squamous Cell Carcinoma Prognostic Signature. *Front. Oncol.* 10, 603864. doi:10.3389/fonc.2020.603864

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. *The Lancet* 389 (10066), 299–311. doi:10.1016/S0140-6736(16)30958-8

Jafarzadeh, M., Tavallaie, M., Soltani, B. M., Hajipoor, S., and Hosseini, S. M. (2020). LncRNA HSPC324 Plays Role in Lung Development and Tumorigenesis. *Genomics* 112 (3), 2615–2622. doi:10.1016/j.ygeno.2020.02.012

Jiang, A., Liu, N., Bai, S., Wang, J., Gao, H., Zheng, X., et al. (2021). Identification and Validation of an Autophagy-Related Long Non-coding RNA Signature as a Prognostic Biomarker for Patients with Lung Adenocarcinoma. *J. Thorac. Dis.* 13 (2), 720–734. doi:10.21037/jtd-20-2803

Jin, D., Song, Y., Chen, Y., and Zhang, P. (2020). Identification of a Seven-lncRNA Immune Risk Signature and Construction of a Predictive Nomogram for Lung Adenocarcinoma. *Biomed. Res. Int.* 2020, 1–17. doi:10.1155/2020/7929132

Kumar, V., and Gabrilovich, D. I. (2014). Hypoxia-inducible Factors in Regulation of Immune Responses in Tumour Microenvironment. *Immunology* 143 (4), 512–519. doi:10.1111/imm.12380

Labiano, S., Palazon, A., and Melero, I. (2015). Immune Response Regulation in the Tumor Microenvironment by Hypoxia. *Semin. Oncol.* 42 (3), 378–386. doi:10.1053/j.seminoncol.2015.02.009

Lemjabbar-Alaoui, H., Hassan, O. U., Yang, Y.-W., and Buchanan, P. (2015). Lung Cancer: Biology and Treatment Options. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1856 (2), 189–210. doi:10.1016/j.bbcan.2015.08.002

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.* 49 (22), e129. doi:10.1093/nar/gkab829

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. *Nucl. Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Li, J.-P., Li, R., Liu, X., Huo, C., Liu, T.-T., Yao, J., et al. (2020). A Seven Immune-Related lncRNAs Model to Increase the Predicted Value of Lung Adenocarcinoma. *Front. Oncol.* 10, 560779. doi:10.3389/fonc.2020.560779

Li, J., Meng, H., Bai, Y., and Wang, K. (2016). Regulation of lncRNA and its Role in Cancer Metastasis. *Oncol. Res.* 23 (5), 205–217. doi:10.3727/096504016X14549667334007

Li, Y., Shen, R., Wang, A., Zhao, J., Zhou, J., Zhang, W., et al. (2021). Construction of a Prognostic Immune-Related LncRNA Risk Model for Lung Adenocarcinoma. *Front. Cel Dev. Biol.* 9, 648806. doi:10.3389/fcell.2021.648806

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an Updated Platform for Analyzing DNA, RNA and Protein Sequences at Sequence Level and Residue Level Based on Machine Learning Approaches. *Nucleic Acids Res.* 47 (20), e127. doi:10.1093/nar/gkz740

Liu, R., Ju, C., Zhang, F., Tang, X., Yan, J., Sun, J., et al. (2020). LncRNA GSEC Promotes the Proliferation, Migration and Invasion by Sponging miR-588/EIF5A2 axis in Osteosarcoma. *Biochem. Biophysical Res. Commun.* 532 (2), 300–307. doi:10.1016/j.bbrc.2020.08.056

Ma, C., Wo, L.-L., Wang, D.-F., Zhou, C.-X., Li, J.-C., Zhang, X., et al. (2021). Hypoxia Activated Long Non-coding RNA HABON Regulates the Growth and Proliferation of Hepatocarcinoma Cells by Binding to and Antagonizing HIF-1 Alpha. *RNA Biol.* 18, 1791–1806. doi:10.1080/15476286.2020.1871215

Ma, T., Wang, X., Meng, L., Liu, X., Wang, J., Zhang, W., et al. (2021). An Effective N6-Methyladenosine-Related Long Non-coding RNA Prognostic Signature for Predicting the Prognosis of Patients with Bladder Cancer. *BMC Cancer* 21 (1), 1256. doi:10.1186/s12885-021-08981-4

Odintsov, I., Mattar, M. S., Lui, A. J. W., Offin, M., Kurzatkowski, C., Delasos, L., et al. (2021). Novel Preclinical Patient-Derived Lung Cancer Models Reveal Inhibition of HER3 and MTOR Signaling as Therapeutic Strategies for NRG1 Fusion-Positive Cancers. *J. Thorac. Oncol.* 16 (7), 1149–1165. doi:10.1016/j.jtho.2021.03.013

Peng, F., Wang, J.-H., Fan, W.-J., Meng, Y.-T., Li, M.-M., Li, T.-T., et al. (2018). Glycolysis Gatekeeper PDK1 Reprograms Breast Cancer Stem Cells under Hypoxia. *Oncogene* 37 (8), 1062–1074. doi:10.1038/onc.2017.368

Peng, W.-X., Koirala, P., and Mo, Y.-Y. (2017). LncRNA-mediated Regulation of Cell Signaling in Cancer. *Oncogene* 36 (41), 5661–5667. doi:10.1038/onc.2017.184

Rami-Porta, R., Asamura, H., Travis, W. D., and Rusch, V. W. (2017). Lung Cancer - Major Changes in the American Joint Committee on Cancer Eighth Edition Cancer Staging Manual. *CA: A Cancer J. Clinicians* 67 (2), 138–155. doi:10.3322/caac.21390

Sanchez-Martinez, D., Allende-Vega, N., Orecchioni, S., Talarico, G., Cornillon, A., Vo, D.-N., et al. (2018). Expansion of Allogeneic NK Cells with Efficient Antibody-dependent Cell Cytotoxicity against Multiple Tumors. *Theranostics* 8 (14), 3856–3869. doi:10.7150/thno.25149

Shi, J., Hua, X., Zhu, B., Ravichandran, S., Wang, M., Nguyen, C., et al. (2016). Somatic Genomics and Clinical Features of Lung Adenocarcinoma: A Retrospective Study. *Plos Med.* 13 (12), e1002162. doi:10.1371/journal.pmed.1002162

Shi, R., Bao, X., Unger, K., Sun, J., Lu, S., Manapov, F., et al. (2021). Identification and Validation of Hypoxia-Derived Gene Signatures to Predict Clinical Outcomes and Therapeutic Responses in Stage I Lung Adenocarcinoma Patients. *Theranostics* 11 (10), 5061–5076. doi:10.7150/thno.56202

Tao, J., Yang, G., Zhou, W., Qiu, J., Chen, G., Luo, W., et al. (2021). Targeting Hypoxic Tumor Microenvironment in Pancreatic Cancer. *J. Hematol. Oncol.* 14 (1), 14. doi:10.1186/s13045-020-01030-w

Wang, J., He, Z., Xu, J., Chen, P., and Jiang, J. (2021). Long Noncoding RNA LINC00941 Promotes Pancreatic Cancer Progression by Competitively Binding miR-335-5p to Regulate ROCK1-Mediated LIMK1/Cofilin-1 Signaling. *Cell Death Dis* 12 (1), 36. doi:10.1038/s41419-020-03316-w

Wang, Y., Zhou, W., Ma, S., Guan, X., Zhang, D., Peng, J., et al. (2020). Identification of a Glycolysis-Related LncRNA Signature to Predict Survival in Diffuse Glioma Patients. *Front. Oncol.* 10, 597877. doi:10.3389/fonc.2020.597877

Wu, L., Wen, Z., Song, Y., and Wang, L. (2021). A Novel Autophagy-related lncRNA Survival Model for Lung Adenocarcinoma. *J. Cel Mol Med* 25 (12), 5681–5690. doi:10.1111/jcmm.16582

You, J., Fang, W., Zhao, Q., Chen, L., Chen, L., and Chen, F. (2021). Identification of a RNA-Seq Based Prognostic Signature with Seven Immune-Related lncRNAs for Lung Adenocarcinoma. *Clin. Lab.* 67 (3), 663. doi:10.7754/Clin.Lab.2020.200663

Zhang, H., Qin, C., Liu, H. W., Guo, X., and Gan, H. (2021). An Effective Hypoxia-Related Long Non-coding RNAs Assessment Model for Prognosis of Clear Cell Renal Carcinoma. *Front. Oncol.* 11, 616722. doi:10.3389/fonc.2021.616722

Zhang, J., Shan, B., Lin, L., Dong, J., Sun, Q., Zhou, Q., et al. (2021). Dissecting the Role of N6-Methylandenosine-Related Long Non-coding RNAs Signature in Prognosis and Immune Microenvironment of Breast Cancer. *Front. Cel Dev. Biol.* 9, 711859. doi:10.3389/fcell.2021.711859

Zhang, L., Tang, X., Wan, J., Zhang, X., Zheng, T., Lin, Z., et al. (2021). Construction of a Novel Signature and Prediction of the Immune Landscape in Soft Tissue Sarcomas Based on N6-Methylandenosine-Related LncRNAs. *Front. Mol. Biosci.* 8, 715764. doi:10.3389/fmolb.2021.715764

Zhou, P., Lu, Y., Zhang, Y., and Wang, L. (2021). Construction of an Immune-Related Six-lncRNA Signature to Predict the Outcomes, Immune Cell Infiltration, and Immunotherapy Response in Patients with Hepatocellular Carcinoma. *Front. Oncol.* 11, 661758. doi:10.3389/fonc.2021.661758

Zhu, N., Hou, J., Wu, Y., Liu, J., Li, G., Zhao, W., et al. (2018). Integrated Analysis of a Competing Endogenous RNA Network Reveals Key lncRNAs as Potential Prognostic Biomarkers for Human Bladder Cancer. *Medicine (Baltimore)* 97 (35), e11887. doi:10.1097/MD.0000000000011887

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.

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/fgene.2022.865204

Genome-wide Exploration of a Pyroptosis-Related Long Non-Coding RNA Signature Associated With the Prognosis and Immune Response in Patients With Bladder Cancer

Xin Gao[1,2,3] *and Jianping Cai*[1,2]*

[1]Graduate School of Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China, [2]The Key Laboratory of Geriatrics, Beijing Institute of Geriatrics, Beijing Hospital, National Center of Gerontology, National Health Commission, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China, [3]Clinical Laboratory, The First People's Hospital of Huaihua / The Fourth Affiliated Hospital of Jishou University, Huaihua, China

Background: Bladder cancer (BLCA) is a malignant tumor with a complex molecular mechanism and high recurrence rate in the urinary system. Studies have shown that pyroptosis regulates tumor cell proliferation and metastasis and affects the prognosis of cancer patients. However, the role of pyroptosis-related (PR) genes or long non-coding RNAs (lncRNAs) in BLCA development is not fully understood.

Methods: We comprehensively analyzed the molecular biological characteristics of PR genes in BLCA, including copy number variation, mutations, expression and prognostic value based on TCGA database. We then identified PR lncRNAs with prognostic value based on the expression of PR genes and performed a consistent clustering analysis of 407 BLCA patients according to the expression of prognosis-related PR lncRNAs and identified two clusters. The least absolute shrinkage and selection operator (LASSO) regression was used to establish a PR lncRNA signature and calculate the risk score associated with the prognosis of patients with BLCA. Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Set Enrichment Analysis (GSEA) were used to evaluate the possible functions of PR lncRNA signature. We also evaluated the relationship between the risk score and tumor immune microenvironment (TIME).

Results: A total of 33 PR genes were obtained in our study and 194 prognosis-related PR lncRNAs were identified. We also constructed a signature consisting of eight-PR-lncRNAs and divided patients into high- and low-risk groups. The overall survival rate of patients with a high risk was significantly lower than patients with a low risk. The risk score was significantly correlated with the degree of infiltration of multiple immune cell subtypes and positively correlated with multiple immune checkpoint genes expression in BLCA. Enrichment analyses showed that these lncRNAs are involved in human immune regulatory functions and immune-related pathways.

Conclusion: Our study comprehensively studied the molecular biological characteristics of PR genes BLCA, and the eight-PR-lncRNA signature we identified might play a crucial role in tumor

Edited by:
Joel Correa Da Rosa,
Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Milad Ashrafizadeh,
Sabancı University, Turkey
Qi Shengcai,
Shanghai Stomatology Prevention Hospital, China

Correspondence:
Jianping Cai
caijp61@vip.sina.com

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *29 January 2022*
Accepted: *04 April 2022*
Published: *27 April 2022*

Citation:
Gao X and Cai J (2022) Genome-wide Exploration of a Pyroptosis-Related Long Non-Coding RNA Signature Associated With the Prognosis and Immune Response in Patients With Bladder Cancer.
Front. Genet. 13:865204.
doi: 10.3389/fgene.2022.865204

immunity and may be able to predict the prognosis of BLCA patients, providing a theoretical basis for an in-depth study of the relationship between the prognosis and TIME.

Keywords: bladder cancer, pyroptosis, tumor immune microenvironment, lncRNA, prognosis

INTRODUCTION

Bladder cancer (BLCA) is the second-most common cause of death from urological tumors, and the incidence is still on the rise (Siegel et al., 2019). Non-muscle-invasive bladder cancer (NMIBC) accounts for about 75% of all primary bladder cancers. Unfortunately, 25% of cases have already developed into muscle-invasive bladder cancer (MIBC) by the time of the initial diagnosis (Kaufman et al., 2009). According to the pathological characteristics of BLCA patients, the main clinical treatments are surgery, radiotherapy, chemotherapy, bladder irrigation therapy and combination therapy (Ghandour et al., 2019). However, 10–30% of patients with NMIBC progress to MIBC after recurrence (Malmström et al., 2017), which has a high risk of metastasis and a poor prognosis, with only a minority of patients surviving more than 5 years (Chou et al., 2016). Immunotherapy is an emerging approach to oncology. Immune checkpoint inhibitors (ICIs) effectively block the escape of cancer cells from immune system surveillance, and these agents have begun to change the treatment strategy for BLCA. Recent studies have shown that infiltration of different immune cells may affect the response to ICIs (Benitez et al., 2020). Moreover, long non-coding RNAs (lncRNAs) are closely related to the effect of immunotherapy in BLCA, e.g., knockdown of lncRNA UCA1 significantly enhances the effect of immune checkpoint PD-1 blockers (Zhen et al., 2018). BLCA has a complex molecular biological mechanism, which is one of the main reasons for the poor efficacy of most therapies, lncRNA plays an important biological role in the progression, cell proliferation and metastasis of BLCA, for example, LINC00958 can promote BLCA by targeting miR-490-3p and AURKA (Zhen et al., 2021). lncRNAs can also act as competitive endogenous RNAs (ceRNAs) targeting snuclear factor-kappaB (NF-κB)-activated miRNAs to promote tumor development (Mirzaei et al., 2021). In addition, lncRNAs are involved in BLCA drug resistance and progression through various pathways (e.g. NF-κB, PI3K/Akt, Wnt, FOXC2 and EZH2), which has important implications for the treatment and prognosis of BLCA patients (Barth et al., 2020; Ashrafizaveh et al., 2021; Mirzaei et al., 2022). Therefore, the important role of lncRNA in BLCA has also been gradually emphasized in recent years. The development of genome sequencing and bioinformatics can help identify many molecular biomarkers to guide the treatment of BLCA patients, but only a few of these can be applied in a clinical setting (Zhang et al., 2021). Therefore, identifying the drivers and inhibitors of bladder carcinogenesis and understanding their mechanisms are essential for detecting new therapeutic targets and prolonging the survival of BLCA patients.

Pyroptosis, also known as inflammatory necrosis, is a form of programmed cell death involving cellular swelling until the cell membrane ruptures. The release of cellular contents leads to an intense inflammatory response (Loveless et al., 2021). Pyroptosis is also an essential part of the body's natural immune response and plays a vital role in the fight against infection (Shi et al., 2017). Gasdermin D (GSDMD) is a key effector molecule in the occurrence process of pyroptosis. Under stimulation with foreign substances, the intracellular pattern recognition receptor (nucleotide-binding domain leucine-rich repeat containing [NLR]) binds to the precursor of caspase-1 through the junction protein ASC and then forms a multi-protein complex to activate caspase-1. The activated caspase-1 then cleaves GSDMD to form a peptide containing the active domain of GSDM-NT, which induces the release of contents, cell membrane perforation and cell rupture, causing an inflammatory response. It also activates IL-1β and IL-18, which are released from the cell to recruit inflammatory cells and expand the inflammatory response (Broz et al., 2020; Liu et al., 2021). Pyroptosis may participate in the formation and development of tumors, and different tissues and genetic backgrounds of pyroptosis may have different effects on cancer. It can inhibit tumors but form a microenvironment suitable for the growth of tumor cells and then promote tumor growth (Xia et al., 2019). Studies have shown that pyroptosis can impact tumor cell proliferation, invasion and metastasis and further affect the cancer prognosis (Al Mamun et al., 2021). The expression of GSDMD in gastric cancer cells is lower than that in non-cancer cells, and the low expression of GSDMD promotes the proliferation of gastric cancer cells (Fang et al., 2020). Abnormally up-regulated GSDMB can also enhance the growth and invasive ability of bladder cancer cells (He et al., 2021). In addition, pyroptosis regulates the tumor immune microenvironment (TIME) and is involved in the body's immune response to tumors (Xi et al., 2019; Zhang et al., 2020). It has been proven that tumor pyroptosis can enhance tumor immunogenicity by attracting more anti-tumor lymphocytes and reconstruct the local or systemic anti-tumor immunity by reversing the immunosuppressive microenvironment around tumor cells (Tan et al., 2021). Therefore, 'inducing tumor pyroptosis' is considered a potential cancer treatment strategy. Interestingly, lncRNAs are also mediators of cancer pyroptosis (Chen et al., 2020; Tang et al., 2021). However, the clinical significance of most pyroptosis-related (PR) lncRNAs has not been clearly investigated.

With the deepening of research in pyroptosis, an increasing number of PR genes have been identified. A PR signature has also been identified in various types of tumors, such as ovarian cancer (Ye et al., 2021), gastric cancer (Shao et al., 2021) and lung adenocarcinoma (Lin et al., 2021). Moreover, PR genes signatures have been established to predict the prognosis of patients with BLCA (Chen et al., 2021; Fu and Wang, 2022). Several studies have recently suggested that PR long non-coding RNAs (lncRNAs) may also participate in the formation and development of tumors. miRNA-214 was reported to inhibit the occurrence of glioma cells by directly targeting caspase-1 (Jiang et al., 2017). LncRNA GAS5 overexpression may also induce caspase-1 upregulation and promote pyroptosis in ovarian cancer cells (Li et al., 2018). At present, PR lncRNAs signatures also have been gradually developed in tumor research to predict the TIME changes and prognosis of tumor patients. Fada et al. (Xia et al., 2021) developed a 15 prognostic PR lncRNAs risk model to predict colon adenocarcinoma patients' prognosis and TIME changes; Similar models have been developed in other tumor types, such as hepatocellular carcinoma (Wu et al., 2021) and kidney renal clear cell carcinoma (Tang et al., 2021). However, at present, few published PR lncRNAs signatures can be used to predict TIME changes and prognosis in patients with BLCA. The role and prognostic value of PR lncRNAs in BLCA have not been clarified.

In the present study, we comprehensively evaluated the molecular characteristics of these PR genes in BLCA and then identified PR lncRNAs with prognostic value based on the expression of PR genes. We performed a consistent clustering analysis of BLCA patients according to the expression of prognosis-related PR lncRNAs and identified two clusters. Based on these findings, the least absolute shrinkage and selection operator (LASSO) regression was used to establish a PR lncRNA signature and calculate the risk score associated with the prognosis of patients with BLCA. We also evaluated the relationship between the risk score and TIME. This study assesses the link between pyroptosis and TIME in BLCA, as well as provides a new reference to predict the prognosis of BLCA patients and identify personalized treatment strategies.

MATERIALS AND METHODS

Acquisition of Data From Patients With Bladder Cancer

We obtained BLCA transcriptome, gene mutation data and clinical data from the TCGA database (https://portal.gdc.cancer.gov/). The mRNA and lncRNA expression profile data were derived from 414 BLCA tissues and 19 normal tissues, and gene mutation data samples were derived from 411 BLCA tissues. Gene copy number variation data were obtained from the UCSC database (https://xenabrowser.net/datapages/), including 413 BLCA samples. The clinical data are shown in **Supplementary Table S1**, we extracted clinical data from 412 patients. Samples without complete clinical information will be excluded in the subsequent clinical correlation analysis.

Analyses of Molecular Characteristics of Pyroptosis-Related Genes in Bladder Cancer

The 33 PR genes were shown in **Supplementary Table S2,** and these genes have been proved to be associated with pyroptosis in previously published studies (Man and Kanneganti, 2015; Wang and Yin, 2017; Karki and Kanneganti, 2019; Xia et al., 2019; Chen et al., 2021). We extracted the expression data of 33 PR genes from the BLCA transcriptome data. Using the limma package for the differential expression analysis, we extracted the copy number variation data of 33 PR genes from the data obtained from the UCSC database. We then counted the frequency of copy number variation of these genes in all samples. The RCircos package was used to visualize the change information of gene copy numbers. Similarly, we used the maftools package to analyze the mutation data of 33 PR genes from the mutation data obtained from the TCGA database and counted the mutation frequencies.

We used the Search Tool for Interaction Genes (STRING) database (https://string-db.org/cgi/input.pl) to construct PPI networks for differentially expressed PR genes and used the OncoLnc online analysis tool (http://www.oncolnc.org/) to perform a prognostic analysis of these genes. The OncoLnc tool can be used to analyze the correlation between mRNA, miRNA or lncRNA expression and the prognosis of patients with specific types of tumors based on the prognostic data of the TCGA database (Anaya, 2016).

Identification of Pyroptosis-Related lncRNAs

We removed the samples with incomplete survival data, and 407 BLCA samples remained after merging with the PR lncRNA expression matrix. The co-expression method based on the expression of 33 PR genes was used to identify PR lncRNAs. A total of 812 PR lncRNAs were identified according to the criteria |correlation coefficient| > 0.4 and $p < 0.01$. The Igraph package was used to visualize the co-expression network. A univariate Cox regression analysis was performed to screen prognosis-related PR lncRNAs at $p < 0.05$.

Analyzing the Correlation Between Tumor Clusters and Clinical Features

The ConsensusClusterPlus packet is an algorithm that can identify cluster members and their number in datasets (such as microarray gene expression profiles) (Wilkerson and Hayes, 2010). A consistent clustering analysis was used to determine the optimal number of clusters (k) and verify the clustering rationality by a resampling-based approach to assess the stability of the clusters. We used this package to perform a consistent clustering analysis based on the prognosis-related PR lncRNA expression matrix and then performed a prognostic correlation analysis of BLCA clusters. The degree of immune cell infiltration in BLCA was evaluated using the CIBERSORT algorithm (Newman et al., 2015). The results of the correlation analysis between the BLCA clusters and immune cell infiltration were considered significant at $p < 0.05$.

Construction of Pyroptosis-Related lncRNA Signature

The BLCA patients were randomly divided into training and testing groups in a 1:1 ratio using the caret R package. The expression matrix of PR lncRNAs was combined with the prognosis data of the patients. A LASSO regression analysis was used to develop a PR lncRNA signature in the training group. The testing and entire groups were used to verify the established signature. The risk score of each BLCA patient was calculated according to the following formula:

Risk score = coefficient ($lncRNA_1$) × expression ($lncRNA_1$) + coefficient ($lncRNA_2$) × expression ($lncRNA_2$) + coefficient ($lncRNA_3$) × expression ($lncRNA_3$) + ... + coefficient ($lncRNA_n$) × expression ($lncRNA_n$).

The BLCA patients in all groups were then identified as high- and low-risk patients based on the median risk score obtained from the training group.

Analyzing the Prognostic Efficacy of the lncRNA Signature in Bladder Cancer

To determine whether or not the prognosis of the signature was independent of other clinical variables, univariate Cox and multivariate Cox regression analyses were used to calculate the values of the risk and other clinical features in predicting the prognosis of patients. The time-dependent receiver operating characteristic (ROC) curve was plotted using the survROC package. The area under the curve (AUC) at one, three and 5 years was calculated to determine the accuracy and specificity of the signature in predicting the prognosis.

Analyzing the Correlation Between Risk Score and Other Factors

We analyzed the correlation between the patients' clinical characteristics (including age, gender, grade and stage), tumor clusters and risk score. The expression of tumor immune checkpoint genes (ICGs) PD-1, PD-L1, PD-L2, CTLA-4, LAG3, CD47, CD4, CD8A and IDO1 in BLCA was obtained from the expression profile. The correlation between the risk score and ICGs was then analyzed. The principal component analysis (PCA) of risk in all BLCA patients was performed using the Rtsne R package to determine whether or not the signature could distinguish between high- and low-risk patients based on the expression of eight lncRNAs.

A Gene Set Enrichment Analysis and Gene Enrichment Analysis

To understand the pathways that differ between the two clusters of BLCA in this study, a GSEA analysis among BLCA clusters was performed using the GSEA 4.1.0 software program, and the results of the pathway analysis were considered significant at a false discovery rate (FDR) of <0.05. To understand the functions and pathways that may be involved in differentially expressed genes between high- and low-risk BLCA, the samples were divided into high- and low-risk groups and then subjected to a gene differential expression analysis. The screening criteria for differentially expressed genes (DEGs) were FDR <0.05 and |log fold change (FC)| > 1. After obtaining DEGs, the DAVID 6.8 database (https://david.ncifcrf.gov/) was used to perform GO and KEGG analyses. All analysis results were considered significant at FDR <0.05.

Statistical Analysis

Kaplan-Meier method was used to analyze the prognosis, and the Log rank test was used to determine the difference. The correlation between the two variables was tested by Spearman correlation analysis. Wilcoxon test was used to analyze the differences between high- and low-risk groups. The results of the above statistical analysis were considered significant at $p < 0.05$. Statistical analyses were performed using R software (version 4.1.2).

RESULTS

Molecular Characterization of Pyroptosis-Related Genes and Identification of Pyroptosis-Related lncRNAs in Bladder Cancer

We extracted the expression data of 33 PR genes and analyzed the differences in the expression between normal and tumor tissues. We found that ELANE, IL6, NLRP1 and NLRP3 had a low differential expression in BLCA; however, AIM2, CASP3, CASP5, CASP6, CASP8, GPX4, GSDMB, GSDMD, NLRP2, NLRP7, PLCG1 and PYCARD had a high differential expression in BLCA (**Figure 1A**). The univariate Cox regression analysis results showed that GSDMB, CASP9, AIM2, CASP6, CASP8, CASP1 and GSDMD were significantly correlated with the prognosis and were protective factors (**Figure 1B**). The copy number variation analysis results showed that the copy number changes were consistent with their expression (**Figure 1C**), with the main copy number changes of AIM2, GSDMC, GSDMD, NLRP7 and NLRP2 showing amplication (gain), and these genes were also highly expressed in BLCA. A mutation analysis identified the three genes (SCAF11, NLRP2 and NLRP7) with the highest mutation rates (**Figure 1D**).

To clarify the relationships between the roles of pyroptosis genes, we performed a PPI network analysis. We found that PYCARD had the most network nodes, suggesting a possible crucial regulatory role of PYCARD in BLCA (**Supplementary Figures S1A-S1B**). A Kaplan-Meier survival analysis showed that the expression of GSDMB and GSDMD was significantly correlated with the survival of patients, and the prognosis of patients with a high expression was better than that of patients with a low expression (**Figures 1E,F**).

According to the criteria |correlation coefficient| > 0.4 and $p < 0.01$, a total of 812 PR lncRNAs were identified from the TCGA BLCA expression profile data, and the co-expression network of PR genes/lncRNAs was plotted (**Supplementary Figure S1C**). The prognosis-related PR lncRNAs were screened using a univariate Cox regression analysis, and 194 prognosis-related PR lncRNAs

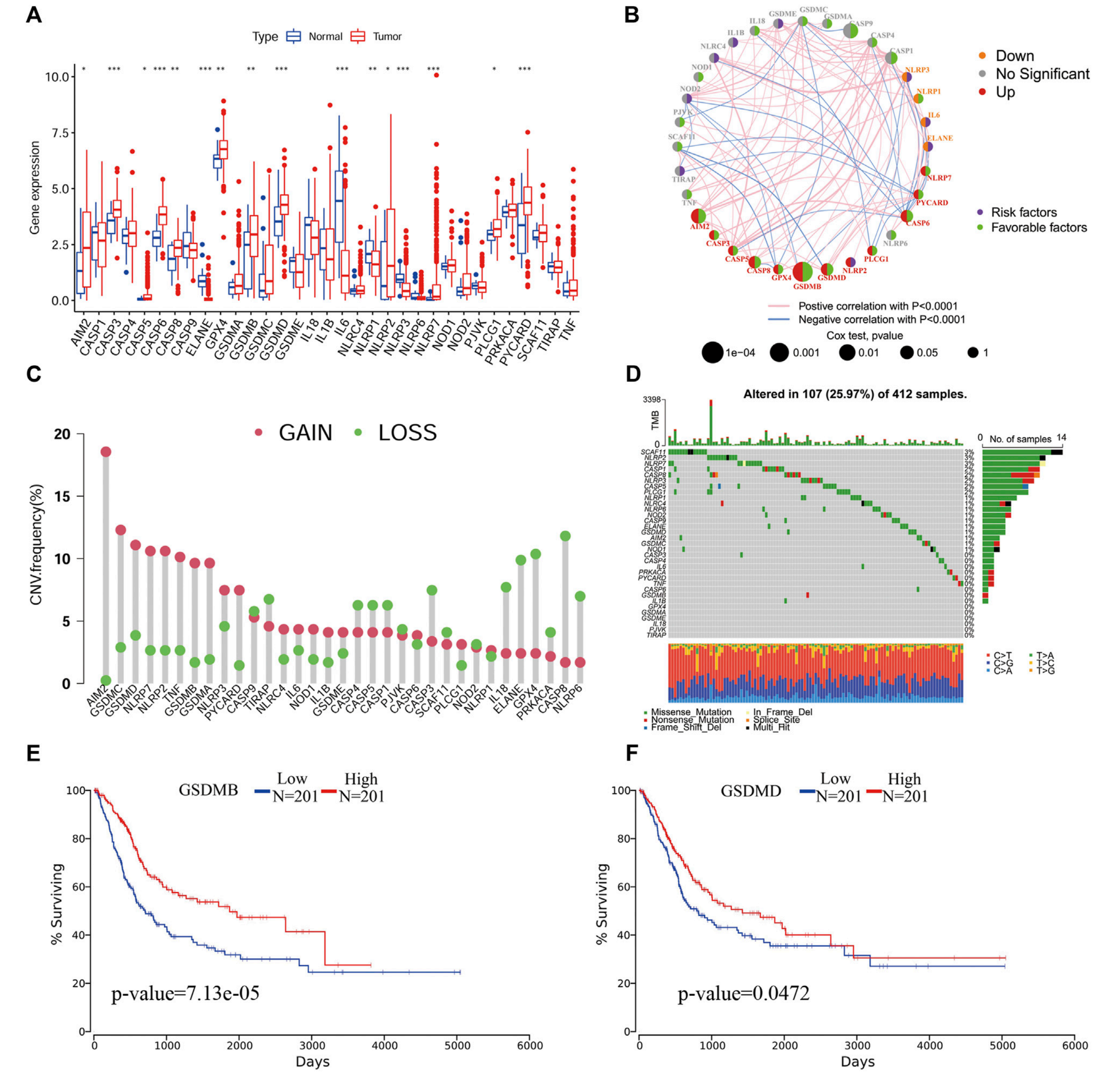


FIGURE 1 | An analysis of the molecular characteristics of PR genes in BLCA. **(A)**, An expression analysis of PR genes in tumor and normal tissues $p < 0.05$(*), $p < 0.01$(**) and $p < 0.001$(***). **(B)**, Co-expression and univariate Cox regression analyses of PR genes in BLCA. **(C)** CNV analysis of PR genes in BLCA. **(D)** Mutation frequency analysis of PR genes in BLCA. **(E)** Kaplan-Meier survival analysis of GSDMB in BLCA. **(F)** Kaplan-Meier survival analysis of GSDMD in BLCA.

were obtained (**Supplementary Figure S1D**). These prognosis-related PR lncRNAs were identified for subsequent research.

Results of Consistent Clustering Analysis of BLCA Based on Pyroptosis-Related lncRNAs

A consensus clustering algorithm was used to classify groups of BLCA patients based on the expression of prognosis-related PR lncRNAs. The k = 2-9 cumulative distribution function (CDF) representing the clustering counts. k = 2 was determined as the optimal clustering parameter based on the similarity of the expression of prognosis-related PR lncRNAs and the ratio of the fuzzy clustering metric. The 407 BLCA patients with complete survival information were divided into 2 clusters: cluster 1 (n = 122) and cluster 2 (n = 285) (**Figure 2A**; **Supplementary Table S3**).

The infiltration level of 23 immune cell subtypes in each sample of BLCA was calculated using the CIBERSORT algorithm. The

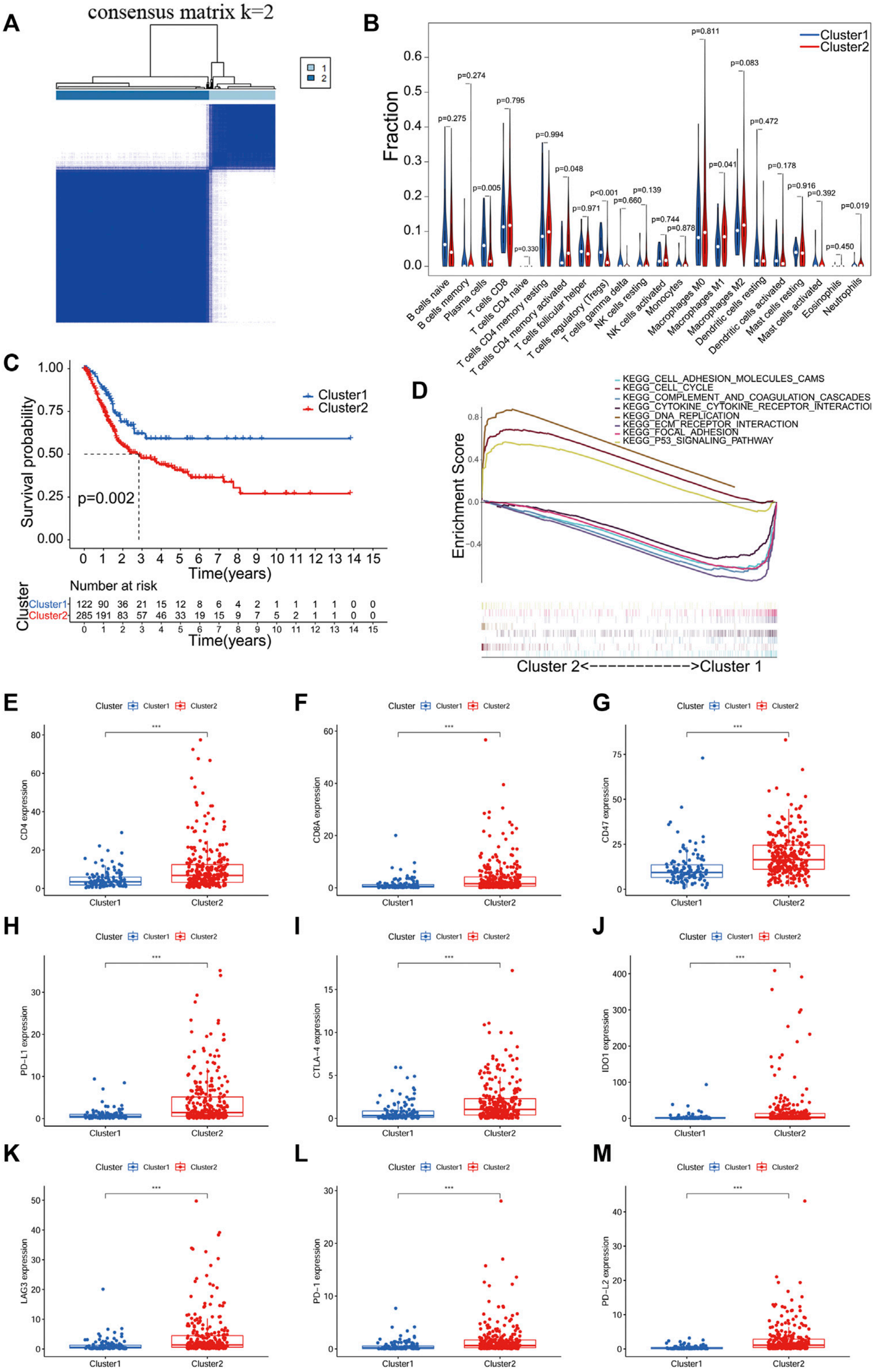


FIGURE 2 | Consistent clustering analysis based on PR lncRNA of BLCA. **(A)**, The TCGA BLCA cohort divided into two clusters at $k = 2$. **(B)**, An analysis of the relationship between clusters of BLCA and immune cell infiltration. **(C)** Kaplan-Meier survival analysis of patients with two clusters of BLCA. **(D)** Gene set enrichment analysis (GSEA) predicted potential functions and pathways between the two clusters. **(E-M)**, Expression analysis of immune checkpoint genes in two clusters of BLCA. $p < 0.05$(*), $p < 0.01$(**) and $p < 0.001$(***).

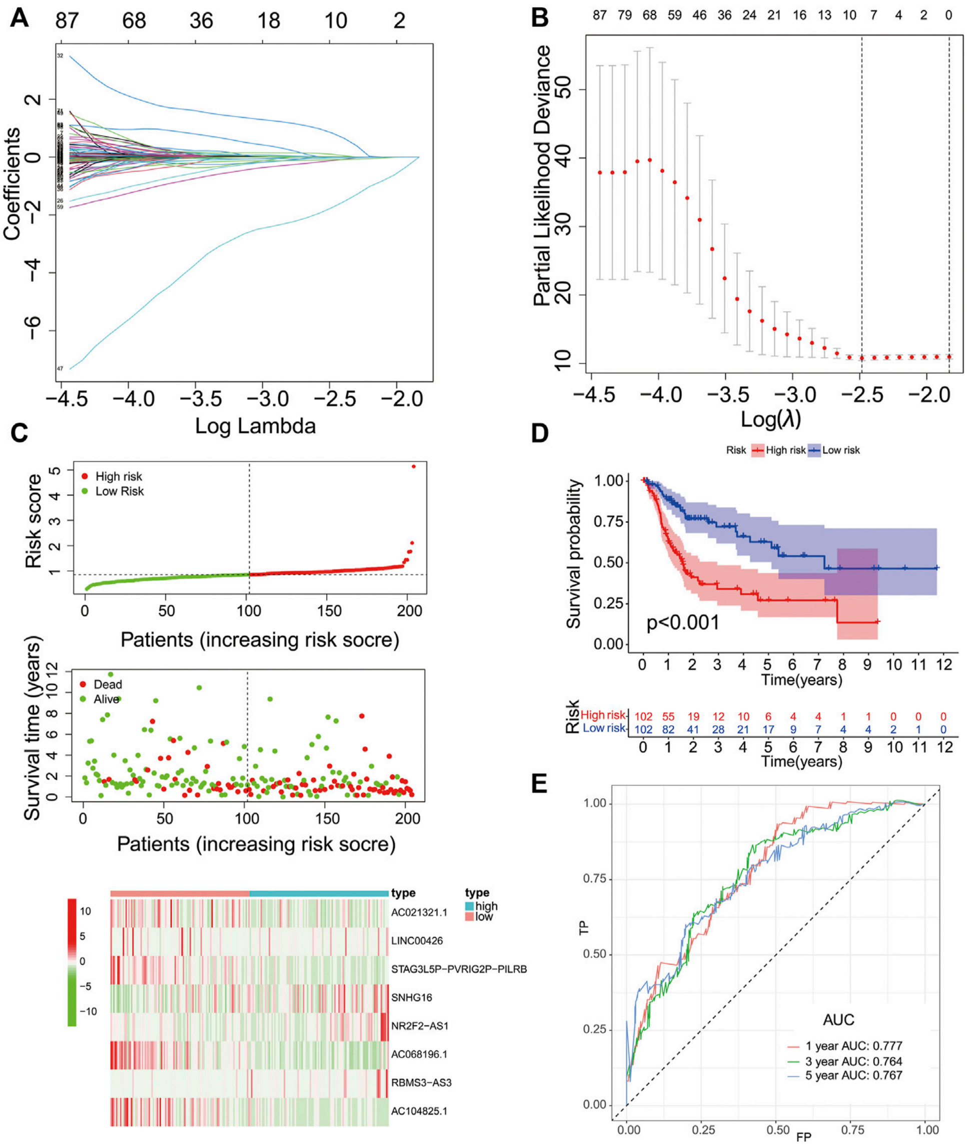


FIGURE 3 | Construction of the PR lncRNA signature in the training group. **(A-B)**, The adjustment parameter (λ) selected in the LASSO model was cross-validated by a factor of 10 of the minimum criterion. **(C)**, The survival status and lncRNA expression heat map. **(D)**, An analysis of the overall survival of high- and low-risk patients in the training group. **(E)**, ROC curves of sensitivity and specificity of the signature for predicting the prognosis.

correlation analysis results between BLCA subtypes and infiltration level of immune cells showed significant differences in T cells CD4$^+$ memory activated, T cells regulatory (Tregs), Plasma cells, Macrophages M1 and Neutrophils between different clusters ($p < 0.05$, **Figure 2B**). The overall survival of both clusters was calculated by the Kaplan-Meier method, and cluster one had a better prognosis than cluster 2 ($p = 0.002$, **Figure 2C**). In the GSEA analysis, we used FDR <0.05 as a filter and found that mainly the following pathways were activated between the two clusters: cell adhesion molecules cams, cell cycle, complement and coagulation cascades, cytokine-cytokine receptor interaction, DNA replication, ECM receptor interaction, focal adhesion and the p53 signaling pathway (**Figure 2D**).

In addition, we also analyzed the expression of ICGs among different clusters. The expression of all ICGs in cluster two was significantly higher than in cluster 1 (**Figures 2E-M**). It means that patients in cluster two are more likely to benefit from immunotargeted therapy.

Development of a PR lncRNA Signature

We then evaluated the reliability of PR lncRNAs for predicting the prognosis of patients. The BLCA patients were randomly divided into training (n = 204) and testing groups (n = 203). Eight significant lncRNAs were identified in the training group using a LASSO regression analysis: AC021321.1, LINC00426, STAG3L5P-

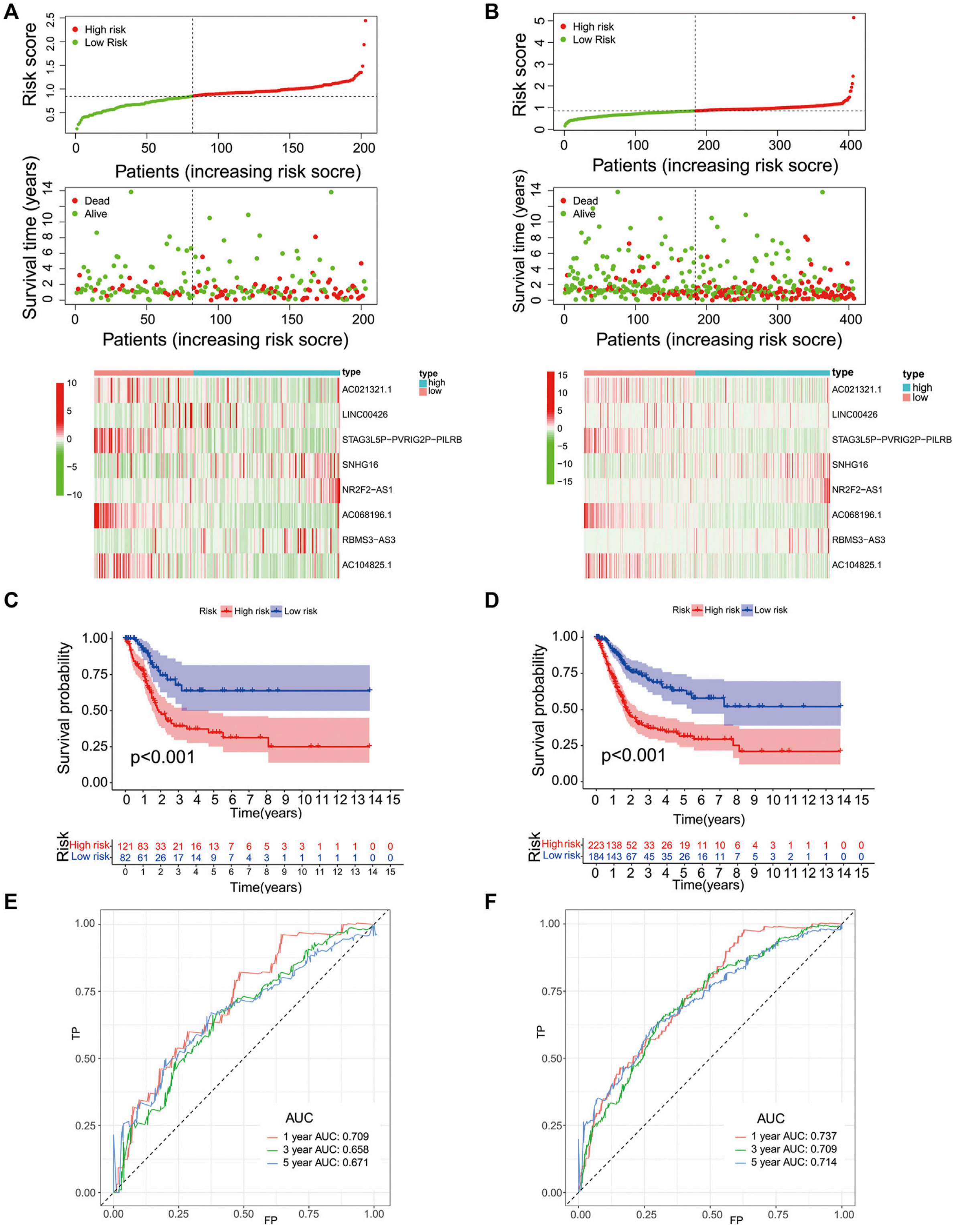


FIGURE 4 | Performance validation of the eight-PR-lncRNA signature. **(A)**, A heat map of the survival status and lncRNA expression in high- and low-risk patients in the testing group. **(B)**, A heat map of the survival status and lncRNA expression in high- and low-risk patients in the entire group. **(C)**, An analysis of the overall survival of high- and low-risk patients in the testing group. **(D)**, An analysis of the overall survival of high- and low-risk patients in the entire group. **(E)**, An assessment of the sensitivity and specificity of the prognostic prediction of the eight-PR-lncRNA signature in the testing group. **(F)**, An assessment of the sensitivity and specificity of the prognostic prediction of the eight-PR-lncRNA signature in the entire group.

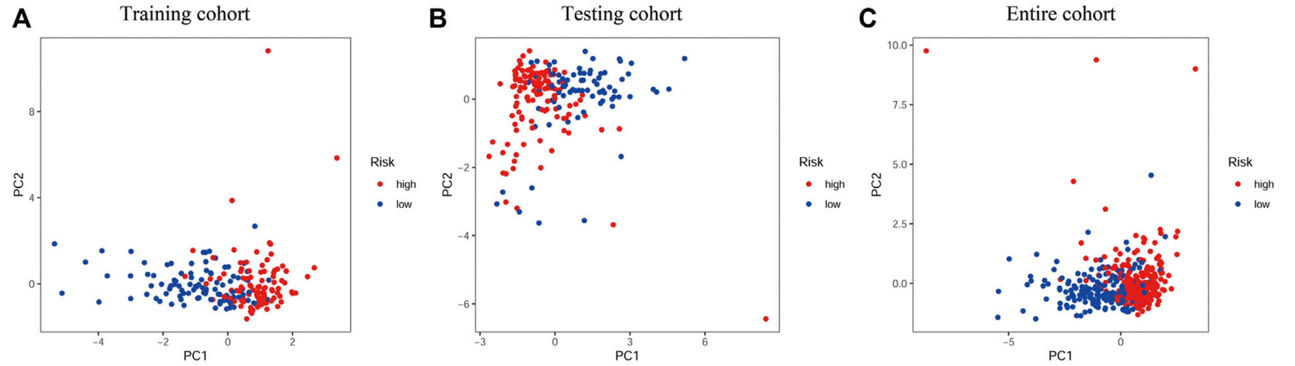


FIGURE 5 | PCA analysis of the different distribution patterns of eight PR lncRNAs on genome-wide expression profiles. **(A)**, Training cohort. **(B)**, Testing cohort. **(C)**, Entire cohort.

PVRIG2P-PILRB, SNHG16, NR2F2-AS1, AC068196.1, RBMS3-AS3 and AC104825.1. The corresponding coefficient for each lncRNA was then obtained (**Figures 3A,B**). Risk scores were calculated for the training, testing and entire groups, as follows:

Risk score = -0.006334916 × expr (AC021321.1) - 0.123481702 × expr (LINC00426) - 0.095912859 × expr (STAG3L5P-PVRIG2P-PILRB) + 0.0163094 × expr (SNHG16) + 0.835268684 × expr (NR2F2-AS1) - 1.751229658 × expr (AC068196.1) + 0.095237679 × expr (RBMS3-AS3) - 0.042537256 × expr (AC104825.1).

The lncRNAs with positive coefficients in the formula are risk factors (SNHG16, NR2F2-AS1 and RBMS3-AS3), while the lncRNAs with negative coefficients are protective factors (AC021321.1, LINC00426, STAG3L5P-PVRIG2P-PILRB, AC068196.1 and AC104825.1).

The training group's median risk score (0.8531) was used as the cut-off value, and patients were identified as low- and high-risk patients based on this cut-off value. The results in **Figure 3C** revealed that the patients with a high risk might have a poor prognosis. The OS analysis of the two groups showed that the OS of the high-risk group was significantly lower than that of the low-risk group (p < 0.001, **Figure 3D**). We used a time-dependent ROC curve to test the sensitivity and specificity of the diagnostic risk characteristics. In the training group, the AUC for predicting the patient survival at 1 year was 0.777, the AUC for predicting the survival at 3 years was 0.764, and the AUC for predicting the survival at 5 years was 0.767 (**Figure 3E**).

Validation of the Signature in Other Groups

We then validated the predictive efficacy of the eight-PR-lncRNA signature in the testing and entire groups. The patients in these two groups were identified as high- and low-risk patients using the same methods. **Figure 4A** and **Figure 4B** show the relationship between the risk score and survival status in the two groups, respectively, and all results were consistent with the training group. **Figures 4C-D** show the prognostic differences between the high- and low-risk patients in the testing and entire groups, respectively. These results were also consistent with the training group. The overall survival of the high-risk group was significantly lower than that in the low-risk group (p < 0.001). The time-dependent ROC curve in the testing group is shown in **Figure 4E**, and the time-dependent ROC curve in the entire group is shown in **Figure 4F**; all of them obtained an ideal AUC value.

We used a PCA analysis to examine the distribution patterns of the eight PR lncRNAs based on the expression profiles of all BLCA patients. The PCA analysis results suggested that the eight-PR-lncRNA signature could divide BLCA patients into high- and low-risk populations (**Figures 5A–C**).

Correlation Between the Eight-PR-lncRNA Signature and Clinical Features

The univariate and multivariate Cox analyses were used to analyze the performance of the signature in the training, testing and entire groups to identify independent factors for the overall survival (OS). The results of the three groups showed that risk was an independent factor associated with a poor prognosis in BLCA patients (p < 0.05; **Figures 6A–F**). The same analysis was performed in the entire group. The heat map visualized the differences in the expression of eight selected PR lncRNAs between the high- and low-risk groups and annotated clinical information (**Figure 6G**). Cluster two had a significantly higher risk than cluster 1 (p < 0.001, **Figure 6H**), consistent with the previous OS analysis results. In addition, the risk score of a high grade for BLCA was significantly higher than that of low-grade disease (**Figure 6I**). The same results were also obtained for the stage (p < 0.001, **Figure 6J**), T stage (p < 0.001, **Figure 6L**) and N stage (p = 0.0015, **Figure 6M**). However, there was no significant difference between the high- and low-ImmuneScore groups (p = 0.051, **Figure 6K**).

A prognostic analysis of high- and low-risk patients in specific clinical characteristics subgroups (age, gender, grade, stage, T, M and N) showed that the prognosis of high-risk patients was poor in all clinical characteristics subgroups, except for the low-grade and M1 subgroups (**Figure 7**).

Correlation Between Risk Score and Tumor Immunity

To understand the relationship between the risk score and the TIME of BLCA, we analyzed the correlation between the risk score and the infiltration level of 23 immune cell subtypes, with the results shown

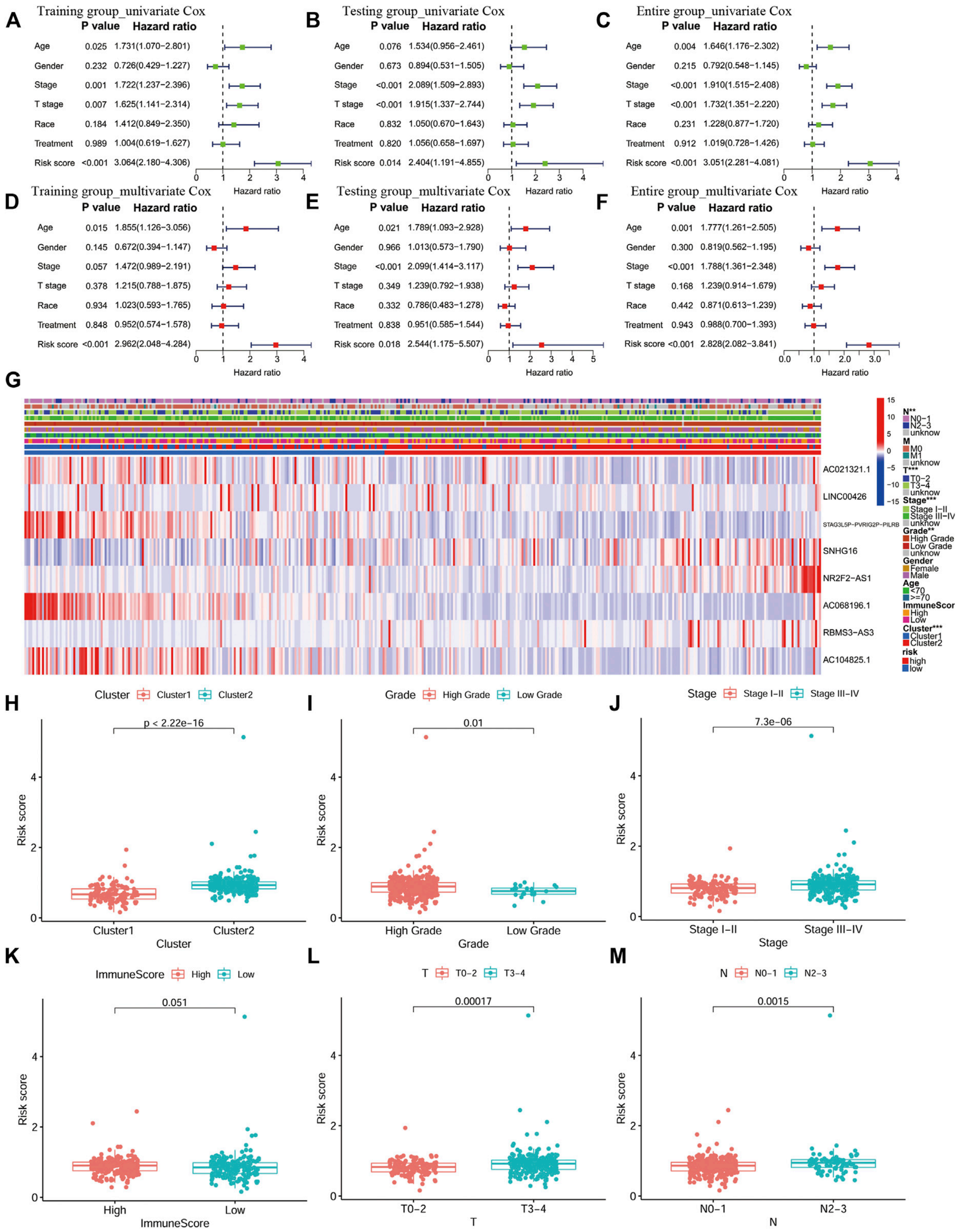


FIGURE 6 | An independent prognostic analysis of the eight-PR-lncRNA signature and a correlation analysis between the risk score and clinical characteristics. **(A)** Univariate Cox regression analysis in the training group. **(B)**, Multivariate Cox regression analysis in the training group. **(C)** Univariate Cox regression analysis in the testing group. **(D)** Multivariate Cox regression analysis in the testing group. **(E)** Univariate Cox regression analysis in the entire group. **(F)** Multivariate Cox regression analysis in the entire group. **(G)** Heat map of the lncRNA expression and clinicopathological features in high- and low-risk patients. $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***). **(H)**, The distribution of risk score in the two groups of consistent clustering results. **(I-J)**, The distribution of risk score by grade and stage of BLCA. **(K)**, The distribution of risk score in the ImmuneScore-high and ImmuneScore-low groups. **(L)**, The distribution of risk score by T stage. **(M)**, The distribution of risk score by N stage.

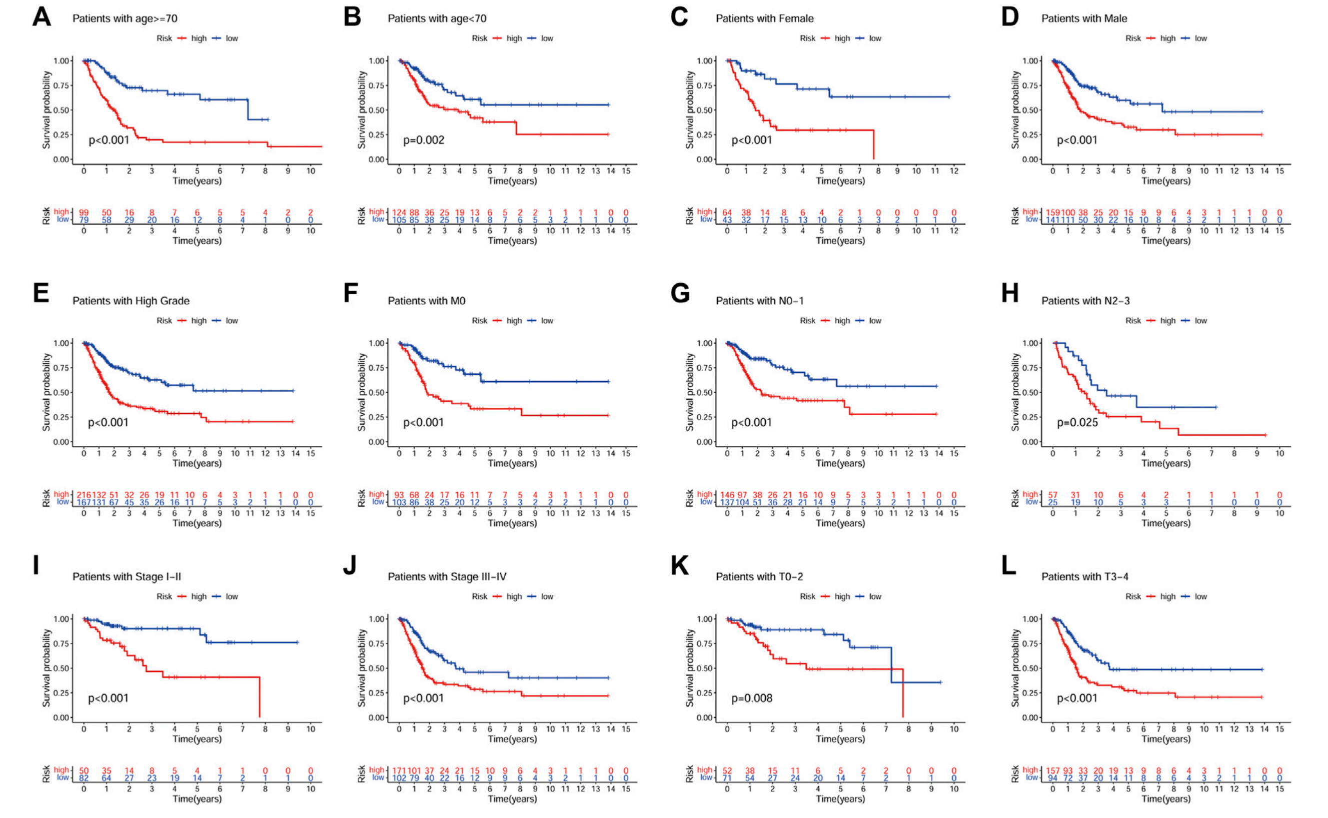


FIGURE 7 | Prognostic analysis of high- and low-risk patients in different clinical characteristics subgroups. **(A)**, Age ≥ 70. **(B)**, Age < 70. **(C)**, Female. **(D)**, Male. **(E)**, High grade. **(F)**, M0. **(G)**, N0-N1. **(H)**, N2-N3. **(I)**, Stage I-II. **(J)**, Stage III-IV. **(K)**, T0-T2. **(L)**, T3-T4.

in **Figure 8A**. Interestingly, there was a degree of heterogeneity in the levels of B-cell, T-cell, NK-cell and Dendritic cell infiltration between the high-risk and low-risk groups. We also examined the correlation between the risk score and the expression of ICGs, and the results showed that the risk score was significantly positively correlated with multiple ICGs ($p < 0.05$; **Figures 8B–J**).

Expression and Function Analysis of the Eight lncRNAs in the Signature

We also evaluated the expression of eight lncRNAs in BLCA. The results showed that the expression of LINC00426, NR2F2-AS1, RBMS3-AS3 and AC104825.1 in BLCA tissue was lower than that in normal tissues, while the expression of AC021321.1, STAG3L5P-PVRIG2P-PILRB, SNHG16 and AC068196.1 in BLCA tissue was higher than that in normal tissues (**Figures 9A,B**). **Figure 9C** demonstrates the regulatory relationship between these lncRNAs and PR genes. In addition, we also analyzed the expression correlation between the ICGs and lncRNAs in the signature. We found that AC021321.1, AC104825.1, AC068196.1 had a negative correlation with all ICGs, while LINC00426 had a positive correlation with all ICGs ($p < 0.05$; **Figure 9D**). To understand the possible function and mechanism of these eight lncRNAs in BLCA, we used a co-expression method to find the protein-coding genes (PCGs) of these eight lncRNAs, and the screening criteria were |Pearson correlation coefficient| > 0.4 and $p < 0.001$ (Gao et al., 2019). A total of 3141 PCGs were obtained, and these PCGs were submitted to the functional enrichment analysis using the DAVID database with FDR <0.05. GO enrichment results showed that these PCGs were mainly enriched in human immune response functions, such as the immune response (BP), inflammatory response (BP), T cell costimulation, regulation of immune response (BP), MHC class II protein complex (CC), T cell receptor complex (CC), immunological synapse (CC), cytokine receptor activity (MF) and MHC class II receptor activity (MF) (**Figure 9E**). The KEGG pathway enrichment analysis showed that these genes were also mainly enriched in immunomodulatory pathways, such as cytokine–cytokine receptor interaction, T cell receptor signaling pathway, B cell receptor signaling pathway and natural killer cell mediated cytotoxicity (**Figure 9F**).

Results of a Functional Analysis Between High- and Low-Risk Groups and Construction of a Nomogram

We also analyzed the functions and pathways involved in the DEGs in high- and low-risk groups. According to the screening criteria |logFC| > 1 and FDR <0.05, a total of 1017 DEGs were screened. Immune-related functions were found in the GO analysis

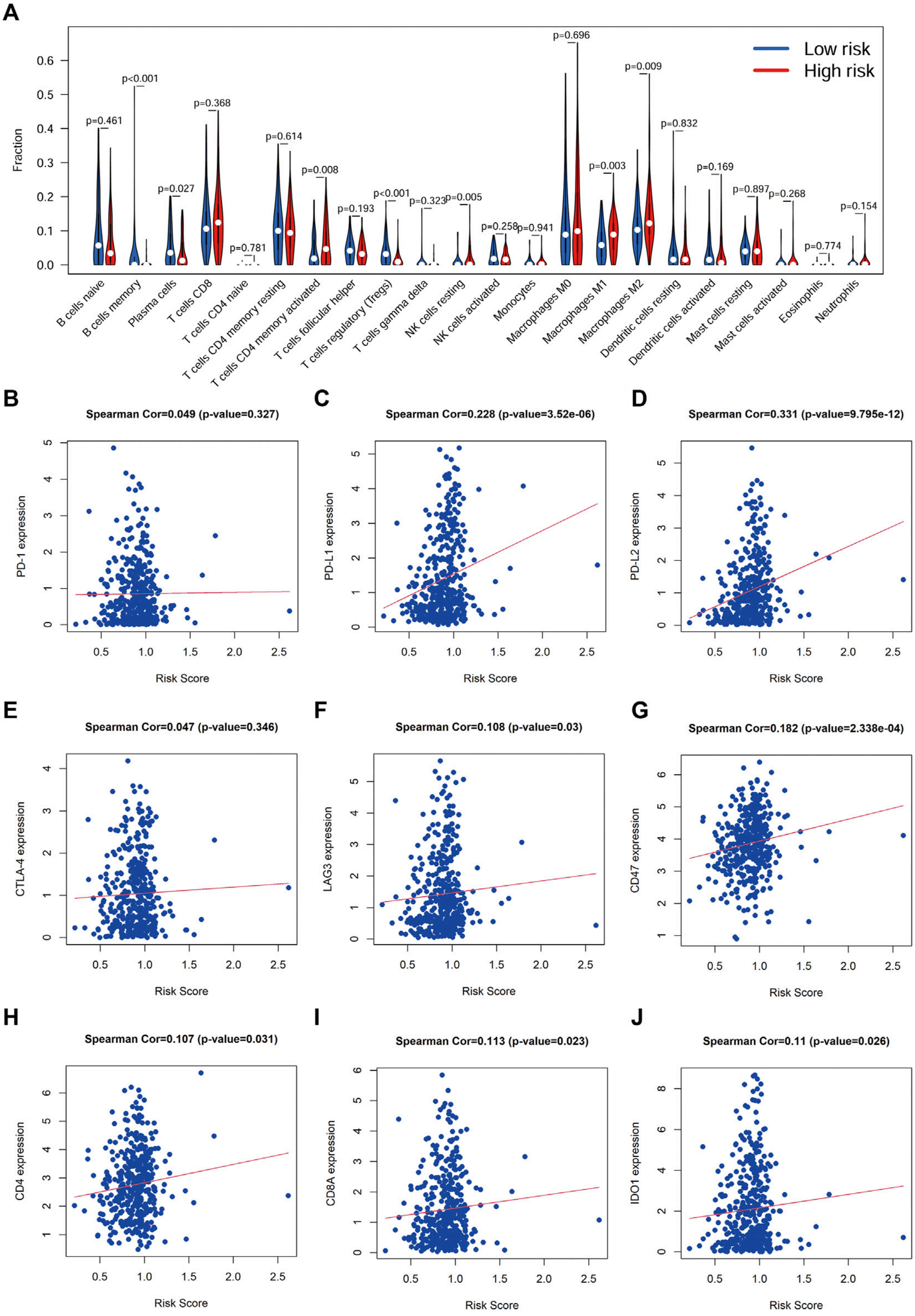


FIGURE 8 | Correlation between risk score and immune cell infiltration and ICGs **(A)**, Correlation analysis between risk score and immune cell infiltration. **(B-J)**, Correlation analysis between the risk score and immune checkpoint genes.

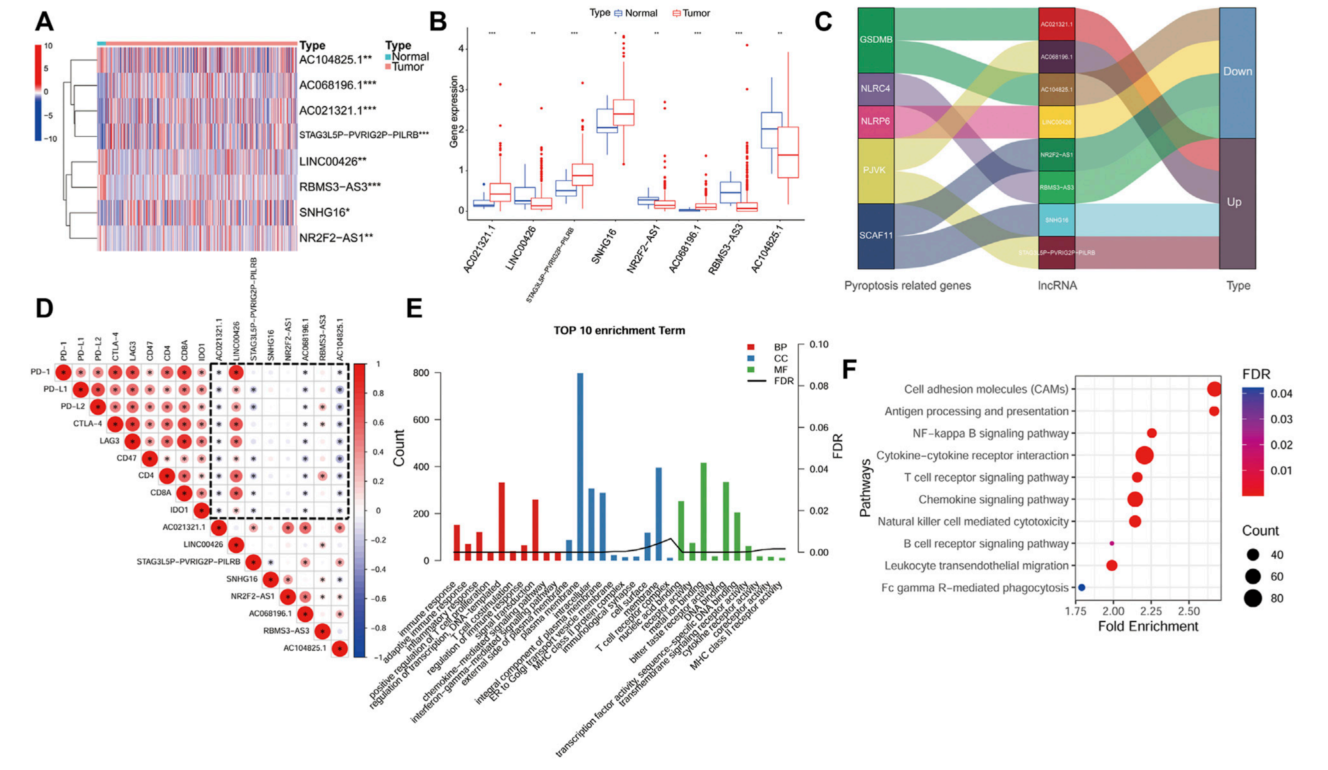


FIGURE 9 | Expression and functional analyses of lncRNAs in the signature. **(A,B)**, An expression analysis of the eight lncRNAs in BLCA tissues and normal tissues $p < 0.05$(*), $p < 0.01$(**) and $p < 0.001$(***). **(C)**, The regulatory relationship between the lncRNAs in the signature and PR genes. **(D)**, An analysis of the correlation between the immune checkpoint genes and the eight lncRNA expression (dotted frame). **(E)**, Results of a GO enrichment analysis of the protein-coding genes (PCGs). **(F)**, Results of a KEGG analysis of PCGs.

results, including inflammatory response (BP) (**Supplementary Figure S2A**), and the KEGG enrichment analysis also identified immune-related pathways, such as cytokine-cytokine receptor interaction (**Supplementary Figure S2B**).

To facilitate the clinical use of our signature to predict the prognosis of BLCA patients, we also developed a nomogram including risk classification and clinical risk characteristics to predict the one-, three- and 5-year OS (**Figure 10A**). The risk scores of the prognostic signature had superior predictive power to other clinical factors. The calibration plots showed that the observation and prediction rates of the OS had ideal consistency (**Figures 10B–D**).

DISCUSSION

BLCA is a tumor of the urinary system with a high incidence. Due to its complex pathogenesis, there are several different genetic subtypes of tumors, and these subtypes may have different therapeutic responses to the same treatment. If not correctly treated, BLCA can have a high morbidity and mortality (Kamat et al., 2016).

Cell death is a common topic in life science. Tumor cells have the ability to escape cell death contributes to the origin of tumors. This ability also plays a crucial role in acquiring treatment resistance, developing recurrence and metastasizing (Hanahan and Weinberg, 2011). Pyroptosis is a type of programmed cell death in inflammation mediated by GSDM (Lu et al., 2021). Our findings found that patients with high GSDMD and GSDMB expression had a better prognosis, and the results of the GSDMD analysis were consistent with previously published studies (Fang et al., 2020). However, the better prognosis of patients with high GSDMB expression seems to contradict previous studies finding that high expression of this gene in bladder cancer promotes tumor cell proliferation (He et al., 2021). Currently, there is controversy regarding the role of GSDMB in tumors. GSDMB is also involved in pyroptosis, it can promote atypical pyroptosis by enhancing the activity of caspase-4 and has the function of inhibiting the proliferation of tumor cells (Li et al., 2020). It is still not clear whether the GSDMB protein cleaved by caspase-3/-6/-7 is involved in pyroptosis. Our results further confirm that the role of GSDMB in tumorigenesis is controversial, indicating that GSDMB has great research value in future research.

Human genome sequencing data has shown that most RNA transcripts of non-protein-coding origin are transcribed from more than 90% of the human genome (Mattick and Makunin, 2006). With further research, more studies have shown that lncRNAs also play an essential role in the development and malignant progression of BLCA (Li et al., 2020). It has been reported that lncRNAs are involved in the pathological processes of various diseases through direct or indirect actions on proteins related to the pyroptosis signaling pathway (He

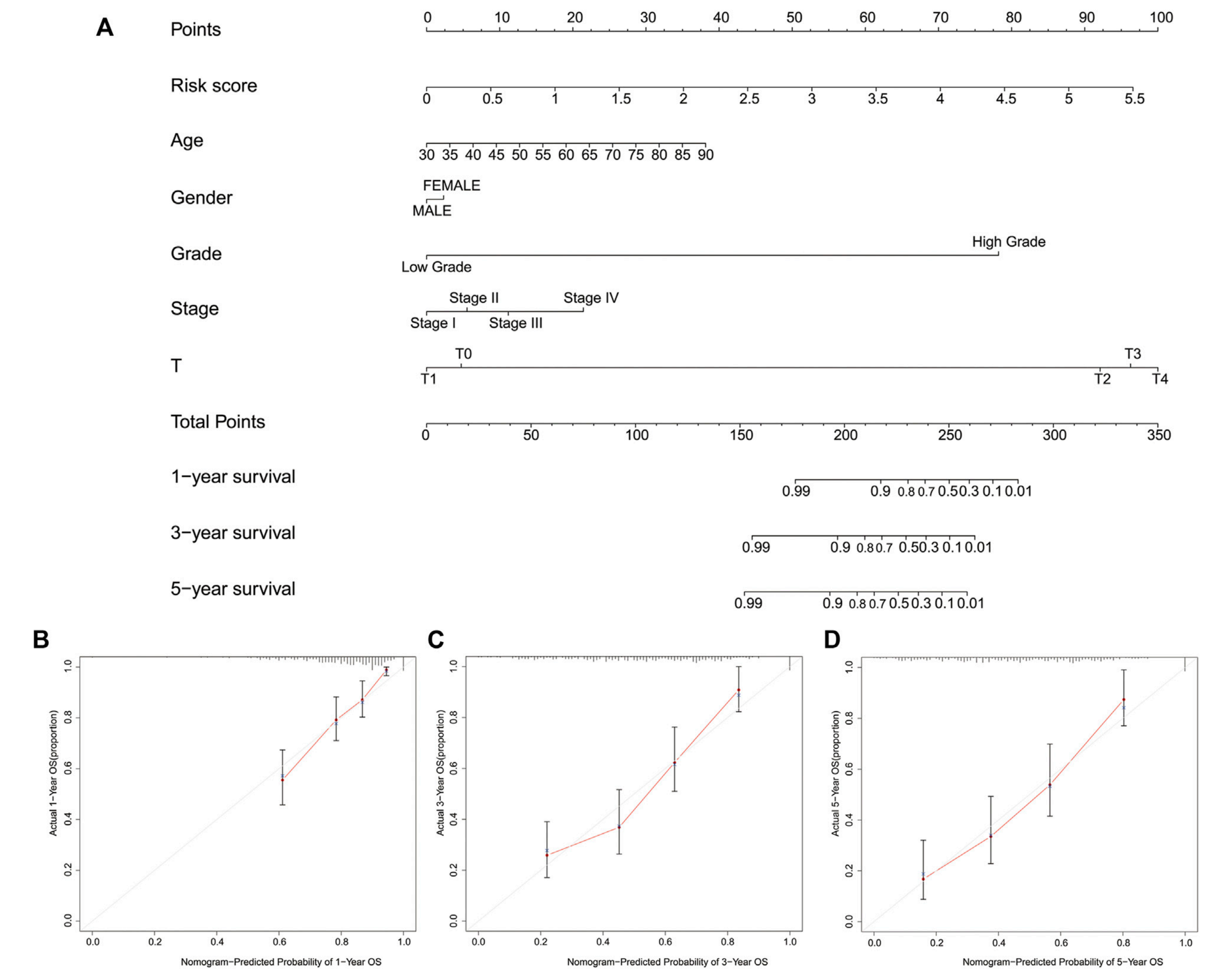


FIGURE 10 | Construction of a nomogram predicting the one-, three- and 5-year overall survival. **(A)**, A nomogram of the probability of predicting prognosis. **(B-D)**, The calibration plot of the nomogram.

et al., 2020). The release of cytokines produced by pyroptosis changes the TIME and promotes the growth of tumors by evading immune surveillance (Loveless et al., 2021). However, at present, there are few PR lncRNA signatures have been developed for BLCA.

We identified 812 PR lncRNAs based on the expression of 33 PR genes, and 194 prognosis-related PR lncRNAs were screened by a univariate Cox regression analysis. The BLCA cohort was then divided into two clusters based on the prognosis-related PR lncRNAs expression using consistent clustering. We found that the degree of infiltration of some immune cells differed significantly among clusters. The expression of the ICGs in cluster two was considerably higher than that in cluster 1, suggesting that patients in cluster two were more likely to have tumor immune escape and benefit from ICI therapy. In addition, the OS of cluster one was better than that of cluster 2, and the tumor grade of cluster one was also lower than that of cluster 2. The results of a GSEA analysis suggested that the following pathways were related to tumor development and metastasis: cell adhesion molecules cams (Cohen et al., 1997; Zhou et al., 2021), cell cycle (Li et al., 2021), cytokine-cytokine receptor interaction (Tang et al., 2020), focal adhesion (Tong et al., 2022) and p53 signaling pathway (Jiao et al., 2020). These results suggest a potential relationship between PR lncRNAs and the progression of BLCA. Consistent cluster analyses based on the PR lncRNA expression may help improve the efficacy of immunotherapy for BLCA.

We next applied LASSO regression to the training group to construct eight-PR-lncRNA signature (including AC021321.1, LINC00426, STAG3L5P-PVRIG2P-PILRB, SNHG16, NR2F2-AS1, AC068196.1, RBMS3-AS3 and AC104825.1). LncRNAs play an integral role in human epigenetic regulatory mechanisms. They participate in biological processes through epigenetic, transcriptional, post-transcriptional and translation regulatory targets, including cell growth, metastasis and apoptosis (Mirzaei et al., 2021, 2022). Their dysfunction is closely related to tumorigenesis (Han et al., 2020; Shigeyasu et al., 2020). Previous studies have shown that LINC00426 and SNHG16 can promote tumor development and participate in the regulation of TIME (Chen et al., 2020; Tao et al., 2020). LINC00426 and SNHG16 play an important role in the occurrence and development of tumors (Li et al., 2020; Wan et al., 2022). NR2F2-AS1 can down-regulate the expression of PDCD4 and inhibit the development of gastric cancer through competitive binding with miR-320b, it can also inhibit miR-494 methylation to regulate oral squamous cell carcinoma cells proliferation (Liang et al., 2022; Luo et al., 2022). Overexpression of RBMS3-AS3 inhibits cell proliferation, migration, invasion, angiogenesis and tumorigenicity of prostate cancer by up-regulating VASH1 (Jiang et al., 2020). The published evidence mentioned above suggests that these PR lncRNAs we identified are indeed associated with tumor development. While other lncRNAs AC021321.1, STAG3L5P-PVRIG2P-PILRB, AC068196.1 and AC104825.1 in our signature have not been reported in any published tumor studies, all were studied for the first time in our study. Our findings may provide evidence for future studies of these lncRNAs.

In the verification group, the signature also showed the same predictive performance as the training group. The OS analysis results indicated that an eight-PR-lncRNA signature could predict the survival rate of BLCA patients to some extent. We also found that risk score from eight-PR-lncRNA signature was an independent prognosis factor for BLCA patients. Patients with high-grade disease had a higher risk score than those with low-grade disease, and the same results were obtained between clusters 1 and 2, which was consistent with the conclusion that the OS of cluster one was better than that of cluster 2. The results of the risk score and ICGs correlation analysis suggested that patients with a high risk were more likely to experience tumor immune escape and benefit more from ICI therapy than others (Gao et al., 2020). Our results were consistent with the published results that pyroptosis can also increase the efficiency of tumor immunotherapy by recruiting immune cells and activating the immune system, its anti-tumor effect is also closely related to multiple ICGs (such as PD-1 or PD-L1) (Li et al., 2021).

To understand the possible function of these eight lncRNAs in BLCA, we used the co-expression method to find the co-expressed PCGs of the eight lncRNAs. The results of PCGs functional enrichment suggested that these eight lncRNAs may have immunomodulatory functions. Similarly, the enrichment analysis of genes that were differentially expressed between the high- and low-risk groups also found immune-related processes and pathways, such as inflammatory response (BP) and cytokine-cytokine receptor interaction (KEGG) (Bao and Cao, 2016). Furthermore, we also developed a nomogram containing risk classification and clinical risk characteristics to facilitate the clinical development and utilization of our findings (Iasonos et al., 2008). All these findings establish a close association between PR lncRNAs and the prognosis of BLCA patients as well as changes in TIME. The shortcoming of this study was the lack of lncRNA expression data from other sources for external validation, this is because we cannot find a suitable dataset containing these eight lncRNAs probes in other source datasets. Therefore, further external validation is needed to verify the reliability of the signature, and experimental validation of the role of these lncRNAs in BLCA cells should also be performed in the future.

CONCLUSIONS

Our study systematically evaluated the molecular biological characteristics and prognostic value of PR genes/lncRNAs in BLCA and identified an eight-PR-lncRNA signature (including AC021321.1, LINC00426, STAG3L5P-PVRIG2P-PILRB, SNHG16, NR2F2-AS1, AC068196.1, RBMS3-AS3 and AC104825.1) related to the prognosis of BLCA patients. We also analyzed the role of this signature in the TIME and its potential regulatory mechanisms, which provides an essential basis for future studies concerning the relationship between PR lncRNAs and BLCA immunity. Our findings will also help identify novel prognostic biomarkers and therapeutic targets for BLCA.

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

XG collected and assembled of data, analysed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft. JC conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

FUNDING

This work was supported by the National Key R&D Program of China (2018YFC2000300), the Hunan Provincial Natural Science Foundation of China (2021JJ40427) and The School-level scientific research project of Jishou University (Jdlc2021).

ACKNOWLEDGMENTS

We thank the Cancer Genome Atlas (TCGA) database for its public datasets used in our study. XG would like to thank himself for his efforts over the years, and even if he is walking all alone in his doctoral career, he will make it through the darkness and live a life that's filled with the stories to be told.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.865204/full#supplementary-material

REFERENCES

Al Mamun, A., Mimi, A. A., Aziz, M. A., Zaeem, M., Ahmed, T., Munir, F., et al. (2021). Role of Pyroptosis in Cancer and its Therapeutic Regulation. *Eur. J. Pharmacol.* 910, 174444. doi:10.1016/j.ejphar.2021.174444

Anaya, J. (2016). OncoLnc: Linking TCGA Survival Data to mRNAs, miRNAs, and lncRNAs. *PeerJ Comp. Sci.* 2, e67. doi:10.7717/peerj-cs.67

Ashrafizaveh, S., Ashrafizadeh, M., Zarrabi, A., Husmandi, K., Zabolian, A., Shahinozzaman, M., et al. (2021). Long Non-coding RNAs in the Doxorubicin Resistance of Cancer Cells. *Cancer Lett.* 508, 104–114. doi:10.1016/j.canlet.2021.03.018

Bao, Y., and Cao, X. (2016). Epigenetic Control of B Cell Development and B-Cell-Related Immune Disorders. *Clinic Rev. Allerg Immunol.* 50, 301–311. doi:10.1007/s12016-015-8494-7

Barth, D. A., Juracek, J., Slaby, O., Pichler, M., and Calin, G. A. (2020). lncRNA and Mechanisms of Drug Resistance in Cancers of the Genitourinary System. *Cancers* 12, 2148. doi:10.3390/cancers12082148

Benitez, J. C., Remon, J., and Besse, B. (2020). Current Panorama and Challenges for Neoadjuvant Cancer Immunotherapy. *Clin. Cancer Res.* 26, 5068–5077. doi:10.1158/1078-0432.CCR-19-3255

Broz, P., Pelegrín, P., and Shao, F. (2020). The Gasdermins, a Protein Family Executing Cell Death and Inflammation. *Nat. Rev. Immunol.* 20, 143–157. doi:10.1038/s41577-019-0228-2

Chen, W., Jiang, T., Mao, H., Gao, R., Zhang, H., He, Y., et al. (2020). SNHG16 Regulates Invasion and Migration of Bladder Cancer through Induction of Epithelial-To-Mesenchymal Transition. *Hum. Cel* 33, 737–749. doi:10.1007/s13577-020-00343-9

Chen, W., Zhang, W., Zhou, T., Cai, J., Yu, Z., and Wu, Z. (2021). A Newly Defined Pyroptosis-Related Gene Signature for the Prognosis of Bladder Cancer. *Ijgm* Vol. 14, 8109–8120. doi:10.2147/IJGM.S337735

Chen, X., Chen, H., Yao, H., Zhao, K., Zhang, Y., He, D., et al. (2021). Turning up the Heat on Non-immunoreactive Tumors: Pyroptosis Influences the Tumor Immune Microenvironment in Bladder Cancer. *Oncogene* 40, 6381–6393. doi:10.1038/s41388-021-02024-9

Chen, Z., He, M., Chen, J., Li, C., and Zhang, Q. (2020). Long Non-coding RNA SNHG7 I-nhibits NLRP3-dependent P-yroptosis by T-argeting the miR-34a/SIRT1 axis in L-iver C-ancer. *Oncol. Lett.* 20, 893–901. doi:10.3892/ol.2020.11635

Chou, R., Selph, S. S., Buckley, D. I., Gustafson, K. S., Griffin, J. C., Grusing, S. E., et al. (2016). Treatment of Muscle-Invasive Bladder Cancer: A Systematic Review. *Cancer* 122, 842–851. doi:10.1002/cncr.29843

Cohen, M. B., Griebling, T. L., Ahaghotu, C. A., Rokhlin, O. W., and Ross, J. S. (1997). Cellular Adhesion Molecules in Urologic Malignancies. *Am. J. Clin. Pathol.* 107, 56–63. doi:10.1093/ajcp/107.1.56

Dyugay, I. A., Lukyanov, D. K., Turchaninova, M. A., Serebrovskaya, E. O., Bryushkova, E. A., Zaretsky, A. R., et al. (2022). Accounting for B-Cell Behavior and Sampling Bias Predicts Anti-PD-L1 Response in Bladder Cancer. *Cancer Immunol. Res.* 10, 343–353. doi:10.1158/2326-6066.CIR-21-0489

Fang, Y., Tian, S., Pan, Y., Li, W., Wang, Q., Tang, Y., et al. (2020). Pyroptosis: A New Frontier in Cancer. *Biomed. Pharmacother.* 121, 109595. doi:10.1016/j.biopha.2019.109595

Fu, J., and Wang, Y. (2022). Identification of a Novel Pyroptosis-Related Gene Signature for Predicting Prognosis in Bladder Cancer. *Cancer Invest.* 40, 134–150. doi:10.1080/07357907.2021.1991944

Gao, X., Yang, J., and Chen, Y. (2020). Identification of a Four Immune-related Genes Signature Based on an Immunogenomic Landscape Analysis of clear Cell Renal Cell Carcinoma. *J. Cell Physiol.* 235, 9834–9850. doi:10.1002/jcp.29796

Gao, X., Zhang, S., Chen, Y., Wen, X., Chen, M., Wang, S., et al. (2019). Development of a Novel Six-long Noncoding RNA Signature Predicting Survival of Patients with Bladder Urothelial Carcinoma. *J. Cel. Biochem.* 120, 19796–19809. doi:10.1002/jcb.29285

Ghandour, R., Singla, N., and Lotan, Y. (2019). Treatment Options and Outcomes in Nonmetastatic Muscle Invasive Bladder Cancer. *Trends Cancer* 5, 426–439. doi:10.1016/j.trecan.2019.05.011

Han, W., Yu, F., and Guan, W. (2020). Oncogenic Roles of lncRNA BLACAT1 and its Related Mechanisms in Human Cancers. *Biomed. Pharmacother.* 130, 110632. doi:10.1016/j.biopha.2020.110632

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

He, D., Zheng, J., Hu, J., Chen, J., and Wei, X. (2020). Long Non-coding RNAs and Pyroptosis. *Clinica Chim. Acta* 504, 201–208. doi:10.1016/j.cca.2019.11.035

He, H., Yi, L., Zhang, B., Yan, B., Xiao, M., Ren, J., et al. (2021). USP24-GSDMB Complex Promotes Bladder Cancer Proliferation via Activation of the STAT3 Pathway. *Int. J. Biol. Sci.* 17, 2417–2429. doi:10.7150/ijbs.54442

Iasonos, A., Schrag, D., Raj, G. V., and Panageas, K. S. (2008). How to Build and Interpret a Nomogram for Cancer Prognosis. *Jco* 26, 1364–1370. doi:10.1200/JCO.2007.12.9791

Jiang, Z., Yao, L., Ma, H., Xu, P., Li, Z., Guo, M., et al. (2017). miRNA-214 Inhibits Cellular Proliferation and Migration in Glioma Cells Targeting Caspase 1 Involved in Pyroptosis. *Oncol. Res.* 25, 1009–1019. doi:10.3727/096504016X14813859905646

Jiang, Z., Zhang, Y., Chen, X., Wu, P., and Chen, D. (2020). Long Noncoding RNA RBMS3-AS3 Acts as a microRNA-4534 Sponge to Inhibit the Progression of Prostate Cancer by Upregulating VASH1. *Gene Ther.* 27, 143–156. doi:10.1038/s41434-019-0108-1

Jiao, F., Sun, H., Yang, Q., Sun, H., Wang, Z., Liu, M., et al. (2020). Identification of FADS1 through Common Gene Expression Profiles for Predicting Survival in Patients with Bladder Cancer. *Cmar* Vol. 12, 8325–8339. doi:10.2147/CMAR.S254316

Kamat, A. M., Hahn, N. M., Efstathiou, J. A., Lerner, S. P., Malmström, P.-U., Choi, W., et al. (2016). Bladder Cancer. *The Lancet* 388, 2796–2810. doi:10.1016/S0140-6736(16)30512-8

Karki, R., and Kanneganti, T.-D. (2019). Diverging Inflammasome Signals in Tumorigenesis and Potential Targeting. *Nat. Rev. Cancer* 19, 197–214. doi:10.1038/s41568-019-0123-y

Kaufman, D. S., Shipley, W. U., and Feldman, A. S. (2009). Bladder Cancer. *The Lancet* 374, 239–249. doi:10.1016/S0140-6736(09)60491-8

Li, H. J., Gong, X., Li, Z. K., Qin, W., He, C. X., Xing, L., et al. (2021). Role of Long Non-coding RNAs on Bladder Cancer. *Front Cel Dev Biol* 9, 672679. doi:10.3389/fcell.2021.672679

Li, H., Mu, Q., Zhang, G., Shen, Z., Zhang, Y., Bai, J., et al. (2020). Linc00426 Accelerates Lung Adenocarcinoma Progression by Regulating miR-455-5p as a Molecular Sponge. *Cell Death Dis* 11, 1051. doi:10.1038/s41419-020-03259-2

Li, J., Yang, C., Li, Y., Chen, A., Li, L., and You, Z. (2018). LncRNA GAS5 Suppresses Ovarian Cancer by Inducing Inflammasome Formation. *Biosci. Rep.* 38. doi:10.1042/BSR20171150

Li, L., Jiang, M., Qi, L., Wu, Y., Song, D., Gan, J., et al. (2021). Pyroptosis, a New Bridge to Tumor Immunity. *Cancer Sci.* 112, 3979–3994. doi:10.1111/cas.15059

Li, L., Li, Y., and Bai, Y. (2020). Role of GSDMB in Pyroptosis and Cancer. *Cmar* Vol. 12, 3033–3043. doi:10.2147/CMAR.S246948

Li, Y., Li, G., Guo, X., Yao, H., Wang, G., and Li, C. (2020). Non-coding RNA in Bladder Cancer. *Cancer Lett.* 485, 38–44. doi:10.1016/j.canlet.2020.04.023

Liang, Y., Wu, X., Lee, J., Yu, D., Su, J., Guo, M., et al. (2022). lncRNA NR2F2-AS1 Inhibits the Methylation of miR-494 to Regulate Oral Squamous Cell Carcinoma Cell Proliferation. *Arch. Oral Biol.* 134, 105316. doi:10.1016/j.archoralbio.2021.105316

Lin, W., Chen, Y., Wu, B., Chen, Y., and Li, Z. (2021). Identification of the Pyroptosis-related P-rognostic G-ene S-ignature and the A-ssociated R-egulation axis in L-ung A-denocarcinoma. *Cell Death Discov.* 7, 161. doi:10.1038/s41420-021-00557-2

Liu, X., Xia, S., Zhang, Z., Wu, H., and Lieberman, J. (2021). Channelling Inflammation: Gasdermins in Physiology and Disease. *Nat. Rev. Drug Discov.* 20, 384–405. doi:10.1038/s41573-021-00154-z

Loveless, R., Bloomquist, R., and Teng, Y. (2021). Pyroptosis at the Forefront of Anticancer Immunity. *J. Exp. Clin. Cancer Res.* 40, 264. doi:10.1186/s13046-021-02065-8

Lu, X., Guo, T., and Zhang, X. (2021). Pyroptosis in Cancer: Friend or Foe? *Cancers* 13, 3620. doi:10.3390/cancers13143620

Luo, M., Deng, S., Han, T., Ou, Y., and Hu, Y. (2022). LncRNA NR2F2-AS1 Functions as a Tumor Suppressor in Gastric Cancer through Targeting miR-320b/PDCD4 Pathway. *Histol. Histopathol* 20, 18429. doi:10.14670/HH-18-429

Malmström, P.-U., Agrawal, S., Bläckberg, M., Boström, P. J., Malavaud, B., Zaak, D., et al. (2017). Non-muscle-invasive Bladder Cancer: a Vision for the Future. *Scand. J. Urol.* 51, 87–94. doi:10.1080/21681805.2017.1283359

Man, S. M., and Kanneganti, T.-D. (2015). Regulation of Inflammasome Activation. *Immunol. Rev.* 265, 6–21. doi:10.1111/imr.12296

Mattick, J. S., and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* 151, R17–R29. doi:10.1093/hmg/ddl046

Mirzaei, S., Gholami, M. H., Hushmandi, K., Hshemi, F., Zabolian, A., Canadas, I., et al. (2022). The Long and Short Non-coding RNAs Modulating EZH2 Signaling in Cancer. *J. Hematol. Oncol.* 15, 18. doi:10.1186/s13045-022-01235-1

Mirzaei, S., Zarrabi, A., Hashemi, F., Zabolian, A., Saleki, H., Ranjbar, A., et al. (2021). Regulation of Nuclear Factor-KappaB (NF-Kb) Signaling Pathway by Non-coding RNAs in Cancer: Inhibiting or Promoting Carcinogenesis? *Cancer Lett.* 509, 63–80. doi:10.1016/j.canlet.2021.03.025

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Shao, W., Yang, Z., Fu, Y., Zheng, L., Liu, F., Chai, L., et al. (2021). The Pyroptosis-Related Signature Predicts Prognosis and Indicates Immune Microenvironment Infiltration in Gastric Cancer. *Front. Cel Dev. Biol.* 9, 676485. doi:10.3389/fcell.2021.676485

Shi, J., Gao, W., and Shao, F. (2017). Pyroptosis: Gasdermin-Mediated Programmed Necrotic Cell Death. *Trends Biochem. Sci.* 42, 245–254. doi:10.1016/j.tibs.2016.10.004

Shigeyasu, K., Toden, S., Ozawa, T., Matsuyama, T., Nagasaka, T., Ishikawa, T., et al. (2020). The PVT1 lncRNA Is a Novel Epigenetic Enhancer of MYC, and a Promising Risk-Stratification Biomarker in Colorectal Cancer. *Mol. Cancer* 19, 155. doi:10.1186/s12943-020-01277-4

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

Tan, Y., Chen, Q., Li, X., Zeng, Z., Xiong, W., Li, G., et al. (2021). Pyroptosis: a New Paradigm of Cell Death for Fighting against Cancer. *J. Exp. Clin. Cancer Res.* 40, 153. doi:10.1186/s13046-021-01959-x

Tang, W., Zhu, S., Liang, X., Liu, C., and Song, L. (2021). The Crosstalk between Long Non-coding RNAs and Various Types of Death in Cancer Cells. *Technol. Cancer Res. Treat.* 20, 153303382110330. doi:10.1177/15330338211033044

Tang, X., Zhang, A., Feng, Y., Su, Y., Wang, X., Jiang, F., et al. (2021). A Novel Pyroptosis-Related lncRNAs Signature for Predicting the Prognosis of Kidney Renal Clear Cell Carcinoma and its Associations with Immunity. *J. Oncol.* 2021, 1–15. doi:10.1155/2021/9997185

Tang, Y., Hu, Y., Wang, J., and Zeng, Z. (2020). A Novel Risk Score Based on a Combined Signature of 10 Immune System Genes to Predict Bladder Cancer Prognosis. *Int. Immunopharmacology* 87, 106851. doi:10.1016/j.intimp.2020.106851

Tao, Y., Li, Y., and Liang, B. (2020). Comprehensive Analysis of Microenvironment-Related Genes in Lung Adenocarcinoma. *Future Oncol.* 16, 1825–1837. doi:10.2217/fon-2019-0829

Tong, S., Yin, H., Fu, J., and Li, Y. (2022). Niban Apoptosis Regulator 1 Promotes Gemcitabine Resistance by Activating the Focal Adhesion Kinase Signaling Pathway in Bladder Cancer. *J. Cancer* 13, 1103–1118. doi:10.7150/jca.66248

Wan, L., Gu, D., and Li, P. (2022). LncRNA SNHG16 Promotes Proliferation and Migration in Laryngeal Squamous Cell Carcinoma via the miR-140-5p/NFAT5/Wnt/β-Catenin Pathway axis. *Pathol. - Res. Pract.* 229, 153727. doi:10.1016/j.prp.2021.153727

Wang, B., and Yin, Q. (2017). AIM2 Inflammasome Activation and Regulation: A Structural Perspective. *J. Struct. Biol.* 200, 279–282. doi:10.1016/j.jsb.2017.08.001

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

Wu, Z.-H., Li, Z.-W., Yang, D.-L., and Liu, J. (2021). Development and Validation of a Pyroptosis-Related Long Non-coding RNA Signature for Hepatocellular Carcinoma. *Front. Cel Dev. Biol.* 9, 713925. doi:10.3389/fcell.2021.713925

Xi, G., Gao, J., Wan, B., Zhan, P., Xu, W., Lv, T., et al. (2019). GSDMD Is Required for Effector CD8+ T Cell Responses to Lung Cancer Cells. *Int. Immunopharmacology* 74, 105713. doi:10.1016/j.intimp.2019.105713

Xia, F., Yan, Y., and Shen, C. (2021). A Prognostic Pyroptosis-Related lncRNAs Risk Model Correlates with the Immune Microenvironment in Colon Adenocarcinoma. *Front. Cel Dev. Biol.* 9, 811734. doi:10.3389/fcell.2021.811734

Xia, X., Wang, X., Cheng, Z., Qin, W., Lei, L., Jiang, J., et al. (2019). The Role of Pyroptosis in Cancer: Pro-cancer or Pro-"host"? *Cel Death Dis* 10, 650. doi:10.1038/s41419-019-1883-8

Ye, Y., Dai, Q., and Qi, H. (2021). A Novel Defined Pyroptosis-Related Gene Signature for Predicting the Prognosis of Ovarian Cancer. *Cel Death Discov.* 7, 71. doi:10.1038/s41420-021-00451-x

Zhang, Y., Chen, X., Lin, J., and Jin, X. (2021). Biological Functions and Clinical Significance of Long Noncoding RNAs in Bladder Cancer. *Cel Death Discov.* 7, 278. doi:10.1038/s41420-021-00665-z

Zhang, Z., Zhang, Y., Xia, S., Kong, Q., Li, S., Liu, X., et al. (2020). Gasdermin E Suppresses Tumour Growth by Activating Anti-tumour Immunity. *Nature* 579, 415–420. doi:10.1038/s41586-020-2071-9

Zhen, H., Du, P., Yi, Q., Tang, X., and Wang, T. (2021). LINC00958 Promotes Bladder Cancer Carcinogenesis by Targeting miR-490-3p and AURKA. *BMC Cancer* 21, 1145. doi:10.1186/s12885-021-08882-6

Zhen, S., Lu, J., Chen, W., Zhao, L., and Li, X. (2018). Synergistic Antitumor Effect on Bladder Cancer by Rational Combination of Programmed Cell Death 1 Blockade and CRISPR-Cas9-Mediated Long Non-coding RNA Urothelial Carcinoma Associated 1 Knockout. *Hum. Gene Ther.* 29, 1352–1363. doi:10.1089/hum.2018.048

Zhou, W., Ouyang, J., Li, J., Liu, F., An, T., Cheng, L., et al. (2021). MRPS17 Promotes Invasion and Metastasis through PI3K/AKT Signal Pathway and Could Be Potential Prognostic Marker for Gastric Cancer. *J. Cancer* 12, 4849–4861. doi:10.7150/jca.55719

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.

ORIGINAL RESEARCH
published: 03 June 2022
doi: 10.3389/fgene.2022.833989

A Robust Immuno-Prognostic Model of Non-Muscle-Invasive Bladder Cancer Indicates Dynamic Interaction in Tumor Immune Microenvironment Contributes to Cancer Progression

Xiaomeng Sun[1,2†], Huilin Xu[1†], Gang Liu[1†], Jiani Chen[3], Jinrong Xu[4], Mingming Li[3]* and Lei Liu[1]**

[1]Institutes of Biomedical Sciences and School of Basic Medical Sciences, Fudan University, Shanghai, China, [2]Research Institute, GloriousMed Clinical Laboratory Co., Ltd., Shanghai, China, [3]Department of Pharmacy, Second Affiliated Hospital of Naval Medical University, Shanghai, China, [4]Department of Electronic Engineering, Taiyuan Institute of Technology, Taiyuan, China

Non-muscle-invasive bladder cancer (NMIBC) accounts for more than 70% of urothelial cancer. More than half of NMIBC patients experience recurrence, progression, or metastasis, which essentially reduces life quality and survival time. Identifying the high-risk patients prone to progression remains the primary concern of risk management of NMIBC. In this study, we included 1370 NMIBC transcripts data from nine public datasets, identified nine tumor-infiltrating marker cells highly related to the survival of NMIBC, quantified the cells' proportion by self-defined differentially expressed signature genes, and established a robust immuno-prognostic model dividing NMIBC patients into low-risk versus high-risk progression groups. Our model implies that the loss of crosstalk between tumor cells and adjacent normal epithelium, along with enriched cell proliferation signals, may facilitate tumor progression. Thus, evaluating tumor progression should consider various components in the tumor immune microenvironment instead of the single marker in a single dimension. Moreover, we also appeal to the necessity of using appropriate meta-analysis methods to integrate the evidence from multiple sources in the feature selection step from large-scale heterogeneous omics data such as our study.

Keywords: non-muscle-invasive bladder cancer, tumor immune microenvironment, tumor progression, collagen family, cancer-associated fibroblasts

Edited by:
*Jianjiong Gao,
Memorial Sloan Kettering Cancer Center, United States*

Reviewed by:
*Xin Gao,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Bing Shen,
Shanghai General Hospital, China*

***Correspondence:**
*Jinrong Xu
275919672@qq.com
Mingming Li
limingming@smmu.edu.cn
Lei Liu
liulei_sibs@163.com*

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *12 December 2021*
Accepted: *28 April 2022*
Published: *03 June 2022*

Citation:
*Sun X, Xu H, Liu G, Chen J, Xu J, Li M and Liu L (2022) A Robust Immuno-Prognostic Model of Non-Muscle-Invasive Bladder Cancer Indicates Dynamic Interaction in Tumor Immune Microenvironment Contributes to Cancer Progression.
Front. Genet. 13:833989.
doi: 10.3389/fgene.2022.833989*

Abbreviations: AUA, American Urological Association; AUCs, areas under curves; BCG, Bacillus Calmette-Guerin; CAFs, cancer-associated fibroblasts; CIS or Tis, carcinoma *in situ;* DCs, dendritic cells; DEGs, differentially expressed genes; DFS, disease-free survival; EAU, European Association of Urology; ECM, extracellular matrix; EMT, epithelial-to-mesenchymal transition; EORTC, European Organisation for Research and Treatment of Cancer; FDR, false discovery rate; GEP, gene expression profile; GO, gene ontology; HRs, hazard ratios; IPS, immune prognostic signature; KEGG, Kyoto Encyclopedia of Genes and Genomes; MIBC, muscle-invasive bladder cancer; NK cells, natural killer; NMIBC, non-muscle-invasive bladder cancer; OS, overall survival; PFS, progression-free survival; ROCs, receiver operating characteristic curves; ssGSEA, single-sample gene set enrichment analysis; sTIL, stromal tumor-infiltrating lymphocyte; SUO, Society of Urologic Oncology; TIME, tumor immune microenvironment.

INTRODUCTION

Bladder cancer contributed to 573,278 new cases and 212,536 deaths worldwide (Sung et al., 2021) in 2020. It is one of the cancers with the most longitudinal costs and consumed resources. Approximately 70–75% of newly diagnosed primary bladder cancers are non-muscle-invasive bladder cancer (NMIBC) (Lenis et al., 2020; Ottley et al., 2020). Up to 21–53% of them eventually progress to life-threatening muscle-invasive bladder cancer (MIBC) (Cookson et al., 1997; van den Bosch and Alfred Witjes, 2011), depending on the stage and grade. Identifying the NMIBC patients with a high progression potential at the early treatment stage remains the primary object of bladder cancer clinical practice.

Several risk classification frameworks have been suggested and applied in NMIBC risk management. European Association of Urology (EAU) prognostic factor risk groups updated the EAU NMIBC Guidelines Panel in 2021 by dividing NMIBC patients into four risk groups: low-, intermediate-, high-, and a new, very high-risk group, with the probability of progression at 5-year of <1%, 3.6–4.9%, 9.6–11%, and >40% (Sylvester et al., 2021). Clinicopathological features employed in the panel included: tumor stage, the World Health Organization (WHO) 1973 or 2004/2016 grade, concomitant carcinoma *in situ* (CIS or Tis), number of tumors, tumor size, and age. American Urological Association (AUA) and Society of Urologic Oncology (SUO) also amended the AUA/SUO Joint Guideline in 2020 by classifying NMIBC patients into low-, intermediate-, and high-risk groups (Chang et al., 2016; Chang et al., 2020). Apart from the clinical features used in the EAU Panel, AUA risk stratification also took variant histology, preceding recurrent disease, Bacillus Calmette-Guerin (BCG) treatment failure, and involvement of prostatic urethral into consideration. Although such frameworks essentially help the risk management of NMIBC patients and are readily used in bedside patient care, a more precise solution is always in need.

To fulfill the need, molecular subtyping and gene expression modeling based on the omics analysis have become mainstream in clinical decision support scenarios like diagnosis, treatment response prediction, and prognostic stratification. The UROMOL project, a European multicenter prospective study of NMIBC spanning from 2008 to date, identified high-risk class 2a tumors at the transcriptomic level and high-risk class GC3 tumors at the genomic level (Lindskrog et al., 2021). They also revealed that higher immune cell infiltration strongly correlated with lower recurrence rates. However, the association between immune cell infiltration and cancer progression remained unknown. Since there were too few progression events for evaluating its effect on progression-free survival (PFS), Zheng and colleagues developed an immune prognostic signature (IPS) based on 14 overall survival (OS) associated immune genes. Then they proved that high-risk patients assessed by the IPS score had worse OS than those with low-risk scores in validation datasets (Zheng et al., 2020). Ottley et al. studied the correlations between 11 antibodies relating to molecular subtypes or epithelial-to-mesenchymal transition (EMT) and prognosis in high-risk non-muscle-invasive (HGT1) bladder cancer. They found that both stromal tumor-infiltrating lymphocyte (sTIL) levels in noninvasive papillary urothelial carcinoma areas and increased expression of the luminal markers FOXA1 and SCUBE2 are significantly associated with better disease-free survival (DFS), but no EMT markers showed any trend. They suggested that molecular subtype markers, rather than EMT markers, might be preferable to study biomarkers of HGT1 urothelial carcinoma (Ottley et al., 2020). Rouanne et al. focused on stromal lymphocyte infiltration by evaluating the percentage of stromal area infiltrated by mononuclear inflammatory cells over the total intratumoral stromal area (Rouanne et al., 2019). Similarly, a high density of stromal TILs was associated with the tumor invasion depth in pT1 NMIBC, implying tumor aggressiveness was associated with an increased adaptive immune response, but no association between the level of TILs and survival outcome was observed. A clear clue has shown that the activated tumor immune microenvironment (TIME) could prevent NMIBC tumors from progressing. However, additional integration and refinement of these findings are required to provide a robust immuno-prognostic model for predicting progression in NMIBC patients.

In this study, we reported an integrated analysis using a total of 1370 transcriptome data of NMIBC patients from nine public datasets. Candidate tumor-infiltrating immune cells relating to the well-established prognostic risk factors and survival were filtered by a non-weighted voting system of six deconvolution methods and the survival analysis. Differentially expressed genes (DEGs) representing the candidate immune cells were identified. We used the selected DEGs as predefined signature genes in the single-sample gene set enrichment analysis (ssGSEA) to achieve unbiased quantification of the tumor-infiltrating immune cells. Finally, we developed a robust immune-prognostic model based on the immune cell matrix for evaluating the progression of NMIBC patients.

MATERIALS AND METHODS

Transcriptomic Profiles Analyzed

We searched for public datasets using combined keywords of "NMIBC", "expression profile", and "human" through GEO (Barrett et al., 2013), ArrayExpress (Athar et al., 2019), and PubMed® databases. Exclusion criteria of ineligible datasets were as follows: 1) datasets lacking cancer grade or TNM stage metadata; 2) datasets with only expression profiles of muscle-invasive bladder cancer (MIBC) samples; 3) datasets providing only processed data with negative expression values. Then we de-duplicated the same samples collected from multiple sources. Notably, our study allowed for the inclusion of datasets sequenced by RNA-Seq and microarray platforms. We also allowed sampling of tumors from both primary and recurrent lesions.

Deconvolution of Tumor-Infiltrating Immune Cells

We employed six in silico deconvolution methods to estimate cell composition in 1370 human transcriptome data. The xCell (Aran et al., 2017) performed an enrichment analysis of 64 immune and stromal cell types, illustrating whether a particular type of cell was present. The immunedeconv (Sturm et al., 2019), an integrated deconvolution tool, implemented the other four cell-type quantification algorithms, including quanTIseq (Finotello et al., 2019), TIMER (Li et al., 2016), MCPCounter (Becht et al., 2016), and EPIC (Racle et al., 2017). Moreover, ESTIMATE (Yoshihara et al., 2013) was used to estimate combined immune, stromal, and ESTIMATE scores without giving any single cell-type proportion. In summary, we assessed 64 tumor-infiltrating immune cell scores and six immune infiltration biomarker scores for each processed sample. Names of the cells and biomarkers with their corresponding alias in the six deconvolution methods are provided in **Supplementary Table S3**.

Correlations Between Clinicopathological Features and Immune Cells

To avoid methodological bias, we adopted an unweighted voting system to discover tumor-infiltrating immune cells significantly related to the well-established prognostic risk factors of NMIBC patients. In datasets providing age, sex, stage, grade, tumor size, European Organisation for Research and Treatment of Cancer (EORTC) risk score, and CIS in disease course status data, we compared the distribution of 64 tumor-infiltrating cell deconvolution scores across different levels of the risk factors. Student's t-test and box plots were performed by the "ggplot2" (Wickham, 2016) package of R language (R Core Team, 2021). A cell type in a specific dataset deconvoluted by a particular algorithm with a false discovery rate (FDR) adjusted *p*-value of student's t-test in more than two levels less than 0.05 was counted as one vote for the cell. All votes were categorized into 64 cell types to reveal the tumor immune microenvironment that would predict survival (**Supplementary Table S4**).

Identification of Differentially Expressed Genes of Candidate Immune Cells

The "limma" (Ritchie et al., 2015) package of R language (R Core Team, 2021) was used to identify differentially expressed genes (DEGs) of each candidate immune cell type. Log2-transformed fold changes (log2FC), *p*-values, and FDR adjusted *p*-values of every "source dataset—deconvolution method—immune cell—gene name" sets are provided in **Supplementary Table S5**. Only genes with absolute log2FCs larger than one and FDR *p*-values less than 0.05 were defined as DEGs for corresponding cell types. Furthermore, we defined candidate " cell-gene" combinations by the wFisher (Yoon et al., 2021) *p*-value in all evaluable sets, along with the number of datasets in which the combination was evaluable (**Supplementary Table S6**). The gene with a mean absolute log2FC larger than 0.2 for NK cells and 0.3 for other cells, a wFisher combined *p*-value less than 1.151e-6 (0.05/number of genes 43,440), and identified as significant DEGs in more than three databases were defined as representative gene of the immune cell. The "metapro" (Yoon et al., 2021) package in R (R Core Team, 2021) was used to calculate the combined wFisher *p* values.

Identification of Immune-Cell-Specific DEGs Related to Survival

Faced with dozens to hundreds of DEGs representing one immune cell type, we further narrowed the list by conducting survival analyses in the Kaplan-Meier curve and the forest plot to remove genes that contribute less to survival risk. Divided by the median of candidate genes' expression, we compared the PFS of E-MTAB-4321, DFS of GSE32894, and OS of GSE13507 in low expressed versus high expressed groups (results provided in **Supplementary Table S7**). The Kaplan-Meier curve was fitted by the "survfit" function and visualized by the "ggsurvplot" function. The forest plot was fitted by the "coxph" function and visualized by the "ggforest" function. DEGs with log-rank *p*-values of both analyses less than 0.05 and hazard ratios (HRs) of Cox's proportional hazards models larger than 2.5 or less than 0.5 were defined as the final biomarker genes of the candidate immune cells. All survival analyses were implemented by the "survival" package (Therneau and Grambsch, 2000; Therneau, 2021) and visualized by the "ggplot2" (Wickham, 2016, 2) package in R (R Core Team, 2021). The "ComplexHeatmap" (Gu et al., 2016) package in R (R Core Team, 2021) was used to generate expression heatmaps of the final gene list.

Gene Ontology and Pathway Enrichment of Candidate DEGs

We conducted Gene Ontology (GO) (Ashburner et al., 2000; Gene Ontology Consortium, 2021) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2021) pathway enrichment analyses of the selected immune-cell-specific DEGs by the "clusterProfiler" (Yu et al., 2012; Wu et al., 2021) package in R (R Core Team, 2021).

Calculation of ssGSEA and Z-Score Based Cell Enrichment Scores

Inspired by previous studies (Barbie et al., 2009; Motzer et al., 2020), we employed two methods to evaluate the nine candidate immune cells using gene lists generated by previous steps. The ssGSEA analysis (Subramanian et al., 2005) was performed on the logged expression matrix by the "GSVA" (Hänzelmann et al., 2013) package in R (R Core Team, 2021), and z-score statistics were performed on the non-logged expression matrix by in-house scripts.

Correlations Between Tumor-Infiltrating Immune Cell Score and Survival

Patients in each dataset were divided by the median of enriched immune cell scores into high and low immune infiltrated groups. Survival analyses and log-rank tests of PFS, DFS, and OS in high versus low immune cell infiltrated groups were conducted by the "survfit" function of the "survival" (Therneau and Grambsch, 2000; Therneau, 2021) package. Kaplan-Meier curves were visualized by the "ggsurvplot" function of the "ggplot2" (Wickham, 2016, 2) package in R (R Core Team, 2021). *p* values of both analyses and hazard ratios of high infiltrated groups are provided in **Supplementary Table S8**.

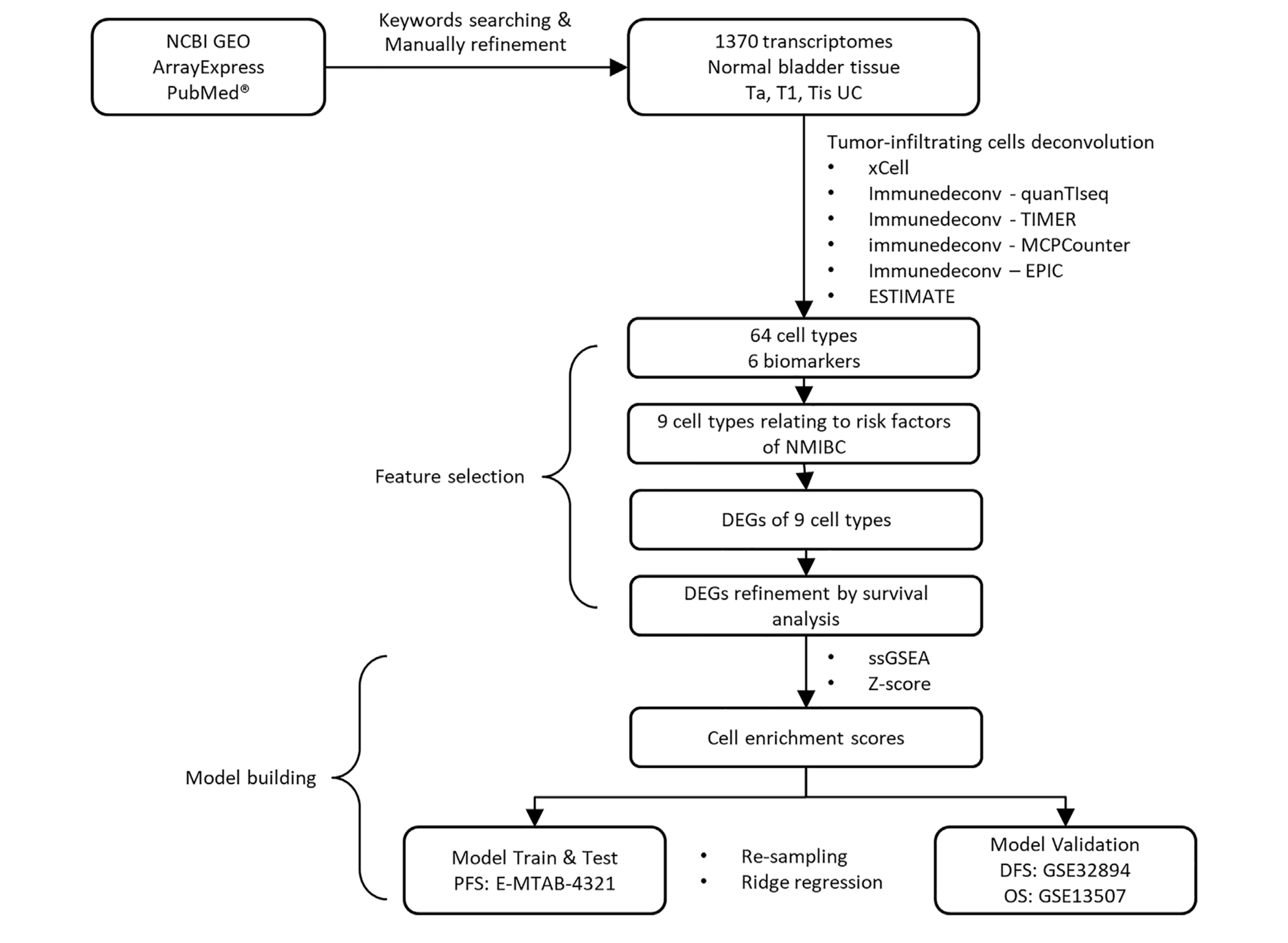


FIGURE 1 | Overview of study design.

Establishment of the Immuno-Prognostic Model

Using 454 samples from E-MTAB-4321 with evaluable PFS records, we randomly re-sampled 5000 times to build training and test sets in a 1:1 ratio. In each sampling scenario, we established a ridge regression model with an estimated enrichment score matrix of the nine tumor-infiltrating immune cells to predict the risk of progression. In each modeling process, tenfold cross-validation was used to select the optimal fitted model. The prediction performance of the models was evaluated by areas under curves (AUCs) of receiver operating characteristic curves (ROCs) in training and test sets. In R language (R Core Team, 2021), the "glmnet" (Friedman et al., 2010) package was used to build the models, and the "pROC" (Robin et al., 2011) package was used to visualize the results.

Statistical Analysis

p-Values less than 0.05 were considered significant in this study unless otherwise specified.

RESULTS

Summary of Datasets and Basic Workflow

The study design and workflow to develop our model are illustrated in **Figure 1**. After keyword searching and manual refinement, we brought nine datasets into this study, including 1370 human transcriptome profiles spanning normal bladder tissues, Ta, T1, and CIS urothelial cancers. Metadata of all the datasets and clinicopathological information of all the samples are provided in **Table 1**; **Supplementary Tables S1,S2**.

With the 1370 transcriptomic profiles, we initially screened nine candidate immune cells associated with the well-established NMIBC prognostic risk factors and then identified the differentially expressed genes (DEGs) representing these cells by significance and differentiation. Using the DEGs' expression matrix, we estimated the proportions of tumor infiltrated immune cells by the gene set enrichment analysis. Using the estimated immune cell score matrix, we established the immune-prognostic model by repeated random sampling, ridge regression modeling, and optimal cutoff confirming.

TABLE 1 | Demographic and disease characteristics of the 1,370 samples included in this study. Data are median (total number of assessable samples; range; IQR) or n (%). IQR: interquartile range. PFS: progression-free survival. DFS: disease-free survival. OS: overall survival.

Characteristics	Value
Age (years)	69 (862; 20–96; 61–76.5)
Age category (years)	—
20–60	198 (14%)
61–80	542 (40%)
> =80	122 (9%)
Not available	508 (37%)
Sex	—
Male	797 (58%)
Female	219 (16%)
Not available	354 (26%)
Tumor Stage	—
T0	91 (6%)
Ta	696 (51%)
Ta-T1	24 (2%)
T1	547 (40%)
CIS/Tis	12 (1%)
WHO 1973 Grade	—
G1	58 (4%)
G2	199 (15%)
G3	285 (21%)
G0/Gx/Not available	828 (60%)
WHO 2004–2016 Grade	—
Low	427 (31%)
High	289 (21%)
Not available	654 (48%)
CIS in the disease course	—
CIS-	472 (34%)
CIS+	103 (8%)
Not available	795 (58%)
Tumor size	—
<=3 cm	311 (23%)
>3 cm	83 (6%)
Not available	976 (71%)
EORTC risk score	—
0	286 (21%)
1	174 (13%)
Not available	910 (66%)
Recurrence	—
FALSE	127 (9%)
TRUE	57 (4%)
Not available	1186 (87%)
Progression beyond the T2 stage	—
FALSE	711 (52%)
TRUE	66 (5%)
Not available	593 (43%)
PFS (months)	33 (460; 0–74.9; 24–42.8)
Cancer-specific survival	—
FALSE	271 (20%)
TRUE	6 (~0%)
Not available	1093 (80%)
DFS (months)	37.9 (173; 0.2–104.4; 21.2–60.2)
Vital status	—
FALSE	144 (11%)
TRUE	42 (3%)
Not available	1184 (86%)
OS (months)	55.3 (104; 2.1–137; 26.4–80.3)

Tumor-Infiltrating Immune Cells Related to Key NMIBC Prognostic Factors

Several risk factors have been proven to be significantly related to the prognosis of NMIBC patients (Liu et al., 2015; Douglas et al., 2021). Tumor size greater than 3 cm, multifocal lesions, concurrent CIS, more advanced cancer stage, higher histological grade, higher EORTC risk score, and higher frequency of prior recurrences were known risks implying higher rates of recurrence or progression. We first conducted a comparative analysis between these risk factors and 64 deconvoluted tumor-infiltrating cell types in each dataset, then employed an unweighted voting schema to identify top cell types that might contribute to NMIBC prognosis. As shown in **Figure 2A**, the top voted and most significant tumor-infiltrating cells included cancer-associated fibroblasts (CAFs), B cells, CD4$^+$ T cells, CD8$^+$ T cells, natural killer (NK) cells, dendritic cells (DCs), macrophages, neutrophils, and endothelial cells. Since xCell is typically used to determine the presence or absence of a specific cell type, rather than to calculate the cell proportion, we only used the sum of votes from the other five methods to filter the most relevant cell types (**Supplementary Table S4**). CD4$^+$ T cells ranked first, being voted in five, three, and six of nine eligible datasets by TIMER (Li et al., 2016), quanTIseq (Finotello et al., 2019), and EPIC (Racle et al., 2017), respectively. Followed by CD4$^+$ T cells, B cells, and CAFs.

Biomarker Genes Representing the Candidate Tumor-Infiltrating Immune Cells

After targeting candidate tumor-infiltrating cells, we wished to ascertain a set of biomarker genes that were representative of the cells and that were also strongly associated with the survival of NMIBC patients. In identifying differentially expressed genes (DEGs) of the nine candidate immune cells, a total of 2757 “cell-DEG” pairs were recognized as repetitive patterns and included in the following analysis (**Supplementary Table S6**). We then analyzed all 972 nonredundant genes in the 2757 “cell-DEG” pairs with forest plot and Kaplan-Meier (KM) curve survival analyses against PFS in E-MTAB-4321, DFS in GSE32894, and OS in GSE13507 (**Figure 2B**). After this, we narrowed the list to 149 unique genes as protective or risk factors of PFS or OS in NMIBC patients. These genes with the cells they represented comprised 368 unique “cell-DEG” pairs (**Table 2**), of which 254 pairs were associated with PFS and 114 pairs with OS (**Supplementary Table S7**). DCs and CAFs were the top two cell types, with more than sixty percent (92/149, 91/149) of the biomarker genes associated with them (**Table 2**).

The expression of 110 PFS-related and 41 OS-related biomarker DEGs was visualized in **Figures 3A,B**. All 99 biomarker DEGs of nine candidate tumor-infiltrating immune cells were subjected to KEGG pathway, GO-biological process

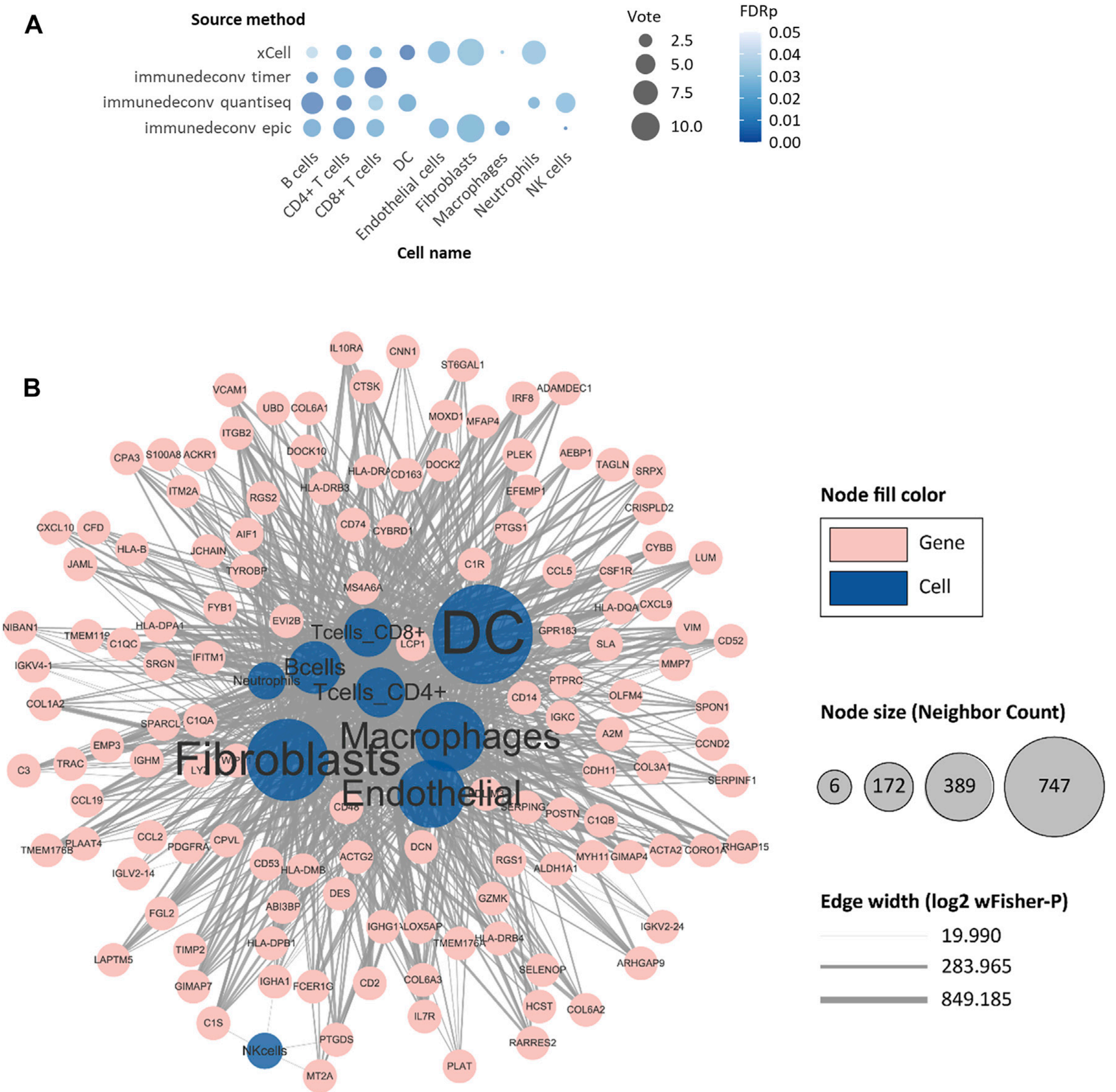


FIGURE 2 | Identification of progression-risk-related tumor-infiltrating cells and differentially expressed genes representing them. **(A)** The non-weighted voting results of Student's t-tests between tumor-infiltrating cells and well-established clinical progression risk factors. Tumor-infiltrating cell scores evaluated by six immune deconvolution methods were used. Only significant results were counted as valid votes shown in the figure. **(B)** The network of differentially expressed genes (DEGs) with their representing tumor-infiltrating cells. The blue circles refer to cell types. The pink circles refer to selected DEGs. The size of blue circles indicates the number of DEGs. The thickness of lines indicates the negative log2 of wFisher combined *p*-value of differential expression testing. Only nodes with more than six adjacent neighbors are shown.

(BP), GO-cellular component (CC), and GO-molecular function (MF) terms enrichment analyses (**Figures 3C–F**). As expected, we found strong evidence pointing to the crosstalk between tumor cells and adjacent normal epithelium, represented by focal adhesion and extracellular matrix (ECM)-receptor interaction. Aberration of these pathways would directly affect the steadiness of tumor cells and thereby cause progression. We also found enriched cell proliferation signals like protein digestion and absorption and the PI3K-Akt signaling pathway. They acted either as energy suppliers or as signal transduction factors to trigger or facilitate the cascade of invasive tumor progression. The chemokine signaling pathway, on the other hand, would help to recruit leukocytes to the site of the inflammation area.

Enrichment of Tumor-Infiltrating Immune Cell Scores

Since the datasets included in our study differed in their transcriptome profiling technologies, we cautiously practiced the enrichment analyses with the logarithmic matrix of original expression data. 43,440 transcripts in 1,370 samples with and without log2-transformation were used to proceed

TABLE 2 | List of biomarker genes representing the nine tumor-infiltrating candidate immune cells.

Bcells	DC	Endothelial	Fibroblasts	Macrophages	Neutrophils	NK cells	T cells_CD4+	T cells_CD8+
CD74	ADTRP	ADCY4	AKAP12	AP1S2	CD74	ANXA10	CASP1	CASP1
COL3A1	AP1S2	AP1S2	ANXA10	BTBD16	IGKV1-17	BTBD16	CD74	CD74
CXCL13	APOL3	BGN	AP1S2	C12orf75	MMP7	CLCA4	CFH	CXCL13
DES	ATF3	CD74	BGN	CAT	RARRES1	CRTAC1	COL1A1	DES
GIMAP7	ATP8B4	CLEC14A	BMP5	CD74	S100A8	ENTPD3	COL3A1	ENPP2
HCLS1	BMP5	CLIC4	BTBD16	CFH		FABP4	GIMAP7	FCER1A
IGHV1-69	CASP1	CLIP3	CCL11	CLIC4		FGFR3	GMFG	GDF15
IGKV1-17	CCL18	COL18A1	CD74	CNN3		RAB4A	IGKV1-17	GIMAP7
MMP7	CCL8	COL18A1	CLIC4	COL1A1		TMPRSS4	MMP7	IGKV1-17
POSTN	CD3G	COL1A1	CLIP3	COL3A1		TP63	MXRA5	SELENOP
RAC2	CD4	COL3A1	COL18A1	COL5A2			POSTN	SPINK1
RARRES1	CD74	COL4A1	COL18A1	CPQ			RAC2	SYNM
S100A8	CFH	COL4A2	COL1A1	CTSE			S100A8	TCF21
SELENOP	CLIC4	COL5A2	COL3A1	DEGS1			TRIM22	TRIM22
SERPINE2	CLIP3	COL8A1	COL4A1	DES			VCAN	
TRIM22	COL1A1	CRTAC1	COL4A2	DKK3			XAF1	
	COL3A1	CYGB	COL5A2	DOCK11				
	COL5A2	DEGS1	COL8A1	DSE				
	COL8A1	DES	CRTAC1	ELOVL5				
	CSF2RB	DKK3	CTSE	ENPP2				
	CSRP1	EDNRA	CXCL13	FBLN1				
	CXCL11	ENPP2	CYGB	FCER1A				
	CXCL13	FBLN1	DEGS1	FERMT2				
	DEGS1	FBN1	DES	FILIP1L				
	DES	FERMT2	DKK3	FSTL1				
	DKK3	FILIP1L	DOCK11	GIMAP7				
	DOCK11	FN1	DSE	GLT8D2				
	DSE	FSTL1	EDNRA	HCLS1				
	EDNRA	GEM	EFHD1	LITAF				
	ENPP2	GIMAP7	FABP6	LRIG1				
	FBN1	GLT8D2	FAM174B	MMD				
	FCER1A	GUCY1A1	FAM3B	MMP7				
	FERMT2	HCLS1	FBLN1	MXRA5				
	FGD2	ITGA1	FBN1	NUPR1				
	FGR	LAMA4	FCER1A	PLSCR4				
	FILIP1L	LRRC32	FERMT2	PODN				
	FN1	MFNG	FILIP1L	POSTN				
	FPR1	NEURL1B	FN1	PRDX3				
	FSTL1	NID1	FSTL1	RARRES1				
	GEM	NID2	GEM	RGS5				
	GIMAP7	NREP	GIMAP7	RPL17				
	GLT8D2	OLFML1	GLT8D2	S1PR3				
	GMFG	OLFML2A	GPX8	SELENOP				
	GPX8	PCDH17	GUCY1A1	SERPINE2				
	GUCY1A1	PDGFRB	HCLS1	SGCE				
	HCLS1	PLAC9	HOXB6	SH3BGRL				
	HLA-DQB2	PODN	IGFBP6	SLC9A9				
	HLA-E	POSTN	ITGA1	STEAP1				
	IGFBP6	PRRX1	LAMA4	SULF1				
	IGHV1-69	RBPMS2	LRIG1	SYNM				
	IGKV1-17	RGS5	LRRC32	TCF21				
	INPP5D	S100A8	MMD	TM4SF1				
	LAMA4	S1PR3	MMP7	TM4SF1				
	LITAF	SELENOP	MRVI1	TMED7				
	LRIG1	SERPINE2	MXRA5	TMEM45A				
	LRRC32	SGCE	NEURL1B	TNC				
	MAF	SULF1	NID1	TRIM22				
	MFNG	SYNM	NID2	TSPAN7				
	MMD	TCF21	NREP	VCAN				
	MMP7	TM4SF1	NUPR1	WDR72				
	MXRA5	TM4SF1	OLFML1					
	NEK6	TNC	OLFML2A					
	NUPR1	TSPAN7	PDGFRB					
	NXN	VCAN	PLAC9					

(Continued on following page)

TABLE 2 | (*Continued*) List of biomarker genes representing the nine tumor-infiltrating candidate immune cells.

Bcells	DC	Endothelial	Fibroblasts	Macrophages	Neutrophils	NK cells	T cells_CD4+	T cells_CD8+
	OLFML1		PLN					
	PDGFRB		PLSCR4					
	PLSCR4		PODN					
	PLXDC2		POSTN					
	PODN		PRRX1					
	POSTN		RAC2					
	PRRX1		RBPMS2					
	RAC2		RGS5					
	S100A8		S100A8					
	SELENOP		S1PR3					
	SERPINA3		SELENOP					
	SERPINB9		SERPINA3					
	SERPINE2		SERPINE2					
	SGCE		SGCE					
	SP110		SMTN					
	SULF1		SULF1					
	SYNM		SYNM					
	TCF21		TCF21					
	TM4SF1		TEAD2					
	TM4SF1		TM4SF1					
	TMEM45A		TM4SF1					
	TNC		TMEM45A					
	TRIM22		TNC					
	TSPAN7		TPST1					
	VCAN		TSPAN7					
	XAF1		VCAN					
	ZFP36		VSIG2					
	ZG16B							

with ssGSEA and z-score-based immune cell enrichment analyses. With the biomarker DEGs listed in **Table 2** as priori-defined sets of immune cell-specific genes, we quantified the infiltration of all nine tumor-infiltrating immune cells in the tumor microenvironment. Enrichment of the cell scores by ssGSEA in all 1370 NMIBC transcriptomes is shown in **Figure 4A**.

To assess the nine immune cells' ability to distinguish NMIBC patients with poor prognosis, we explored correlations between PFS, DFS, and OS with every tumor-infiltrating immune cell score calculated by ssGSEA and z-score methods. The survival analysis (**Supplementary Table S8**) showed that B cells, DCs, endothelial cells, CAFs, $CD4^+$ T cells, and $CD8^+$ T cells enriched by the ssGSEA method were significantly related to PFS in E-MTAB-4321 (**Figure 4B**). Macrophages and $CD8^+$ T-cells enriched by the ssGSEA method were significantly related to OS in GSE13507 (Plots not shown). No cell types were significantly related to DFS in GSE32894.

Robust Immuno-Prognostic Model

To achieve a robust prognostic model independent of the heterogeneous clinical information in eligible datasets, we used the score matrix of all nine candidate immune cells to build our model, although only some subsets of the cells were significantly related to PFS or OS. Since the primary goal of this study was to predict prognosis and risk of progression by key immune features, a total of 454 NMIBC samples from E-MTAB-4321 with assessable progression beyond T2 staging and PFS records were used. With the data, we repeatedly built training and test sets by randomly sampling 5000 times with a 1:1 ratio, fitted immune-prognostic models with the ridge regression, determined the optimal model with the minimum lambda, and evaluated the models with AUCs of ROC curves. Although immune cell enrichment score matrices calculated by both ssGSEA and z-score methods were used in building the immuno-prognostic model, only models built by ssGSEA matrices showed generally higher AUCs (data not shown). The formula of the final model was as follows:

Immuno-Prognostic score = - 0.4111588 + 2.5025813 * Bcells_score - 1.8274560 * DC_score + 6.7589250 * Endothelial_score + 2.6983895 * Fibroblasts_score - 0.1725197 * Macrophages_score + 1.0256969 * Neutrophils_score - 1.8221146 * NKcells_score - 6.0485265 * Tcells_CD4+_score—9.4937697 * Tcells_CD8+_score.

We visualized the prediction effect of the optimal model in **Figure 5A**, the AUCs were 0.827, 0.888, and 0.947 in the training set ($n = 228$), test set with all the other samples ($n = 226$), and test set with balanced progression and non-progression patients ($n = 30$), respectively. The sampling groups of our optimal model are recorded in the last three columns in **Supplementary Table S2**. The optimal cutoff of the Immuno-Prognostic score dividing low-risk and high-risk patients was 0.109. In **Figures 5A,B** conspicuous differentiation of PFS ($p < 0.0001$, log-rank test) was observed in patients with different predicted outcomes. We also expanded our validation of the model in predicting other types of clinical outcomes. The same trend has been observed, but it showed less significance in predicting DFS ($p = 0.21$, log-rank

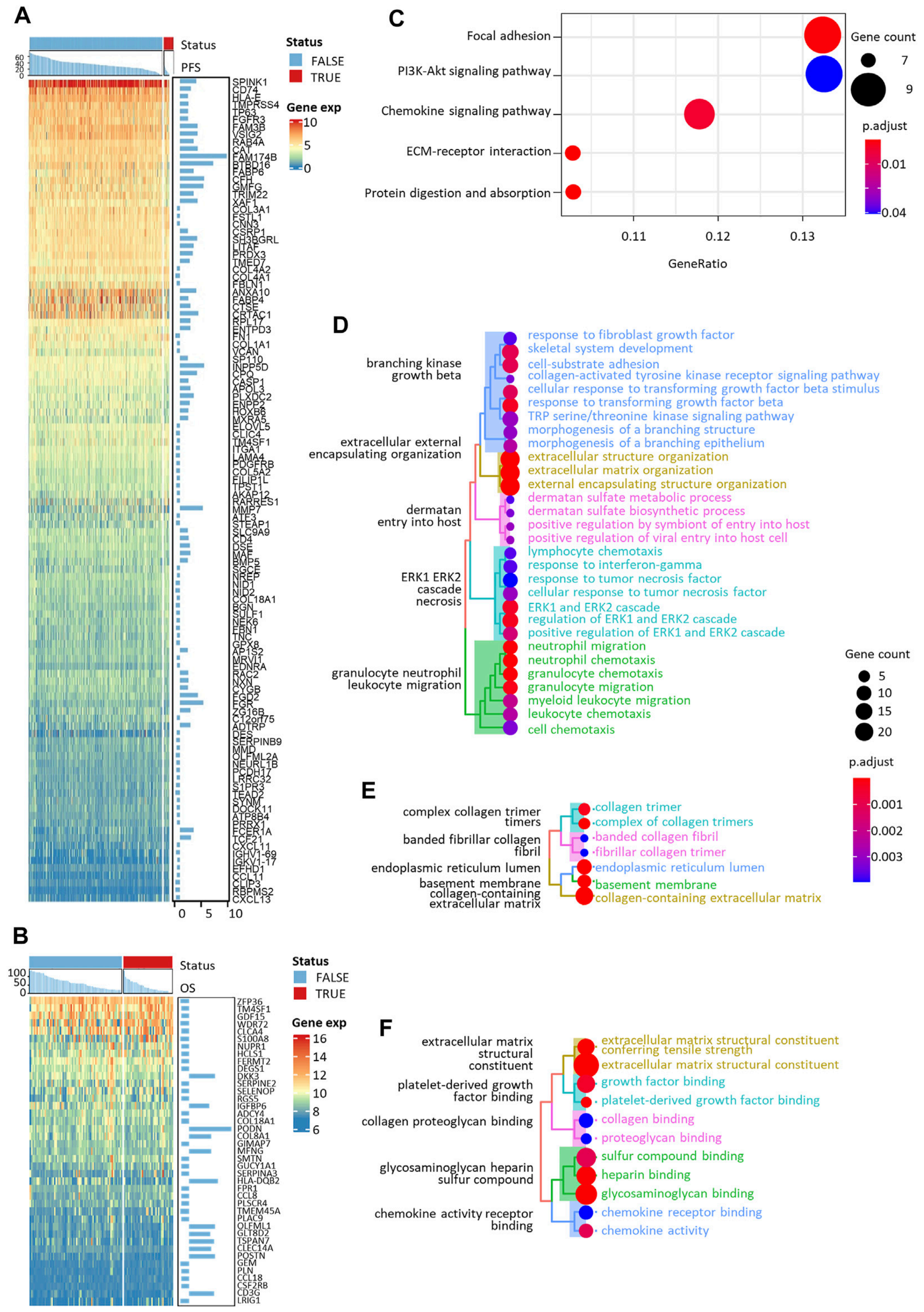


FIGURE 3 | Expression heatmaps and functional enrichment analyses of PFS- and OS-related immune cell-specific DEGs. Expression heatmaps of **(A)** PFS-related and **(B)** OS-related DEGs. KEGG **(C)**, GO-biological process **(D)**, GO-cellular component **(E)**, and GO-molecular function **(F)** enrichment of all the selected DEGs.

test) and OS (p = 0.027, log-rank test). Furthermore, to test the correlation between our model and the well-established survival risk factors of NMIBC, we compared distributions of the predicted immuno-prognostic scores against different levels of CIS in the disease course, EORTC risk score, WHO 1973 or 2004/2016 grade, recurrence, sex, tumor stage, and tumor size. All

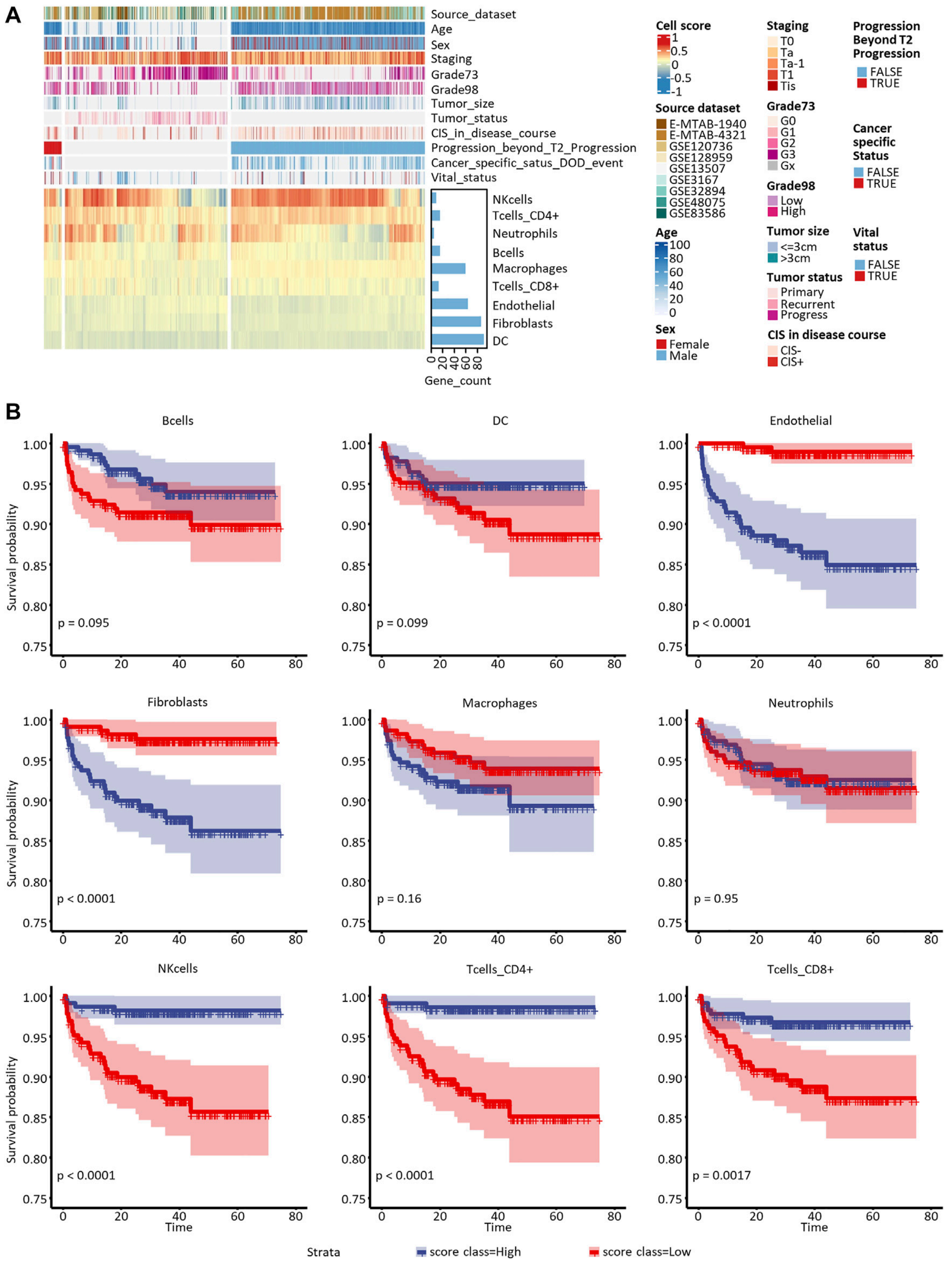


FIGURE 4 | Proportion assessment and prognostic value of the nine candidate tumor-infiltrating cells. **(A)** Heatmap of candidate cells and clinical features of all the eligible 1,370 samples included in this study. Grade73 and Grade98 refer to the WHO 1973 and WHO 2004/2016 Classification Systems for Urothelial Carcinoma, respectively. **(B)** Kaplan-Meier curves of univariate Cox regression in low- versus high-infiltrated groups divided by the nine candidate immune or stromal cells.

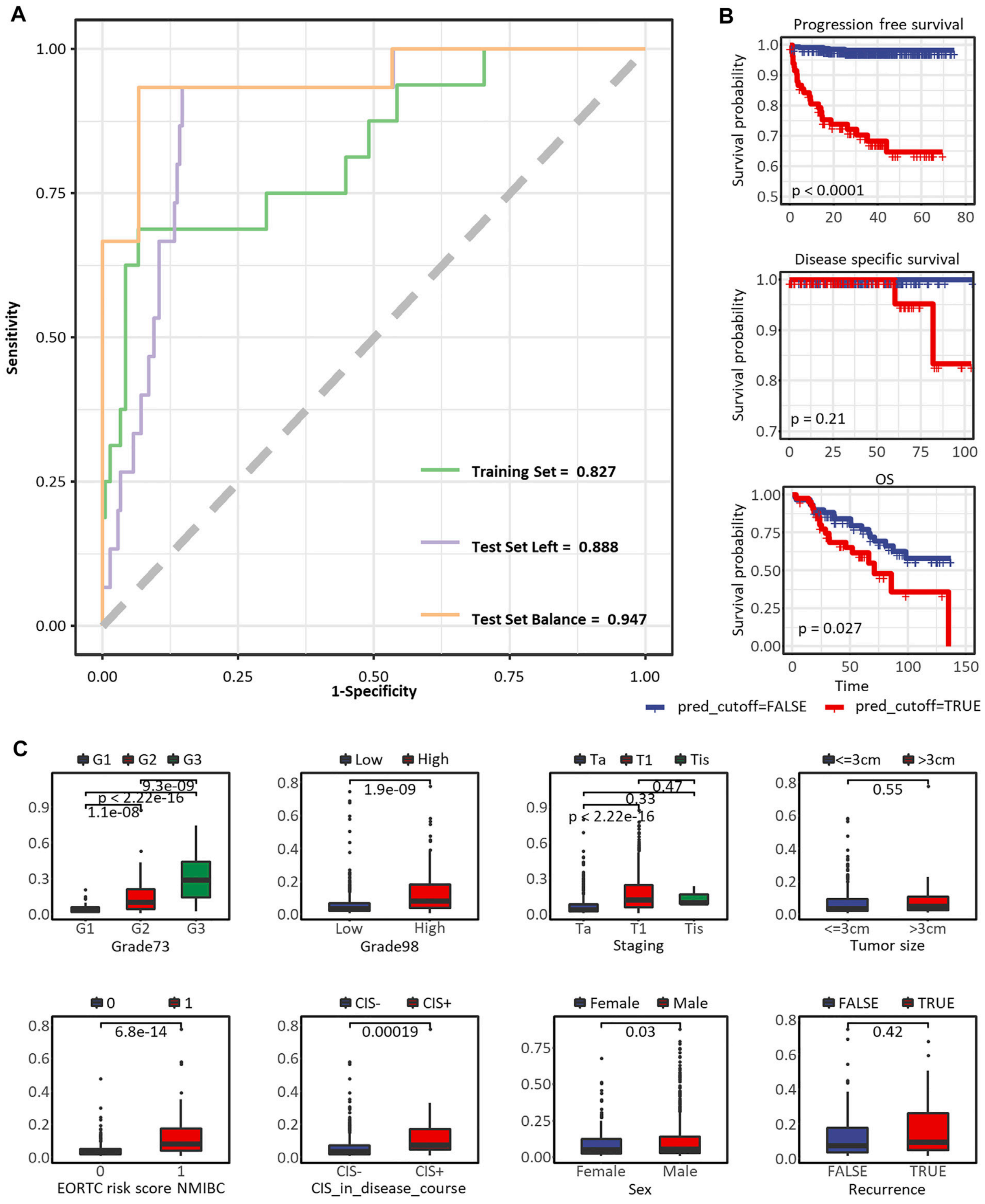


FIGURE 5 | Predictive performance of the immuno-prognostic model. **(A)** The ROC curve to predict PFS in the training set, test set with all the other samples, and test set with balanced progressed and non-progressed samples. **(B)** The Kaplan-Meier curve to predict PFS, DFS, and OS*. **(C)** Box plots comparing risk scores assessed by the immuno-prognostic model in different groups of clinical prognostic risk factors. * In the nine eligible datasets, PFS status was assessed in E-MTAB-4321, GSE13507, and GSE32894, while only E-MTAB-4321 provided survival time. DFS status was assessed in GSE32894, GSE13507, and GSE48075, while only GSE32894 provided survival time. OS status was assessed in GSE13507 and E-MTAB-1940, while only GSE13507 provided survival time. As we plotted here, the survival analyses were only applicable to datasets E-MTAB-4321, GSE32894, and GSE13507.

comparisons showed higher immuno-prognostic scores in higher risk levels, but the trends were insignificant in recurrence status and tumor size. In summary, our model could predict the risk to the progression of NMIBC patients by evaluating the tumor-infiltrating microenvironment. The immuno-prognostic score well reflected the degree of progression risk.

DISCUSSION

With the assumption that cancer progression was associated with immune cell infiltrating, we performed an integrated analysis for developing a robust immuno-prognostic model to evaluate progression risk in NMIBC patients. We identified nine critical tumor-infiltrating cell types: innate immune cells including macrophages, neutrophils, DCs, and NK cells; adaptive immune cells including B cells, CD4$^+$ T cells, and CD8$^+$ T cells; and sentinel cells including CAFs and endothelial cells. The quantification of these immune cells was conducted by ssGSEA using the DEGs recognized from all eligible datasets. Univariate Cox regression supported that some cells could independently distinguish patients with high progression risk. Based on this, we achieved a more robust model using the enrichment matrix of all the nine tumor-infiltrating immune cells and then validated its performance in predicting different types of survival. The predicted risk scores and survival status showed a high correlation with the actual clinical outcomes; however, considering the precision and significance, we suggested using our model in predicting the PFS of NMIBC patients instead of DFS or OS.

We included nine immune cells in our model, even though some showed no independent prognostic value, since we thought their combination would better reflect the coordinated interaction between innate and adaptive immune systems in preventing the normal tissue from aggressive progression. For one thing, many genes were identified as the DEGs for more than one type of immune cells (**Figure 2**; **Table 2**); for another, the functional enrichment analysis of the full set of signature DEGs showed strong evidence of underlying drivers of tumor progression. The collagen family genes, for instance, were independently related to the survival of NMIBC and were simultaneously recognized as the DEG of tumor-infiltrating B cells, CD4$^+$ T cells, CD8$^+$ T cells, DCs, CAFs, macrophages, and endothelial cells. Xu and colleagues reviewed the mechanisms underlying this result (Xu et al., 2019). The complex reticular structure composed of collagen-rich extracellular matrices (ECM) and multiple stromal cells formed dense stromal fibrosis and thereby induced focal hypoxia, leading to increased tumor proliferation and compromised immunotherapy effectiveness (Daniel et al., 2019). The enriched KEGG pathways, including focal adhesion and ECM-receptor interaction (**Figure 3C**), were consistent with the previous description. The extensive interaction between stromal/immune cells and cancer cancers depicted the complexity of the tumor microenvironment, which was why we used cells instead of genes to build our model.

Another detail of our study was that we emphasized the selection of appropriate meta-analysis methods in the feature selection step and the careful use of renormalization methods. Toro-Domínguez and colleagues reviewed the three main types of meta-analysis strategies based on effect sizes, *p*-values combination, and rank combination (Toro-Domínguez et al., 2021). We chose wFisher (Yoon et al., 2021), a modified *p*-value combination method, to filter the DEGs representing candidate immune cells. The wFisher method was suitable for studies from different platforms or conditions. In our case, combining the analysis of nine transcriptomic datasets sequenced by both RNA-Seq and microarray platforms fit the method's usage characteristics. The method also allowed combining results from heterogeneous analyses without rigorous renormalization. This feature elicited the second focus of our discussion: the renormalization of integrated transcriptomic data. Normalization of bulk RNA data included quantifying transcripts and standardizing data from different sources. The former was thoroughly discussed in the review of RNA sequencing technology (Stark et al., 2019). Here we mainly discussed the latter scenario, as the complexity of cancer biology required integrative studies with combined data from different researches. Shen and Wulff published their evaluations of various normalization methods for integrating large-scale metabolomics data, yielding the same conclusion that choosing the proper normalization method according to the data scale and downstream analysis would vastly improve the confidence of research results (Shen et al., 2016; Wulff and Mitchell, 2018). For transcriptome data, most studies still focused on the transcripts quantification question in the single-source dataset (Dillies et al., 2013; Li et al., 2015), while some of them also evaluated sophisticated frameworks and proposed a protocol to deal with raw RNA-Sequencing (RNA-Seq) data (Sahraeian et al., 2017). We found that few discussion has been made on the systematic renormalization of transcript data from multiple sources by multiple sequencing technologies, but some attempts were separately made and recommended in previous studies (Mooney et al., 2013; Risso et al., 2014; Ayers et al., 2017; Danaher, 2018; Liu et al., 2019). After modeling with both renormalized and non-normalized data (results shown in our Github or Gitee repositories listed in the Data Availability Statement section), we believed the renormalization method combining RNA-Seq and microarray data was still not well-established. We built our model for predicting PFS in NMIBC patients based on RNA-Seq data alone. We suggested that any further applications of our model should consider using RNA-Seq data rather than microarrays.

In conclusion, we identified nine critical tumor-infiltrating immune cells, quantified the cells' proportion in the tumor immune microenvironment with self-defined signature genes, and established a robust immune-prognostic model for predicting the progression of NMIBC patients. Our study showed system-wide coordination of the immune and stromal cells in defending aberrant cell proliferation and aggressive tumor growth and invasion. Thus, modeling strategies regarding the tumor microenvironment as a whole system may be optimal in clinical decision support applications, which we believe is why multi-omics and integrative studies were replacing single biomarker and single dimension studies. In previous studies, single dimension data, such as the density of stromal TILs evaluated by H&E-stained slides, failed to predict survival outcomes independently (Rouanne et al., 2019; Ottley et al., 2020). Rouanne and colleagues only proved that the stromal TILs were associated with the tumor invasion depth in pT1 NMIBCs. Ottley and colleagues combined the sTILs levels with IHC and ISH biomarkers to improve the prognostic potential. In this shift to complex modeling with multiple dimension data, we raised the importance of appropriate data preprocessing procedures, including but not limited to the selection of appropriate meta-analysis methods. Moreover, some limitations of our research had to be mentioned here. With the inspiration from the UROMOL2021 study (Lindskrog et al., 2021), we initiated our investigation with the hypothesis that dynamic interactions in tumor immune microenvironment would reflect not only the progression risk but also the response to local treatment like intravesical instillation of chemotherapeutic or immunotherapeutic agents. Several efficient predictive biomarkers have been developed and widely evaluated in pan-cancer scenarios, such as the 18-gene gene expression profile (GEP) score (Ayers et al., 2017) has a high discriminatory value in predicting the response to pembrolizumab in Keynote-001, Keynote-012, and Keynote-028. Unfortunately, we did our research and failed to get enough high-quality response data to therapies in NMIBC patients. In the current study, we validated only the prognostic value of our model. Nevertheless, we wish to expand its usage in prognostic and predictive conditions in the future.

DATA AVAILABILITY STATEMENT

Transcriptomic data of all datasets used in this study were available in public databases at the following URLs: E-MTAB-1940, microarray, ArrayExpress: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1940/; E-MTAB-4321, RNA-Seq, ArrayExpress: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4321/; GSE12073, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12073; GSE128959, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128959; GSE13507, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13507; GSE3167, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3167; GSE32894, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32894; GSE48075, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48075; GSE83586, microarray, GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83586. All codes to reproduce the results and figures in this article and point-by-point responses were published on GitHub (https://github.com/XiaomengSun315/NMIBC_immuno-prognostic) repository.

AUTHOR CONTRIBUTIONS

Study concept and design: LL, ML, and JX; methodology, analysis, and interpretation of data: XS, HX, GL, JC, and ML; drafting of the manuscript: XS and ML; critical revision of the manuscript for important intellectual content: ML; statistical analysis: XS, HX, GL, and JC. All authors read and approved the final manuscript.

FUNDING

This study is supported by the National Key Research and Development Program of China (Grant No. 2021YFC2500403), the S&T Program of Hebei (Grant No. 21377734D), and the Shanghai Committee of Science and Technology, China (Grant No. 18511102704) to LL.

ACKNOWLEDGMENTS

The authors thank the reviewers for their insightful comments, which have greatly helped improve this article's scientific rigor.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.833989/full#supplementary-material

REFERENCES

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18, 220. doi:10.1186/s13059-017-1349-1

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress Update - from Bulk to Single-Cell Expression Data. *Nucleic Acids Res.* 47, D711–D715. doi:10.1093/nar/gky964

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). IFN-γ-related mRNA Profile Predicts Clinical Response to PD-1 Blockade. *J. Clin. Investigation* 127, 2930–2940. doi:10.1172/JCI91190

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

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for Functional Genomics Data Sets--Update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., et al. (2016). Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression. *Genome Biol.* 17, 218. doi:10.1186/s13059-016-1070-5

Chang, S. S., Boorjian, S. A., Chou, R., Clark, P. E., Daneshmand, S., Konety, B. R., et al. (2016). Diagnosis and Treatment of Non-muscle Invasive Bladder Cancer: AUA/SUO Guideline. *J. Urology* 196, 1021–1029. doi:10.1016/j.juro.2016.06.049

Chang, S. S., Boorjian, S. A., and Chou, R. (2020). Diagnosis and Treatment of Non-muscle Invasive Bladder Cancer: AUA/SUO Joint Guideline. *Review* 196 (4), 1021–1029. Available at: https://www.auanet.org/guidelines/guidelines/bladder-cancer-non-muscle-invasive-guideline#x2563.

Cookson, M. S., Herr, H. W., Zhang, Z.-F., Soloway, S., Sogani, P. C., and Fair, W. R. (1997). The Treated Natural History of High Risk Superficial Bladder Cancer: 15-year Outcome. *J. Urology* 158, 62–67. doi:10.1097/00005392-199707000-00017

Danaher, P. (2018). Pan-cancer Adaptive Immune Resistance as Defined by the Tumor Inflammation Signature (TIS): Results from the Cancer Genome Atlas (TCGA). *J. Immunother. Cancer* 17, 63. doi:10.1186/s40425-018-0367-1

Daniel, S. K., Sullivan, K. M., Labadie, K. P., and Pillarisetty, V. G. (2019). Hypoxia as a Barrier to Immunotherapy in Pancreatic Adenocarcinoma. *Clin. Transl. Med.* 8, 10. doi:10.1186/s40169-019-0226-9

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis. *Briefings Bioinforma.* 14, 671–683. doi:10.1093/bib/bbs046

Douglas, J., Struss, W., and Williams, S. (2021). "Risk Stratification of Patients: Risk Tables and Assessment - NMIBC and MIBC," in *Bladder Cancer: A Practical Guide*. Editors A. M. Kamat and P. C. Black (Cham: Springer International Publishing), 41–52. doi:10.1007/978-3-030-70646-3_5

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., et al. (2019). Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data. *Genome Med.* 11, 34. doi:10.1186/s13073-019-0638-6

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

Gene Ontology Consortium (2021). The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Res.* 49, D325–D334. doi:10.1093/nar/gkaa1113

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data. *Bioinformatics* 32, 2847–2849. doi:10.1093/bioinformatics/btw313

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

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. doi:10.1093/nar/gkaa970

Lenis, A. T., Lec, P. M., Chamie, K., and Mshs, M. (2020). Bladder Cancer: A Review. *JAMA* 324, 1980–1991. doi:10.1001/jama.2020.17598

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, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the Normalization Methods for the Differential Analysis of Illumina High-Throughput RNA-Seq Data. *BMC Bioinforma.* 16, 347. doi:10.1186/s12859-015-0778-7

Lindskrog, S. V., Prip, F., Lamy, P., Taber, A., Groeneveld, C. S., Birkenkamp-Demtröder, K., et al. (2021). An Integrated Multi-Omics Analysis Identifies Prognostic Molecular Subtypes of Non-muscle-invasive Bladder Cancer. *Nat. Commun.* 12, 2301. doi:10.1038/s41467-021-22465-w

Liu, S., Hou, J., Zhang, H., Wu, Y., Hu, M., Zhang, L., et al. (2015). The Evaluation of the Risk Factors for Non-muscle Invasive Bladder Cancer (NMIBC) Recurrence after Transurethral Resection (TURBt) in Chinese Population. *PLOS ONE* 10, e0123617. doi:10.1371/journal.pone.0123617

Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., et al. (2019). Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review. *Front. Bioeng. Biotechnol.* 7, 358. doi:10.3389/fbioe.2019.00358

Mooney, M., Bond, J., Monks, N., Eugster, E., Cherba, D., Berlinski, P., et al. (2013). Comparative RNA-Seq and Microarray Analysis of Gene Expression Changes in B-Cell Lymphomas of *Canis familiaris*. *PLoS One* 8, e61088. doi:10.1371/journal.pone.0061088

Motzer, R. J., Banchereau, R., Hamidi, H., Powles, T., McDermott, D., Atkins, M. B., et al. (2020). Molecular Subsets in Renal Cancer Determine Outcome to Checkpoint and Angiogenesis Blockade. *Cancer Cell* 38, 803–817. doi:10.1016/j.ccell.2020.10.011

Ottley, E. C., Pell, R., Brazier, B., Hollidge, J., Kartsonaki, C., Browning, L., et al. (2020). Greater Utility of Molecular Subtype rather Than Epithelial-to-mesenchymal Transition (EMT) Markers for Prognosis in High-risk Non-muscle-invasive (HGT1) Bladder Cancer. *J. Pathol. Clin. Res.* 6, 238–251. doi:10.1002/cjp2.167

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data. *eLife* 6, e26476. doi:10.7554/eLife.26476

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples. *Nat. Biotechnol.* 32, 896–902. doi:10.1038/nbt.2931

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinforma.* 12, 77. doi:10.1186/1471-2105-12-77

Rouanne, M., Betari, R., Radulescu, C., Goubar, A., Signolle, N., Neuzillet, Y., et al. (2019). Stromal Lymphocyte Infiltration Is Associated with Tumour Invasion Depth but Is Not Prognostic in High-Grade T1 Bladder Cancer. *Eur. J. Cancer* 108, 111–119. doi:10.1016/j.ejca.2018.12.010

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., et al. (2017). Gaining Comprehensive Biological Insight into the Transcriptome by Performing a Broad-Spectrum RNA-Seq Analysis. *Nat. Commun.* 8, 59. doi:10.1038/s41467-017-00050-4

Shen, X., Gong, X., Cai, Y., Guo, Y., Tu, J., Li, H., et al. (2016). Normalization and Integration of Large-Scale Metabolomics Data Using Support Vector Regression. *Metabolomics* 12, 89. doi:10.1007/s11306-016-1026-5

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

Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., et al. (2019). Comprehensive Evaluation of Transcriptome-Based Cell-type Quantification Methods for Immuno-Oncology. *Bioinformatics* 35, i436–i445. doi:10.1093/bioinformatics/btz363

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

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

Sylvester, R. J., Rodríguez, O., Hernández, V., Turturica, D., Bauerová, L., Bruins, H. M., et al. (2021). European Association of Urology (EAU) Prognostic Factor Risk Groups for Non-muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel. *Eur. Urol.* 79, 480–488. doi:10.1016/j.eururo.2020.12.033

Therneau, T. M., and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Therneau, T. M. (2021). A Package for Survival Analysis in R. Available at: https://CRAN.R-project.org/package=survival. (July 24, 2021).

Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E., and Carmona-Sáez, P. (2021). A Survey of Gene Expression Meta-Analysis: Methods and Applications. *Brief. Bioinform* 22, 1694–1705. doi:10.1093/bib/bbaa019

van den Bosch, S., and Alfred Witjes, J. (2011). Long-term Cancer-specific Survival in Patients with High-Risk, Non-muscle-invasive Bladder Cancer and Tumour Progression: a Systematic Review. *Eur. Urol.* 60, 493–500. doi:10.1016/j.eururo.2011.05.045

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* 2, 100141. doi:10.1016/j.xinn.2021.100141

Wulff, J. E., and Mitchell, M. W. (2018). A Comparison of Various Normalization Methods for LC/MS Metabolomics Data. *ABB* 09, 339–351. doi:10.4236/abb.2018.98022

Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., et al. (2019). The Role of Collagen in Cancer: from Bench to Bedside. *J. Transl. Med.* 17, 309. doi:10.1186/s12967-019-2058-1

Yoon, S., Baik, B., Park, T., and Nam, D. (2021). Powerful P-Value Combination Methods to Detect Incomplete Association. *Sci. Rep.* 11, 6980. doi:10.1038/s41598-021-86465-y

Yoshihara, K., Shahmoradgoli, M., Martínez, 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 A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zheng, Z., Mao, S., Zhang, W., Liu, J., Li, C., Wang, R., et al. (2020). Dysregulation of the Immune Microenvironment Contributes to Malignant Progression and Has Prognostic Value in Bladder Cancer. *Front. Oncol.* 10, 542492. doi:10.3389/fonc.2020.542492

Conflict of Interest: Author XS was employed by GloriousMed Clinical Laboratory Co., Ltd.

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

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

ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/fgene.2022.852159

The Great Majority of Homologous Recombination Repair-Deficient Tumors Are Accounted for by Established Causes

Paula Štancl[1], Nancy Hamel[2], Keith M. Sigel[3], William D. Foulkes[2,4,5], Rosa Karlić[1] and Paz Polak[3]*†*

[1]Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia, [2]Cancer Research Program, Research Institute of the McGill University Health Centre, Montreal, QC, Canada, [3]Icahn School of Medicine at Mount Sinai, New York, NY, United States, [4]Department of Human Genetics, McGill University Montreal, Montreal, QC, Canada, [5]Cancer Axis, Lady Davis Institute, Jewish General Hospital, Montreal, QC, Canada

Background: Gene-agnostic genomic biomarkers were recently developed to identify homologous recombination deficiency (HRD) tumors that are likely to respond to treatment with PARP inhibitors. Two machine-learning algorithms that predict HRD status, CHORD, and HRDetect, utilize various HRD-associated features extracted from whole-genome sequencing (WGS) data and show high sensitivity in detecting patients with *BRCA1/2* bi-allelic inactivation in all cancer types. When using only DNA mutation data for the detection of potential causes of HRD, both HRDetect and CHORD find that 30–40% of cases that have been classified as HRD are due to unknown causes. Here, we examined the impact of tumor-specific thresholds and measurement of promoter methylation of *BRCA1* and *RAD51C* on unexplained proportions of HRD cases across various tumor types.

Methods: We gathered published CHORD and HRDetect probability scores for 828 samples from breast, ovarian, and pancreatic cancer from previous studies, as well as evidence of their biallelic inactivation (by either DNA alterations or promoter methylation) in HR-related genes. ROC curve analysis evaluated the performance of each classifier in specific cancer. Tenfold nested cross-validation was used to find the optimal threshold values of HRDetect and CHORD for classifying HR-deficient samples within each cancer type.

Results: With the universal threshold, HRDetect has higher sensitivity in the detection of biallelic inactivation in *BRCA1/2* than CHORD and resulted in a higher proportion of unexplained cases. When promoter methylation was excluded, in ovarian carcinoma, the proportion of unexplained cases increased from 26.8 to 48.8% for HRDetect and from 14.7 to 41.2% for CHORD. A similar increase was observed in breast cancer. Applying cancer-type-specific thresholds led to similar sensitivity and specificity for both methods. The cancer-type-specific thresholds for HRDetect reduced the number of unexplained cases from 21 to 12.3% without reducing the 96% sensitivity to known events. For CHORD, unexplained cases were reduced from 10 to 9% while sensitivity increased from 85.3 to 93.9%.

Edited by:
Rodrigo Drummond,
A.C.Camargo Cancer Center, Brazil

Reviewed by:
James M Ford,
Stanford University, United States
Grainne O'Kane,
University Health Network, Canada

****Correspondence:***
Paz Polak
pazpolak@gmail.com
Rosa Karlić
rosa@bioinfo.hr

†Present address:
Paz Polak,
C2i Genomics, New York (HQ), New York, NY, United States

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *10 January 2022*
Accepted: *03 May 2022*
Published: *17 June 2022*

Citation:
Štancl P, Hamel N, Sigel KM, Foulkes WD, Karlić R and Polak P (2022) The Great Majority of Homologous Recombination Repair-Deficient Tumors Are Accounted for by Established Causes. Front. Genet. 13:852159. doi: 10.3389/fgene.2022.852159

Conclusion: These results suggest that WGS-based HRD classifiers should be adjusted for tumor types. When applied, only ~10% of breast, ovarian, and pancreas cancer cases are not explained by known events in our dataset.

Keywords: homologous recombination deficiency, HRDetect, CHORD, whole-genome sequencing, promoter methylation

INTRODUCTION

The recognition of biallelic germline or somatic mutations in *BRCA1/2* is, to date, one of the most clinically relevant and frequently used genetic biomarkers of homologous recombination repair deficiency (HRD) in the clinics (Dougherty et al., 2017; Hoppe et al., 2018). Patients harboring germline pathogenic variants (GPVs) in *BRCA1/2* have a higher risk of developing breast and/or ovarian cancer (Mersch et al., 2015). Patients with germline or somatic mutations have an enhanced benefit from targeted therapies such as platinum-based chemotherapy or poly (ADP-ribose) polymerase inhibitors (PARPi) (Hennessy et al., 2010). The terms "BRCAness" or "HRD phenotype" refer to tumors with similar clinicopathological and molecular characteristics to tumors with *BRCA1* and *BRCA2* GPVs (Lord and Ashworth, 2016). Gene alterations occurring in other homologous recombinant associated genes, such as *PALB2* (Tischkowitz et al., 2007; Thomas and Brown, 2015) and *RAD51C/D* (Kondrashova et al., 2017; Polak et al., 2017), have been linked to the HRD phenotype. Inactivation through promoter methylation of *BRCA1* and *RAD51C* has also been found to result in HRD tumors (Ruscito et al., 2014; Polak et al., 2017; Staaf et al., 2019), and these tumors also demonstrate increased sensitivity to PARPi and platinum (Kondrashova et al., 2018).

Advances in tumor sequencing resulted in the development of methods to identify HRD tumors independently of identifying the cause. Cancer genomes of patients with *BRCA1/2* mutations are enriched with particular mutational patterns as well as a high number of distinct LOH regions. In addition, *BRCA1/2*-deficient tumors include small deletions with ≥4 bp flanking homology. Several structural variations are typical of *BRCA1/2*-deficient cancer genomes, including deletions up to 100 kb, unclustered tandem duplications of ~10 kb associated with *BRCA1* mutations (Willis et al., 2017), and deletions up to 1-10 kb in cancers are found in patients with *BRCA2* mutations (Degasperi et al., 2020). A specific single-base substitution signature (also known as single-nucleotide variants), referred to as COSMIC signature 3, is strongly associated with *BRCA1/2* deficiency (Polak et al., 2017).

Whole-genome sequencing (WGS) data enable the detection of different genomic alterations such as base substitutions, indels, rearrangements, and copy number aberrations, which are the result of homologous recombination deficiency. There are two HRD classifiers that are based on features extracted from WGS data. HRDetect (Davies et al., 2017) is a weighted logistic regression model based on six input features: the proportion of small deletions with microhomology at the breakpoint junction, HRD index based on genomic scars, COSMIC signatures 3 and 8, and two rearrangement signatures 3 and 5. This model was trained on *BRCA1/2*-null breast cancers. The classifier of Homologous Recombination Deficiency (CHORD) (Nguyen et al., 2020) is a random forest model that uses relative counts of somatic mutation contexts from WGS data.

Both classifiers classify >90% of tumors with biallelic inactivation via DNA mutation of *BRCA1/2* as HRD and have generally high accuracy as measured by AUC~0.98 (area under the curve) (Davies et al., 2017; Nguyen et al., 2020). Mutations in *PALB2*, *RAD51C/D,* and *BARD1* are associated with HRD signatures (Polak et al., 2017; Matis et al., 2021) and account for a small fraction of non-*BRCA1/2*-mutated HRD cases (Golan et al., 2021). Nguyen et al. (2020), in the paper that introduced CHORD, reported that a substantial proportion (~40%) of cancer samples identified as HR-deficient did not harbor any mutation in known HR-related genes (Nguyen et al., 2020), while Davies et al. (2017) reported more than 30% of these cases. These findings indicate that conventional testing for mutations in HR genes will miss a considerable number of HRD tumors where HRD is caused by unknown reasons.

The possible source of high unexplained cases could be either technical or biological. Both HRDetect and CHORD are continuous scores, designed to determine if a tumor exhibits HRD. Both use a universal threshold that was not optimized for specific cancer types. HRDetect threshold was developed based on the breast cancer dataset but this cut-off has been used for other cancer types. The CHORD study used a 0.5 cut-off. In addition, *BRCA1/RAD51C* promoter methylation is not measured in most WGS studies or on only one subset of these samples.

Here, we aim to examine the range of missing proportions of HRD samples across various three tumors where HRD is frequently reported (breast, ovarian and pancreas cancer) and determined the impact of cancer-type-specific thresholds as well as of promoter methylation *BRCA1/RAD51C* for an available subset of cases. To do so, we used published CHORD and HRDetect scores for these three cancers (Davies et al., 2017; Degasperi et al., 2020; Nguyen et al., 2020), as well as published HRDdetect scores for pancreas cancer (Golan et al., 2021) and CHORD scores that we calculated. In the case of ovarian and breast cancers, we limited our study to the subset of patients with available data for the methylation status of the *BRCA1/RAD51C* promoter. We determined the proportion of unexplained cases if we use cancer-type specific thresholds (for pancreas, ovarian, and breast cancer) and promoter methylation status (for ovarian and breast cancers).

MATERIALS AND METHODS

Datasets. Studies that performed homologous-recombination deficiency detection analysis on the same samples using the

CHORD (Nguyen et al., 2020) and HRDetect (Davies et al., 2017; Degasperi et al., 2020) classifiers were selected. From the selected studies, we made the largest unique intersection of sample names containing prediction scores of HR deficiency for both classifiers, CHORD and HRDetect. The dataset was divided into four major groups of HR-related cancers: breast, pancreatic, and ovarian (**Supplementary Table S1**), while other cancer types were put into a separate category (**Supplementary Table S2**) due to the low number of biallelic events and samples labeled as HRD. We included only breast and ovarian cancer samples that had verified *BRCA1*/2 with respect to methylation (Davies et al., 2017). The methylation status of HR-related gene promotors was considered to be an important underlying cause of HRD in tumors and we wanted to include only samples with validated methylation status to perform the downstream analysis. For the pancreatic dataset, we used 391 pancreatic samples whose data alongside the HRDetect classifier results were provided by Golan et al. (2021). For pancreatic samples, we ran the CHORD classifier using the default setting as it was previously described (Nguyen et al., 2020). The final combined dataset consisted of discrete datasets of 1) 371 breast cancers, 2) 66 ovarian cancers, 3) 391 pancreatic cancers, and 4) 1 238 samples belonging to other cancer types. For each sample in selected studies, we extracted the available methylation status of *BRCA1/2* genes for the breast and ovarian cancer samples alongside biallelic and monoallelic alternations in HR-related genes for all cancer types. We considered biallelic germline inactivation to be present when a germline pathogenic variant (GPV) was the first hit with the second hit being loss-of-heterozygosity (LOH) or somatic mutation. Somatic biallelic inactivation was considered where at least one hit was a somatic mutation, while promoter hypermethylation biallelic inactivation was defined as when one hit was promoter methylation and the other one was somatic or LOH. Monoallelic inactivation was considered when only one gene had any mutation other than LOH. Samples carrying biallelic inactivation in HR-related genes were considered to be true HR-deficient tumors. A detailed summary of all biallelic and monoallelic alterations in analyzed HR-related genes can be found in **Supplementary Table S1** and **Supplementary Table S2**, alongside the source of information regarding these gene alterations.

Assessment of the accuracy of CHORD and HRDetect classifiers through ROC and precision-recall curves. To assess the accuracy of each classifier for each of the four major cancer types, we calculated receiver operating characteristics (ROCs) using the R function "roc" from package "pROC" (Robin et al., 2011) and precision-recall (PR) curves using R function "pr.curve" from package "PRROC" (Grau et al., 2015) by comparing CHORD and HRDetect probability scores against samples carrying biallelic inactivation in HR-related genes. Bootstrapping (2000 samples) was performed to estimate the 95% CI of the area under the ROC curve (AUC). Additionally, we compared the performance of these classifiers when no methylation data are available for breast and ovarian cancers to highlight the importance of promoter hypermethylation in HR-deficient tumors.

Determining the optimal threshold. We applied a tenfold nested cross-validation approach to find the optimal threshold values of HRDetect and CHORD for classifying samples as HR-deficient or -proficient within breast, pancreatic, and ovarian cancers. The inner tenfolds were used to calculate the average optimal threshold, while the outer folds in the cross-validation process containing 10% of test data were used to assess the accuracy of the classification of HR-deficient samples. The reported optimal threshold for each classifier was calculated as the mean of all the average thresholds in outer loops for each cancer type.

Statistical analysis. Probabilistic scores from CHORD and HRDetect classifiers were compared with Spearman correlation (Spearman, 1987) using R functions cor () or cor. test (). The one-sided partially overlapping samples z-test for dichotomous variables with R function "Prop.test" from package "Partiallyoverlapping" (Derrick, 2018) was used to determine the statistically significant differences in the proportion of samples classified as HRD samples with and without evidence between CHORD and HRDetect. An one-sided Fisher's exact test using the R function "pairwise_fisher_test" from package "rstatix" (Kassambara, 2021) was used for testing the differences of explained and unexplained classifications between cancer types within each classifier. For comparison of the different ROC curves, we used the DeLong's test (two-sided, paired-samples) for two correlated ROC curves using the R function "roc.test" from package "pROC" (Robin et al., 2011). Corrections for multiple hypothesis testing were done using Bonferroni correction, and adjusted *p*-values were reported. All the analyses were carried out in R statistical programming language version 4.1.0.

RESULTS

Large Proportion of Homologous Recombination Repair Deficiency Classified Tumors Is Explained by Biallelic Inactivation of *BRCA1/2*

To investigate the performance of CHORD and HRDetect classifiers on the same samples, we utilized the classifiers' results from previous studies (Davies et al., 2017; Degasperi et al., 2020; Nguyen et al., 2020; Golan et al., 2021) across 2,066 samples from 10 cancer types. Here, we have focused on comparing HRDetect and CHORD scores for a total of 828 tumors, composed of three cancers associated with HR deficiency: breast (n = 371), pancreatic (n = 391), ovarian (n = 66) (**Figure 1A**), while the remaining seven cancers are shown in the supplementary (**Supplementary Figure S2**, **Supplementary Table S2**). When comparing the probability score of a tumor possessing HRD for each sample, we see that CHORD and HRDetect have similar probability scores (**Supplementary Figure S1**, Spearman correlation of 0.67). Of the total 828 samples belonging to the three important HRD-related cancers, biallelic alterations (either somatic, germline, deep deletion, or promoter hypermethylation) of HR-related genes

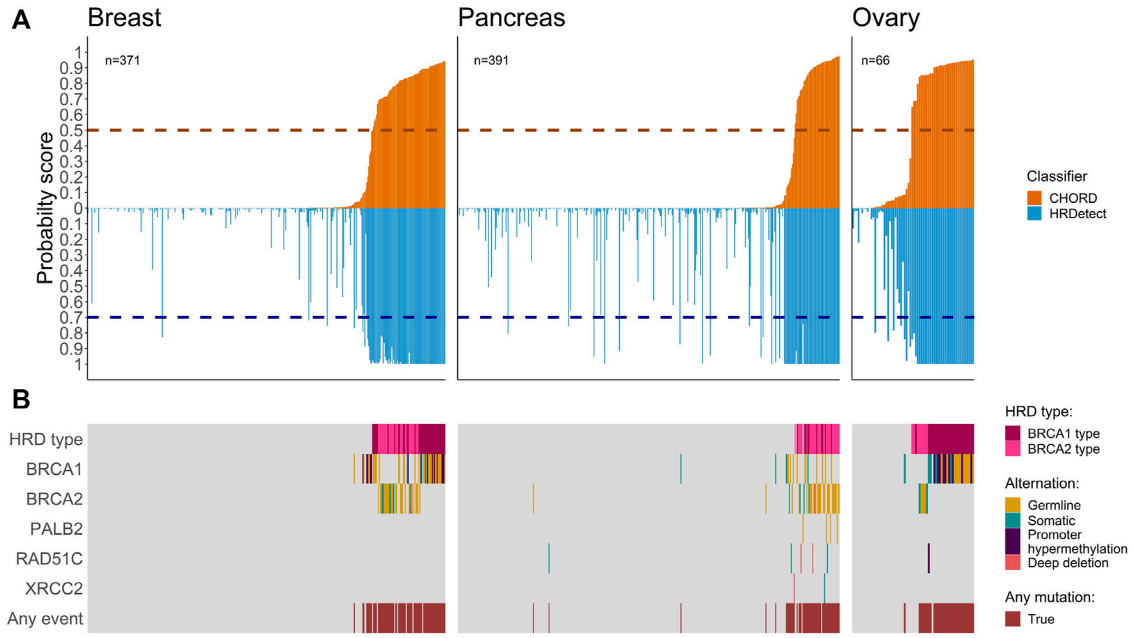


FIGURE 1 | Co-mutation plots for breast, pancreas, and ovarian cancers. **(A)** Mirror bar plot showing the probability score of CHORD (orange) and HRDetect (blue) classifiers for each sample alongside the default threshold value for each classifier (horizontal dashed line, 0.5 for CHORD and 0.7 for HRDetect). Samples are ordered by the CHORD probability score from the lowest to the highest. **(B)** The biallelic inactivation in genes related to HR deficiency. HRD types (*BRCA1* and *BRCA2* types) were assigned by the CHORD classifier.

were found in 163 samples. As expected, samples with higher HRD probability scores (in both classifiers) had a higher number of biallelic inactivation events in *BRCA1/2* genes compared to samples with lower scores (**Figure 1B**). Somatic homozygous deletion, labeled as deep deletion, were observed in *BRCA2* in a single breast cancer patient and in pancreatic cancer (*RAD51B* (n = 2), *RAD51C* (*n* = 2) and *XRCC2* (*n* = 1)) (**Supplementary Table S1**).

Among other cancer types, we observed four prostate samples with high HRD scores from both classifiers containing biallelic inactivation in *BRCA1/2* and one biliary sample with a germline *BRCA1* alteration where both HRD scores were above default (**Supplementary Figure S1**). Due to lack of evidence for HR deficiency in other cancers and the smaller sample size of identified HRD samples, other cancers were excluded for the downstream analysis and we only benchmarked results for breast, ovarian, and pancreatic cancer samples.

We proceeded to compare the fraction of HRD classified cases that are explained by the different types of biallelic inactivation in *BRCA1/2* based on HRDetected and CHORD. The most abundant biallelic inactivation patterns in the dataset included g*BRCA1/2* (n = 52 + 54) and s*BRCA1/2* mutations (*n* = 12 + 11) (**Supplementary Table S1**). The *BRCA1* promoter methylation status was available only for breast and ovarian (*n* = 23) and it accounted for a significant number of the total biallelic events (23 out of 175, 12.8% (95% CI [8.7–19.3])). Nearly all of the cases with known biallelic inactivation (157 out of 163, 96.4% (95% CI [91.8–98.5])) were in tumors that are above the default threshold of either of the classifiers.

The largest proportion of unexplained HRD cases was observed in ovarian cancer (14.7%, 95% CI [5.5–31.8]) using CHORD and in pancreatic samples (28.2%, 95% CI [59.7–81.6]) using HRDetect (**Table 1**; **Figure 2**). Larger fractions of unexplained cases were obtained using HRDetect compared to CHORD with the default threshold value (one-sided *z*-test for partially overlapping samples, *p*-value < 10^{-13}) (**Figure 2**), ranging from around 10 to 28% depending on the cancer type. When looking at each classifier closely, we see that the highest difference is between breast and pancreatic cancers and HRD unexplained cases for HRDetect (one-sided Fisher's exact test, *p*-value = 0.0375). Multiple biallelic inactivation events can occur in HR genes in the same patients; for instance, one ovarian sample contained *sBRCA1* and promoter hypermethylation of *RAD51C*, while a pancreatic sample had a somatic deep deletion of both *RAD51B* and *RAD51C* (**Supplementary Table S1**).

Performance of CHORD and HRDetect Classifiers

As previously reported, both classifiers, CHORD and HRDetect, achieved exceptional performance in identifying biallelic events in breast and ovarian cancer types as shown by the high area under the ROC curve (AUC) above 0.96 and 0.9, respectively (**Figure 3**). In addition, we calculated the area under the precision-recall curve (AUPRC) that was high and well above 90% across all cancer types. No statistically significant difference was detected between CHORD and HRDetect AUC values (*p* > 0.05, DeLong's test).

TABLE 1 | Summary of tumor samples classified to possess HRD by CHORD and HRDetect within an individual cancer type.

	Evidence of Biallelic Inactivation		No Evidence of Bi-allelic Inactivation	
	Count	Proportion (95% CI)	Count	Proportion (95% CI)
CHORD				
Breast	69	90.8 (81.4-95.9)	7	9.2 (4.1-18.6)
Ovary	29	85.3 (68.2-94.5)	5	14.7 (5.5-31.8)
Pancreas	41	89.1 (75.6-95.9)	5	10.9 (4.1-24.4)
HRDetect				
Breast	76	87.4 (78.1-93.2)	11	12.6 (6.8-21.9)
Ovary	30	73.2 (56.8-85.2)	11	26.8 (14.8-43.2)
Pancreas	51	71.8 (59.7-81.6)	20	28.2 (18.4-40.3)

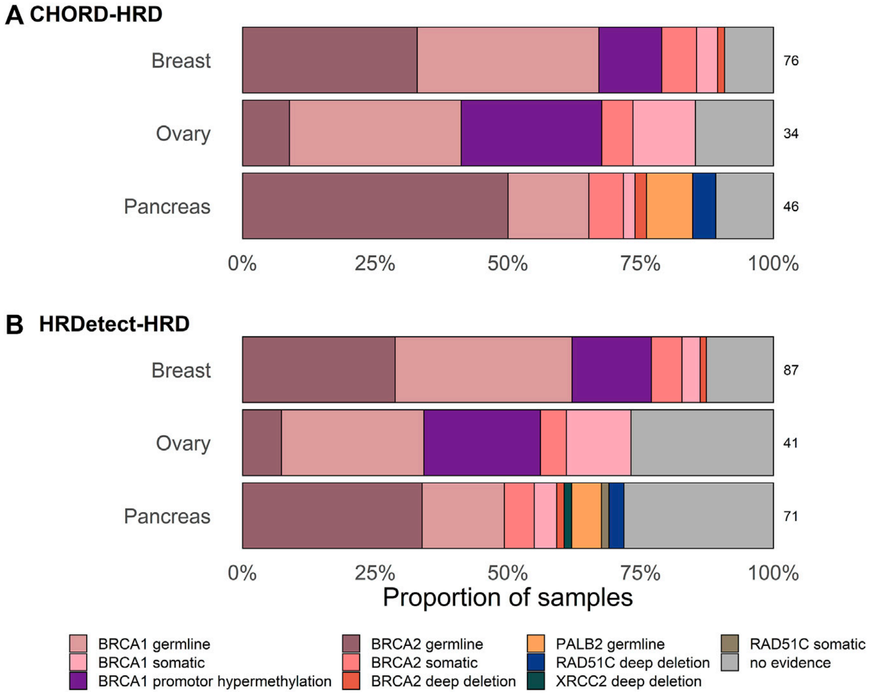


FIGURE 2 | Proportion of samples with and without biallelic alteration in HR-genes classified as HR-deficient with default threshold of **(A)** HRDetect of 0.7 and **(B)** CHORD classifiers of 0.5. Only one alteration in the gene is shown per sample based on the hierarchical order of genes as follows: *BRCA1*, *BRCA2*, *RAD51C*, *PALB2,* and *XRCC2*.

Impact of Exclusion of Promoter Methylation on the Performance

To assess the importance of promoter methylation in the evaluation of HRD classifiers' performance, we removed the methylation data of *BRCA1/RAD51C* promoters in breast and ovarian the only cancer types for which methylation data were available. We observed a significant drop in classifiers' performance for breast and ovarian samples (**Figure 3**). In ovarian cancer, the drop in AUC values was significantly affected, falling from 0.987 to 0.873 for CHORD (*p*-value = 0.044, DeLong's test) and from 0.987 to 0.828 for HRDetect (*p*-value = 0.011, DeLong's test). In contrast, the breast cancer AUC values were still above 0.96 for both classifiers (*p*-value = 0.057 for CHORD and 0.055 for HRDetect, DeLong's test) and AUPRC values were slightly above 0.7 compared to 0.949 when methylation status was included.

Revisiting Threshold Values for Homologous Recombination Repair Deficiency Classification of Different Cancer Types

The current threshold of HRDetect(0.7) was determined based on the breast dataset, while CHORD 0.5 was arbitrarily chosen. Considering

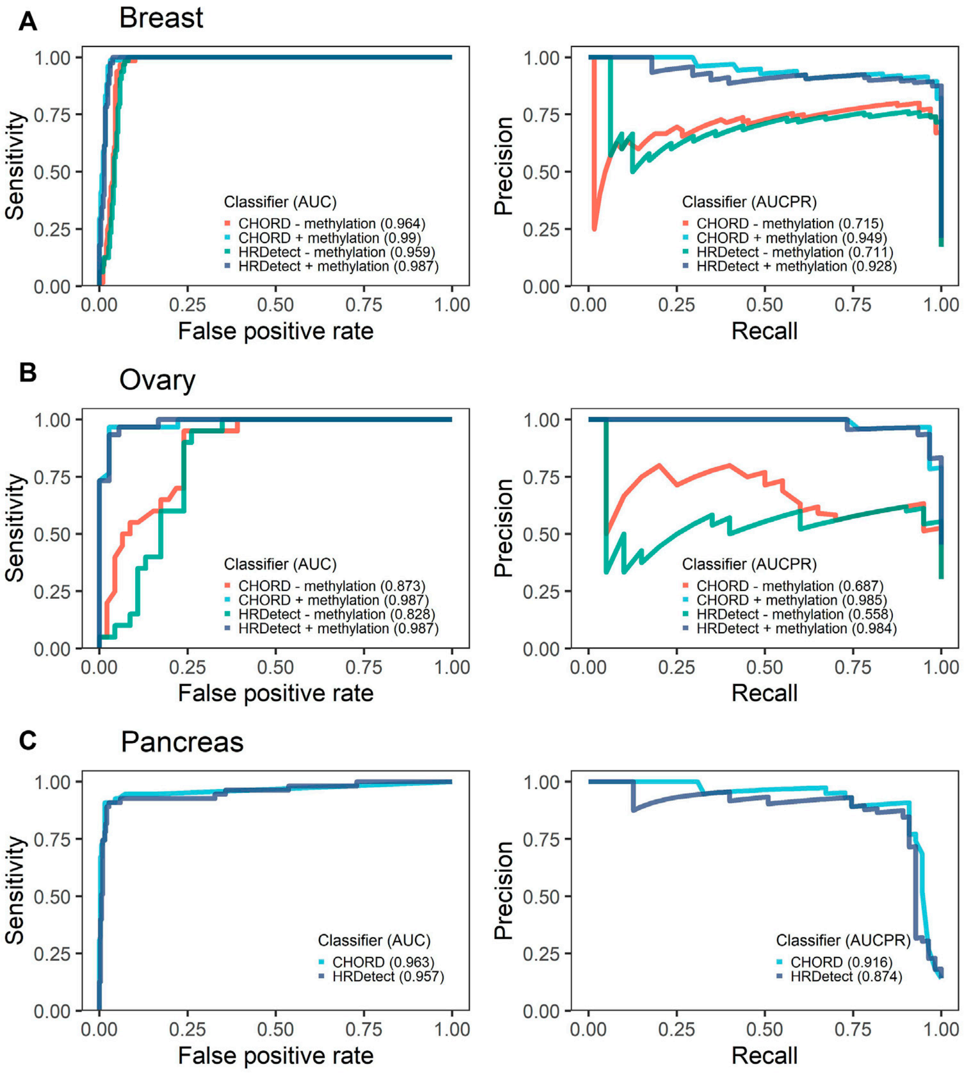


FIGURE 3 | Receiver operating characteristics (ROCs) with the respective area under the curve (AUC) and precision-recall curves (PR) with the area under the precision-recall curve (AUCPR) showing the performance of CHORD and HRDetect classifier with and without methylation data for breast **(A)** and ovarian cancers **(B)**. Pancreas **(C)** cancer data do not have methylation data.

different machine-learning algorithms underlying CHORD and HRDetect for classifying HRD in samples and different training data, we sought to determine an optimal threshold value for the individual cancer types in our cohort. For each cancer type and classifier, we performed 10-fold nested cross-validation to calculate the optimal threshold value (given in detail in the Methods section). The accuracy of both classifiers with default threshold values was similar across cancers, while the most considerable increase was detected in ovarian cancer (accuracy CHORD 0.91 and HRDetect 0.83) (**Table 2**). Cancer-type-specific (optimal) threshold values differ from the classifiers' default ones, but the overall accuracy improves slightly or remains the same. The only exception is the optimal value of HRDetect in ovarian cancer where the accuracy improved by 12%. The number of samples with evidence of bi-allelic alterations in HR-related genes and classification as HR deficient by the classifiers were more abundant in optimal values of the CHORD classifier in breast and pancreatic cancers compared to the default threshold in the same cancer type. The proportion of classified HRD cases in the dataset without known biallelic evidence decreased for both CHORD, 10.9–8.9%, and HRDetect, 21.1–12.3%. Monoallelic mutations were found in pancreatic cancer (**Supplementary Figure S3**). Using default threshold values, the majority of monoallelic mutation in HR-related genes occurs in homologous recombination proficient (HRP) samples, where HRDetect has more HRD unexplained cases and two monoallelic mutations in HR-related genes. The monoallelic alterations were detected in HRD-labeled samples only with HRDetect with default and an optimal threshold value.

TABLE 2 | Summary table of confusion matrix results with accuracy for default and optimal (cancer-type-specific) threshold values of CHORD and HRDetect classifiers for classifying samples as homologous recombination deficient (HRD) or homologous recombination proficient (HRP).

	CHORD						HRDetect					
	HRD		HRP				HRD		HRP			
Bi-allelic Evidence	Yes	No	Yes	No	Threshold	Accuracy	Yes	No	Yes	No	Threshold	Accuracy
Breast												
default	69	7	9	286	0.50	0.96	76	11	2	282	0.70	0.96
optimal	75	9	3	284	0.10	0.97	77	11	1	282	0.68	0.97
Ovary												
default	29	5	1	31	0.50	0.91	30	11	0	25	0.70	0.83
optimal	29	1	1	35	0.84	0.97	29	2	1	34	0.99	0.95
Pancreas												
default	41	5	14	331	0.50	0.95	51	20	4	316	0.70	0.94
optimal	49	5	6	331	0.13	0.97	50	9	5	327	0.98	0.96

[a] *HRD and HRP categories are determined by CHORD and HRDetect based on default or optimal (cancer-type-specific) thresholds.*

DISCUSSION

Our study shows an integrated overview of detecting homologous recombination deficiency in cancers using CHORD and HRDetect classifiers. Here, we have mainly focused on three cancers most commonly associated with HRD: breast, ovarian, and pancreatic cancers. We observed that biallelic inactivation of genes explains a large fraction of samples possessing HRD when using a universal default threshold, as was demonstrated in previous studies (Davies et al., 2017; Nguyen et al., 2020; Golan et al., 2021). However, around 10–28% of patients without known underlying causes were detected by these classifiers despite their high performance based on the default threshold. In this study, we found that by applying a cancer-type-specific threshold the number of unexplained cases reduced to around 8.9–12.3% without decreasing the sensitivity of 96%. We estimate that in this dataset up to ~10% of HRD cases are caused by types of alterations that still have not been associated with HRD and therefore gene-centric testing for mutations in HR genes will likely fail to identify them. Similar results apply to the analysis of other cancer types in which the HRD cancers are rarer in comparison to the four well-known HRD cancers. The low number of HRD mutations in prostate samples and other cohorts did not allow the determination of a reliable cancer-type-specific threshold. The small fraction of unexplained cases is consistent with our previous proposal (Foulkes and Polak, 2019; Matis et al., 2021) that if alterations in novel genes lead to HRD, taken together, they will all account for only a very small proportion of all HRD cases.

The different cut-offs that we observed may be due to subtle differences across cancer in the mutational landscape even for tumors with different same gene defects, especially in mutational signatures (Degasperi et al., 2020). Furthermore, as it was highlighted by Nguyen et al. (2020), additional threshold optimization and validations are also required when applying classifiers to WGS data generated by other variant calling pipelines. Our cohort contained data generated by various pipelines for CHORD and HRDetect in each cancer type, which may affect the overall comparison of results between these classifiers. In addition to the threshold value, it is important to investigate other features affecting the mutation landscape of tumors, such as deficiency in mismatch repair (MMR), which may have a negative impact on the overall performance of classifiers in specific tumors. It was noted by Golan et al. (2021) that one pancreatic sample with biallelic inactivation in *BRCA2* and *PMS2* (responsible for MMR) was misclassified by HRDetect and CHORD classifier and had both scores near zero.

In addition to cancer-type-specific thresholds that reduce the number of unexplained cases, we demonstrated the importance of including the promoter methylation status of *BRCA1* and *RAD51C* in order to evaluate the fraction of HRD cases that are explained by known causes. In breast and ovarian cancers, for which methylation analysis is most often conducted, promoter methylation of *BRCA1* accounts for at least 20% of explained biallelic inactivation cases of HRD, labeled by either of the classifiers, and lack of methylation data significantly affects the performance of classifiers. The proportion of unexplained cases in other cancer types may have been reduced if methylation analysis data existed, especially in pancreatic cancer where some detected monoallelic PVs could have other events such as promoter methylation that would explain their HRD. These observations highlight the advantage of using these classifiers alongside conventional testing for patient selection and stratification in clinics, as was already suggested by several studies (Zhao et al., 2017; Staaf et al., 2019; Chopra et al., 2020). The relationship between the presence of HRD and response to therapies such as PARP inhibitors is not precise and there is currently no "ground truth" for measuring HRD. Resistance to PARP inhibitors can co-exist with HRD (Dias et al., 2021), so the presence of HRD is not by itself a direct predictor of response to PARP inhibitors and other drugs such as platinum that cause double-strand DNA breaks. Combinations of different approaches such as

WGS-based, FDA-approved assays, and newer functional assays such as the RAD51 foci assay (Pellegrino et al., 2022) will ultimately lead to a better selection of HRD patients for appropriate therapies. Hence, our review re-analysis emphasizes the power of both CHORD and HRDetect in the stratification of patients possessing HRD phenotype across various cancers, as well as the importance of identification and further validation of new unrevealed oncogenic mutations.

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

PP and RK conceptualized and supervised the study. PŠ collected and analyzed the data. PŠ, PP, and RK wrote the manuscript. WF, NH, and KS participated in manuscript revision. All authors contributed to the article and approved the submitted version.

FUNDING

PŠ and RK were funded by grants from the Croatian National Science Foundation Project PREDI-COO (A statistical modeling approach to predict the cell-of-origin and investigate mechanisms of cancer development) (Project number: IP-2019-04-9308). WF and NH are supported by grants from the Canadian Institutes of Health Research (FDN-148390) and the Canadian Cancer Society (#706389).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.852159/full#supplementary-material

REFERENCES

Chopra, N., Tovey, H., Pearson, A., Cutts, R., Toms, C., Proszek, P., et al. (2020). Homologous Recombination DNA Repair Deficiency and PARP Inhibition Activity in Primary Triple Negative Breast Cancer. *Nat. Commun.* 11, 1–12. doi:10.1038/s41467-020-16142-7

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., et al. (2017). HRDetect is a Predictor of BRCA1 and BRCA2 Deficiency Based on Mutational Signatures. *Nat. Med.* 23, 517–525. doi:10.1038/nm.4292

Degasperi, A., Amarante, T. D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., et al. (2020). A Practical Framework and Online Tool for Mutational Signature Analyses Show Intertissue Variation and Driver Dependencies. *Nat. Cancer* 1, 249–263. doi:10.1038/s43018-020-0027-5

Derrick, B. (2018). Partiallyoverlapping: Partially Overlapping Samples Tests. R package version 2.0. 2018. Available at: https://cran.r-project.org/.

Dias, M. P., Moser, S. C., Ganesan, S., and Jonkers, J. (2021). Understanding and Overcoming Resistance to PARP Inhibitors in Cancer Therapy. *Nat. Rev. Clin. Oncol.* 18, 773–791. doi:10.1038/s41571-021-00532-x

Dougherty, B. A., Lai, Z., Hodgson, D. R., Orr, M. C. M., Hawryluk, M., Sun, J., et al. (2017). Biological and Clinical Evidence for Somatic Mutations in BRCA1 and BRCA2 as Predictive Markers for Olaparib Response in High-Grade Serous Ovarian Cancers in the Maintenance Setting. *Oncotarget* 8, 43653–43661. doi:10.18632/oncotarget.17613

Foulkes, W. D., and Polak, P. (2019). Journey's End: The Quest for BRCA-Like Hereditary Breast Cancer Genes Is Nearly over. *Ann. Oncol.* 30, 1023–1025. doi:10.1093/annonc/mdz152

Golan, T., O'Kane, G. M., Denroche, R. E., Raitses-Gurevich, M., Grant, R. C., Holter, S., et al. (2021). Genomic Features and Classification of Homologous Recombination Deficient Pancreatic Ductal Adenocarcinoma. *Gastroenterology* 160, 2119–2132. e9. doi:10.1053/j.gastro.2021.01.220

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

Hennessy, B. T. J., Timms, K. M., Carey, M. S., Gutin, A., Meyer, L. A., Flake, D. D., et al. (2010). Somatic Mutations in BRCA1 and BRCA2 Could Expand the Number of Patients that Benefit from Poly (ADP Ribose) Polymerase Inhibitors in Ovarian Cancer. *J. Clin. Oncol.* 28, 3570–3576. doi:10.1200/JCO.2009.27.2997

Hoppe, M. M., Sundar, R., Tan, D. S. P., and Jeyasekharan, A. D. (2018). Biomarkers for Homologous Recombination Deficiency in Cancer. *J. Natl. Cancer Inst.* 110, 704–713. doi:10.1093/jnci/djy085

Kassambara, A. (2021). Rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. 2021. Available at: https://cran.r-project.org/package=rstatix.

Kondrashova, O., Nguyen, M., Shield-Artin, K., Tinker, A. V., Teng, N. N. H., Harrell, M. I., et al. (2017). Secondary Somatic Mutations Restoring RAD51C and RAD51D Associated with Acquired Resistance to the PARP Inhibitor Rucaparib in High-Grade Ovarian Carcinoma. *Cancer Discov.* 7, 984–998. doi:10.1158/2159-8290.CD-17-0419

Kondrashova, O., Topp, M., Topp, M., Nesic, K., Lieschke, E., Ho, G.-Y., et al. (2018). Methylation of All BRCA1 Copies Predicts Response to the PARP Inhibitor Rucaparib in Ovarian Carcinoma. *Nat. Commun.* 9, 3970. doi:10.1038/s41467-018-05564-z

Lord, C. J., and Ashworth, A. (2016). BRCAness Revisited. *Nat. Rev. Cancer* 16, 110–120. doi:10.1038/nrc.2015.21

Matis, T. S., Zayed, N., Labraki, B., de Ladurantaye, M., Matis, T. A., Camacho Valenzuela, J., et al. (2021). Current Gene Panels Account for Nearly All Homologous Recombination Repair-Associated Multiple-Case Breast Cancer Families. *npj Breast Cancer* 7, 109. doi:10.1038/s41523-021-00315-8

Mersch, J., Jackson, M. A., Park, M., Nebgen, D., Peterson, S. K., Singletary, C., et al. (2015). Cancers Associated with BRCA1 and BRCA2 Mutations Other Than Breast and Ovarian. *Cancer* 121, 269–275. doi:10.1002/cncr.29041

Nguyen, L., Martens, W. M. J., Van Hoeck, A., and Cuppen, E. (2020). Pan-Cancer Landscape of Homologous Recombination Deficiency. *Nat. Commun.* 11, 1–12. doi:10.1038/s41467-020-19406-4

Pellegrino, B., Herencia-Ropero, A., Llop-Guevara, A., Pedretti, F., Moles-Fernández, A., Viaplana, C., et al. (2022). Preclinical *In Vivo* Validation of the RAD51 Test for Identification of Homologous Recombination-Deficient Tumors and Patient Stratification. *Cancer Res.* 82 (8), 1646–1657. doi:10.1158/0008-5472.CAN-21-2409

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., et al. (2017). A Mutational Signature Reveals Alterations Underlying Deficient Homologous Recombination Repair in Breast Cancer. *Nat. Genet.* 49, 1476–1486. doi:10.1038/ng.3934.A

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 Bioinforma.* 12, 77. doi:10.1186/1471-2105-12-77

Ruscito, I., Dimitrova, D., Vasconcelos, I., Gellhaus, K., Schwachula, T., Bellati, F., et al. (2014). BRCA1 Gene Promoter Methylation Status in High-Grade Serous Ovarian Cancer Patients - A Study of the Tumour Bank Ovarian Cancer (TOC) and Ovarian Cancer Diagnosis Consortium (OVCAD). *Eur. J. Cancer* 50, 2090–2098. doi:10.1016/j.ejca.2014.05.001

Spearman, C. (1987). The Proof and Measurement of Association between Two Things. By C. Spearman, 1904. *Am. J. Psychol.* 100, 441–471. doi:10.2307/1422689

Staaf, J., Glodzik, D., Bosch, A., Vallon-christersson, J., Reuterswärd, C., and Häkkinen, J. (2019). Whole-Genome Sequencing of Triple-Negative Breast Cancers in a Population-Based Clinical Study. *Nat. Med.* 25, 1526–1533. doi:10.1038/s41591-019-0582-4

Thomas, P. S., and Brown, P. H. (2015). Breast-Cancer Risk in Families with Mutations in PALB2. *Breast Dis.* 26, 206–208. doi:10.1016/j.breastdis.2015.07.017

Tischkowitz, M., Xia, B., Sabbaghian, N., Reis-Filho, J. S., Hamel, N., Li, G., et al. (2007). Analysis of PALB2/FANCN -Associated Breast Cancer Families. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6788–6793. doi:10.1073/pnas.0701724104

Willis, N. A., Frock, R. L., Menghi, F., Duffey, E. E., Panday, A., Camacho, V., et al. (2017). Mechanism of Tandem Duplication Formation in BRCA1-Mutant Cells. *Nature* 551, 590–595. doi:10.1038/nature24477

Zhao, E. Y., Shen, Y., Pleasance, E., Kasaian, K., Leelakumari, S., Jones, M., et al. (2017). Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. *Clin. Cancer Res.* 23, 7521–7530. doi:10.1158/1078-0432.CCR-17-1941

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

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

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 30 June 2022
doi: 10.3389/fgene.2022.865313

RNA-SSNV: A Reliable Somatic Single Nucleotide Variant Identification Framework for Bulk RNA-Seq Data

Qihan Long[1,2,3], *Yangyang Yuan*[1,2,3] *and Miaoxin Li*[1,2,3,4,5]*

[1]Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China, [2]Center for Precision Medicine, Sun Yat-Sen University, Guangzhou, China, [3]Center for Disease Genome Research, Sun Yat-Sen University, Guangzhou, China, [4]Guangdong Provincial Key Laboratory of Biomedical Imaging and Guangdong Provincial Engineering Research Center of Molecular Imaging, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China, [5]Key Laboratory of Tropical Disease Control (SYSU), Ministry of Education, Guangzhou, China

The usage of expressed somatic mutations may have a unique advantage in identifying active cancer driver mutations. However, accurately calling mutations from RNA-seq data is difficult due to confounding factors such as RNA-editing, reverse transcription, and gap alignment. In the present study, we proposed a framework (named RNA-SSNV, https://github.com/pmglab/RNA-SSNV) to call somatic single nucleotide variants (SSNV) from tumor bulk RNA-seq data. Based on a comprehensive multi-filtering strategy and a machine-learning classification model trained with comprehensively curated features, RNA-SSNV achieved the best precision–recall rate (0.880–0.884) in a testing dataset and robustly retained 0.94 AUC for the precision–recall curve in three validation adult-based TCGA (The Cancer Genome Atlas) datasets. We further showed that the somatic mutations called by RNA-SSNV tended to have a higher functional impact and therapeutic power in known driver genes. Furthermore, VAF (variant allele fraction) analysis revealed that subclonal harboring expressed mutations had evolutionary selection advantage and RNA had higher detection power to rescue DNA-omitted mutations. In sum, RNA-SSNV will be a useful approach to accurately call expressed somatic mutations for a more insightful analysis of cancer drive genes and carcinogenic mechanisms.

Keywords: cancer, somatic mutation, RNA, RNA-Seq, machine learning, RNA-SSNV

OPEN ACCESS

Edited by:
Joel Correa Da Rosa,
Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Rodrigo Gularte Mérida,
Memorial Sloan Kettering Cancer Center, United States
Yiyang Wu,
Vanderbilt University Medical Center, United States

***Correspondence:**
Miaoxin Li
limiaoxin@mail.sysu.edu.cn

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *29 January 2022*
Accepted: *17 May 2022*
Published: *30 June 2022*

Citation:
Long Q, Yuan Y and Li M (2022) RNA-SSNV: A Reliable Somatic Single Nucleotide Variant Identification Framework for Bulk RNA-Seq Data. Front. Genet. 13:865313. doi: 10.3389/fgene.2022.865313

INTRODUCTION

Cancer is the leading cause of death and an important barrier to increasing life expectancy (Sung et al., 2021). According to GLOBOCAN 2020 estimates of cancer incidence and mortality, 19.3 million new cancer cases and 10.0 million cancer deaths occurred in 2020 (Sung et al., 2021). Somatic mutations are usually induced by environmental factors, and it is well known that their accumulation with aging and evolution in human cells will lead to malignant transformation and eventually cancer (Watson et al., 2013). Thus, comprehensive somatic mutation identification in cancer such as the Catalogue Of Somatic Mutations In Cancer (Tate et al., 2019) (COSMIC database) can help characterize its genomic complexities (Watson et al., 2013) and discover oncogenic mutations and driver genes which significantly influence cancer development (Bailey et al., 2018). Furthermore, person-level somatic mutations also have their own oncogenic and therapeutic implications in multiple cancers (lung cancer (Skoulidis and Heymach,2019), bladder cancer (Cazier et al., 2014; Wen et al., 2021), and glioblastoma (Lin et al., 2021; McDonald et al., 2015)), targeting the corresponding mutant proteins or pathways. Currently, most somatic mutation identification studies were based on DNA-level, actionable practices in somatic mutation detection within whole-genome or whole-exome sequencing data have been developed to facilitate precision oncology (Xiao et al., 2021).

Mutations within exons are supposed to be transcribed into RNA, and be reflected in the translated protein. However, many DNA mutations within exons were not found in RNA because they were located in the non-transcribed allele or had no or low expression (O Brien et al., 2015). Yizhak et al. (2019) reported that 65% of DNA somatic mutations within 243 TCGA tumor samples were not detected in RNA. Rashid et al.(2014) found that only 27% of mutated alleles got expressed in multiple myeloma. The significant lack of DNA mutations in RNA indicated that not all DNA mutations have certain effects finally. RNA can be a reliable source to distinguish mutations that have been expressed to affect cellular functions. Although RNA-seq is mainly used for gene expression and fusion discoveries in clinical oncology (Wang et al., 2020), previous studies showed that calling genomic variants in expressed exons using RNA-seq data was feasible and cost-effective (Chepelev et al., 2009; Cirulli et al., 2010; Gonorazky et al., 2019; Piskol et al., 2013; Quinn et al., 2013). The advantages included making the most abundant RNA-seq data resources and discovering rare somatic mutations with the low-level DNA allele fraction at higher sequencing depths in sufficiently expressed genes (Chepelev et al., 2009; Cirulli et al., 2010; Gonorazky et al., 2019; Piskol et al., 2013; Quinn et al., 2013; Liu et al., 2014). However, calling somatic mutations within RNA-seq data was challenging compared with calling variants in WES data. The main challenge was the high false-positive rate, deriving from errors during reverse transcription, misalignment near splicing junctions (exon ends), RNA editing, and modification during post-transcriptional processing (Cirulli et al., 2010; Xu, 2018). Multiple RNA somatic mutation calling tools and pipelines have been developed to remove these false-positive calls, which can be divided into two categories: statistical filtering strategy-based (García-Nieto et al., 2019; Neums et al., 2018; Yizhak et al., 2019) and machine learning–based approaches (Muyas et al., 2020; Sheng et al., 2016). For instance, GLMVC (Sheng et al., 2016) calls RNA somatic mutations based on a bias-reduced generalized linear model trained by the characteristics of RNA-seq data. VaDiR (Neums et al., 2018) integrates results from three variant callers and produced higher precision results through consensus combination but sacrificed sensitivity. RNA-MuTect (Yizhak et al., 2019) comprehensively filtered mutations within artifact sites and achieved optimal performance. RF-RNAMut (Muyas et al., 2020) utilized a machine learning model to distinguish somatic variants from germline variants identified in RNA-seq data. Although existed tools have their advantages and highlights, they had their limitations: (1) unsatisfying precision–recall performance with the maximum reported precision–recall to be 0.87–0.72 (Yizhak et al., 2019), (2) required restricted resources such as DNA and RNA panel of normal (PoN) calls from ~6500 GTEx samples to achieve a desired result (Yizhak et al., 2019), and (3) model not specifically trained to recognize excessive artifacts in RNA but to identify germline mutations as negative (Muyas et al., 2020).

Here, we introduce a framework named RNA-SSNV (https://github.com/pmglab/RNA-SSNV). It is a unified framework containing a universal pipeline to call RNA somatic single nucleotide variants from the combination of tumor RNA-seq and normal WES data, a multi-filtering strategy to remove doubtful calls with little loss of sensitivity and a supervised machine learning model to identify somatic mutations and artifacts. Our framework achieved the best overall performance for precision and recall, requiring only public reference resources. To validate the generalization performance of our framework, we utilized RNA-SSNV within TCGA lung squamous cell carcinoma (LUSC), bladder urothelial carcinoma (BLCA), and glioblastoma multiforme (GBM) independent datasets. RNA-SSNV achieved similar performance in the area under curve (AUC) for the precision–recall curve with 0.94 for all three datasets. Given its high precision–recall performance, RNA-SSNV will help exploit expressed somatic variants, further extend the range of RNA-seq applications and make full use of abundant RNA-seq data resources.

MATERIALS AND METHODS

Framework Overview

Our RNA somatic single nucleotide variant identification framework (RNA-SSNV) consists of three major components, including a RNA somatic mutation calling step, a multi-step filtering process and a machine-learning based prediction (**Figure 1**). The underlying hypothesis of RNA-SSNV is that RNA-specific mutations have unique biological and technique features; thus, a comprehensive filtration process and a machine learning model based on these features can substantially improve the accuracy of RNA somatic mutation calling.

Datasets

Our datasets were retrieved from GDC, which had harmonized pipelines (https://docs.gdc.cancer.gov/Data/Introduction/) to generate RNA-seq and DNA-seq data. All RNA-seq datasets were aligned to GRCh38 build using a two-pass method with STAR, which required preprocessing before mutation calling. All DNA-seq datasets were aligned to the GRCh38 reference using bwa (Li and Durbin, 2009) and co-cleaned using the GATK toolkit (McKenna et al., 2010), which can directly be utilized in mutation calling.

We chose the TCGA lung adenocarcinoma (LUAD) cohort as the training dataset that contained the largest patient scale (511) compared with other available cancer cohorts. Our training dataset comprised paired tumor RNA-seq and tumor/normal WES data derived from 511 LUAD patients, which simultaneously generated DNA and RNA somatic mutations. Our independent validation datasets comprised paired tumor RNA-seq and normal WES data derived from 498 LUSC, 441 BLCA, and 198 GBM patients, for which we called RNA somatic mutations to get validating records.

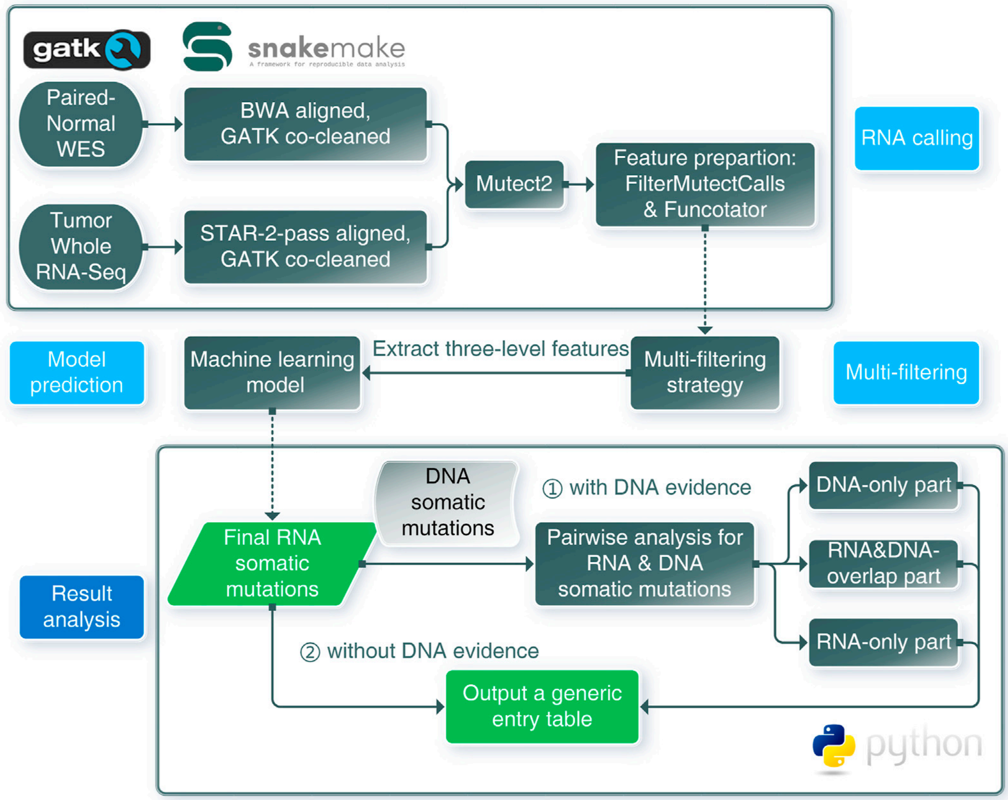

gatk
snakemake
A framework for reproducible data analysis
Paired-Normal WES
BWA aligned, GATK co-cleaned
Tumor Whole RNA-Seq
STAR-2-pass aligned, GATK co-cleaned
Mutect2
Feature prepration: FilterMutectCalls & Funcotator
RNA calling
Model prediction
Machine learning model
Extract three-level features
Multi-filtering strategy
Multi-filtering
DNA somatic mutations
① with DNA evidence
DNA-only part
Final RNA somatic mutations
Pairwise analysis for RNA & DNA somatic mutations
RNA&DNA-overlap part
Result analysis
② without DNA evidence
RNA-only part
Output a generic entry table
python

FIGURE 1 | Schematic overview of the framework for RNA somatic mutation identification. RNA calling: RNA-seq and WES data were aligned and co-cleaned accordingly. Mutect2 was used to conduct RNA somatic calling with paired tumor RNA-seq and normal WES data. Features were extracted from outputs of FilterMutectCalls and Funcotator. Multi-filtering: multi-filtering strategy was conducted in Mutect2 called mutations by removing multiallelic, RNA-editing, immunoglobin, and HLA sites. Model prediction: using the trained model, mutations with extracted features were predicted as positive or negative, only positives were regarded as reliable mutations. Result analysis: pairwise analysis can be conducted when DNA evidence was available. RNA-SSNV will output a generic entry table containing all features and predicting information to facilitate downstream analysis.

Mutation Calling

Theoretically, calling somatic mutations within RNA-seq data can be easily conducted using callers designed for DNA. Haplotype-based callers (GATK Mutect2 (Benjamin et al., 2019; Cibulskis et al., 2013), TNscope) had been proven to outperform position-based variant callers due to their inherent technical advantage in complex variants and high mutation loading regions (Pei et al., 2020; Xu, 2018). In addition, we queried the TCGA helpdesk and learned that our RNA-seq data (TCGA LUAD, LUSC, GBM, and BLCA projects) were sequenced by the UNC center using poly-T mRNA enriching strategy, which indicated that only transcribed exon regions (GENCODE v22 annotated exon regions) within mature mRNA can be sequenced (Kukurba and Montgomery, 2015) and our paired normal targeted capture exome sequencing (WES) data had a canonical target region (Agilent SureSelect TargetInterval). Thus, we chose to utilize Mutect2 to perform somatic variant calling and only retain mutations within targeted coding regions (overlap of exons and WES targets).

Normally, our STAR-2-pass aligned RNA-seq data required a co-cleaning process to conduct mutation calling. Following GATK recommended procedures (RNAseq Best Practice), our aligned RNA-seq bam was passed to the MarkDuplicates tool to identify duplicate reads and help remove PCR-related artifacts. Next, SplitNCigarReads hard-clipped and reformat some alignments which span introns causing large-scale mistaken indels. Finally, it shall undergo base quality recalibration conducted by BaseRecalibrator and ApplyBQSR to detect and correct patterns of systematic errors in the base quality scores.

After obtaining analysis-ready bam files, we utilized Mutect2 to call RNA somatic mutations from paired tumor RNA-seq and normal WES data, DNA somatic mutations from tumor and normal WES data. For the TCGA LUAD training set, we called RNA and DNA somatic mutations to help construct the training dataset. For TCGA LUSC, GBM, and BLCA validation sets, calling RNA somatic mutations were sufficient to validate our framework's performance. For DNA somatic mutations omitted in RNA which required verification, we applied the force-calling mode in Mutect2 to retrieve their RNA mutational status. Finally,

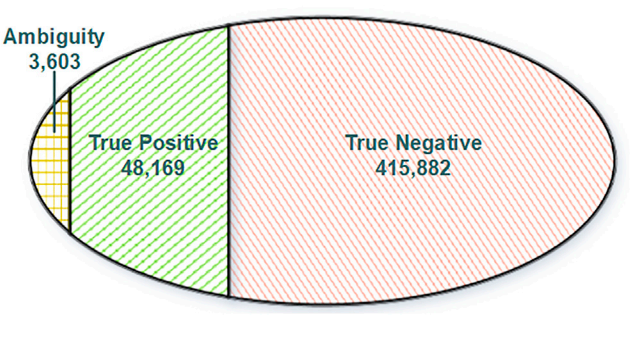


FIGURE 2 | Venn diagram of training dataset categories. True positive: RNA somatic mutations overlapping with GDC mutations. Ambiguity: RNA somatic mutations overlapping with GDC omitted somatic mutations. True negative: RNA somatic mutations without DNA support.

we utilized FilterMutectCalls to generate quality information as training features and assess the performance for Mutect2's default filtering, Funcotator to annotate variants and facilitate downstream analysis.

Multi-Filtering Strategy

Before model training or predicting, RNA somatic mutations shall be comprehensively filtered to remove known possible artifacts (García-Nieto et al., 2019; Yizhak et al., 2019). Our multi-filtering strategy included removing multi-allelic mutations, RNA-editing sites, IgG, and HLA regions. For multi-allelic mutations, we removed mutations containing three or more allele types to avoid misaligning artifacts. For RNA editing events, we combined A-to-I RNA editing information from the REDIportal (Mansi et al., 2021) database and further editing information from the DARNED (Kiran et al., 2013) database. We removed all mutations which located in the union set of RNA editing events to prevent these false-positive calls. For IgG regions, we removed mutations falling into IgG genes to avoid noisy alignments (Ye et al., 2013). For HLA regions, we removed the HLA mutations in chromosome 6 which contained a high density of germline variants (Buhler and Sanchez-Mazas, 2011).

Construct a Training Dataset

For all TCGA projects involved in our study, the GDC Data Portal (https://portal.gdc.cancer.gov/) already provided open-access DNA somatic mutations detected by four different callers MuSE, MuTect2, SomaticSniper, and VarScan (Ellrott et al., 2018) with stringent thresholds. Using the GDC MAF Concatenation Tool (https://github.com/wwysoc2/gdc-maf-tool), we combined the curated mutations from four callers, and constructed a union set of all available DNA somatic mutations for each cancer type (LUAD, LUSC, BLCA, and GBM) to maximize the sensitivity. In addition, given that GDC somatic variant calling pipeline had strict criteria leading to the loss of some true positive somatic mutations, we called our own DNA somatic mutations using raw sequencing data and retrieved GDC-omitted DNA somatic mutations.

Normally, variations in DNA will be passed and presented in RNA through transcription. Reciprocally, any RNA somatic mutations presented in DNA should be true positive since they have got evidence from DNA. Moreover, other RNA somatic mutations lacking support from DNA will be regarded as true negative. To construct a reliable training dataset for model training, we split our RNA somatic mutations into three categories (**Figure 2**) based on evidence from the GDC database and GDC-omitted DNA somatic mutations. Finally, based on the information from FilterMutectCalls output and annotation information of Funcotator, we systematically extracted features for each training record with three categories: variant, genotype, and annotation levels (**Supplementary Table S1**).

Performance Metrics

Due to the extreme distribution bias for true positive and true negative classes (TP : TN = 1:8), our main purpose was to identify true positive RNA somatic mutations correctly. We chose precision, recall, F1 scores, and areas under the precision–recall curve (PR-AUC) as major performance metrics in our study because they are insensitive to class imbalance. Other metrics derived from the confusion matrix (**Table 1**) were also introduced for evaluation.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}},$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}},$$

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$\text{False positive rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}},$$

$$\text{False negative rate} = \frac{\text{False Negative}}{\text{False Negative} + \text{True Negative}},$$

$$\text{True negative rate} = \frac{\text{True Negative}}{\text{False Negative} + \text{True Negative}}.$$

Model Training and Validation

Records within the training dataset were split into training and testing subsets (9:1). We utilized the training subset for model parameter tuning, feature selection, and model training. For the testing subset, we utilized them for testing the model's generalization performance.

To handle the imbalanced distribution for TP and TN classes, we chose a weighted random forest classifier (RandomForestClassifier, scikit-learn 0.24.2) to reduce the bias

TABLE 1 | Confusion matrix demonstration.

	Predicted condition		
	Label	**Positive**	**Negative**
True condition	Positive	True positive	False negative
	Negative	False positive	True negative

by assigning inversely proportioned weights to different classes (Zhu and Pierskalla, 2016). First, we utilized recursive feature elimination with 10-fold cross-validation (RFECV, scikit-learn 0.24.2) to select optimal features. Second, we utilized a 10-fold cross-validated grid-search over a parameter grid (GridSearchCV, scikit-learn 0.24.2) to fine-tune optimal parameters (max_depth, min_samples_split, min_samples_leaf, max_features, etc.). Finally, we constructed a machine learning model for RNA somatic mutation identification with optimal features and parameters, and applied it in testing subset to assess its generalization performance.

Following the procedures mentioned earlier, we conducted somatic single nucleotide variants calling in LUSC, BLCA, and GBM cohorts, utilized a multi-filtering strategy and built validation datasets based on extracted features. We applied our discriminant model in these validation datasets and retrieved assessing metrics to further demonstrate the generalization performance.

Also, we validated the necessity of introducing a new training dataset from another cancer type. We added the GBM dataset into the initial training dataset and constructed a new random-forest classifier. After retrieving and assessing metrics for the new random-forest classifier within LUSC and BLCA independent validation datasets, we compared them with our initial model's performance.

Model Interpretation and Visualization

We utilized impurity-based feature importance for tree-based machine learning models to help interpret features' contributions within our model. The higher its contribution, the more important the feature. Impurity-based feature importance (Gini importance) is computed as the total reduction of the criterion brought by that feature and retrieved through our model's attribute feature_importances_. Because traditional feature importance mainly focused on overall model interpretation, we also introduced the SHAP (SHapley Additive exPlanations, https://github.com/slundberg/shap) (Lundberg and Lee, 2017) python package to help visualize prediction (Lundberg et al., 2018) and provide local explanations (Lundberg et al., 2020). We provided feature contributions calculated by SHAP for predicted probability and conducted a single prediction's visualization by invoking the force_plot function. We also investigated the feature contributions of the training dataset. We calculated and visualized the sum of SHAP value magnitudes by summary_plot function in SHAP to show the distribution of each feature's impacts on the model output (lift or lower prediction probability).

Whole Framework Implementation

We built our whole framework using Snakemake (Köster and Rahmann, 2012) and class-oriented python scripts. Snakemake (https://github.com/snakemake/snakemake) was applied to manage standard bioinformatic workflows involved in this study (co-cleaning, calling, and annotation) and conduct task auto-management without complicating shell scripts. Function-oriented python scripts contained feature extraction, model training and testing, and model utilizing function. Both Snakemake-based workflows and python scripts were available within our project repository (https://github.com/pmglab/RNA-SSNV), which helped create reproducible analysis.

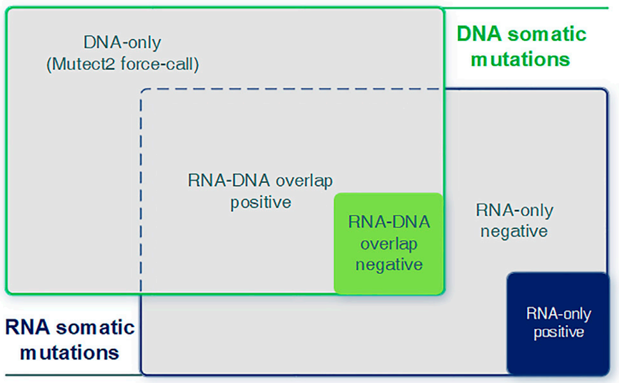


FIGURE 3 | Graphical introduction for the DNA-only, DNA–RNA overlap, and RNA-only parts. Graphical introduction for detailed combination of RNA and DNA somatic mutations. DNA-only: DNA somatic mutations not detected (expressed) in RNA. RNA–DNA overlap: somatic mutations detected in both RNA and DNA. RNA-only: RNA somatic mutations without any DNA evidence.

Analyze RNA Mutations With DNA Evidence

We integrated predicted RNA somatic mutations with known DNA mutations to analyze the relevance between RNA and DNA. We examined their intersectionality and split them into three parts (RNA–DNA overlap, DNA-only, and RNA-only) and two sub-categories (positive and negative class, **Figure 3**). Each part and sub-category had its biological implication and interpretation requiring further investigation. The RNA–DNA overlap part stood for RNA mutations with DNA evidence support. DNA-only part stood for DNA mutations not detected in RNA, and we utilized the Mutect2 force-call mode to inspect their coverage status in RNA. RNA-only part stood for RNA mutations not detected in DNA, and most of them were artifacts due to lack of DNA evidence or low sequence qualities.

Cancer driver genes were under positive selection during tumorigenesis (Martinez-Jimenez et al., 2020). Here, we focused on cancer-specific driver genes (https://www.intogen.org/) to explore their enrichment patterns (number distribution, functional impact, and therapeutic power) between expressed (RNA–DNA overlap part) and un-expressed (DNA-only part) somatic mutation panels. For pathogenicity prediction, Combined Annotation–Dependent Depletion (CADD) (Rentzsch et al., 2019), Eigen Principal Components (Eigen-PC) (Ionita-Laza et al., 2016), Polymorphism Phenotyping version 2 (PolyPhen-2) (Adzhubei et al., 2010), Protein Variation Effect Analyzer (PROVEAN) (Choi et al., 2012), UMD-Predictor (Ioannidis et al., 2016), Rare Exome Variant Ensemble Learner (REVEL) (Frederic et al., 2009), and Sorting Intolerant From Tolerant (SIFT) (Ng, 2003) were top-performing prediction tools on somatic variants

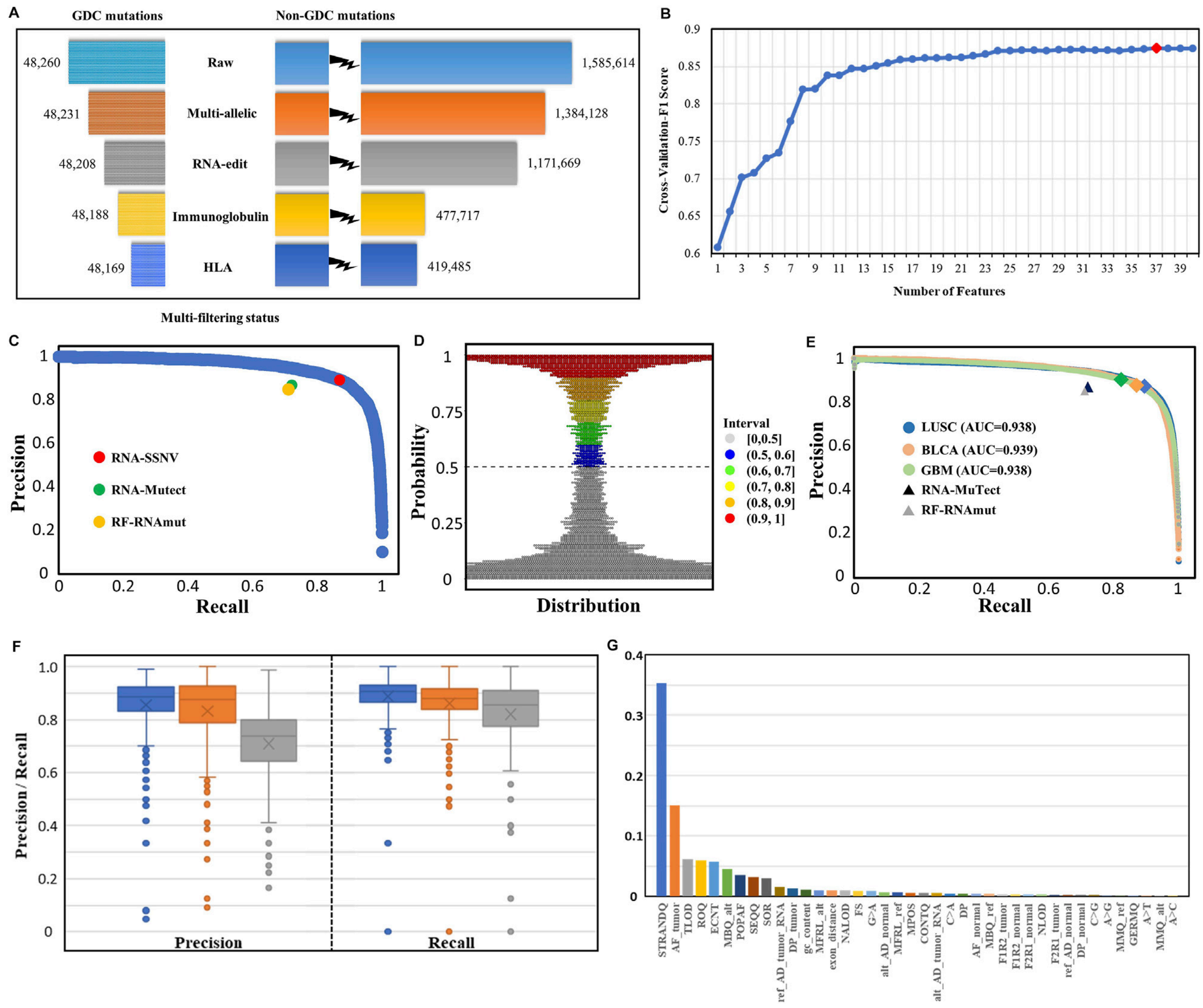


FIGURE 4 | Multi-filtering strategy and machine-learning model performance in testing and validation datasets. **(A)** Loss of GDC mutations (true positive) and non-GDC mutations after the removal of multiallelic, RNA-editing, immunoglobulin, and HLA sites. **(B)** Change in cross-validated F1 score with the number of features decreasing using the Recursive Feature Elimination with Cross-Validation (RFECV) method. Initial number of features was 40 and each iteration removed one least important feature. **(C)** P–R (blue) curve for the testing dataset. RNA-SSNV achieved 0.880 precision and 0.884 recall rate (red point) in the testing dataset under the default 0.5 threshold. RNA-Mutect (green point) and RF-RNAmut (orange point) had reported precision–recall with 0.87–0.72 and 0.85–0.71, respectively. **(D)** Probability distribution of the predicted scores for the testing dataset. Most somatic mutation records were at the upper or lower ends of the plot, conforming a clear classification boundary. **(E)** P–R curves for independent validation datasets. P–R curves for LUSC (blue), BLCA (orange), and GBM (green) had identical 0.94 AUC. The peaks meant slightly different P–R performances for our model using the default 0.5 threshold in three datasets: LUSC (0.872–0.894), BLCA (0.876–0.870), and GBM (0.902–0.825). P–Rs for RNA-Mutect and RF-RNAmut were also used for comparison. **(F)** Precision and recall distribution for each case across three types of cancer (LUSC, BLCA, and GBM). Box plots showed median, 25th and 75th quantiles, outliers were presented as dots. **(G)** Relative importance distribution for each feature. Gini impurity-based feature importance values were normalized to sum to one.

(Suybeng et al., 2020). Thus, we used the dbNSFP v4.1a (Liu et al., 2020) database to annotate missense mutations with the aforementioned prediction scores. The chi-squared test was used to calculate the significance (p-value) of enriched distribution and odds ratio (OR). A two-sided independent t-test was used to determine the significance (p-value) of the difference between the means of two prediction groups.

We also conducted an analysis of transcriptome-wide allele-specific expression (ASE) to identify ASE events in somatic mutations and their impacts on gene expression which affected carcinogenesis. We chose cases containing both tumor and paired-normal RNA-seq data from LUSC and BLCA cohorts (LUSC: 49 cases, BLCA: 19 cases), and curated their gene expression profiles from the UCSC Xena database (https://xena.ucsc.edu/). Then, we chose only heterozygous SNVs in both tumor RNA-seq and WES data (RNA–DNA overlap part), and implemented chi-squared tests on the RNA and DNA allelic depths with a significance cutoff of p-value 0.01 to

TABLE 2 | Confusion matrix for the holdout testing dataset.

		Predicted condition	
	Label	Positive	Negative
True condition	Positive	4,165	546
	Negative	566	41,129

identify somatic SNV ASEs (Heap et al., 2010; Liu et al., 2016). Finally, we compared the TPM value of tumor and paired-normal samples of cases harboring the somatic SNV ASEs to examine the alteration of total gene expression, and defined the TPM fold change (FC) of 2 and 1/2 as the thresholds of upregulated and downregulated genes (Liu et al., 2018).

RESULTS

General Performance of the Framework

After the initial RNA somatic mutation calling and multi-filtering step, we collected 467,654 mutations in the LUAD training dataset and 721,234, 323,323, and 126,449 mutations in LUSC, BLCA, and GBM independent validation datasets, respectively. To evaluate the effectiveness of multi-filtering strategy, we validated the loss of GDC mutations in the LUAD training dataset (**Figure 4A**) and LUSC, BLCA, and GBM independent validation datasets (**Supplementary Figure S1**). We found that the loss was negligible (0.1%), whereas the reduction of possible artifact calls was rather significant (70%); such preprocessing guaranteed a relatively pure mutation set for training and predicting. Furthermore, our framework's built-in machine learning model was trained and fine-tuned by 10-fold cross-validation. In total, 37 features from three categories were kept for model training after feature selection conducted in the initial 40 features (**Figure 4B**). Finally, our framework achieved 88.0% precision and 88.4% recall rate within the testing dataset (**Figure 4C**), and other assessing metrics (**Table 2**) were also satisfying. For example, the false-positive rate was 0.014, the false-negative rate was 0.013, and the true-negative rate was 0.987. Moreover, most RNA somatic mutations were at the upper or lower ends of the bay plan plot according to the predicted probability distribution of the testing dataset (**Figure 4D**), which suggested a clear classification result.

To inspect the generalization performance of our framework, we applied our RNA somatic mutation discriminant model to three independent validation datasets. As a result, RNA-SSNV successfully discriminated GDC high confidence somatic variants from WES-targeted coding RNA mutations with significantly higher precision, recall, and PR-AUC (LUSC P–R: 0.872–0.894, BLCA P–R: 0.876–0.870, and GBM P–R: 0.902–0.825, **Figure 4E**), compared with other RNA somatic detection tools such as RNA-Mutect (Yizhak et al., 2019) (precision: 0.87, recall: 0.72) and RF-RNAmut (Muyas et al., 2020) (precision: 0.85, recall: 0.71). Specially, RNA somatic mutations within cancer-specific driver genes had better performance (LUSC P–R: 0.924–0.921, BLCA P–R: 0.929–0.896, and GBM P–R: 0.921–0.883) and they had higher coverages than total RNA somatic mutations (median sequencing coverages—LUSC overall: 42, driver: 60, two-sided independent *t*-test *p*-value: 1.06e-7; BLCA overall: 41, driver: 44, *p*-value: 7.35e-7; GBM overall: 46, driver: 76, *p*-value: 1.12 e-8). Thus, critical mutations within cancer driver genes can be reliably identified in RNA-seq data, which also guarantees our framework's clinical value.

For case-level performance, as expected, LUSC and BLCA retained a median precision of 0.885 and 0.876 across cases, but GBM only reached 0.739 median precision (**Figure 4F**), contradicting its general precision of 0.902. Such contradiction was caused by four high-mutation-rate (harbored more than 100 DNA mutations) cases having high precision (>0.950). In contrast, most GBM cases had extremely low somatic mutation rates with less than 30 DNA mutations transcribed in RNA. Thus, some less identifiable RNA editing events and novel mutations rescued by RNA can easily twist GBM's case-level precision but are hard to affect GBM's general precision. In addition, LUSC, BLCA, and GBM reached a median recall of 0.905, 0.880, and 0.857, concordant with their general recall. Also, RNA somatic mutation counts were highly correlated with DNA (Pearson correlation coefficient: LUSC: 0.905, BLCA: 0.937, and GBM: 0.607, **Supplementary Figure S2**) after excluding outlier cases with extreme mutation counts, suggesting the high accuracy of our framework.

We investigated the contributions of 37 features using an importance plot based on Gini impurity (**Figure 4G**) which showed that STRANDQ was the most important feature for discriminating RNA somatic mutations, followed by AF_tumor, TLOD, ROQ, and ECNT with nontrivial feature importance scores. In addition, features containing other sequencing qualities and population allele frequencies also played a role in prediction because they represented mutations' reliability and germline evidence. We found that the prevalent RNA editing allelic changes "A>G" came at the bottom of the importance list, which indicated that our multi-filtering strategy adequately removed these editing sites and reduced their influence. Furthermore, we, in detail, illustrated the effects of 37 features on the prediction model by SHAP (Muyas et al., 2020) and ascertained whether their variations lowered or lifted the predicted probability (**Supplementary Figure S3**). After feature selection, we excluded "A>C," "A>T," and "MMQ_alt" features. Among all allelic change features, "A>G," "C>A," "C>G," and "G>A" were retained. Out of which, "A>G" and "G>A" represented A-to-I (Wang et al., 2021) and C-to-U (Lerner et al., 2019) RNA editing events, and their existence had negative impacts on the model output. On the contrary, "C>A" and "C>G" represented RNA-editing exclusive allelic changes that exhibited positive impacts. Interestingly, we also found that high tumor allele depth for reference base and alternative base had opposite impacts, which indicated that RNA somatic mutations with high reference allele depth or low alternative allele depth in the tumor sample tended to be artifacts.

Applications

Evaluation With Known DNA Evidence

We compared RNA-level somatic mutations with DNA-level to investigate the biological mechanisms for their intersection and uniqueness. As a result, we made a tabular overview (**Table 3**) and Venn diagrams (**Supplementary Figure S4**) to illustrate detailed

TABLE 3 | Overview of RNA somatic mutations combined with DNA.

Cancer type	RNA initial	RNA DNA overlap		RNA only		DNA only	P–R
		Positive	Negative	Positive	Negative		
LUSC	721,234	49,527	5,873	6,963	658,871	105,644	0.877–0.894
BLCA	323,323	43,945	6,557	6,206	266,615	71,614	0.876–0.870
GBM	126,449	9,153	1,947	970	114,379	16,104	0.904–0.825

Notes: Cancer type—LUSC: lung squamous cell carcinoma, BLCA: bladder urothelial carcinoma, GBM: glioblastoma multiforme.
DNA only—Counts of mutations only observed in the GDC DNA mutation set (not in RNA).
RNA–DNA overlap—Counts of mutations observed in both GDC DNA mutation set and RNA mutation set.
RNA only—Counts of mutations only observed in the RNA somatic mutation set (not in DNA).
RNA total—Counts of mutations observed in the total RNA somatic mutation set.
P–R—Precision–recall metric for RNA somatic mutations with GDC mutations as a golden standard dataset.

distribution for the combination of RNA and DNA-level somatic mutations. Here, our framework successfully identified authentic mutations from the RNA-only part (which got ignored/not covered in WES data) to increase information gain and improve diagnostic yield. For all three parts, the RNA-only part had the largest mutation counts. The vast majority were labeled as negative (97.7–99.2%), indicating that our framework had successfully identified most artifacts in RNA because all these negative calls shall be filtered in the final output. Interestingly, when comparing mutation counts of the DNA-only part with the RNA–DNA overlap part, we found that less than 1/3 DNA somatic mutations got expressed in RNA. Such phenomenon was concordant with another study, mainly due to insufficient sequence coverage in low-expression or un-expression genes (Yizhak et al., 2019). Further analysis was listed in the following section for elaborate explanations.

Variably RNA-Expressed Mutations Harbored a Special Enrichment Pattern

In detail, we explored the RNA expression ratios (number of expressed DNA somatic mutations/number of all DNA somatic mutations) for each case of three cancer types (**Figure 5A**), median expression ratios for LUSC, BLCA, and GBM were 0.312, 0.349, and 0.256, respectively. Highly variable expression ratios (0.000–0.632) in three types of cancer suggested that different DNA somatic mutations had various expression statuses in RNA. Notably, although the brain has a high number of expressed genes than other human tissues (Naumova et al., 2013), expression ratios of GBM were still significantly lower than those of LUSC or BLCA. These results indicated that DNA somatic mutations might be variably expressed or not expressed at all, and RNA somatic mutations were important to evaluate possible expression status.

To investigate whether the RNA-expressed somatic mutations tended to have larger functional impacts than those that only existed in DNA, probably resulting from the positive selection of cancer subclonal, we compared the impact scores of mutations within cancer-specific driver genes (Martinez-Jimenez et al., 2020) between RNA–DNA overlap and DNA-only parts. Interestingly, the cancer driver genes' mutations were enriched in the RNA–DNA overlap part (LUSC: OR = 2.01, p = 1.14 e-68, BLCA: OR = 2.57, p = 9.89 e-119, GBM: OR = 2.70, p = 1.73 e-16. **Supplementary Table S2**), even though the DNA-only part had excessive mutation counts than the RNA–DNA overlap part (DNA-only/RNA–DNA overlap: ~2/1). Moreover, we compared the predicted pathogenicity scores for missense mutations located within cancer driver genes between RNA–DNA overlap and DNA-only parts, and found that all RNA–DNA overlap parts had significantly higher pathogenicity scores across three cancer types and seven prediction tools (p-value < 1 e-5, **Figure 5B**). The significantly higher prediction scores implied that predicted damaging mutations tended to be selectively expressed in driving tumorigenesis, and our RNA-level somatic mutation identification framework effectively enriched the functional mutations.

Furthermore, we want to explore whether actionable mutations tend to get expressed in RNA and exhibit clinical effects. Thus, we assessed the therapeutic power for mutations in cancer driver genes between RNA–DNA overlap and DNA-only variants using the OncoKB database (Chakravarty et al., 2017) (https://www.oncokb.org/, **Supplementary Table S3**). Therapeutic sites within the RNA–DNA overlap part were far more than DNA-only across three cancer types (LUSC: OR = 13.34, p = 8.35e-19, BLCA: OR = 3.27, p = 4.26e-16, GBM: OR = 4.26, p = 3. 66e-4, **Table 4**), indicating that the RNA-level somatic mutations calling can enrich clinical therapeutic variants. Notably, we observed that some therapeutic mutations from the OncoKB database also occurred in the DNA-only part. For example, except for 52 RNA–DNA overlap somatic mutations in BLCA, PIK3CA also had 12 DNA-only somatic mutations with "Level_3B″ OncoKB annotation (Chakravarty et al., 2017). We found that even if the 12 TCGA BLCA cases containing 12 DNA-only somatic mutations had sufficient expression level for the PIK3CA gene (TPM: 23.7–51.7, curated from UCSC Xena(Goldman et al., 2020) dataset), the 12 mutations' alternative allele still got un-expressed (median alt allele-depth: 0) leading to unlikely benefit from certain targeted therapies. Therefore, although PIK3CA is a valuable therapeutic target for inhibitors of PI3K/AKT/mTOR pathways in advanced bladder cancer (Ross et al., 2016; Willis et al., 2020), the detailed expression status of the mutations should be carefully evaluated when the targeted therapy is considered. Such phenomenon was opposed to the assumption that mutations located within sufficiently

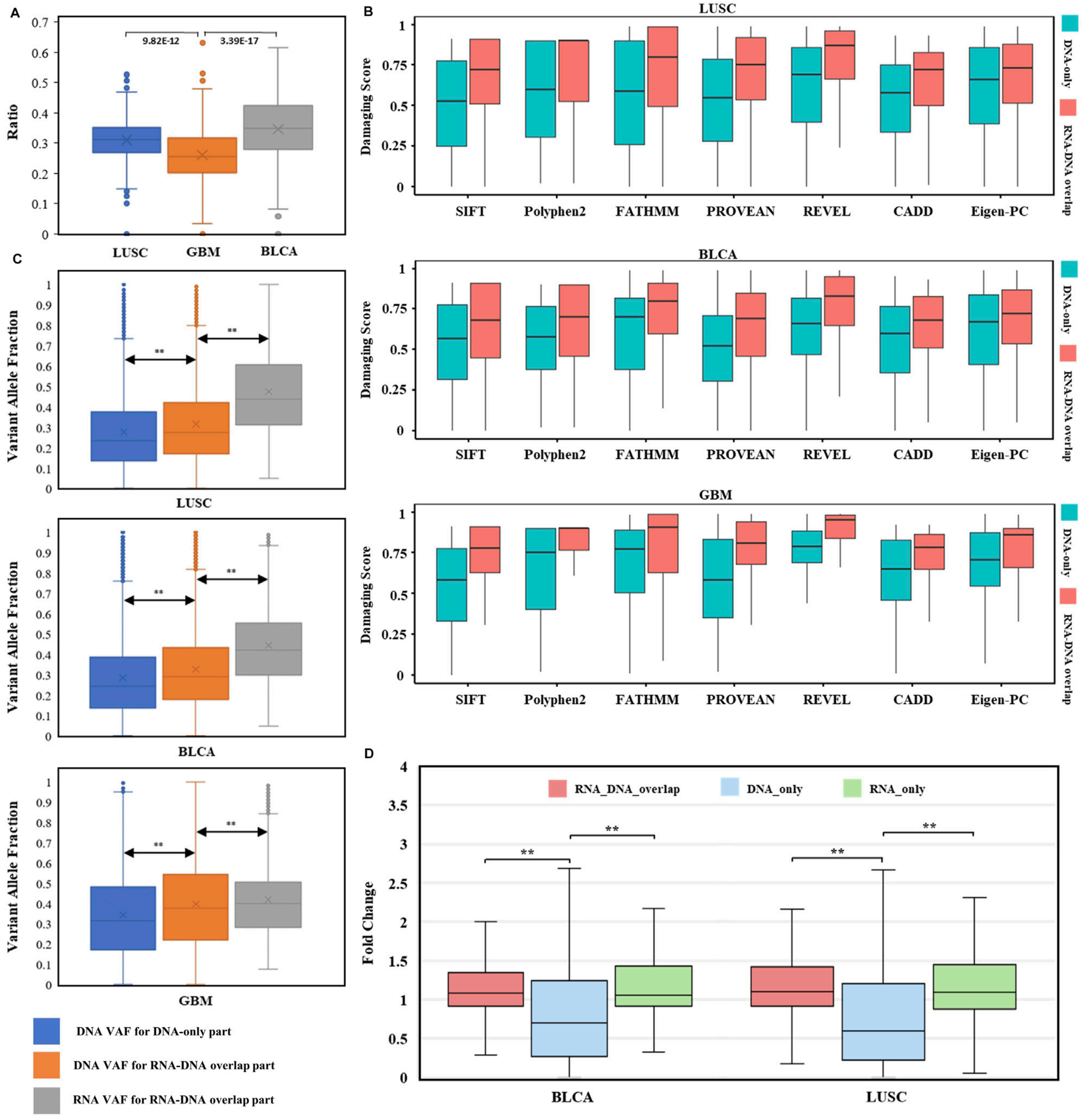


FIGURE 5 | Evaluation of RNA somatic mutations and integrative analysis with DNA evidence. **(A)** Distribution of RNA expression ratios for known DNA somatic mutations across three types of cancer. Box plots' heights ranged from 0.000 to 0.632. The comparisons utilized two-sided independent *t*-test with *p*-value < 1e-5. **(B)** Distributions of seven pathogenicity prediction scores for missense mutations within cancer driver genes across three cancer types (LUSC, BLCA, and GBM). DNA-only and RNA–DNA overlap parts in each cancer type were used for comparison (all comparisons passed two-sided independent *t*-test with *p*-value < 1e-5). **(C)** Variant allele fraction (VAF) distributions of DNA-only and RNA–DNA overlap parts within three cancer types. Left box: VAF distribution for DNA somatic mutations in DNA-only part. Middle box: VAF distribution for DNA somatic mutations in the RNA–DNA-overlap part. Right box: VAF distribution for RNA somatic mutations in the RNA–DNA-overlap part. The comparisons utilized two-sided independent *t*-test with *p*-value < 1e-5. **(D)** TPM fold change (FC) distributions for BLCA and LUSC. The comparisons utilized the Wilcoxon rank-sum test with *p*-value < 1e-5.

expressed genes had undoubtful effects making them potential therapeutic targets, and RNA-level mutations were required to validate these targets' transcription status.

Given that RNAs were enriched with mutations of higher functional impact and therapeutic value, we assessed the performance of RNA-level somatic mutations for discovering

TABLE 4 | Overview of therapeutic mutation distribution in three types of cancer.

Therapeutic level	LUSC		BLCA		GBM	
	RNA–DNA overlap	DNA only	RNA–DNA overlap	DNA only	RNA–DNA overlap	DNA only
Level_1 FDA-approved drug	0	0	43	1	0	0
Level_2 standard care	0	0	0	0	1	0
Level_3 clinical evidence	58	3	140	40	13	1
Level_4 biological evidence	85	10	96	22	28	7
Counts sum	143	13	279	63	42	8
Total	1,240	1,333	1,565	1,014	175	116
OR (*p*_value)	13.34 (8.35 e-19)		3.27 (4.26 e-16)		4.26 (3.66 e-4)	

Notes: Level_1: FDA-recognized biomarker predictive of response to an FDA-approved drug.

Level_2: Standard care biomarker recommended by the NCCN predictive of response to an FDA approved drug.

Level_3–3A: Compelling clinical evidence supports the biomarker as being predictive of response to a drug; 3B: standard care or investigational biomarker predictive of response to an FDA-approved or investigational drug.

Level_4: Compelling biological evidence supports the biomarker as being predictive of response to a drug.

Counts sum: Sum of therapeutic mutation counts.

Total: Total counts for mutations located within cancer-specific driver genes.

cancer driver genes by other statistical methods. Here, WITER (Jiang et al., 2019) was adopted to test the enrichment of somatic mutations due to positive selection in tumorigenesis (Jiang et al., 2019; Martinez-Jimenez et al., 2020). We compared the significant genes based on RNA-level somatic mutations to those based on the DNA-level somatic mutations in three cancer datasets. Among all significant genes ($FDR < 0.1$), the RNA somatic mutations led to a higher proportion of known cancer driver genes from the Intogen database (Martinez-Jimenez et al., 2020) in two of the three datasets than DNA (LUSC: 6/7 *vs*. 6/9 and GBM: 5/5 *vs*. 5/18, see details in **Supplementary Tables S4, S5**) with identical cancer-driver genes, another cancer type (BLCA: 12/18 *vs*. 15/19) also had a similar proportion. This result suggested that the RNA-level may lead to fewer false-positive estimations for driver genes than DNA-level.

In addition to known cancer driver genes, other significant genes based on the RNA-level somatic mutations, though un-registered in the Intogen database, were also functionally important to cancer development. CTNNB1, for example, had a significant *q*-value of 0.081 in BLCA. CTNNB1's mutations have been found to cause aberrant WNT/CTNNB1 signaling and are associated with the susceptibility and prognosis of breast, endometrial, and gastric cancers (Kurnit et al., 2017; van Schie and van Amerongen, 2020; Wang et al., 2012). CHEK2 ($q = 0.087$ in LUSC, $q = 0.052$ in BLCA) played an important role in the repair of DNA damage, and its heterozygous mutations had been found to be causing genetic susceptibility to lung cancer (Wang et al., 2014) and bladder cancer (Złowocka et al., 2008). Although our detected CHEK2 somatic mutations were not inherited or passed on, their heterozygosity was similar and induced cancer risk. In a word, RNA can also prioritize potential cancer driver genes.

RNA Increased Mutation Detection Power.

VAF (variant allele fraction) was the fraction of sequencing reads harboring the mutation when performing NGS (Friedlaender et al., 2021), measuring the subclonal prevalence of specific mutations (Benard et al., 2021). We compared the DNA VAF distribution for DNA-only and RNA–DNA overlap parts within three cancer types to examine the subclonal selection advantage for expressed mutations. Higher DNA VAF was observed in expressed DNA somatic mutations (**Figure 5C** left comparison, $p < 1$ e-5), indicating the trend of cancer evolution for subclonal harboring RNA somatic mutations. Interestingly, RNA VAF was significantly higher than DNA VAF within expressed mutations of RNA–DNA overlap part (**Figure 5C** right comparison, $p < 1$ e-5), suggesting an expression tendency for the mutant allele. The common cancer WES study has a mutation limit of detection (LoD) at 5% VAF, and reporting these subclonal mutations incurs the risk of sequencing error–induced false positives (Yan et al., 2021). For these low-VAF (<0.05) DNA somatic mutations, their RNA VAFs were much higher, with median values of 0.374 in LUSC, 0.342 in BLCA, and 0.241 in GBM. Therefore, RNA somatic mutations exhibited subclonal selection superiority and increased the power for low-VAF mutation detection.

Here, we, in detail, demonstrated the recovery of DNA-omitted mutations for our framework. For the RNA-only part, we found that our framework helped rescue ~10% of mutations (**Table 3**) which were missed based on DNA sequencing data. Most of the rescued mutations had low alternative allele depth (median: 0–1) or alternative allele fraction (median: 0–0.03) in WES data but opposite situations (median alt allele depth: 8–10, median alt allele fraction: 0.31–0.67) in RNA-seq data. There were also 102, 120, and 8 mutations located within cancer driver genes out of 6,997, 6,233, and 969 positive mutations from LUSC, BLCA, and GBM, respectively (**Supplementary Table S6**). Furthermore, we discovered biologically important cancer variants within these overlooked "driver" mutations using the DoCM database (Ainscough et al., 2016) (http://docm.info). We found that 17 out of 102, 14 out of 120, and 2 out of 8 DNA-overlooked "driver" mutations in LUSC, BLCA, and GBM had literature support from one or more publications (**Supplementary Table S7**). For example, TCGA-FD-A5BS had TP53 p.R282W mutation rescued by RNA with its reference-alternative allele depth in DNA: 19-1, RNA: 17-14. The R282W mutant had been found to cause the gain of novel oncogenic functions (GOF) in p53 proteins and associate with

TABLE 5 | RNA somatic mutation within cancer-driver genes in TCGA-90-6837.

Mutation	Gene	RNA		DNA		Protein change
		RefDepth	AltDepth	RefDepth	AltDepth	
chr4:186633790 T>C	*FAT1*	4	27	95	0	K1406R
chr8:116866708 G>A	*RAD21*	45	36	36	0	L8F
chr12:49051078 C>A	*KMT2D*	12	7	73	1	E869*
chr17:7673793 G>C	*TP53*	36	64	23	0	A276G
chr19:33026624 G>A	*RHPN2*	18	6	87	0	T65I
chr22:41178035 G>A	*EP300*	78	51	54	0	Q2108Q

poorer cancer outcomes with a more prominent GOF effect (Zhang et al., 2016).

Low tumor purity can bias somatic mutation detection with the positive correlation between mutation numbers and tumor purities (Cheng et al., 2020). For example, TCGA-90-6837 in LUSC with its CPE (Aran et al., 2015) (consensus measurement of purity estimations) lower than average (0.56 vs. 0.68) had no official DNA mutation (WES failed to detect), we investigated its RNA somatic mutations identified by our framework to confirm its mutational status. We found that out of its 192 RNA somatic mutations, six mutations fell within cancer driver genes, and their existence had been ignored by WES (**Table 5**). Among these mutations, KMT2D is a lung tumor suppressor gene (Alam et al., 2020), and its mutation was one of the most significant prognostic factors in LUSC(Ardeshir-Larijani et al., 2018). We found that KMT2D p.E869* mutation could cause its truncation leading to tumor progression. In addition, TP53 p.A276G mutation had been found to locate within the DNA binding domain of the TP53 protein and presumably have deleterious impacts on protein functions (Chang et al., 2021) with pathogenic ClinVar database (Landrum et al., 2020) interpretation (Accession: VCV000185319.3). These findings confirmed that RNA-seq data could provide valuable supplementary information useful for clinical decisions and improve diagnostic yield in extreme cases when DNA failed to detect actionable mutations.

Transcriptome-Wide Allele-Specific Expression Analysis

We calculated the TPM fold change (FC) to measure gene differential expression status. After excluding infinite FC values, we found that the median gene FC for RNA-expressed mutations was significantly higher than unexpressed mutations (**Figure 5D**). Thus, genes harboring RNA-expressed somatic mutations tended to have higher expression level in tumor samples than in paired normal samples.

We detected somatic SNV-level ASEs, and found that 24.8% of 3876 and 23.2% of 1700 somatic mutations exhibited ASE events in LUSC and BLCA RNA–DNA overlap parts. As expected, most (~90%) ASE somatic mutations had over-expressed mutant alleles. The results showed that certain expressed somatic mutations had higher expression superiority in the mutant allele than the wild allele, which further enhanced the mutation detection power in RNA. Furthermore, we curated gene lists for 10 signaling pathways in cancer (Sanchez-Vega et al., 2018) and explored the functional alteration on signaling pathways for ASE somatic mutations. Ideally, if the ASE somatic mutation is functional, the direction of ASE event for the mutant allele should be the same as the direction of gene expression alteration for tumor vs. paired-normal samples (Liu et al., 2018). Thus, we mapped ASE somatic mutations to genes involving cancer signaling pathways with identical expression change direction. Finally, we identified several pathways (cell cycle, HIPPO, RTK RAS, TGF-Beta, and WNT) containing heavily altered genes with ASE events (**Supplementary Table S8**). Interestingly, seemly "benign" synonymous mutations also contained ASE events and altered gene expression level. For example, NF1 is a tumor suppressor that negatively regulates RAS signaling (Redig et al., 2016). NF1 p.L43L mutation in TCGA-39-5040 had an over-expressing mutant allele (DNA VAF: 0.32, RNA VAF: 0.63) and showed an upregulated gene expression (tumor/paired-normal fold change: 2.53), which activated NF1 function to under-regulate the RAS signaling pathway and suppressed carcinogenesis.

DISCUSSION

Although common somatic mutation detection practices come with WES, important and actionable mutations are often conserved in RNA-seq. Therefore, we developed RNA-SSNV, an integrative framework to identify RNA somatic single nucleotide variants called within tumor RNA-seq and paired-normal WES data. To maximize performance, we combined multi-filtering strategies and a machine-learning model. For the multi-filtering strategy, we found that it removed massive artifacts (~70%) while omitting few true positive calls (~0.1%). Before constructing the classification model, we also evaluated the performance of the GATK-recommended filtering tool (FilterMutectCalls) for the LUAD training dataset and LUSC, BLCA, and GBM validating datasets using precision–recall metrics. The result showed that FilterMutectCalls achieved a satisfying recall but a low precision rate (LUAD P–R: 0.380–0.865, LUSC P–R: 0.399–0.871, BLCA P–R: 0.442–0.886, and GBM P–R: 0.540–0.881), which may lead to large false-positive calls. Because FilterMutectCalls was originally designed based on DNA somatic mutation filtering strategy, which may not be fully compatible with RNA, we adopted a machine learning model with comprehensive features to conduct classification. For model training, we adopted various techniques to ensure its reliability. To construct a high-quality training dataset, we used GDC DNA mutations as the golden standard and self-called DNA mutations as important supplementary information to separate pure true positive and true negative sets from multi-filtered RNA mutations. In a comparison of using two data sources (RNA mutations and golden-standard DNA mutations) to construct the training dataset, the introduction of self-called DNA mutations significantly improved our machine model's performance (increased precision–recall from 0.843-0.875 to current 0.883–0.885 by 4%). We also conducted feature selection and fine-tuning to improve the model's performance. Eventually, our trained model achieved superior performance of 88.0% precision and 88.4% recall rate in the testing dataset compared with other state-of-art RNA somatic mutation detection tools such as RNA-Mutect (Yizhak et al., 2019) (precision: 0.87, recall: 0.72) and RF-RNAmut (Muyas et al., 2020) (precision: 0.85, recall: 0.71).

When utilized in independent validation datasets (TCGA LUSC, BLCA, and GBM), RNA-SSNV achieved similar performance as in the testing dataset, which had 0.871–0.895, 0.876–0.871, and 0.902–0.830 precision–recall rate, respectively. Not only can our framework reliably detect RNA somatic mutations, but it also can conduct pairwise analysis with provided DNA mutations. Although our framework achieved satisfying performance within somatic RNA single-nucleotide variants' identification, limited scenarios in which only RNA somatic mutations can be retrieved such as the GTEx project (Lonsdale et al., 2013) (contained RNA-seq data from ~6700 samples across 29 normal tissues). Common RNA-seq practices involving research always included DNA-seq data which generated somatic DNA mutations simultaneously; thus, the investigation for the relationship between DNA-level and RNA-level somatic mutations was essential. Multiple studies have found that combining DNA-level and RNA-level somatic mutation can achieve maximum performance for mutational investigation (Krug et al., 2018; Newman et al., 2021; Wilkerson et al., 2014; Zhang et al., 2020). Thus, we split DNA and RNA somatic mutations into three parts: DNA–RNA overlap part, DNA-only part, and RNA-only part; and each part had positive and negative sub-parts representing our model's classifications. The DNA–RNA overlap part represented orthogonal validated DNA and RNA mutations; its positive sub-part contained reliable cancer somatic mutations with clinical usage, but its negative sub-part contained false-negative calls misclassified by our model. When using SHAP to analyze these false-negative calls (**Supplementary Figure S5**), we found that G>A mutant status had significant impacts, which implicated that APOBEC-mediated C-to-U RNA editing events (Lerner et al., 2019) contributed to misclassification and current RNA editing resources were insufficient to filter C>U editing sites. DNA-only part represented DNA mutations omitted in RNA somatic mutation calling, and we found that some DNA mutations' reference allele got selectively expressed while their alternative allele got silenced. To explore how many DNA-only somatic mutations got selectively expressed, we calculated the selective expression ratios (number of mutations with reference allelic depth>10/number of DNA somatic mutations not identified in RNA) for DNA-only parts across three cancer types (**Supplementary Figure S6**). The median mutation selective expression ratios for LUSC, BLCA, and GBM were 0.134, 0.120, and 0.154, respectively, confirming that DNA somatic mutations within GBM had higher selective expression tendency than LUSC (p = 0.003) and BLCA (p = 5.63 e-6), possibly due to innate upregulation of DNA repair mechanisms (Ferri et al., 2020). We retrieved their information in RNA using Mutect2's force-calling mode and utilized our model to classify them. Most of them were predicted negative as expected, but a small portion (1.9%) was predicted as positive, suggesting that our selected caller (Mutect2) might have a little neglection. We also observed that mutations' reference allele–specific expression within driver genes leads to doubtful translation effects. In addition, most mutations located within collagen-related genes (COL11A1, COL6A3, COL5A2, etc.) were found silenced while these genes got sufficiently expressed in RNA (**Supplementary Table S9**). Interestingly, the proteome database (Human Cancer Proteome Variation Database) also contains nearly no evidence for mutant collagen proteins across three cancer types which were abnormal because massive DNA somatic mutations had been found in these genes. The RNA-only part represented RNA mutations without DNA evidence support. Its negative sub-part was artifacts, but its positive sub-part included RNA-rescued mutations missing in DNA that contained mutations within cancer driver genes (1.4%) to provide more therapeutic targets and help with clinical decisions. A major shortcoming of WES is uneven coverage of sequence reads over the exome targets contributing to many low-coverage regions (Wang et al., 2017; Xiao et al., 2021), and substantial inter-individual variation in coverage of medically implicated genes caused false-negative mutation calls due to low coverage (Barbitoff et al., 2020; Kong et al., 2018). Although using replicate exome-sequencing can improve WES coverage by 4.3–12.7% (Cherukuri et al., 2015), improve variant calling accuracy (Zhang et al., 2014), and enhance clinical interpretation, information redundancy and excess costs limited its usage. Compared with replicate exome-sequencing, RNA-seq has improved somatic single nucleotide variants, and clinically actionable mutations are often conserved in RNA.

We also examined the potential of improving our model's performance by introducing additional training data from different cancer types. After adding GBM cancer–type data into the training dataset, we only observed a slight improvement within the testing dataset (recall rate increased 1.3%) and the AUC for P–R curves for TCGA LUSC, BLCA–independent validation datasets remained stable at 0.94 (**Supplementary Figure S7**). The unchanged performance suggested that our model trained with LUAD datasets probably has already contained key features of RNA somatic mutation in cancer cells and is applicable for other cancers. Although the general performance for our model was identical across three validation datasets, performances under default threshold (0.5) slightly differed and a dynamic shift of threshold according to different aims (prefer higher precision or recall) was required. In addition, due to insufficient C-to-U RNA editing database resources, the current model sacrificed high recall to ensure removing editing events for the G>A mutation type. The high distribution of G>A mutations (52.3%) in false-negative sets of TCGA LUSC–independent validation dataset reflected this imperfection. Therefore, we recommended that users manually review predicted-negative G>A mutations within known driver genes to improve diagnosis. To facilitate user to inspect predictions, we provided codes to visualize the contribution of important features using SHAP library and a canonical table to exhibit all useful information for user-specified records. A major limitation of our framework was that it was designed to identify RNA somatic mutations only from tumor RNA-seq and paired-normal WES data. Future works will include extending RNA somatic mutation identification scope into other sequencing data types (single-cell RNA-seq or whole-genome DNA-seq).

For cancer research involving both WES and RNA-seq data, the conventional analysis strategy uses WES data to call somatic mutations and then validates whether somatic mutations exist in RNA-seq data. However, the conventional strategy may still omit some somatic mutations in RNA-seq data. Our study significantly improved the capability to call RNA somatic mutations and further revealed the association between somatic mutations derived from RNA and DNA, providing valuable supplementary information for conventional cancer somatic mutation analysis.

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

QL: conceptualization, methodology, software, investigation, formal analysis, and writing–original draft; YY: visualization, writing–original draft; ML: conceptualization, funding acquisition, resources, supervision, and writing–review and editing.

FUNDING

This work was funded by the National Natural Science Foundation of China (31970650 and 32170637), Science and Technology Program of Guangzhou (201803010116), and Guangdong project (2017GC010644).

ACKNOWLEDGMENTS

The results shown here are in part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and its Genomic Data Commons (GDC, https://gdc.cancer.gov/).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.865313/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., and Bork, P. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods.* 7(4), 248–249. doi:10.1038/nmeth0410-248

Ainscough, B. J., Griffith, M., Coffman, A. C., Wagner, A. H., Kunisaki, J., Choudhary, M. N., et al. (2016). DoCM: A Database of Curated Mutations in Cancer. [Letter; Research Support, N.I.H., Extramural]. *Nat. Methods.* 13 (10), 806–807. doi:10.1038/nmeth.4000

Alam, H., Tang, M., Maitituoheti, M., Dhar, S. S., Kumar, M., Han, C. Y., et al. (2020). KMT2D Deficiency Impairs Super-enhancers to Confer a Glycolytic Vulnerability in Lung Cancer. *Cancer Cell* 37 (4), 599–617. doi:10.1016/j.ccell.2020.03.005

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic Pan-Cancer Analysis of Tumour Purity. *Nat. Commun.* 6 (1). doi:10.1038/ncomms9971

Ardeshir-Larijani, F., Bhateja, P., Lipka, M. B., Sharma, N., Fu, P., and Dowlati, A. (2018). KMT2D Mutation Is Associated with Poor Prognosis in Non–small-cell Lung Cancer. *Clin. Lung Cancer.* 19 (4), e489–e501. doi:10.1016/j.cllc.2018.03.005

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173 (2), 371–385. doi:10.1016/j.cell.2018.02.060

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., et al. (2020). Systematic Dissection of Biases in Whole-Exome and Whole-Genome Sequencing Reveals Major Determinants of Coding Sequence Coverage. *Sci. Rep.-UK* 10 (1). doi:10.1038/s41598-020-59026-y

Benard, B. A., Leak, L. B., Azizi, A., Thomas, D., Gentles, A. J., and Majeti, R. (2021). Clonal Architecture Predicts Clinical Outcomes and Drug Sensitivity in Acute Myeloid Leukemia. *Nat. Commun.* 12 (1). doi:10.1038/s41467-021-27472-5

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054. doi:10.1101/861054

Buhler, S., and Sanchez-Mazas, A. (2011). HHLA DNA Sequence Variation Among Human Populations: Molecular Signatures of Demographic and Selective Events. *PLoS One* 6 (2), e14643. doi:10.1371/journal.pone.0014643

Cazier, J. B., Rao, S. R., McLean, C. M., Walker, A. K., Wright, B. J., Jaeger, E. E. M., et al. (2014). Whole-genome Sequencing of Bladder Cancers Reveals Somatic CDKN1A Mutations and Clinicopathological Associations with Mutation Burden. *Nat. Commun.* 5 (1). doi:10.1038/ncomms4756

Chakravarty, D., Gao, J., Phillips, S. M., Kundra, R., Zhang, H., Wang, J., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* Alexandria, VA: American Society of Clinical Oncology. doi:10.1200/PO.17.00011

Chang, T., Chen, S., Lin, W., Huang, C., Evans, M. I., Chung, I., et al. (2021). Comparison of Genetic Profiling between Primary Tumor and Circulating Tumor Cells Captured by Microfluidics in Epithelial Ovarian Cancer: Tumor Heterogeneity or Allele Dropout? *Diagn. (Basel)* 11 (6), 1102. doi:10.3390/diagnostics11061102

Cheng, J., He, J., Wang, S., Zhao, Z., Yan, H., Guan, Q., et al. (2020). Biased Influences of Low Tumor Purity on Mutation Detection in Cancer. *Front. Mol. Biosci.* 7, 533196. doi:10.3389/fmolb.2020.533196

Chepelev, I., Wei, G., Tang, Q., and Zhao, K. (2009). Detection of Single Nucleotide Variations in Expressed Exons of the Human Genome Using RNA-Seq. *Nucleic Acids Res.* 37 (16), e106. doi:10.1093/nar/gkp507

Cherukuri, P. F., Maduro, V., Fuentes-Fajardo, K. V., Lam, K., Adams, D. R., Tifft, C. J., et al. (2015). Replicate Exome-Sequencing in a Multiple-Generation Family: Improved Interpretation of Next-Generation Sequencing Data. *BMC Genomics* 16 (1). doi:10.1186/s12864-015-2107-y

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., Chan, A. P., and de Brevern, A. G. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* 7 (10), e46688. doi:10.1371/journal.pone.0046688

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples. *Nat. Biotechnol.* 31 (3), 213–219. doi:10.1038/nbt.2514

Cirulli, E. T., Singh, A., Shianna, K. V., Ge, D., Smith, J. P., Maia, J. M., et al. (2010). Screening the Human Exome: A Comparison of Whole Genome and Whole Transcriptome Sequencing. *Genome Biol.* 11 (5), R57. doi:10.1186/gb-2010-11-5-r57

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* 6 (3), 271–281. doi:10.1016/j.cels.2018.03.002

Ferri, A., Stagni, V., and Barilà, D. (2020). Targeting the DNA Damage Response to Overcome Cancer Drug Resistance in Glioblastoma. *Int. J. Mol. Sci.* 21 (14), 4910. doi:10.3390/ijms21144910

Frederic, M. Y., Lalande, M., Boileau, C., Hamroun, D., Claustres, M., Beroud, C., et al. (2009). UMD-predictor, a New Prediction Tool for Nucleotide Substitution Pathogenicity -- Application to Four Genes: FBN1, FBN2, TGFBR1, and TGFBR2. *Hum. Mutat.* 30 (6), 952–959. doi:10.1002/humu.20970

Friedlaender, A., Tsantoulis, P., Chevallier, M., De Vito, C., and Addeo, A. (2021). The Impact of Variant Allele Frequency in EGFR Mutated NSCLC Patients on Targeted Therapy. *Front. Oncol.* 11. doi:10.3389/fonc.2021.644472

García-Nieto, P. E., Morrison, A. J., and Fraser, H. B. (2019). The Somatic Mutation Landscape of the Human Body. *Genome Biol.* 20 (1). doi:10.1186/s13059-019-1919-5

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nat. Biotechnol.* 38 (6), 675–678. doi:10.1038/s41587-020-0546-8

Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* 104 (3), 466–483. doi:10.1016/j.ajhg.2019.01.012

Heap, G. A., Yang, J. H. M., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., et al. (2010). Genome-wide Analysis of Allelic Expression Imbalance in Human Primary Cells by High-Throughput Transcriptome Resequencing. *Hum. Mol. Genet.* 19 (1), 122–134. doi:10.1093/hmg/ddp473

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* 99 (4), 877–885. doi:10.1016/j.ajhg.2016.08.016

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A Spectral Approach Integrating Functional Genomic Annotations for Coding and Noncoding Variants. *Nat. Genet.* 48 (2), 214–220. doi:10.1038/ng.3477

Jiang, L., Zheng, J., Kwan, J. S. H., Dai, S., Li, C., Li, M. J., et al. (2019). WITER: A Powerful Method for Estimation of Cancer-Driver Genes Using a Weighted Iterative Regression Modelling Background Mutation Counts. *Nucleic Acids Res.* 47 (16), e96. doi:10.1093/nar/gkz566

Kiran, A. M., O'Mahony, J. J., Sanjeev, K., and Baranov, P. V. (2013). Darned in 2013: Inclusion of Model Organisms and Linking with Wikipedia. *Nucleic Acids Res.* 41, D258–D261. Database issue). doi:10.1093/nar/gks961

Kong, S. W., Lee, I., Liu, X., Hirschhorn, J. N., and Mandl, K. D. (2018). Measuring Coverage and Accuracy of Whole-Exome Sequencing in Clinical Context. *Genet. Med.* 20 (12), 1617–1626. doi:10.1038/gim.2018.51

Köster, J., and Rahmann, S. (2012). Snakemake--a Scalable Bioinformatics Workflow Engine. *Bioinformatics* 28 (19), 2520–2522. doi:10.1093/BIOINFORMATICS/BTS480

Krug, A. K., Enderle, D., Karlovich, C., Priewasser, T., Bentink, S., Spiel, A., et al. (2018). Improved EGFR Mutation Detection Using Combined Exosomal RNA and Circulating Tumor DNA in NSCLC Patient Plasma. *Ann. Oncol.* 29 (3), 700–706. doi:10.1093/annonc/mdx765

Kukurba, K. R., and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* 2015 (11), p84970. doi:10.1101/pdb.top084970

Kurnit, K. C., Kim, G. N., Fellman, B. M., Urbauer, D. L., Mills, G. B., Zhang, W., et al. (2017). CTNNB1 (Beta-catenin) Mutation Identifies Low Grade, Early Stage Endometrial Cancer Patients at Increased Risk of Recurrence. *Mod. Pathol.* 30 (7), 1032–1041. doi:10.1038/modpathol.2017.15

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., et al. (2020). ClinVar: Improvements to Accessing Data. *Nucleic Acids Res.* 48 (D1), D835–D844. doi:10.1093/nar/gkz972

Lerner, T., Papavasiliou, F., and Pecori, R. (2019). RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-To-U RNA Editing Machinery and its Implication in Human Disease. *Genes* 10 (1), 13. doi:10.3390/genes10010013

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Lin, W., Qiu, X., Sun, P., Ye, Y., Huang, Q., Kong, L., et al. (2021). Association of IDH Mutation and 1p19q Co-deletion with Tumor Immune Microenvironment in Lower-Grade Glioma. *Mol. Ther. Oncolytics.* 21, 288–302. doi:10.1016/j.omto.2021.04.010

Liu, J., McCleland, M., Stawiski, E. W., Gnad, F., Mayba, O., Haverty, P. M., et al. (2014). Integrated Exome and Transcriptome Sequencing Reveals ZAK Isoform Usage in Gastric Cancer. *Nat. Commun.* 5 (1). doi:10.1038/ncomms4830

Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). DbNSFP V4: A Comprehensive Database of Transcript-specific Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Genome Med.* 12 (1). doi:10.1186/s13073-020-00803-9

Liu, Z., Dong, X., and Li, Y. (2018). A Genome-wide Study of Allele-specific Expression in Colorectal Cancer. *Front. Genet.* 9. doi:10.3389/fgene.2018.00570

Liu, Z., Gui, T., Wang, Z., Li, H., Fu, Y., Dong, X., et al. (2016). CisASE: A Likelihood-Based Method for Detecting Putativecis -regulated Allele-specific Expression in RNA Sequencing Data. *Bioinformatics* 32 (21), 3291–3297. doi:10.1093/bioinformatics/btw416

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The Genotype-Tissue Expression (GTEx) Project. *Nat. Genet.* 45 (6), 580–585. doi:10.1038/ng.2653

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* 2 (1), 56–67. doi:10.1038/s42256-019-0138-9

Lundberg, S. M., and Lee, S. (2017). "A Unified Approach to Interpreting Model Predictions,"in *Advances in Neural Information Processing Systems*. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Reprinted: Curran Associates, Inc.), 30, 4765–4774.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., et al. (2018). Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery. *Nat. Biomed. Eng.* 2 (10), 749–760. doi:10.1038/s41551-018-0304-0

Mansi, L., Tangaro, M. A., Lo Giudice, C., Flati, T., Kopel, E., Schaffer, A. A., et al. (2021). REDIportal: Millions of Novel A-To-I RNA Editing Events from Thousands of RNAseq Experiments. *Nucleic Acids Res.* 49 (D1), D1012–D1019. doi:10.1093/nar/gkaa916

Martinez-Jimenez, F., Muinos, F., Sentis, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., et al. (2020). A Compendium of Mutational Cancer Driver Genes. *Nat. Rev. Cancer.* 20(10), 555–572. doi:10.1038/s41568-020-0290-x

McDonald, K. L., Tabone, T., Nowak, A. K., and Erber, W. N. (2015). Somatic Mutations in Glioblastoma Are Associated with Methylguanine-DNA Methyltransferase Methylation. *Oncol. Lett.* 9 (5), 2063–2067. doi:10.3892/ol.2015.2980

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110

Muyas, F., Zapata, L., Guigó, R., and Ossowski, S. (2020). The Rate and Spectrum of Mosaic Mutations during Embryogenesis Revealed by RNA Sequencing of 49 Tissues. *Genome Med.* 12 (1). doi:10.1186/s13073-020-00746-1

Naumova, O. Y., Lee, M., Rychkov, S. Y., Vlasova, N. V., and Grigorenko, E. L. (2013). Gene Expression in the Human Brain: The Current State of the Study of Specificity and Spatiotemporal Dynamics. *Child. Dev.* 84 (1), 76–88. doi:10.1111/cdev.12014

Neums, L., Suenaga, S., Beyerlein, P., Anders, S., Koestler, D., Mariani, A., et al. (2018). VaDiR: An Integrated Approach to Variant Detection in RNA. *GigaScience* 7 (2). doi:10.1093/gigascience/gix122

Newman, S., Nakitandwe, J., Kesserwan, C. A., Azzato, E. M., Wheeler, D. A., Rusch, M., et al. (2021). Genomes for Kids: The Scope of Pathogenic Mutations in Pediatric Cancer Revealed by Comprehensive DNA and RNA Sequencing. *Cancer Discov.* 11 (12), 3008–3027. doi:10.1158/2159-8290.CD-20-1631

Ng, P. C. (2003). SIFT: Predicting Amino Acid Changes that Affect Protein Function. *Nucleic Acids Res.* 31 (13), 3812–3814. doi:10.1093/nar/gkg509

O Brien, T. D., Jia, P., Xia, J., Saxena, U., Jin, H., Vuong, H., et al. (2015). Inconsistency and Features of Single Nucleotide Variants Detected in Whole Exome Sequencing versus Transcriptome Sequencing: A Case Study in Lung Cancer. *Methods* 83, 118–127. doi:10.1016/j.ymeth.2015.04.016

Pei, S., Liu, T., Ren, X., Li, W., Chen, C., and Xie, Z. (2020). Benchmarking Variant Callers in Next-Generation and Third-Generation Sequencing Analysis. *Brief. Bioinform.* doi:10.1093/bib/bbaa148

Piskol, R., Ramaswami, G., and Li, J. B. (2013). Reliable Identification of Genomic Variants from RNA-Seq Data. *Am. J. Hum. Genet.* 93 (4), 641–651. doi:10.1016/j.ajhg.2013.08.008

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., et al. (2013). Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS One* 8 (3), e58815. doi:10.1371/journal.pone.0058815

Rashid, N. U., Sperling, A. S., Bolli, N., Wedge, D. C., Van Loo, P., Tai, Y. T., et al. (2014). Differential and Limited Expression of Mutant Alleles in Multiple Myeloma. *Blood* 124 (20), 3110–3117. doi:10.1182/blood-2014-04-569327

Redig, A. J., Capelletti, M., Dahlberg, S. E., Sholl, L. M., Mach, S., Fontes, C., et al. (2016). Clinical and Molecular Characteristics ofNF1 -Mutant Lung Cancer. *Clin. Cancer Res.* 22 (13), 3148–3156. doi:10.1158/1078-0432.CCR-15-2377

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: Predicting the Deleteriousness of Variants throughout the Human Genome. *Nucleic Acids Res.* 47 (D1), D886–D894. doi:10.1093/nar/gky1016

Ross, R. L., McPherson, H. R., Kettlewell, L., Shnyder, S. D., Hurst, C. D., Alder, O., et al. (2016). PIK3CA Dependence and Sensitivity to Therapeutic Targeting in Urothelial Carcinoma. *BMC Cancer* 16 (1). doi:10.1186/s12885-016-2570-0

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic Signaling Pathways in the Cancer Genome Atlas. *Cell* 173 (2), 321–337. doi:10.1016/j.cell.2018.03.035

Sheng, Q., Zhao, S., Li, C., Shyr, Y., and Guo, Y. (2016). Practicability of Detecting Somatic Point Mutation from RNA High Throughput Sequencing Data. *Genomics* 107 (5), 163–169. doi:10.1016/j.ygeno.2016.03.006

Skoulidis, F., and Heymach, J. V. (2019). Co-occurring Genomic Alterations in Non-small-cell Lung Cancer Biology and Therapy. *Nat. Rev. Cancer.* 19 (9), 495–509. doi:10.1038/s41568-019-0179-8

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

Suybeng, V., Koeppel, F., Harlé, A., and Rouleau, E. (2020). Comparison of Pathogenicity Prediction Tools on Somatic Variants. *J. Mol. Diagnostics* 22 (12), 1383–1392. doi:10.1016/j.jmoldx.2020.08.007

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: The Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 47 (D1), D941–D947. doi:10.1093/nar/gky1015

van Schie, E. H., and van Amerongen, R. (2020). Aberrant WNT/CTNNB1 Signaling as a Therapeutic Target in Human Breast Cancer: Weighing the Evidence. *Front. Cell Dev. Biol.* 8. doi:10.3389/fcell.2020.00025

Wang, H., Chen, S., Wei, J., Song, G., and Zhao, Y. (2021). A-to-I RNA Editing in Cancer: From Evaluating the Editing Level to Exploring the Editing Effects. *Front. Oncol.* 10. doi:10.3389/fonc.2020.632187

Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S., and Girirajan, S. (2017). Novel Metrics to Measure Coverage in Whole Exome Sequencing Datasets Reveal Local and Global Non-uniformity. *Sci. Rep.-UK* 7 (1). doi:10.1038/s41598-017-01005-x

Wang, S., Tian, Y., Wu, D., Zhu, H., Luo, D., Gong, W., et al. (2012). Genetic Variation of CTNNB1 Gene Is Associated with Susceptibility and Prognosis of Gastric Cancer in a Chinese Population. *Mutagenesis* 27 (6), 623–630. doi:10.1093/mutage/ges027

Wang, Y., Mashock, M., Tong, Z., Mu, X., Chen, H., Zhou, X., et al. (2020). Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. *Front. Oncol.* 10. doi:10.3389/fonc.2020.00447

Wang, Y., McKay, J. D., Rafnar, T., Wang, Z., Timofeeva, M. N., Broderick, P., et al. (2014). Rare Variants of Large Effect in BRCA2 and CHEK2 Affect Risk of Lung Cancer. *Nat. Genet.* 46 (7), 736–741. doi:10.1038/ng.3002

Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013). Emerging Patterns of Somatic Mutations in Cancer. *Nat. Rev. Genet.* 14 (10), 703–718. doi:10.1038/nrg3539

Wen, L., Britton, C. J., Garje, R., Darbro, B. W., and Packiam, V. T. (2021). The Emerging Role of Somatic Tumor Sequencing in the Treatment of Urothelial Cancer. *Asian J. Urology* 8 (4), 391–399. doi:10.1016/j.ajur.2021.06.005

Wilkerson, M. D., Cabanski, C. R., Sun, W., Hoadley, K. A., Walter, V., Mose, L. E., et al. (2014). Integrated RNA and DNA Sequencing Improves Mutation Detection in Low Purity Tumors. *Nucleic Acids Res.* 42 (13), e107. doi:10.1093/nar/gku489

Willis, O., Choucair, K., Alloghbi, A., Stanbery, L., Mowat, R., Charles Brunicardi, F., et al. (2020). PIK3CA Gene Aberrancy and Role in Targeted Therapy of Solid Malignancies. *Cancer Gene Ther.* 27 (9), 634–644. doi:10.1038/s41417-020-0164-0

Xiao, W., Ren, L., Chen, Z., Fang, L. T., Zhao, Y., Lack, J., et al. (2021). Toward Best Practice in Cancer Mutation Detection with Whole-Genome and Whole-Exome Sequencing. *Nat. Biotechnol.* 39 (9), 1141–1150. doi:10.1038/s41587-021-00994-5

Xu, C. (2018). A Review of Somatic Single Nucleotide Variant Calling Algorithms for Next-Generation Sequencing Data. *Comput. Struct. Biotechnol. J.* 16, 15–24. doi:10.1016/j.csbj.2018.01.003

Yan, Y. H., Chen, S. X., Cheng, L. Y., Rodriguez, A. Y., Tang, R., Cabrera, K., et al. (2021). Confirming Putative Variants at ≤ 5% Allele Frequency Using Allele Enrichment and Sanger Sequencing. *Sci. Rep.-UK.* 11 (1). doi:10.1038/s41598-021-91142-1

Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. (2013). IgBLAST: An Immunoglobulin Variable Domain Sequence Analysis Tool. *Nucleic Acids Res.* 41 (W1), W34–W40. doi:10.1093/nar/gkt382

Yizhak, K., Aguet, F., Kim, J., Hess, J. M., Kübler, K., Grimsby, J., et al. (2019). RNA Sequence Analysis Reveals Macroscopic Somatic Clonal Expansion across Normal Tissues. *Science* 364 (6444), w726. doi:10.1126/science.aaw0726

Zhang, G., Tang, X., Liang, L., Zhang, W., Li, D., Li, X., et al. (2020). DNA and RNA Sequencing Identified a Novel Oncogene VPS35 in Liver Hepatocellular Carcinoma. *Oncogene* 39 (16), 3229–3244. doi:10.1038/s41388-020-1215-6

Zhang, Y., Coillie, S. V., Fang, J., and Xu, J. (2016). Gain of Function of Mutant P53: R282W on the Peak? *Oncogenesis* 5 (2), e196. doi:10.1038/oncsis.2016.8

Zhang, Y., Li, B., Li, C., Cai, Q., Zheng, W., and Long, J. (2014). Improved Variant Calling Accuracy by Merging Replicates in Whole-Exome Sequencing Studies. *Biomed. Res. Int.* 2014, 1–7. doi:10.1155/2014/319534

Zhu, J., and Pierskalla, W. P. (2016). Applying a Weighted Random Forests Method to Extract Karst Sinkholes from LiDAR Data. *J. Hydrol.* 533, 343–352. doi:10.1016/j.jhydrol.2015.12.012

Złowocka, E., Cybulski, C., Górski, B., Dębniak, T., Słojewski, M., Wokołorczyk, D., et al. (2008). Germline Mutations in theCHEK2 Kinase Gene Are Associated with an Increased Risk of Bladder Cancer. *Int. J. Cancer.* 122 (3), 583–586. doi:10.1002/ijc.23099

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

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

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 15 July 2022
doi: 10.3389/fgene.2022.863747

A Diagnostic Model Using Exosomal Genes for Colorectal Cancer

Tianxiang Lei[1,2†], *Yongxin Zhang*[1,2†], *Xiaofeng Wang*[1,2†], *Wenwei Liu*[3], *Wei Feng*[1,2] and *Wu Song*[1]*

[1]*Department of Gastrointestinal Surgery, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China,* [2]*Laboratory of General Surgery, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China,* [3]*Center for Digestive Disease, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, China*

Colorectal cancer (CRC) is a leading cause of cancer-related deaths worldwide. Exosomes have great potential as liquid biopsy specimens due to their presence and stability in body fluids. However, the function and diagnostic values of exosomal genes in CRC are poorly understood. In the present study, exosomal data of CRC and healthy samples from the exoRBase 2.0 and Gene Expression Omnibus (GEO) databases were used, and 38 common exosomal genes were identified. Through the least absolute shrinkage and selection operator (Lasso) analysis, support vector machine recursive feature elimination (SVM-RFE) analysis, and logistic regression analysis, a diagnostic model of the training set was constructed based on 6 exosomal genes. The diagnostic model was internally validated in the test and exoRBase 2.0 database and externally validated in the GEO database. In addition, the co-expression analysis was used to cluster co-expression modules, and the enrichment analysis was performed on module genes. Then a protein–protein interaction and competing endogenous RNA network were constructed and 10 hub genes were identified using module genes. In conclusion, the results provided a comprehensive understanding of the functions of exosomal genes in CRC as well as a diagnostic model related to exosomal genes.

Keywords: exosome, diagnostic model, functions, colorectal cancer, bioinformatics analysis

OPEN ACCESS

Edited by:
Liang Zhao,
University of São Paulo, Brazil

Reviewed by:
Ehsan Nazemalhosseini-Mojarad,
Shahid Beheshti University of Medical Sciences, Iran
Laura La Paglia,
Institute for High Performance Computing and Networking (ICAR), Italy

Correspondence:
Wu Song
songwu@mail.sysu.edu.cn

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *27 January 2022*
Accepted: *19 May 2022*
Published: *15 July 2022*

Citation:
Lei T, Zhang Y, Wang X, Liu W, Feng W and Song W (2022) A Diagnostic Model Using Exosomal Genes for Colorectal Cancer.
Front. Genet. 13:863747.
doi: 10.3389/fgene.2022.863747

INTRODUCTION

Colorectal cancer (CRC) is one of the most common malignant tumors of the digestive tract and presents the second-highest mortality rate (9.2%) among all cancers (Arnold et al., 2017; Bray et al., 2018). Tumor metastases and invasion are associated with poor prognosis and lead to a low five-year survival rate in CRC patients (Siegel et al., 2018). CRC is typically detected by colonoscopy, measuring carcinoembryonic antigen (CEA) levels, multitarget stool DNA testing, and the septin 9 gene methylation blood test (Ahluwalia et al., 2021). Although colonoscopy is a highly sensitive method, it is invasive and uncomfortable for patients and its accuracy depends on the skill level and experience of the endoscopist (Schreuders et al., 2015). CEA levels have been widely used as tumor markers for the detection of CRC. However, it is still limited in terms of sensitivity and specificity. Therefore, there is an urgent need to identify more effective and less invasive surrogate CRC-specific diagnostic markers for the rapid, noninvasive, and high sensitivity screening of patients.

The use of exosomes as a new noninvasive approach for diagnosing diseases has attracted growing attention (Mousavi et al., 2019). Exosomes are extracellular vesicles containing messenger RNAs (mRNA), microRNAs (miRNA), long noncoding RNAs (lncRNA), circular RNAs (circRNA), DNA, lipids, and proteins with sizes between 40 and 150 nm and density between 1.13 and 1.19 g/ml (Yáñez-Mó et al., 2015; Shi et al., 2021). Notably, exosomal biomarkers and molecule information remain relatively stable in most body fluids because they are protected from degradation and external impact (Xiao et al., 2019). Therefore, exosomes have great potential as liquid biopsy specimens for various diseases (Wang et al., 2017; Liu et al., 2019). In particular, cancer cells secrete significantly more exosomes than normal cells (Mousavi et al., 2019; Nabariya et al., 2020). Cancer-derived exosomes likely serve as new circulating biomarkers for the early detection of cancer as they carry cargo reflective of genetic or signaling alterations in the cancer cells of origin (Li et al., 2015; Melo et al., 2015). Therefore, exosomes may be an ideal candidate to act as a biomarker for CRC.

In the present study, we identified several differentially expressed exosomal genes from public databases to understand the underlying molecular changes and biological mechanisms. For the diagnosis of CRC patients, we established a 6–exosomal gene diagnosis model using the least absolute shrinkage and selection operator (LASSO), support vector machine recursive feature elimination (SVM-RFE), and logistic regression analyses. This model was verified using a receiver operating characteristic (ROC) curve in internal and external sets. In addition, the co-expression analysis was used to cluster co-expression modules, and the enrichment analysis was performed on module genes. We also established protein–protein interaction (PPI) networks to investigate hub genes and constructed competing endogenous RNA (ceRNA) networks related to serum exosomal genes in CRC. The purpose of the present study was to obtain further insight into the underlying functions of exosomal genes and to identify any potential diagnostic exosomal genes using the bioinformatics analysis in CRC.

MATERIALS AND METHODS

Data Source and Identification of Differentially Expressed Exosomal Genes

The flowchart of this study is presented in **Supplementary Figure S1**. We extracted serum exosomal data relative to mRNA, lncRNA, and circRNA from 35 CRC patients and 118 healthy persons from the exoRBase 2.0 database (Li S. et al., 2018) (http://www.exorbase.org/). We also downloaded serum exosomal data from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). The GSE100063 and GSE100206 datasets contain exosomal data (mRNA) from 12 CRC patients and 32 normal persons. All CRC and healthy exosomal data were available and included in the study. After data normalization and log base 2 transformations, differentially expressed genes between CRC patients and healthy individuals were identified using the limma R package (Ritchie et al., 2015). Differentially expressed genes were defined as those whose expression differences were associated with an adjusted p <0.05. This study was approved by the Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University [Approval number (2021)687].

Identification of the Diagnosis-Related Exosomal Gene Signature Associated With Colorectal Cancer

Genes that were differentially expressed (adjusted p <0.05) in the exoRBase 2.0, GSE100063, and GSE100206 datasets were selected. First, 70% of the samples in exoRBase 2.0 were randomly selected as a training set. The LASSO regression analysis was used to obtain diagnostic exosomal genes from the training set. The SVM-RFE analysis was used simultaneously to screen exosomal genes in the training set for CRC diagnosis. Then, we combined the LASSO and SVM-RFE analyses to obtain the common exosomal genes. Next, we selected common exosomal genes that were regulated (up- or downregulated) in the same direction to build a diagnosis-related exosomal gene signature by the multivariate logistic regression analysis. Finally, the ROC curve analysis and the area under the curve (AUC) were used to estimate the diagnostic value of the diagnosis-related exosomal gene signature using the pROC package in R (Robin et al., 2011).

Validation of the Diagnosis-Related Exosomal Gene Signature

As a validation set, 30% of the samples in exoRBase 2.0, GSE100063, and GSE100206 datasets were selected. To validate whether candidate exosomal genes had an important diagnostic value in patients with CRC, we also measured the ROC curve and the AUC value in the validation datasets. p <0.05 is considered statistically significant.

Cell Culture, Human Plasma Samples, and Isolation of Exosomes

HCoEpiC, HCT116, and SW480 cell lines (purchased from ATCC) were cultured in a DMEM medium (Cellmax) containing 10% fetal calf serum, and 100 U/ml each of penicillin and streptomycin at 37°C with 5% CO_2. HCoEpiC, HCT116, and SW480 cells were cultured with a full medium at 80% confluency and replaced with a fresh medium without fetal bovine serum. After a 48-h culture, the cell medium was harvested. In addition, thirty-two CRC patient serums and seventeen healthy human serums were collected in the First Affiliated Hospital of Sun Yat-sen University in December 2021. The samples were stored at −80 before exosome extraction. The exosomes were collected from the cell culture supernatant by differential centrifugations. In addition, exosome morphology was identified by transmission electron microscopy (TEM), the nanoparticle tracking analysis (NTA), and the expression of exosome surface markers CD9, TSG101, and HSP70 were evaluated by the Western blotting analysis.

Validation of Exosomal Gene Expression Levels

Total RNA from exosomes was extracted using the TRIzol reagent (Invitrogen, NYC, United States) and reverse transcribed with the PrimeScript RT kit (Takara, China). Real-time PCR was carried out using the SYBR PreMix Ex Taq II kit (Takara, China). GAPDH was used as the normalized control of mRNA. The relative expression levels of mRNA in exosomes were calculated by the $2^{-\Delta\Delta CT}$ method. The primer sequences used in this study are listed in **Supplementary Table S1**.

The Co-Expression Analysis

The weighted gene co-expression network analysis (WGCNA) was used to identify the co-expression network of differentially expressed exosomal genes in exoRBase 2.0 using the WGCNA package in R (Langfelder and Horvath, 2008). A weighted adjacency was constructed by calculating using the Pearson correlations of all gene pairs. Soft power β = 7 was selected to construct a standard scale-free network. The similarity matrix, which was constructed using the Pearson's correlation coefficients of all gene pairs, was transformed into a topological overlap matrix (TOM) as well as the corresponding dissimilarity (1-TOM). Then, a hierarchical clustering dendrogram of the 1-TOM matrix was used to classify similar gene expressions into different gene co-expression modules. Afterward, a module-clinical trait association was calculated to identify functional modules in a co-expression network. The brown and gray modules were selected for further analysis.

The Gene Ontology Term and Kyoto Encyclopedia of Genes and Genomes Functional Enrichment Analyses

To further clarify the potential biological functions of the exosomal genes in the modules, we performed the Gene Ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses using the clusterProfiler package in R (Yu et al., 2012). The GO terms and KEGG pathways with p <0.05 were considered significantly enriched.

Protein–Protein Interaction Network Construction and Hub Gene Identification

In this study, the exosomal genes in the brown and gray modules were analyzed. The PPI network was constructed by the online STRING database (https://string-db.org), and interaction with a combined score >0.7 was considered statistically significant. Cytoscape, an open-source bioinformatics software platform, was used to visualize molecular interaction networks. Hub genes were identified using the Degree algorithm of the cytoHubba plugin in Cytoscape.

Exosomal ceRNA Network Construction

The TargetScan database (www.targetscan.org) and miRanda (http://cbio.mskcc.org/microrna_data/miRanda-aug2010.tar.gz) statistical models that predict the effects of miRNAs binding to canonical sites of mRNA were used to accomplish miRNA prediction. miRNA/circRNA and miRNA/lncRNA interactions were predicted using ENCORI (http://starbase.sysu.edu.cn/) and miRcode (http://www.mircode.org/), respectively. The ceRNA regulatory network was constructed according to the ceRNA regulatory mechanism and the differentially expressed lncRNAs, circRNAs, and mRNAs in exosomes, and the network were visualized by Cytoscape.

RESULTS

Identification of Differential Expression of Exosomal Genes

The RNA sequencing exosomal data for CRC and healthy samples were downloaded from the exoRBase 2.0 and GEO databases. A total of 2839 differentially expressed exosomal genes were obtained from the exoRBase 2.0 dataset (**Figure 1A**) and 475 differentially expressed exosomal genes in the GSE100063 and GSE100206 datasets (**Figure 1B**). A total of 38 differentially expressed exosomal genes were common in the two databases (**Figure 1C**).

Development and Verification of the Diagnostic Model

We randomly split all the samples in the exoRBase 2.0 dataset into a training set (70%) and a validation set (30%). We combined the LASSO and SVM-RFE methods to obtain 9 common exosomal genes (**Figures 1D–F**). Next, in the training and validation set, we selected common exosomal genes that were in the same regulated direction to build a diagnosis-related exosomal gene signature using the multivariate logistic regression analysis (**Figures 2A,B**). Therefore, we obtained 6 exosomal genes including *H3F3A*, *MYL6*, *FBXO7*, *TUBA1C*, *MEF2C*, and *BANK1*. Furthermore, these 6 exosomal genes were identified based on the model according to the following formula: index = *H3F3A* × (3.67971578945741) + MYL6 × (2.01995321634272) + FBXO7 × (−0.86536340517915) + TUBA1C × (0.184108184598825) + MEF2C × (−3.86524121779742) + BANK1 × (−5.35388054514927). The AUC for the gene signature was 0.981 in the training set (**Figure 2C**) and 0.923 in the internal test set (**Figure 2D**). Importantly, we also used external datasets to validate the gene signature: it showed good diagnostic ability, with an AUC of 0.995 (**Figure 2E**).

Validation of Exosomal Gene Expression Levels

To further validate the expression levels of these model exosomal genes, we extracted exosomes from HCoEpiC, HCT116, SW480 cells, and human plasmas. The purification of the exosomes was validated by TEM, NTA, and Western blotting analysis. TEM detected double-layer spherical vesicles ranging from 30 to 160 nm in size, which confirmed the presence of exosomes (**Supplementary Figure S2A**). NTA characterized the size and concentration of exosomes (**Supplementary Figure S2B**). Western blotting analysis estimated the quantity and purity of exosomes by detecting exosomal marker proteins (CD9, TSG101, and HSP70) (**Supplementary Figure S2C**). Next, we validated the expression of signature exosomal genes in the exosomes obtained by RT-qPCR. The results showed that *H3F3A*, *MYL6*, and *TUBA1C* were significantly upregulated in CRC cells, and *MEF2C* and *FBXO7* were significantly downregulated in CRC cells. However, no significant differences were observed in *BANK1* expression between HCT116 and SW480 cells and HCoEpic cells (**Supplementary Figure S2D**). Furthermore, we validated the expression levels of the model exosomal genes in the serums from the CRC patients and healthy humans. The results showed that the mRNA levels of three genes (*H3F3A*, *TUBA1C,* and *MYL6*) were significantly elevated in the CRC exosomes, whereas *BANK1*, *MEF2C,* and *FBXO7* were downregulated in the CRC exosomes when compared with those in the healthy human exosomes (**Figure 3**). These results were consistent with the results of the exoRBase 2.0 and GEO databases.

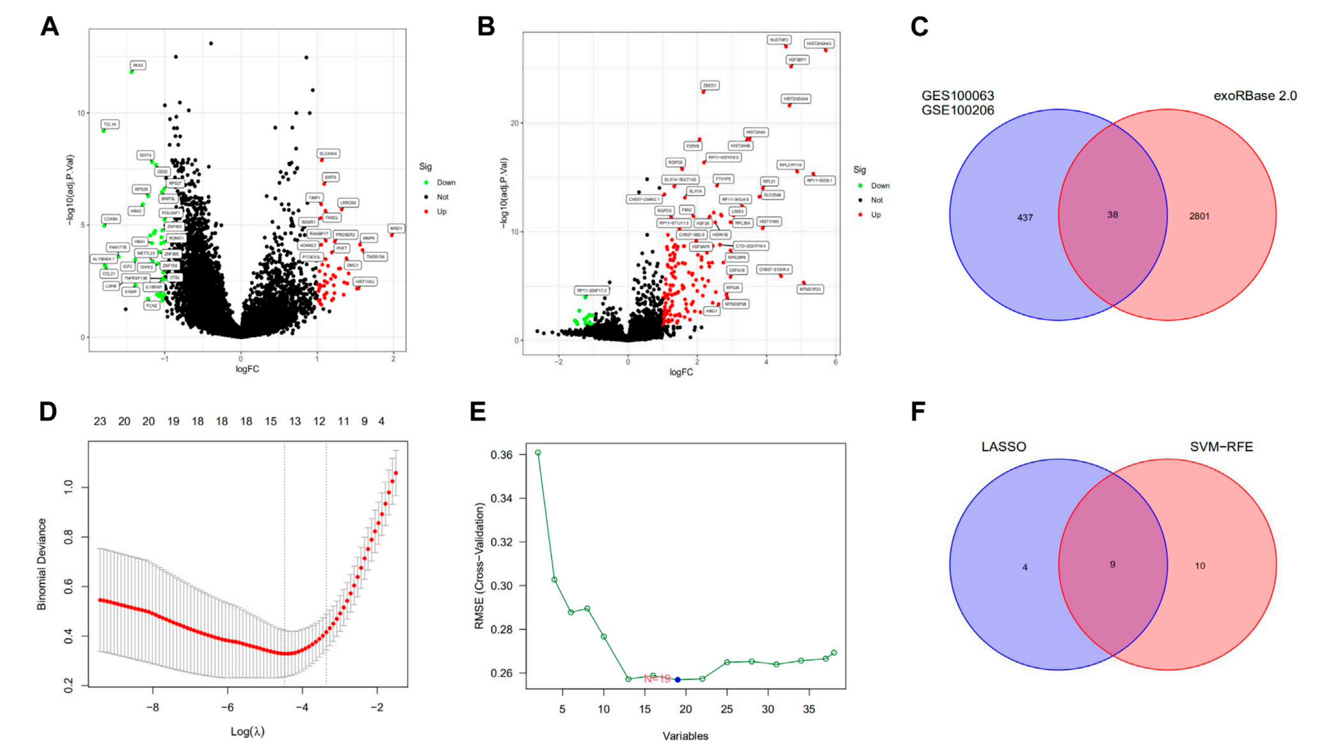


FIGURE 1 | Differentially expressed exosomal genes and identification of diagnostic exosomal genes in CRC. **(A)** Differentially expressed exosomal genes between CRC patients and controls in the exoRBase 2.0 database. **(B)** Exosomal genes differentially expressed between CRC patients and controls in the GSE100063 and GSE100206 datasets. **(C)** The intersection of differentially expressed exosomal genes in the exoRBase 2.0, the GSE100063, and the GSE100206 dataset. **(D)** The LASSO method identified 13 diagnostic exosomal genes. **(E)** The SVM-RFE method identified 19 diagnostic exosomal genes. **(F)** The intersection of diagnostic exosomal genes in the two analyses. CRC, colorectal cancer; LASSO, least absolute shrinkage and selection operator; SVM-RFE, support vector machine recursive feature elimination.

The Weighted Gene Co-Expression Network Analysis and Key Module Identification

To identify exosomal genes associated with CRC, we analyzed differentially expressed exosomal genes between CRC and healthy samples in the exoRBase 2.0 database using WGCNA. The included samples were clustered with the hierarchical average linkage clustering method (**Figure 4A**). The optimal soft power threshold for WGCNA was set to 7 to preserve the scale-free topology and effective connectivity (**Figure 4B**). Four co-expression modules of differentially expressed exosomal genes were established (**Figure 4C**). The brown and gray modules were found to have a positive correlation with CRC.

Enrichment of Module Genes

To explore the potential function of exosomal genes in the brown and gray modules, the enrichment analysis of GO and KEGG was performed. The GO analysis of these differentially expressed exosomal genes revealed that "small molecule catabolic process," "contractile fiber," and "structural constituent of muscle" were the most frequent biological terms for biological process, cellular components, and molecular functions, respectively (**Figure 4D**). The KEGG analysis revealed that these exosomal genes were mainly enriched in the "apelin signaling pathway," "ubiquitin mediated proteolysis," "spinocerebellar ataxia," and "tight junction" (**Figure 4E**).

Protein–Protein Interaction Network Construction and Hub Genes Screening

The PPI network of the exosomal genes in the brown and gray modules was constructed through the STRING database and visualized with Cytoscape. The PPI network and hub genes identified from the network were obtained through the degree algorithm of the CytoHubba plugin. According to degree scores, the top-scoring genes, including *HIST1H3E*, *HIST1H3J*, *HIST1H3A*, *HIST1H2BC*, *SEH1L*, *H3F3A*, *HIST1H2BJ*, *HIST1H2BF*, *HIST1H2BB*, and *RHEB*, were considered the hub genes (**Figure 4F**).

Construction of the Exosomal ceRNA Network

After the differentially expressed exosomal genes were identified in the brown and gray co-expression modules,

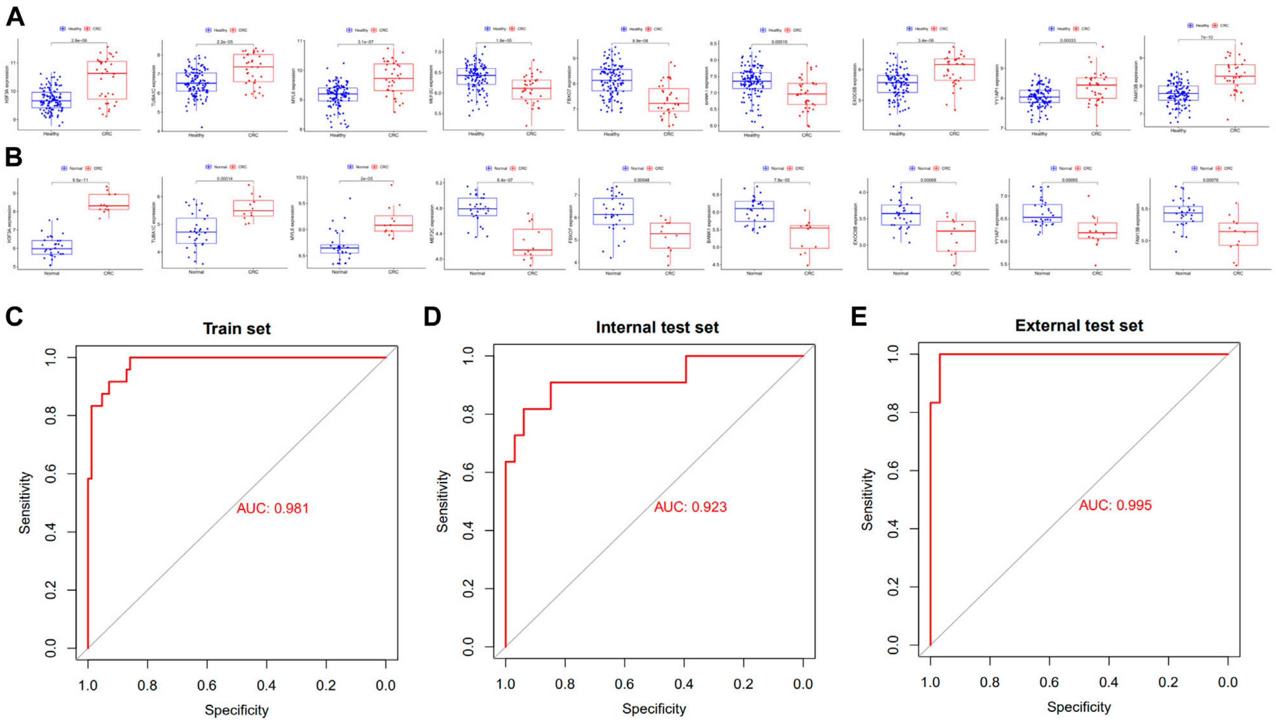


FIGURE 2 | Potential exosomal genes for the diagnosis of CRC. The relative expression level of diagnostic exosomal genes **(A)** between CRC serum and healthy samples in the training set and **(B)** in the external validation dataset. ROC curves of the exosomal gene signature in the **(C)** training set and **(D)** in the internal validation set of the exoRBase 2.0 database. **(E)** ROC curves of the exosomal gene signature in the external validation set of the GSE100063 and GSE100206 database. CRC, colorectal cancer; ROC, receiver operating characteristic.

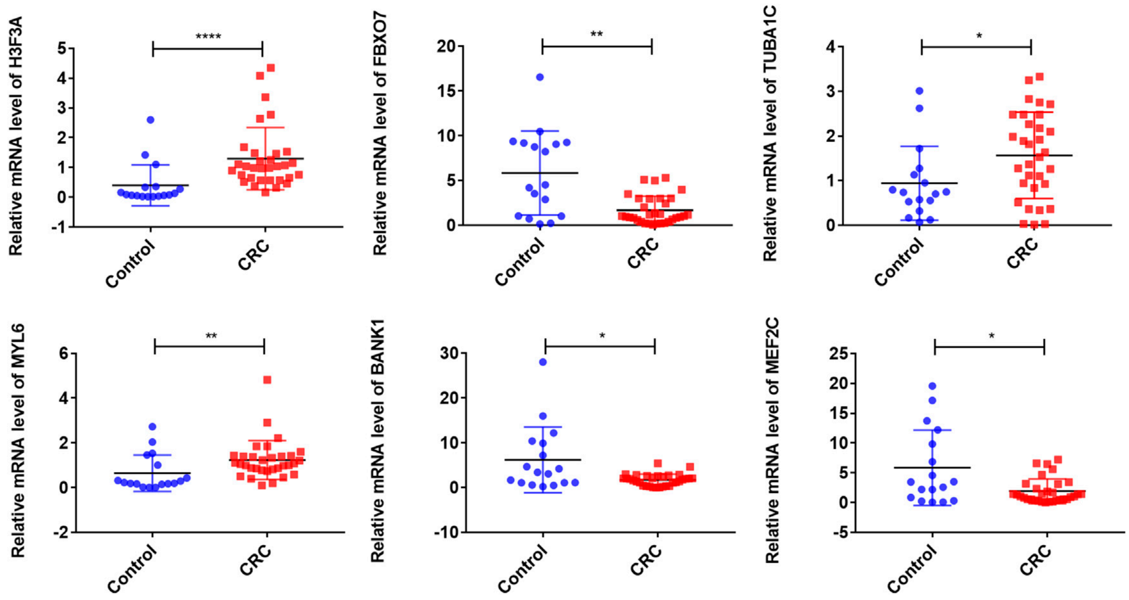


FIGURE 3 | Validation of exosomal gene expression levels in CRC patient serums and controls. ****p < 0.0001; **p < 0.01; *p < 0.05; CRC, colorectal cancer.

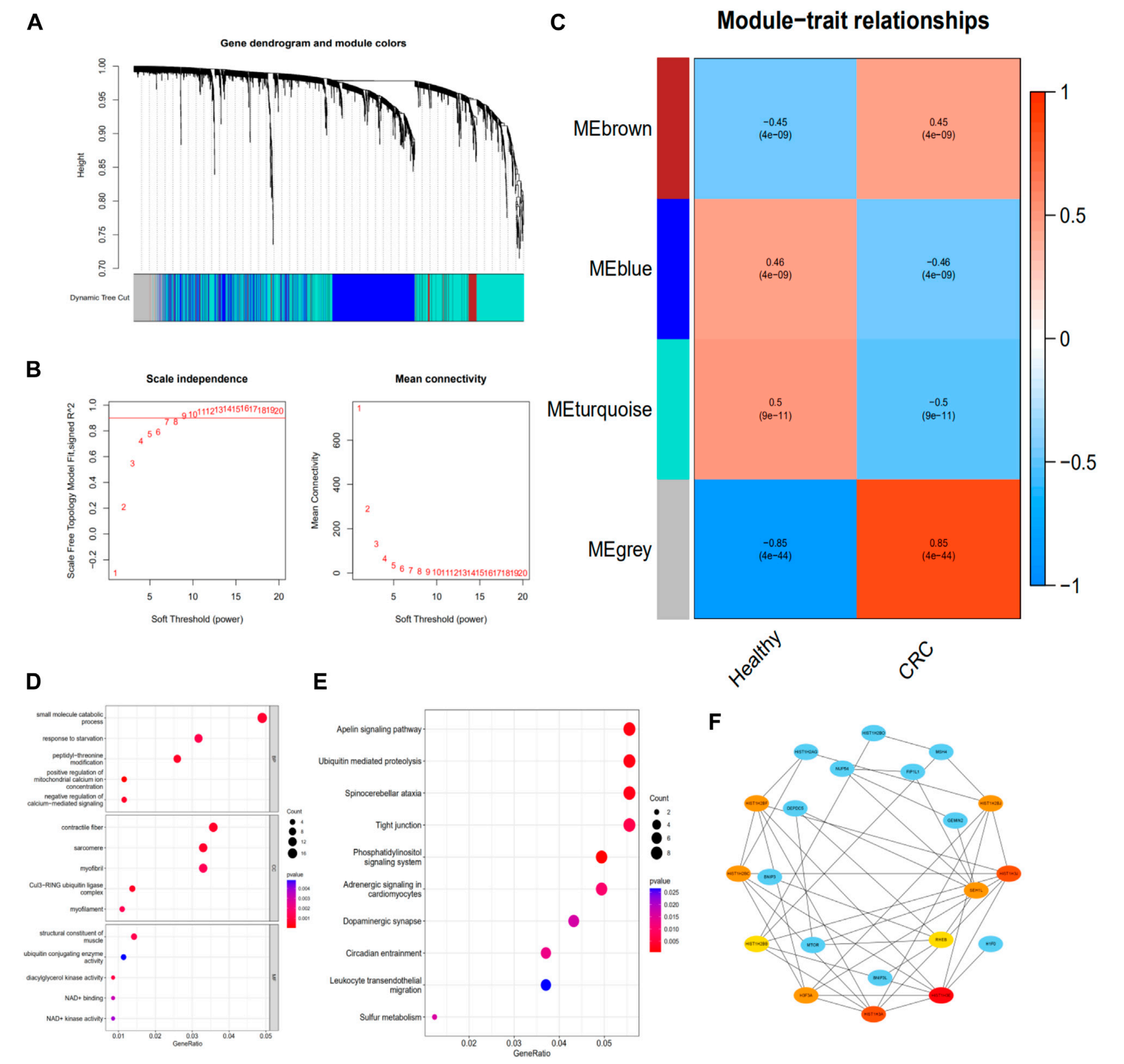


FIGURE 4 | Identifying the functions of CRC-associated exosomal genes. **(A)** Clustering dendrogram. **(B)** Determination of soft-thresholding power in the weighted gene co-expression network analysis. **(C)** Module–trait associations evaluated by correlations between CRC and clinical traits. **(D)** The GO enrichment analysis of the exosomal genes in the brown and gray modules. **(E)** The KEGG pathway of the exosomal genes in the brown and gray modules. **(F)** The PPI network of genes in the brown and gray modules and hub genes screening. CRC, colorectal cancer; PPI, protein–protein interaction

these exosomal genes were used to construct ceRNA networks. Exosomal lncRNA and circRNA that were differentially expressed (adjusted p <0.05) in exoRBase 2.0 were selected. Based on mRNA, lncRNA, circRNA, and the predicted corresponding miRNA, we constructed a ceRNA network. The ceRNA network consisted of 5 circRNA nodes, 2 lncRNA nodes, 40 miRNA nodes, and 72 mRNA nodes (**Figure 5**).

DISCUSSION

CRC is a highly malignant cancer with a poor prognosis. CRC patients are at substantial risk of recurrence and metastasis. Therefore, early diagnosis is very important to improve the clinical prognosis of CRC patients. Liquid biopsy, a recent and hot topic in cancer detection, has been considered for the early diagnosis of cancer (Chang et al., 2019). The functional states of

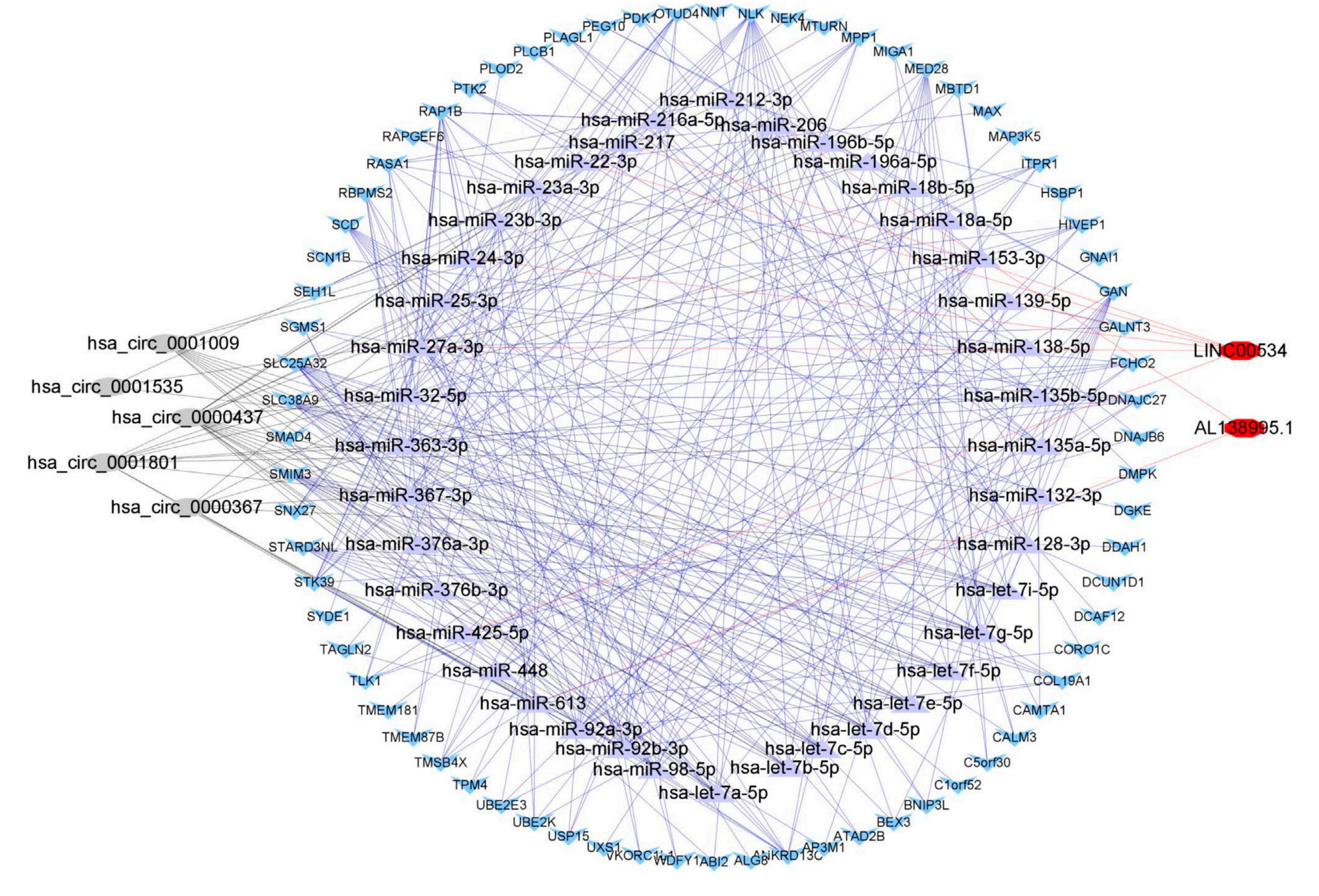


FIGURE 5 | A competing endogenous RNA network associated with exosomal genes.

cancer cells could be assessed by testing the exosome they secreted, which provides the basis for exosome-mediated noninvasive cancer liquid biopsy (Yang et al., 2020). Currently, various studies show that the increased or decreased expression of exosomes plays an important role in different kinds of cancer, including CRC (Hon et al., 2017; Galamb et al., 2019; Xiao et al., 2020). Thus, exosomes are crucial potential candidates for the early detection of CRC.

To date, exosomal gene–based diagnostic models have not been described for CRC, although considerable efforts have been made to develop prognostic signatures based on differentially expressed genes (Chen et al., 2019; Dai et al., 2020; Wang et al., 2020). The present study analyzed differences in exosomal gene expression between CRC patients and healthy humans. Importantly, we identified diagnostic exosomal genes based on a comprehensive analysis that could serve as valuable biomarkers in the clinical setting.

Potential exosomal gene modules related to CRC were identified with the WGCNA analysis. The brown and gray modules were found to have a positive correlation with CRC. To better understand the potential function of exosomal genes among the brown and gray modules, the GO and KEGG enrichment analysis and the PPI network were conducted. The results of the functional and pathway enrichment analyses showed that the exosomal genes in the modules were mainly enriched in the structural components of the muscle and apelin signaling pathway. The PPI network of exosomal genes in the brown and gray modules was constructed and 10 hub genes were selected using the CytoHubba plugin in Cytoscape. Among these genes, H3F3A, as a diagnostic biomarker in the present study, was also identified as a hub gene, and maybe as a promoter of CRC progression and metastasis. The ceRNA network played a critical role in the initiation and progression of CRC (Ke et al., 2019; Ma et al., 2020). In the present study, we further constructed ceRNA networks based on those key exosomal genes. This approach provided a novel view of the RNA–RNA crosstalk in the exosome and indicated the potential diagnostic and therapeutic functions of exosomal ceRNA networks in CRC.

Our exosomal gene–based model highlighted 6 exosomal genes, that is, *H3F3A*, *MYL6*, *FBXO7*, *TUBA1C*, *MEF2C*, and *BANK1*. These genes and their biological functions have been studied in some tumors. *H3F3A* is one of two genes encoding histone H3.3, a noncanonical histone variant, and has been established as a major driver gene of malignant gliomas (Schwartzentruber et al., 2012; Sturm et al., 2012; Wu et al., 2012). *MYL6*, a gene that encodes hexameric ATPase cellular motor protein, is upregulated in circulating tumor cells of many cancers (Yadavalli et al., 2017). *FBXO7* may have a proto-oncogenic role in epithelial tumors (Laman et al., 2005). *TUBA1C*, as a type of tubulin, is associated with tumor cell death and cell proliferation, and *TUBA1C* overexpression is predicted to have a poor prognosis (Li C.-W. et al., 2018; Li et al., 2020). *MEF2C* was traditionally considered a development-associated factor and can inhibit tumor growth *in vitro* and *in vivo* (Bai et al., 2015). *BANK1*, which encodes a protein adaptor that is predominantly expressed in B cells, is a putative tumor suppressor gene in B-cell lymphomagenesis (Yan et al., 2014). Our results showed that diagnostic exosomal genes may participate in the development of human CRC; however, the underlying molecular mechanism of these genes in the prognosis of CRC requires further investigation. Our experimental results showed that the differential expression levels of diagnostic exosomal genes in cell lines were approximately in agreement with the serum exosomal data in these public databases.

To the best of our knowledge, this is the first reported exosomal gene–based model for CRC. However, the potential limitations of this study should also be considered when interpreting the findings. First, we could not explore the association between these exosomal genes and the prognosis of CRC patients due to the lack of therapeutic and prognostic information. Second, because the data we analyzed were obtained from public databases, further experimental studies are necessary to validate our findings.

To conclude, we investigated the potential functions and diagnostic values of exosomal genes in CRC through a comprehensive bioinformatics analysis. A diagnostic exosomal gene model was constructed. This model can assess the value of exosomal genes to diagnose CRC using a noninvasive method and might be useful for the development of individualized treatment for CRC patients, but the feasibility of its use in the population needs to be further validated.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University [Approval number (2021)687]. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

TL and WS conceived the concept, instructed data analysis, and revised the manuscript. YZ and XW conducted most data analysis, prepared figures and tables, and wrote the manuscript draft. WL and WF helped with some analysis and interpretation of data. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by the National Natural Science Foundation of China (grant number: 81871908). Guangdong Provincial Natural Science Foundation (grant number: 2018A030313715). Guangzhou Science and Technology Plan-General Project (grant number: 201904010036). General projects of Guangdong Provincial Natural Science Foundation (grant number: 82070529).

ACKNOWLEDGMENTS

The authors would like to thank the exoRBase 2.0 and GEO databases for the availability of the data and thank Dr. ZeBing Song for his help and guidance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.863747/full#supplementary-material

Supplementary Figure S1 | Overall workflow of the present study.

Supplementary Figure S2 | Identification of exosomes and validation of signature exosomal gene expression levels. **(A)** Morphologies of exosomes in HCoEpic, HCT116, and SW480 cell lines were observed by transmission electron microscopy. **(B)** The size and concentration of exosomes by the nanoparticle tracking analysis. **(C)** Exosomes were analyzed by Western blotting using antibodies against exosomal markers. **(D)** The relative expression fold of diagnostic exosomal genes was determined. *****P*<0.0001; ****P*<0.001; ***P*<0.01.

REFERENCES

Ahluwalia, P., Kolhe, R., and Gahlay, G. K. (2021). The Clinical Relevance of Gene Expression Based Prognostic Signatures in Colorectal Cancer. *Biochim. Biophys. Acta (BBA) - Rev. Cancer* 1875 (2), 188513. doi:10.1016/j.bbcan.2021.188513

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 (4), 683–691. doi:10.1136/gutjnl-2015-310912

Bai, X. L., Zhang, Q., Ye, L. Y., Liang, F., Sun, X., Chen, Y., et al. (2015). Myocyte Enhancer Factor 2C Regulation of Hepatocellular Carcinoma via Vascular Endothelial Growth Factor and Wnt/β-Catenin Signaling. *Oncogene* 34 (31), 4089–4097. doi:10.1038/onc.2014.337

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. Clin.* 68 (6), 394–424. doi:10.3322/caac.21492

Chang, L., Ni, J., Zhu, Y., Pang, B., Graham, P., Zhang, H., et al. (2019). Liquid Biopsy in Ovarian Cancer: Recent Advances in Circulating Extracellular Vesicle Detection for Early Diagnosis and Monitoring Progression. *Theranostics* 9 (14), 4130–4140. doi:10.7150/thno.34692

Chen, L., Lu, D., Sun, K., Xu, Y., Hu, P., Li, X., et al. (2019). Identification of Biomarkers Associated with Diagnosis and Prognosis of Colorectal Cancer

Patients Based on Integrated Bioinformatics Analysis. *Gene* 692, 119–125. doi:10.1016/j.gene.2019.01.001

Dai, G. P., Wang, L. P., Wen, Y. Q., Ren, X. Q., and Zuo, S. G. (2020). Identification of Key Genes for Predicting Colorectal Cancer Prognosis by Integrated Bioinformatics Analysis. *Oncol. Lett.* 19 (1), 388–398. doi:10.3892/ol.2019.11068

Galamb, O., Barták, B. K., Kalmár, A., Nagy, Z. B., Szigeti, K. A., Tulassay, Z., et al. (2019). Diagnostic and Prognostic Potential of Tissue and Circulating Long Non-Coding RNAs in Colorectal Tumors. *World J. Gastroenterol.* 25 (34), 5026–5048. doi:10.3748/wjg.v25.i34.5026

Hon, K. W., Abu, N., Ab Mutalib, N.-S., and Jamal, R. (2017). Exosomes as Potential Biomarkers and Targeted Therapy in Colorectal Cancer: A Mini-Review. *Front. Pharmacol.* 8, 583. doi:10.3389/fphar.2017.00583

Ke, M. J., Ji, L. D., and Li, Y. X. (2019). Explore Prognostic Marker of Colorectal Cancer Based on ceRNA Network. *J. Cell. Biochem.* 120 (12), 19358–19370. doi:10.1002/jcb.28860

Laman, H., Funes, J. M., Ye, H., Henderson, S., Galinanes-Garcia, L., Hara, E., et al. (2005). Transforming Activity of Fbxo7 Is Mediated Specifically Through Regulation of Cyclin D/cdk6. *Embo J.* 24 (17), 3104–3116. doi:10.1038/sj.emboj.7600775

Langfelder, P., and Horvath, S. (2008). WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinf.* 9, 559. doi:10.1186/1471-2105-9-559

Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015). Circular RNA Is Enriched and Stable in Exosomes: A Promising Biomarker for Cancer Diagnosis. *Cell Res.* 25 (8), 981–984. doi:10.1038/cr.2015.82

Li, C.-W., Chiu, Y.-K., and Chen, B.-S. (2018a). Investigating Pathogenic and Hepatocarcinogenic Mechanisms from Normal Liver to HCC by Constructing Genetic and Epigenetic Networks via Big Genetic and Epigenetic Data Mining and Genome-Wide NGS Data Identification. *Dis. Markers* 2018, 1–22. doi:10.1155/2018/8635329

Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., et al. (2018b). exoRBase: A Database of circRNA, lncRNA and mRNA in Human Blood Exosomes. *Nucleic Acids Res.* 46 (D1), D106–d112. doi:10.1093/nar/gkx891

Li, J., Zhou, Y., Yan, Y., Zheng, Z., Hu, Y., and Wu, W. (2020). Sulforaphane-Cysteine Downregulates CDK4/CDK6 and Inhibits Tubulin Polymerization Contributing to Cell Cycle Arrest and Apoptosis in Human Glioblastoma Cells. *Aging* 12 (17), 16837–16851. doi:10.18632/aging.103537

Liu, H., Li, P.-w., Yang, W.-q., Mi, H., Pan, J.-l., Huang, Y.-c., et al. (2019). Identification of Non-Invasive Biomarkers for Chronic Atrophic Gastritis from Serum Exosomal microRNAs. *BMC Cancer* 19 (1), 129. doi:10.1186/s12885-019-5328-7

Ma, Z., Han, C., Xia, W., Wang, S., Li, X., Fang, P., et al. (2020). circ5615 Functions as a ceRNA to Promote Colorectal Cancer Progression by Upregulating TNKS. *Cell Death Dis.* 11 (5), 356. doi:10.1038/s41419-020-2514-0

Melo, S. A., Luecke, L. B., Kahlert, C., Fernandez, A. F., Gammon, S. T., Kaye, J., et al. (2015). Glypican-1 Identifies Cancer Exosomes and Detects Early Pancreatic Cancer. *Nature* 523 (7559), 177–182. doi:10.1038/nature14581

Mousavi, S., Moallem, R., Hassanian, S. M., Sadeghzade, M., Mardani, R., Ferns, G. A., et al. (2019). Tumor-Derived Exosomes: Potential Biomarkers and Therapeutic Target in the Treatment of Colorectal Cancer. *J. Cell. Physiol.* 234 (8), 12422–12432. doi:10.1002/jcp.28080

Nabariya, D. K., Pallu, R., and Yenuganti, V. R. (2020). Exosomes: The Protagonists in the Tale of Colorectal Cancer? *Biochim. Biophys. Acta (BBA) - Rev. Cancer* 1874 (2), 188426. doi:10.1016/j.bbcan.2020.188426

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinf.* 12, 77. doi:10.1186/1471-2105-12-77

Schreuders, E. H., Ruco, A., Rabeneck, L., Schoen, R. E., Sung, J. J. Y., Young, G. P., et al. (2015). Colorectal Cancer Screening: A Global Overview of Existing Programmes. *Gut* 64 (10), 1637–1649. doi:10.1136/gutjnl-2014-309086

Schwartzentruber, J., Korshunov, A., Liu, X.-Y., Jones, D. T. W., Pfaff, E., Jacob, K., et al. (2012). Driver Mutations in Histone H3.3 and Chromatin Remodelling Genes in Paediatric Glioblastoma. *Nature* 482 (7384), 226–231. doi:10.1038/nature10833

Shi, H., Wang, M., Sun, Y., Yang, D., Xu, W., and Qian, H. (2021). Exosomes: Emerging Cell-Free Based Therapeutics in Dermatologic Diseases. *Front. Cell Dev. Biol.* 9, 736022. doi:10.3389/fcell.2021.736022

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer Statistics, 2018. *CA A Cancer J. Clin.* 68 (1), 7–30. doi:10.3322/caac.21442

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D.-A., Jones, D. T. W., Konermann, C., et al. (2012). Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell* 22 (4), 425–437. doi:10.1016/j.ccr.2012.08.024

Wang, J., Yan, F., Zhao, Q., Zhan, F., Wang, R., Wang, L., et al. (2017). Circulating Exosomal miR-125a-3p as a Novel Biomarker for Early-Stage Colon Cancer. *Sci. Rep.* 7 (1), 4150. doi:10.1038/s41598-017-04386-1

Wang, Y., Chen, Y., Xiao, S., and Fu, K. (2020). Integrated Analysis of the Functions and Prognostic Values of RNA-Binding Proteins in Colorectal Cancer. *Front. Cell Dev. Biol.* 8, 595605. doi:10.3389/fcell.2020.595605

Wu, G., Broniscer, A., McEachron, T. A., Lu, C., Paugh, B. S., Becksfort, J., et al. (2012). Somatic Histone H3 Alterations in Pediatric Diffuse Intrinsic Pontine Gliomas and Non-Brainstem Glioblastomas. *Nat. Genet.* 44 (3), 251–253. doi:10.1038/ng.1102

Xiao, Y., Li, Y., Yuan, Y., Liu, B., Pan, S., Liu, Q., et al. (2019). The Potential of Exosomes Derived from Colorectal Cancer as a Biomarker. *Clin. Chim. Acta* 490, 186–193. doi:10.1016/j.cca.2018.09.007

Xiao, Y., Zhong, J., Zhong, B., Huang, J., Jiang, L., Jiang, Y., et al. (2020). Exosomes as Potential Sources of Biomarkers in Colorectal Cancer. *Cancer Lett.* 476, 13–22. doi:10.1016/j.canlet.2020.01.033

Yadavalli, S., Jayaram, S., Manda, S. S., Madugundu, A. K., Nayakanti, D. S., Tan, T. Z., et al. (2017). Data-Driven Discovery of Extravasation Pathway in Circulating Tumor Cells. *Sci. Rep.* 7, 43710. doi:10.1038/srep43710

Yan, J., Nie, K., Mathew, S., Tam, Y., Cheng, S., Knowles, D. M., et al. (2014). Inactivation of BANK1 in a Novel IGH-Associated Translocation T(4;14)(q24;q32) Suggests a Tumor Suppressor Role in B-Cell Lymphoma. *Blood Cancer J.* 4 (5), e215. doi:10.1038/bcj.2014.36

Yáñez-Mó, M., Siljander, P. R.-M., Andreu, Z., Bedina Zavec, A., Borràs, F. E., Buzas, E. I., et al. (2015). Biological Properties of Extracellular Vesicles and Their Physiological Functions. *J. Extracell. Vesicles* 4, 27066. doi:10.3402/jev.v4.27066

Yang, D., Zhang, W., Zhang, H., Zhang, F., Chen, L., Ma, L., et al. (2020). Progress, Opportunity, and Perspective on Exosome Isolation - Efforts for Efficient Exosome-Based Theranostics. *Theranostics* 10 (8), 3684–3707. doi:10.7150/thno.41580

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

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