DISSECTING THE PATHOGENESIS OF COMPLEX DISEASES BASED ON GENOME VARIATION TO PROMOTE THE DEVELOPMENT OF PRECISION MEDICINE

EDITED BY: Peng Wang, Xinyi Liu, Fan Zhang and Hui Zhi

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-037-8
DOI 10.3389/978-2-83250-037-8

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

DISSECTING THE PATHOGENESIS OF COMPLEX DISEASES BASED ON GENOME VARIATION TO PROMOTE THE DEVELOPMENT OF PRECISION MEDICINE

Topic Editors:
Peng Wang, Harbin Medical University, China
Xinyi Liu, University of Illinois at Chicago, United States
Fan Zhang, University of North Texas Health Science Center, United States
Hui Zhi, Harbin Medical University, China

Citation: Wang, P., Liu, X., Zhang, F., Zhi, H., eds. (2022). Dissecting the Pathogenesis of Complex Diseases Based on Genome Variation to Promote the Development of Precision Medicine. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-037-8

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fgene.2021.680413

Identification of a Prognostic Signature for Ovarian Cancer Based on the Microenvironment Genes

Xiao Huo[1†], Hengzi Sun[2†], Shuangwu Liu[3], Bing Liang[2], Huimin Bai[2], Shuzhen Wang[2] and Shuhong Li[2*]*

[1] Peking University Third Hospital Institute of Medical Innovation and Research, Beijing, China, [2] Department of Obstetrics and Gynecology, Beijing Chao-Yang Hospital, Capital Medical University, Beijing, China, [3] School of Medicine, ShanDong University, Jinan, China

OPEN ACCESS

Edited by:
Peng Wang,
Harbin Medical University, China

Reviewed by:
Dong Lu,
Baylor College of Medicine, United States
Sisi Chen,
Mayo Clinic, United States

****Correspondence:***
Shuzhen Wang
darrywang2003@163.com
Shuhong Li
lishuhongcyyy@163.com

† These authors have contributed equally to this work

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

Received: *14 March 2021*
Accepted: *15 April 2021*
Published: *13 May 2021*

Citation:
Huo X, Sun H, Liu S, Liang B, Bai H, Wang S and Li S (2021) Identification of a Prognostic Signature for Ovarian Cancer Based on the Microenvironment Genes. Front. Genet. 12:680413. doi: 10.3389/fgene.2021.680413

Background: Ovarian cancer is highly malignant and has a poor prognosis in the advanced stage. Studies have shown that infiltration of tumor microenvironment cells, immune cells and stromal cells has an important impact on the prognosis of cancers. However, the relationship between tumor microenvironment genes and the prognosis of ovarian cancer has not been studied.

Methods: Gene expression profiles and SNP data of ovarian cancer were downloaded from the TCGA database. Cluster analysis, WGCNA analysis and univariate survival analysis were used to identify immune microenvironment genes as prognostic signatures for predicting the survival of ovarian cancer patients. External data were used to evaluate the signature. Moreover, the top five significantly correlated genes were evaluated by immunohistochemical staining of ovarian cancer tissues.

Results: We systematically analyzed the relationship between ovarian cancer and immune metagenes. Immune metagenes expression were associated with prognosis. In total, we identified 10 genes related to both immunity and prognosis in ovarian cancer according to the expression of immune metagenes. These data reveal that high expression of ETV7 (OS, HR = 1.540, 95% CI 1.023–2.390, $p = 0.041$), GBP4 (OS, HR = 1.834, 95% CI 1.242–3.055, $p = 0.004$), CXCL9 (OS, HR = 1.613, 95% CI 1.080 – 2.471, $p = 0.021$), CD3E (OS, HR = 1.590, 95% CI 1.049 –2.459, $p = 0.031$), and TAP1 (OS, HR = 1.766, 95% CI 1.163 –2.723, $p = 0.009$) are associated with better prognosis in patients with ovarian cancer.

Conclusion: Our study identified 10 immune microenvironment genes related to the prognosis of ovarian cancer. The list of tumor microenvironment-related genes provides new insights into the underlying biological mechanisms driving the tumorigenesis of ovarian cancer.

Keywords: ovarian cancer, microenvironment, immune metagenes, prognosis, TCGA

INTRODUCTION

Cancer seriously endangers human health, and in recent years, the incidence of malignant tumors has increased annually. The World Health Organization reported 18.1 million new cancer cases and 9.6 million cancer-related deaths worldwide in 2018. Ovarian cancer is a common gynecologic malignancy and the fifth leading cause of cancer-related deaths in women (Siegel et al., 2018). The lifetime risk of ovarian cancer in women is 1.3%. The 5-year survival rate ranged from 29 to 93%, depending on the initial diagnosis (Torre et al., 2018). Despite advances in treatment strategies and techniques, the mortality rate of ovarian cancer remains high. The main reason is the lack of obvious symptoms and effective screening for ovarian cancer. Sixty percent of patients were diagnosed with advanced ovarian cancer (Dinh et al., 2008). Standard treatment for advanced ovarian cancer includes tumor cell destruction and standard chemotherapy. However, most patients relapse within 2–3 years after first-line chemotherapy and die as a consequence of chemotherapy resistance (Odunsi, 2017). Thus, new treatment strategies and paradigms are greatly needed for these patients.

Malignant solid tumor tissue is heterogeneous and includes not only tumor cells but also tumor-associated normal epithelial and stromal cells, immune cells and vascular cells. The process of tumor development depends on a variety of complex signaling pathways between tumor cells and the tumor microenvironment (Kreuzinger et al., 2017). With the improvement of understanding the molecular basis of immune recognition and immune regulation in tumor cells, immunotherapy has aroused great interest (Nelson, 2015). Tumor microenvironment cells and the degree of infiltration of immune and stromal cells in tumors have been reported to significantly contribute to the prognosis. In the tumor microenvironment, immune and stromal cells are two main types of non-tumor components and have been proposed to be valuable for the diagnosis and prognosis evaluations of tumors (Senbabaoglu et al., 2016; Winslow et al., 2016; Ovarian Tumor Tissue Analysis (Otta) Consortium, Goode et al., 2017). Many algorithms have been developed to calculate tumor purity using gene expression and DNA methylation data (Carter et al., 2012; Yoshihara et al., 2013; Zheng et al., 2017). The immune and stromal scores calculated based on the ESTIMATE algorithm (Yoshihara et al., 2013) promote the quantitative determination of immune and stromal components in tumors. In this algorithm, the authors calculated immune and stromal scores by analyzing specific gene expression characteristics of immune and stromal cells to predict non-tumor cell infiltration. This algorithm has been applied to prostate cancer (Shah et al., 2017) and breast cancer (Jia et al., 2018), and the results show the effectiveness of this algorithm, but there are no detailed studies on ovarian cancer.

The Cancer Genome Atlas (TCGA) has been established to improve cancer prevention, diagnosis and treatment by applying high-throughput genome analysis techniques to provide a better understanding of cancer (Cancer Genome Atlas Research Network, 2008). To better understand the effect of immune microenvironment-related genes on the prognosis of ovarian cancer, we systematically analyzed the expression profile data in the TCGA database and mined the genes related to the microenvironment of ovarian cancer and poor prognosis. Finally, we obtained a set of microenvironment genes associated with poor prognosis in ovarian cancer patients and validated them with the online tool KMplot[1].

MATERIALS AND METHODS

Data Source and Data Pre-processing

TCGA Data

We used the GDC API to download level 3 data for OC patients from the TCGA database[2] (December 26, 2018). The data included the following: (1) RNA-seq data (n = 379). The Fragment Per Kilobase of transcript per Million mapped reads (FPKM) data of RNA-Seq were downloaded from the TCGA and further converted into Transcript Per Million (TPM) expression profiles and RNA-Seq Count data; (2) Single nucleotide polymorphism (SNP) data (n = 436); and (3) Clinical follow-up information (n = 587) including survival and outcome.

Immune Metagenes Scores

Thirteen kinds of immune metagenes, which correspond to various types of immune cells and reflect various immune functions, were identified from previous reports (Safonov et al., 2017). For each sample, according to the gene expression levels of immune metagenes, we selected the median expression level of each type of immune metagenes and designated these levels as the immune metagenes score for these samples.

Immune Cell Scores

We downloaded the scores of six types of immune cells corresponding to each sample of ovarian cancer from the Tumor Immune Estimation Resource (TIMER)[3] database. The six types of immune cells were B cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils, macrophages and dendritic cells.

Immune Scores and Stromal Scores

Stromal and immune scores were estimated from transcriptomic profiles of the ovarian cancer cohort from TCGA using the ESTIMATE algorithm. We used the R software package estimate to calculate the immune and stromal scores of each sample. ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) is a tool for predicting tumor purity, and the presence of infiltrating stromal/immune cells in tumor tissues using gene expression data. ESTIMATE algorithm is based on single sample Gene Set Enrichment Analysis and generates three scores: stromal score (that captures the presence of stroma in

[1]http://kmplot.com

[2]http://cancergenome.nih.gov

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

tumor tissue), immune score (that represents the infiltration of immune cells in tumor tissue), and estimate score (that infers tumor purity).

Overall Survival Curve and Differential Expression Analysis

The data were processed: (1) KM plots were generated to illustrate the relationship between patients' overall survival and gene expression levels of immune metagenes. The relationship was tested by log-rank test. (2) Weighted gene co-expression network analysis (WGCNA), an R software package (Langfelder and Horvath, 2008; Wang et al., 2019), was used to construct a weighted co-expression network. A soft threshold of 8 was selected to screen the co-expression modules. The protein-protein interaction (PPI) network was retrieved from STRING database (Szklarczyk et al., 2015) and reconstructed via Cytoscape software (Shannon et al., 2003; Wang et al., 2020). (3) The R software package clusterProfiler for KEGG enrichment analysis was used, and a significance of false discover rate (FDR) < 0.05 was selected. (4) Data analysis was performed using the package DESeq2. The log2 (Foldchange)$| > 1$ and FDR < 0.05 were set as the cutoff values to screen for differentially expressed genes.

Immunohistochemical Staining (IHC)

We collected a total of 168 human ovarian cancer tissue samples, which had accompanying follow-up information, from archives of paraffin-embedded tissues between January 2010 and January 2015 at the Department of Pathology of Beijing Chao-Yang Hospital. The follow-up was performed until December 31, 2020. The pathological diagnoses were reconfirmed by a pathologist. The patients included in present study were all (1) Epithelial ovarian cancer, (2) Underwent cytoreductive surgery and subsequent chemotherapy, (3) With follow-up information. The exclusion criteria were (1) Ovarian germ cell tumor, ovarian sex cord stromal tumor or metastatic cancer, (2) Unstandardized treatment, (3) No informed consent, (4) Lost to follow-up, and (5) No enough pathological samples.

The project was approved by the Ethical Committee (Beijing Chao-Yang Hospital), and informed consent was acquired from patients. IHC was performed as previously described (Li et al., 2010). Antibodies against the following were used: ETV7 1:200 abcam ab229832, GBP4 1:50 abcam ab232693, CXCL9 1:100 abcam ab137792, CD3E 1:500 abcam ab237721, TAP1 1:200 abcam ab137013. The scoring details have been described previously (Zhang et al., 2015). The intensity of immunostaining was graded as follows: 1+, weak; 2+, moderate; 3+, strong or 4+, very strong. The area of positive cancer cells in each microscopic field was categorized as follows: 1+, 0–25%; 2+, 25–50%; 3+, 50–75%, or 4+, 75–100%. The sum between 5 and 80 was obtained by multiplying the two scores by 5. A sum from 0 to 42 was assigned as "low expression" and that from 43 to 80 as "high expression." All pathological diagnoses were confirmed in a blinded manner by three expert pathologists.

RESULTS

Correlation Analysis of Immune Metagenes With Immunological Components in the Tumor Microenvironment

To observe the relationship between 13 types of immune metagenes scores, we calculated the correlation between them, as shown in **Supplementary Figure 1A**. The average correlations of natural killer cells (NK), regulatory T cells (Tregs), interferon-inducible genes (IF_I) and major histocompatibility complex class II antigen (MHC2) with other metagenes were the smallest and were 0.08157227, 0.23253018, 0.3120958, and 0.398014, respectively. The other classes of metagenes were highly correlated, which indicates that there is a certain consistency in the expression of metagenes in ovarian cancer. Furthermore, we analyzed the immune metagenes scores and six kinds of immune cells in the tumor microenvironment, as shown in **Supplementary Figure 1B**. We found that in addition to NK, Tregs and IF_I have smaller correlations with the content of six kinds of immune cells, and the scores for other metagenes were >0.4, suggesting that the immune metagenes and immune cells in the immune microenvironment have a significant correlation. Finally, we calculated the correlation between immune metagenes and immune and stromal scores, as shown in **Supplementary Figure 1C**. The correlation of the other 10 types of immune metagenes, except for NK, IF_I and Tregs, was very high, with an average higher than 0.4. In conclusion, the expression of these immune metagenes was closely related to the immune components in the tumor microenvironment.

Relationship Between Immune Metagenes and Clinical Stage

According to the expression levels and stages of immune metagenes in each sample, we calculated the expression level distribution of immune metagenes in different stages, as shown in **Supplementary Figures 2A–M** (the number of Stage I samples was too small to be counted, so we counted only Stages II-IV). Immune metagenes showed a trend of successively declining expression of Stages II-IV, and ImmuneScore, follicular helper T cells (Tfh) and signal transducer and activator of transcription 1 (STAT1) had significant differences in various stages. The prognostic differences in the four stages were further analyzed as shown in **Supplementary Figure 2N**, and different stages had significant prognostic differences. This result suggests that the expression of immune metagenes may be closely related to the prognosis of ovarian cancer.

Prognostic Difference Analysis of Immune Metagenes

To observe the expression and prognosis of the relationship between immune metagenes, we classified as high- and low-expression samples according to the median expression of metagenes. KM plots was used for prognostic difference analysis, as shown in **Figures 1A–M**. In all immune metagenes, the

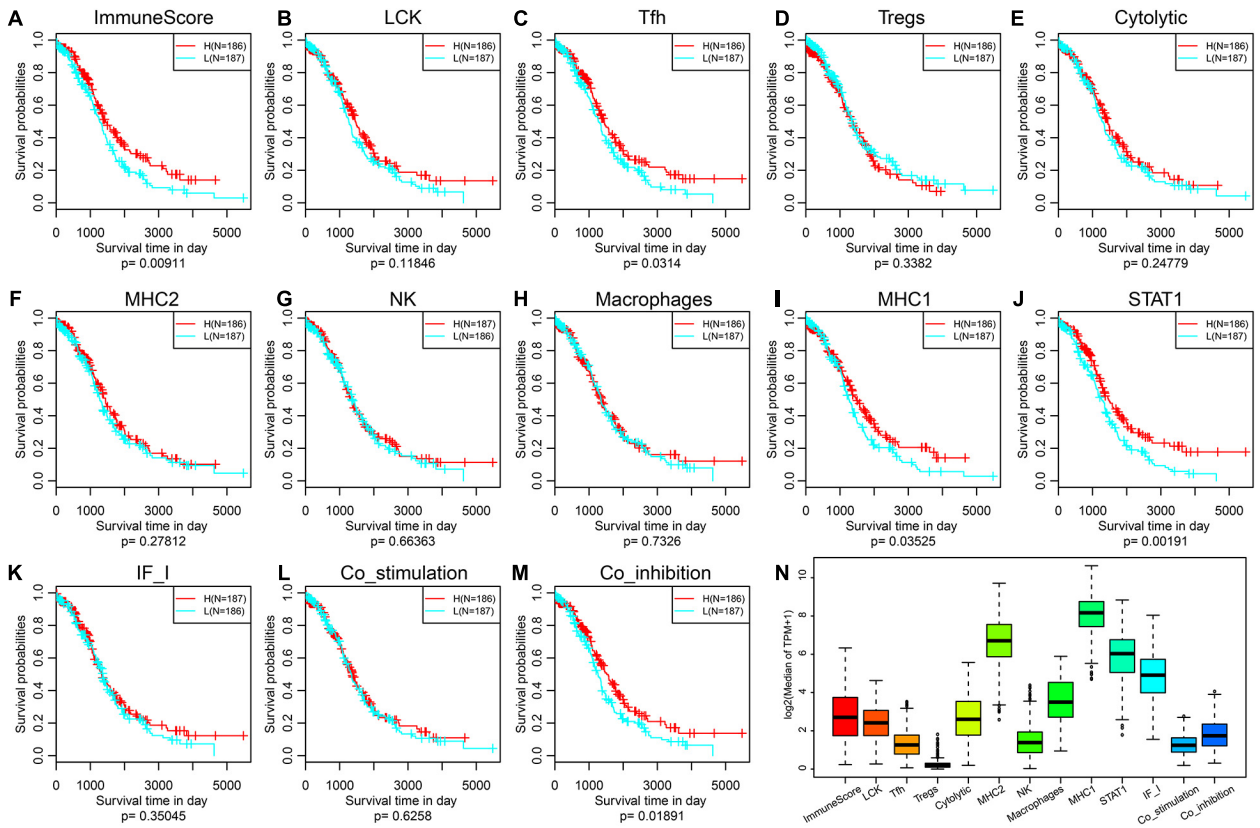


FIGURE 1 | (A–M) KM curve of the expression and prognosis of immune metagenes; group H represents high-expression genes, and group L represents low-expression genes. **(N)** The expression distribution of immune metagenes.

low-expression group had a worse prognosis than the high-expression group, in which ImmuneScore, Tfh, MHC1, STAT1 and Co_inhibition showed significant differences in prognosis, suggesting that the high expression of metagenes was a good prognostic factor. Next, we analyzed the expression distribution of immune metagenes as shown in **Figure 1N**. Except for the low expression of Tregs, the median expression of other types of metagenes was generally high. This result suggests that these immune metagenes are commonly expressed genes in ovarian cancer, indicating the potential of these metagenes as a new prognostic marker.

Relationship Between Immune Metagenes and BRCA Mutations

BRCA genes are tumor suppressor genes that play important roles in cell replication regulation, DNA damage repair and normal cell growth. If BRCA genes are mutated, they will lose their ability to inhibit tumorigenesis. There are hundreds of BRCA mutation types, which are related to the occurrence of many cancers in the human body; among these cancers, breast cancer is the most closely related to *BRCA* mutations, followed by ovarian cancer. Therefore, we analyzed the relationship between these immune metagenes and BRCA1 and BRCA2 mutations. First, MuTect (Cibulskis et al., 2013) was used to process SNP data downloaded from the TCGA and to extract mutation data of BRCA1 and BRCA2. The expression relationship of immune metagenes in the BRCA1 mutation group and wild-type group samples was analyzed as shown in **Supplementary Figures 3A–M**. There were eight immune metagenes with significant expression differences, and the expression of the wild-type group was significantly higher than the mutant group. In addition, Macrophages, MHC1 and STAT1 had no significant differences, but the *P*-value was on the edge of significance. Second, we analyzed the differences in expression for immune metagenes between the BRCA2 mutation group and the wild-type group, as shown in **Supplementary Figures 3N–Z**. There were no significant differences in metagene expression among immune metagenes. This finding is consistent with previous studies and shows that BRCA2 mutations in ovarian cancer have no prognostic significance (Goode et al., 2017).

WGCNA Analysis Mining Immune Metagenes Related Modules

To further excavate the prognosis of ovarian cancer immune microenvironment-related markers, we obtained the expression data for a total of 379 samples. A total of 15,268 transcripts with more than 75% TPM $>$ 1 and median absolute deviation $>$ median was selected from these samples. First,

hierarchical clustering was used for cluster analysis of the samples, as shown in **Figure 2A**. There were some outlier samples. We screened the samples with a distance of more than 47,000 as outlier samples and finally obtained a total of 328 samples. Second, Pearson correlation coefficient was used to calculate the distance between each transcript. WGCNA was used to construct a weighted co-expression network. A soft threshold of 8 was selected to screen the co-expression modules. The research showed that the co-expression network conforms to the scale-free network; that is, the log(k) of the node with connectivity k was negatively correlated with the log(P(k) of the probability of the node, and the correlation coefficient was >0.8. To ensure that the network was scale-free, we select $\beta = 8$ (**Figures 2B,C**). Third, the expression matrix was transformed into an adjacency matrix, and then the adjacency matrix was transformed into a topological matrix. Based on the topology overlap matrix (TOM), we used the average-linkage hierarchical clustering method to cluster the genes. According to the standard of the hybrid dynamic shear tree, the minimum number of genes in each gene network module was set to 30. After determining the gene module by using the dynamic shearing method, we successively calculated the characteristic vector value (eigengenes) of each module and then performed cluster analysis on the module to merge the modules that were close to each other into new modules. Height = 0.25, deepSplit = 2, and minModuleSize = 30 were the set values. A total of 62 modules were obtained (**Figure 2D**). The gray module is the gene set that cannot be aggregated into other modules. The transcript statistics of each module are shown in **Supplementary Table 4**, from which 8,047 transcripts were assigned to 62 co-expression modules. We calculated the correlation between the feature vectors of the 62 modules and the immune metagenes, as shown in **Figure 2E**. The sienna3, yellow, antiquewhite4 and ivory modules have the highest correlations with the immune metagenes, with an average correlation >0.39. The number of transcripts in the four modules was 69, 378, 33, and 54, respectively, containing a total of 534 genes.

We further analyzed the function of genes in the four modules most related to immune metagenes. Among the four modules, the sienna3 module was enriched into 13 pathways. The yellow module was enriched into 54 pathways. The antiquewhite4 module was enriched into 23 pathways. The ivory module was enriched in 20 pathways. The relationship between the pathways enriched by these four modules was analyzed (**Figure 3**); There are 70 pathways enriched by the four modules, of which 31 are enriched by two or more modules, respectively. This result indicates that there are many intersections between the enriched pathways, of which eight are enriched by three modules at the same time (allograft rejection, autoimmune thyroid disease, cell adhesion molecules, Epstein-Barr virus infection, graft-vs.-host disease, herpes simplex infection, human T-cell leukemia virus 1 infection NK cell-mediated cytotoxicity, and type I diabetes mellitus). These pathways are closely related to immunity and cell adhesion.

To select genes associated with immune metagenes, we calculated the correlation between the gene and module and analyzed the correlation distribution of these genes as shown in **Supplementary Figure 5**. These correlation coefficients presented a bimodal distribution. With 0.72 as the critical point, we selected 248 genes with the maximum correlation coefficient >0.72.

Differential Gene Analysis of Immune Differential Samples

Most of the immune metagenes are related to the prognosis, and the most significant type of immune metagenes such as ImmuneScore and STAT1 were selected. First, samples were divided into two groups, high ImmuneScore group and low ImmuneScore group, based on the average according to the ImmuneScore level. Then, the R software package DESeq2 was used to analyze the differentially expressed genes between the two groups of samples. In total, 219 differentially expressed genes were obtained, as shown in **Supplementary Figure 6A**, indicating that the up-regulated genes were significantly larger than the down-regulated genes and that up-regulated multiple genes was larger than the down-regulated multiple genes, in general. The expression profiles of these 219 genes are further visualized in **Supplementary Figure 6B**; there were obvious differences in the expression of differentially expressed genes in the high ImmuneScore group and low ImmuneScore group. Similarly, the samples were divided into two groups, the high STAT1 group and the low STAT1 group, based on the average according to the level of STAT1. Differentially expressed genes were screened by DESeq2, as shown in **Supplementary Figures 6C,D**. The differences in the STAT1 distribution results are similar to those in the ImmuneScore, and the expression levels were significantly higher for high-expression genes than in low-expression genes.

Screening of Immune Microenvironment Genes With Prognostic Value

To further analyze the co-expression relationship between genes with different immune scores and immune metagenes, we integrated 248 genes associated with the four most relevant metagenes modules, 219 genes with differential expression from ImmuneScore and 211 genes with differential expression from STAT1. We selected a total of 70 genes from all three, excluding 24 genes in 13 immune metagenes and resulting in 46 genes, as shown in **Figure 4A**. Next, we used the R software package clusterProfiler for KEGG enrichment analysis of these genes, and the selection threshold FDR < 0.05 is shown in **Figure 4B**. A total of 19 genes were enriched into 12 pathways, and most of these genes are related to immune diseases. The protein network interaction of these 46 genes were analyzed by using the R package STRINGdb. First, the 46 genes were mapped into the STRING database, and the network relationships among these genes were obtained as shown in **Figure 4C**. A total of 104 edges and 40 nodes were obtained. We analyzed the degree distribution of nodes in these networks as shown in **Figure 4D**. From this result, the degree of each node is higher, with an average degree of 5.7, indicating that these genes are closely related.

To screen genes with prognostic value in the immune microenvironment, we first analyzed the relationship between the expression of these 46 genes and prognosis using univariate

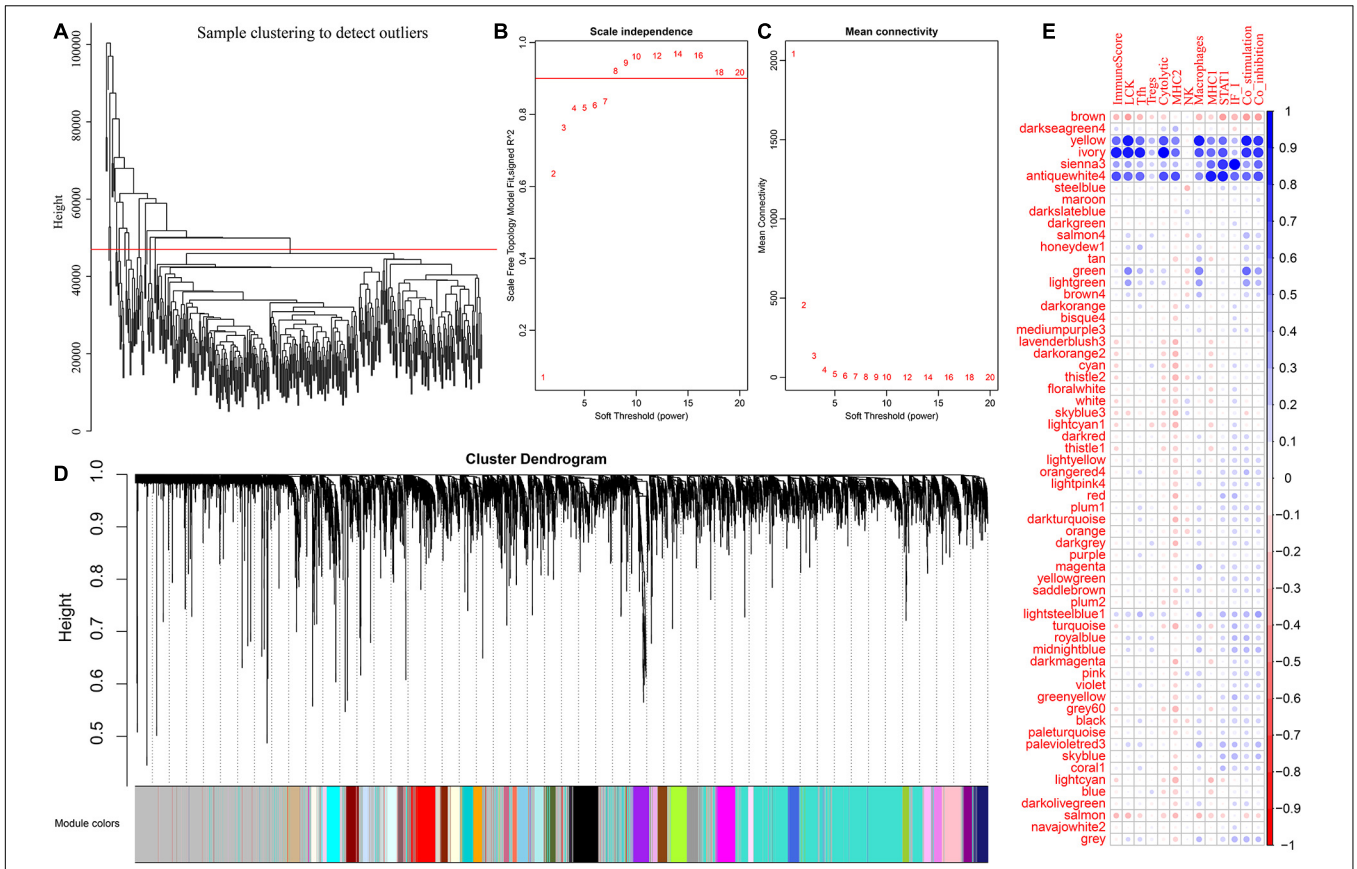


FIGURE 2 | (A) The sample clustering analysis. **(B,C)** Analysis of network topology for various soft-thresholding powers. **(D)** The figure shows gene dendrogram and module colors. **(E)** The correlation between each module and the expression of immune metagenes.

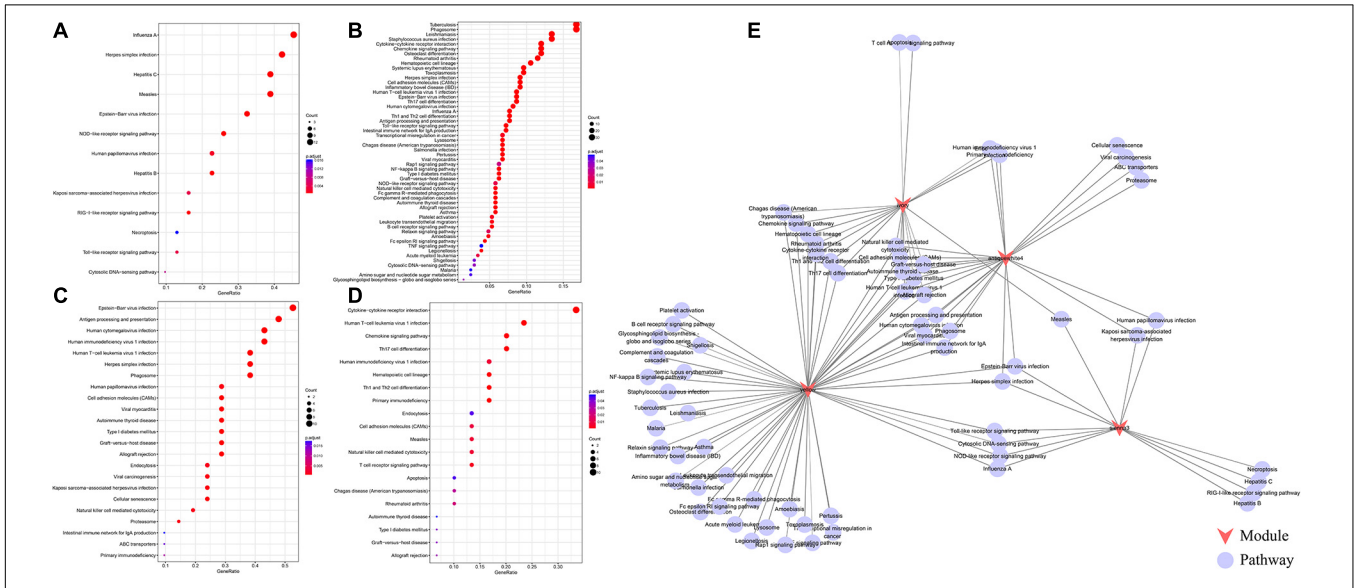


FIGURE 3 | (A) The enrichment results of the sienna3module. The larger the circle is, the more module genes are containedin the pathway. The redder the color is, the more significant the gene was. **(B)** The pathways enriched by the yellow module.**(C)** The pathways enriched by the antiquewhite4 module. **(D)** The pathways enriched by the ivory module. **(D)** An interactive network of pathways enriched by the four modules. **(E)** An interactive network of pathways enriched by the four modules.

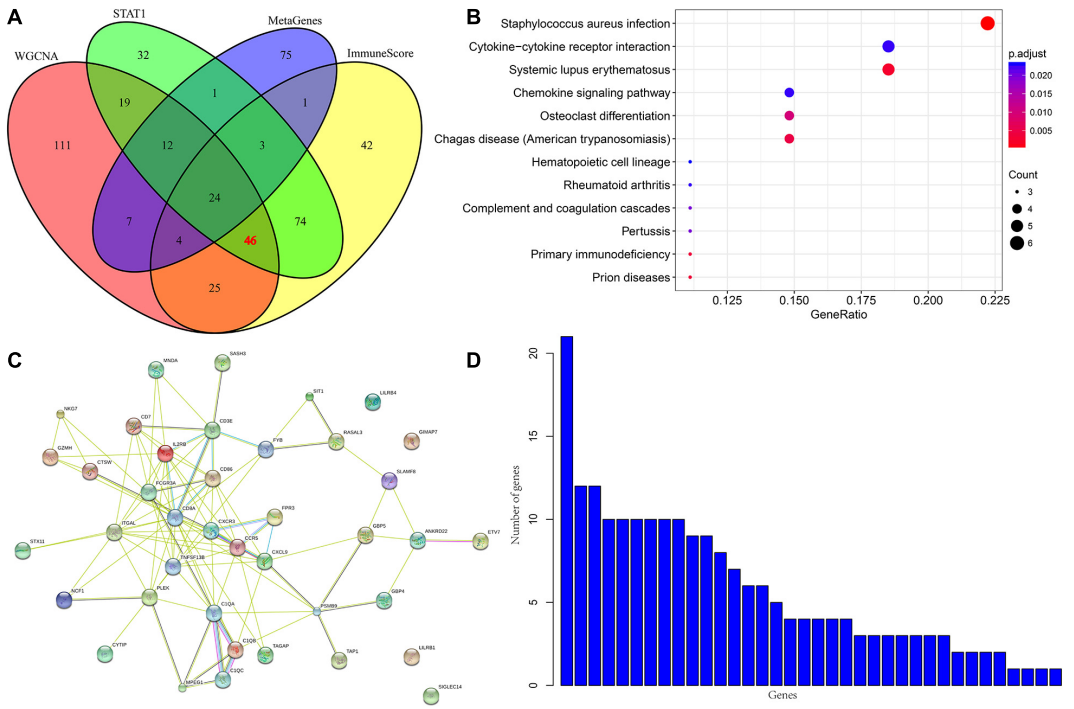


FIGURE 4 | (A) Venn diagram. **(B)** Analysis results of KEGG enrichment. **(C)** Protein-protein interaction network. **(D)** Degree distribution of protein-protein interaction network.

TABLE 1 | Genes with prognostic value.

Genes	*p*-value	HR	Low 95%CI	High 95%CI
ENSG00000225492	0.00023	0.9584	0.936978	0.980312
ENSG00000168394	0.001941	0.995765	0.993095	0.998441
ENSG00000138755	0.002792	0.994868	0.991517	0.998229
ENSG00000240065	0.00507	0.995758	0.992802	0.998723
ENSG00000211753	0.007477	0.975419	0.957792	0.993371
ENSG00000162654	0.008348	0.991075	0.984494	0.997699
ENSG00000211772	0.010813	0.986706	0.976604	0.996913
ENSG00000256262	0.013449	0.951503	0.914724	0.989761
ENSG00000010030	0.019584	0.978497	0.9608	0.996521
ENSG00000198851	0.02893	0.982099	0.966311	0.998146
ENSG00000277734	0.030677	0.990207	0.98141	0.999084
ENSG00000154451	0.031368	0.974711	0.95224	0.997713
ENSG00000152766	0.037442	0.961857	0.927262	0.997742
ENSG00000206337	0.048843	0.994756	0.989566	0.999973

survival analysis based on the prognostic information of the samples. A total of 14 genes were obtained by selecting $p < 0.05$ as the threshold, as shown in **Table 1**. The hazard ratio (HR) of these 14 genes was less than 1, and their high expression was related to good prognosis. Furthermore, we used clinical stages as a covariant to analyze the relationship between these genes and prognosis to exclude the influence of clinical stages and ultimately obtained 10 independent prognostic factors, as shown in **Table 2**.

According to the expression levels of these 10 prognostic genes (CXCL9, ETV7, GBP4, TRBC2, GBP1P1, CD3E, USP30-AS1, TRBV28, TAP1, and PSMB9), we divided the samples into two groups according to the median expression levels. The prognostic differences between the high-expression group and the low-expression group were analyzed. As shown in **Supplementary Figures 7**, **9** of the 10 genes with a high-expression prognosis were significantly better than the low-expression prognosis. There was a significant trend in the TRBV28 gene, but it was not obvious. This may be because the 5-year survival rate is inseparable, but the prognosis is obviously different after 5 years.

To verify the relationship between these 10 genes and prognosis, we used the online tool KMplot to analyze the relationship between these 10 genes and overall survival in ovarian cancer. We retrieved 6 genes from the KMplot platform. The KM curves of these 6 genes (two of which have two probes) are shown in **Figure 5**, and six genes were characterized by a high expression of prognosis as being good. Five of these genes (ETV7, GBP4, CXCL9, CD3E, and TAP1) are significantly correlated with prognosis, which is consistent with our analysis.

Evaluation of the Prognosis of Ovarian Cancer and Hub Genes by IHC

From January 2010 and January 2015, 168 human ovarian tissue samples which had accompanying follow-up information. **Supplementary Table 8** summarizes the characteristics of all patients, including age, disease stage, and tumor grade. We selected the five hub genes to evaluate gene expression values by IHC. The expression of ETV7 (33.13 ± 1.65), GBP4 (28.48 ± 1.48), CXCL9 (23.30 ± 1.30), CD3E (36.52 ± 1.59), and

TABLE 2 | Stages were introduced as covariates to obtain significant prognostic genes.

Genes	*p*-value	HR	Low 95%CI	High 95%CI	Entrezid	Symbol
ENSG00000138755	0.00849	0.995426	0.992033	0.99883	4,283	CXCL9
ENSG00000010030	0.04061	0.981223	0.963579	0.99919	51,513	ETV7
ENSG00000162654	0.022077	0.992104	0.985392	0.998861	115,361	GBP4
ENSG00000211772	0.02085	0.987918	0.977784	0.998157	28638	TRBC2
ENSG00000225492	0.00091	0.962018	0.940256	0.984283	400,759	GBP1P1
ENSG00000198851	0.048451	0.983727	0.967827	0.999888	916	CD3E
ENSG00000256262	0.026533	0.956429	0.919515	0.994825	100,131,733	USP30-AS1
ENSG00000211753	0.013667	0.977354	0.959721	0.995311	28,559	TRBV28
ENSG00000168394	0.005226	0.996155	0.993465	0.998852	6,890	TAP1
ENSG00000240065	0.009633	0.996091	0.993141	0.999049	5,698	PSMB9

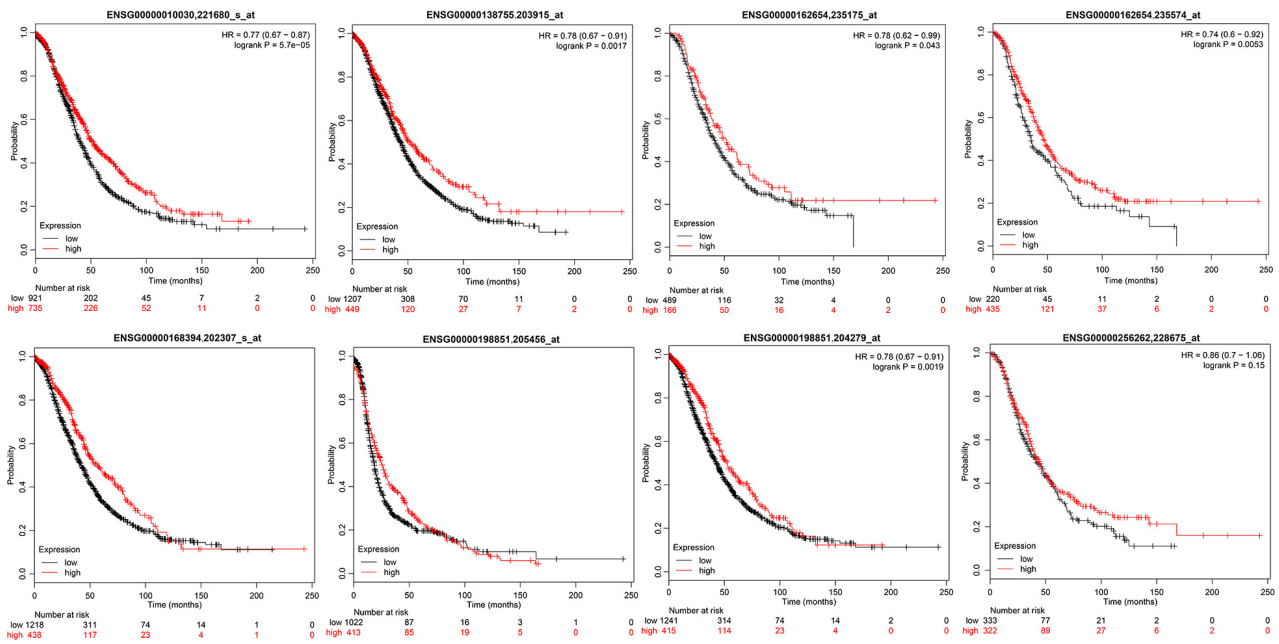


FIGURE 5 | The prognostic KM curve of the seven genes in the KMplot platform.

TAP1 (29.94 ± 1.37) are shown in **Figures 6A–K**. The correlation between expression of these genes and ovarian cancer prognosis is shown in **Figures 6L–P**. These data reveal that high expression of ETV7 (OS, HR = 1.540, 95% CI 1.023–2.390, p = 0.041), GBP4 (OS, HR = 1.834, 95% CI 1.242–3.055, p = 0.004), CXCL9 (OS, HR = 1.613, 95% CI 1.080 –2.471, p = 0.021), CD3E (OS, HR = 1.590, 95% CI 1.049 –2.459, p = 0.031), and TAP1 (OS, HR = 1.766, 95% CI 1.163 –2.723, p = 0.009) are associated with better prognosis in patients with ovarian cancer.

DISCUSSION

Ovarian cancer is the most common cause of death from gynecologic malignancy (Torre et al., 2015). Epithelial ovarian cancer (EOC) is the most common ovarian tumor with a lack of specific clinical symptoms at early stage, 75% of patients were diagnosed with advanced tumors (FIGO III/IV), and the standard of treatment was complete resection of all visible tumor lesions and platinum-based chemotherapy (Ferlay et al., 2015). Although most patients with advanced ovarian cancer respond to standard ovarian cancer therapeutic approaches, 70% of patients will eventually relapse and develop chemotherapy resistance (Hennessy et al., 2009). Therefore, more effective prognostic and therapeutic strategies to reduce the mortality rate of ovarian cancer are being actively explored. Stromal cells, extracellular matrix and exosomes comprise the tumor microenvironment. Intrinsic genes of tumor cells, especially master transcription factors, determine the occurrence, development and evolution of ovarian cancer, but the surrounding microenvironment interacts with tumor cells through secretory interactions, providing an impetus for the invasion and metastasis of tumor cells (Pietras and Ostman, 2010). In recent years, the tumor microenvironment has gradually been considered to play an important role in ovarian cancer metastasis and may become a potential

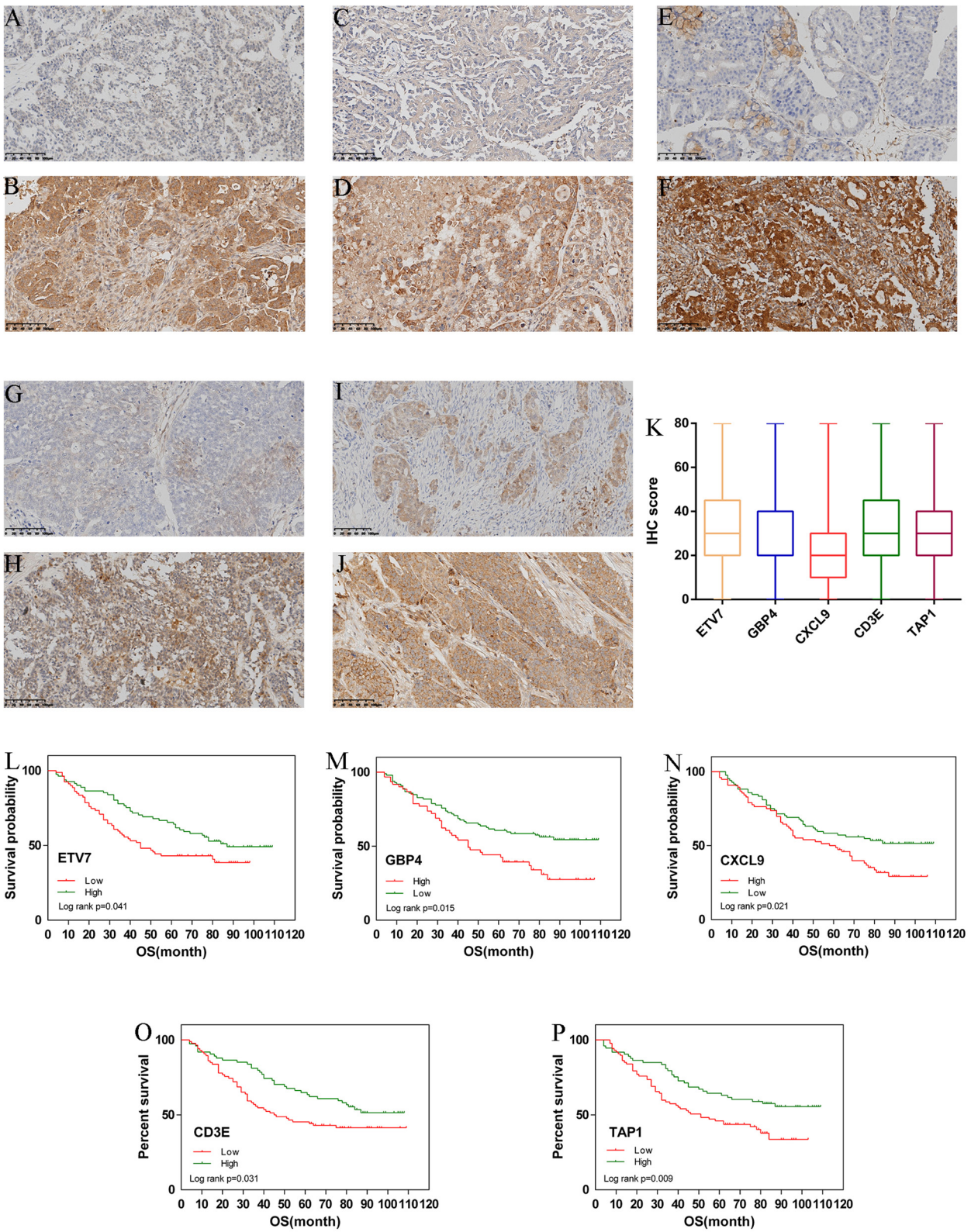


FIGURE 6 | Immunohistochemistry for ETV7, GBP4, CXCL9, CD3E, and TAP1. Samples of ovarian cancer (*N* = 168). Ovarian cancer sample of weak and strong immunostaining score for ETV7 **(A,B)**, GBP4 **(C,D)**, CXCL9 **(E,F)**, CD3E **(G,H)**, and TAP1 **(I,J)**, respectively. Expression of each gene is depicted in **(K)** slides. (X 100). Overall survival (OS) curves for ovarian cancer (*N* = 168) according to ETV7 **(L)**, GBP4 **(M)**, CXCL9 **(N)**, CD3E **(O)**, and TAP1 **(P)** gene expression status (low or high). Geneexpression status was divided according to their median values.

biomarker for the diagnosis and treatment of ovarian cancer patients (Luo et al., 2016). To fully understand the biological behavior of ovarian cancer, it is necessary to consider the environment in which ovarian cancer cells exist and how they are manipulated by their surroundings to promote malignant phenotypes.

In recent years, with the development of sequencing technology, as well as public databases such as TCGA and Gene Expression Omnibus (GEO) database, a large number of studies have been conducted on human cancer gene expression. In ovarian cancer, Men et al. (2018) performed a genome-wide analysis of gene expression profiling in the TCGA and developed an 11 gene expression signature-based risk score that can predict a patient's survival. In another study that used robust Bayesian network modeling, 13 hub genes including ARID1A, C19orf53, CSKN2A1, and COL5A2 signature with a prognostic function in ovarian cancer was established (Zhang et al., 2014; Guo et al., 2020). However, most studies focused on oncogene panels of ovarian cancer.

In present study, we performed a multistep bioinformatics analysis using data from the TCGA database and identified a list of tumor microenvironment-related genes that may contribute to ovarian cancer overall survival. We first used RNA-Seq data of ovarian cancer in the TCGA (379 samples) to systematically analyze the relationship between ovarian cancer and immune metagenes; we found that the expression of immune metagenes was closely related to the immune components in the tumor microenvironment. Next, the expression levels of immune metagenes in different stages were analyzed, and different stages had significant prognostic differences (**Figure 2**). Third, by analyzing the relationship between these immune metagenes and BRCA1 and BRCA2 mutations, the expression of immune metagenes was found to be related to only BRCA1 mutations. Finally, we screened 10 genes related to immunity and prognosis in ovarian cancer according to the expression of immune metagenes. By cross validation with KMplot, an independent cohort of 1,816 ovarian patients, we identified 5 tumor microenvironment-related genes that showed a significant correlation between gene expression and prognosis. Our results may provide new insights into the underlying biological mechanisms driving the tumorigenesis of ovarian cancer.

This study identified tumor microenvironment-related genes, including monokine induced by gamma interferon (MIG or CXCL9), E26 transformation-specific variant 7 (ETV7), guanylate binding protein 4 (GBP4), and CD3 epsilon chain (CD3E). In agreement with a previous study, we found that these genes were differentially expressed in a variety of human tumors and correlated with survival time. For example, CXCL9 is located on human chromosome 4 and is induced by IFN-γ but not by IFN-α/β. CXCL9 predominantly mediates lymphocytic infiltration to the focal sites and suppresses tumor growth (Gorbachev et al., 2007). CXCL9 can predict survival and is regulated by cyclooxygenase inhibition in advanced serous ovarian cancer (Bronger et al., 2016). Wu et al. used the KM method as well as Cox's univariate and multivariate hazard regression models and found that the higher the CXCL9 expression is, the higher the overall survival rate for colorectal carcinoma patients (Wu et al., 2016). In addition, plasma CXCL9 has been found to predict the survival of patients with advanced pancreatic ductal adenocarcinoma receiving chemotherapy, potentially improving treatment outcomes (Qian et al., 2019). In cervical carcinoma, low expression of CD3E was correlated with poor disease-specific and disease-free survival, and high CD3E expression was correlated with improved disease-specific survival (Punt et al., 2015). Moreover, this gene was also considered as a hub gene in head and neck squamous cell carcinoma (Upreti et al., 2016). A high expression of GBP4 was correlated with favorable overall survival in skin (cutaneous) melanoma patients followed for over 30 years (Wang et al., 2018). Therefore, these gene-associated tumor microenvironments may serve as important roles in the pathogenesis of ovarian cancer.

However, our study may have some disadvantages. First, there is a lack of experimental research that can explain the biological significance and molecular mechanism of immune microenvironment genes in ovarian cancer. Second, a small portion of the results are not statistically significant, but there is a trend difference, which may be due to the limited sample size. Third, the prognostic value of these immune microenvironment genes needs to be validated from a large independent cohort before they can be applied to clinical practice. Moreover, the microenvironment gene also significantly associated with the prognosis of other histology types ovarian cancer has been rarely studied in present research. According to histological and pathological morphological differences, ovarian cancer can be divided into various types: serous carcinoma, mucinous carcinoma, endometrioid carcinoma, clear cell carcinoma and other types of tumors. Different types of ovarian cancer have obvious clinical pathological differences and molecular differences (Tone et al., 2008). However, since more than 70% of ovarian epithelial cancer are serous types, there are no enough samples of other types in the dataset from TCGA for effective analysis. In further study, we will pay more attention to the prognosis between the microenvironment genes and other types of ovarian cancer.

CONCLUSION

In conclusion, through a comprehensive analysis of the data of ovarian cancer patients, we found a group of immune microenvironment genes that can be used as potential biomarkers to predict the prognosis of ovarian cancer patients. This study provides a new understanding of the potential relationship between the tumor microenvironment and ovarian cancer prognosis and provides a new molecular target for the development of more effective treatment methods for ovarian cancer. This study will help refine and personalize treatment.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Beijing Chao-Yang Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XH: study design, manuscript writing, and data analysis. HS: data analysis, data collection, manuscript writing, and resources. SL: study design, resources, and data analysis. SW: funding and resources. BL and HB: resources. All authors have read, edited and approved of the final version of the manuscript.

FUNDING

This study was funded by the Key Projects of the Sailing Plan of Beijing Medical Administration (ZYLX201713).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.680413/full#supplementary-material

REFERENCES

Bronger, H., Singer, J., Windmuller, C., Reuning, U., Zech, D., Delbridge, C., et al. (2016). CXCL9 and CXCL10 predict survival and are regulated by cyclooxygenase inhibition in advanced serous ovarian cancer. *Br. J. Cancer* 115, 553–563. doi: 10.1038/bjc.2016.172

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

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30, 413–421. doi: 10.1038/nbt.2203

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, 213–219. doi: 10.1038/nbt.2514

Dinh, P., Harnett, P., Piccart-Gebhart, M. J., and Awada, A. (2008). New therapies for ovarian cancer: cytotoxics and molecularly targeted agents. *Crit. Rev. Oncol. Hematol.* 67, 103–112. doi: 10.1016/j.critrevonc.2008.01.012

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi: 10.1002/ijc.29210

Goode, E. L., Block, M. S., Kalli, K. R., Vierkant, R. A., Chen, W., Fogarty, Z. C., et al. (2017). Dose-response association of CD8+ tumor-infiltrating lymphocytes and survival time in high-grade serous ovarian cancer. *JAMA Oncol.* 3:e173290. doi: 10.1001/jamaoncol.2017.3290

Gorbachev, A. V., Kobayashi, H., Kudo, D., Tannenbaum, C. S., Finke, J. H., Shu, S., et al. (2007). CXC chemokine ligand 9/monokine induced by IFN-gamma production by tumor cells is critical for T cell-mediated suppression of cutaneous tumors. *J. Immunol.* 178, 2278–2286.

Guo, Q., Wang, J., Gao, Y., Li, X., Hao, Y., Ning, S., et al. (2020). Dynamic TF-lncRNA regulatory networks revealed prognostic signatures in the development of ovarian cancer. *Front. Bioeng. Biotechnol.* 8:460. doi: 10.3389/fbioe.2020.00460

Hennessy, B. T., Coleman, R. L., and Markman, M. (2009). Ovarian cancer. *Lancet* 374, 1371–1382. doi: 10.1016/S0140-6736(09)61338-6

Jia, D., Li, S., Li, D., Xue, H., Yang, D., and Liu, Y. (2018). Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. *Aging* 10, 592–605. doi: 10.18632/aging.101415

Kreuzinger, C., Geroldinger, A., Smeets, D., Braicu, E. I., Sehouli, J., Koller, J., et al. (2017). A complex network of tumor microenvironment in human high-grade serous ovarian cancer. *Clin. Cancer Res.* 23, 7621–7632. doi: 10.1158/1078-0432.CCR-17-1159

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

Li, Y. L., Ye, F., Cheng, X. D., Hu, Y., Zhou, C. Y., Lu, W. G., et al. (2010). Identification of glia maturation factor beta as an independent prognostic predictor for serous ovarian cancer. *Eur. J. Cancer (Oxf. Engl. 1990)* 46, 2104–2118. doi: 10.1016/j.ejca.2010.04.015

Luo, Z., Wang, Q., Lau, W. B., Lau, B., Xu, L., Zhao, L., et al. (2016). Tumor microenvironment: the culprit for ovarian cancer metastasis? *Cancer Lett.* 377, 174–182. doi: 10.1016/j.canlet.2016.04.038

Men, C. D., Liu, Q. N., and Ren, Q. (2018). A prognostic 11 genes expression model for ovarian cancer. *J. Cell. Biochem.* 119, 1971–1978. doi: 10.1002/jcb.26358

Nelson, B. H. (2015). New insights into tumor immunity revealed by the unique genetic and genomic aspects of ovarian cancer. *Curr. Opin. Immunol.* 33, 93–100. doi: 10.1016/j.coi.2015.02.004

Odunsi, K. (2017). Immunotherapy in ovarian cancer. *Ann. Oncol.* 28(suppl. 8), viii1–viii7. doi: 10.1093/annonc/mdx444

Ovarian Tumor Tissue Analysis (Otta) Consortium, Goode, E. L., Block, M. S., Kalli, K. R., Vierkant, R. A., Chen, W., et al. (2017). Dose-response association of CD8+ tumor-infiltrating lymphocytes and survival time in high-grade serous ovarian cancer. *JAMA Oncol.* 3:e173290.

Pietras, K., and Ostman, A. (2010). Hallmarks of cancer: interactions with the tumor stroma. *Exp. Cell Res.* 316, 1324–1331. doi: 10.1016/j.yexcr.2010.02.045

Punt, S., Houwing-Duistermaat, J. J., Schulkens, I. A., Thijssen, V. L., Osse, E. M., de Kroon, C. D., et al. (2015). Correlations between immune response and vascularization qRT-PCR gene expression clusters in squamous cervical cancer. *Mol. Cancer* 14:71. doi: 10.1186/s12943-015-0350-0

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

Safonov, A., Jiang, T., Bianchini, G., Gyorffy, B., Karn, T., Hatzis, C., et al. (2017). Immune gene expression is associated with genomic aberrations in breast cancer. *Cancer Res.* 77, 3317–3324. doi: 10.1158/0008-5472.CAN-16-3478

Senbabaoglu, Y., Gejman, R. S., Winer, A. G., Liu, M., Van Allen, E. M., de Velasco, G., et al. (2016). Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* 17:231. doi: 10.1186/s13059-016-1092-z

Shah, N., Wang, P., Wongvipat, J., Karthaus, W. R., Abida, W., Armenia, J., et al. (2017). Regulation of the glucocorticoid receptor via a BET-dependent enhancer drives antiandrogen resistance in prostate cancer. *eLife* 6:e27861. doi: 10.7554/eLife.27861

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

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

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

Tone, A. A., Begley, H., Sharma, M., Murphy, J., Rosen, B., Brown, T. J., et al. (2008). Gene expression profiles of luteal phase fallopian tube epithelium from BRCA mutation carriers resemble high-grade serous carcinoma. *Clin. Cancer Res.* 14, 4067–4078. doi: 10.1158/1078-0432.Ccr-07-4959

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

Torre, L. A., Trabert, B., DeSantis, C. E., Miller, K. D., Samimi, G., Runowicz, C. D., et al. (2018). Ovarian cancer statistics, 2018. *CA Cancer J. Clin.* 68, 284–296. doi: 10.3322/caac.21456

Upreti, D., Zhang, M. L., Bykova, E., Kung, S. K., and Pathak, K. A. (2016). Change in CD3zeta-chain expression is an independent predictor of disease status in head and neck cancer patients. *Int. J. Cancer* 139, 122–129. doi: 10.1002/ijc.30046

Wang, P., Li, X., Gao, Y., Guo, Q., Ning, S., Zhang, Y., et al. (2020). LnCeVar: a comprehensive database of genomic variations that disturb ceRNA network regulation. *Nucleic Acids Res.* 48, D111–D117. doi: 10.1093/nar/gkz887

Wang, P., Li, X., Gao, Y., Guo, Q., Wang, Y., Fang, Y., et al. (2019). LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. *Nucleic Acids Res.* 47, D121–D127. doi: 10.1093/nar/gky1144

Wang, Q., Wang, X., Liang, Q., Wang, S., Xiwen, L., Pan, F., et al. (2018). Distinct prognostic value of mRNA expression of guanylate-binding protein genes in skin cutaneous melanoma. *Oncol. Lett.* 15, 7914–7922. doi: 10.3892/ol.2018.8306

Winslow, S., Lindquist, K. E., Edsjo, A., and Larsson, C. (2016). The expression pattern of matrix-producing tumor stroma is of prognostic importance in breast cancer. *BMC Cancer* 16:841. doi: 10.1186/s12885-016-2864-2

Wu, Z., Huang, X., Han, X., Li, Z., Zhu, Q., Yan, J., et al. (2016). The chemokine CXCL9 expression is associated with better prognosis for colorectal carcinoma patients. *Biomed. Pharmacother.* 78, 8–13. doi: 10.1016/j.biopha.2015.12.021

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

Zhang, Q., Burdette, J. E., and Wang, J. P. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Syst. Biol.* 8:1338. doi: 10.1186/s12918-014-0136-9

Zhang, S. F., Wang, X. Y., Fu, Z. Q., Peng, Q. H., Zhang, J. Y., Ye, F., et al. (2015). TXNDC17 promotes paclitaxel resistance via inducing autophagy in ovarian cancer. *Autophagy* 11, 225–238. doi: 10.1080/15548627.2014.998931

Zheng, X., Zhang, N., Wu, H. J., and Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* 18:17. doi: 10.1186/s13059-016-1143-5

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fgene.2021.747270

Integrating Genetic and Transcriptomic Data to Reveal Pathogenesis and Prognostic Markers of Pancreatic Adenocarcinoma

Kaisong Bai[1,2†], Tong Zhao[3†], Yilong Li[2,4†], Xinjian Li[1,2], Zhantian Zhang[1,2], Zuchao Du[1,2], Zimin Wang[1,2], Yan Xu[1,2], Bei Sun[2,4] and Xuewei Bai[1,2*]*

[1] Department of General Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Key Laboratory of Hepatosplenic Surgery, Ministry of Education, Harbin, China, [3] School of Life Sciences and Technology, Harbin Institute of Technology, Harbin, China, [4] Department of Pancreatic and Biliary Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Xinyi Liu,
University of Illinois at Chicago,
United States

Reviewed by:
Zheng Zhao,
Capital Medical University, China
Jingrun Ye,
Qingdao Agricultural University, China

***Correspondence:**
Xuewei Bai
baixuewei78@163.com
Bei Sun
sunbei70@tom.com

[†] These authors have contributed equally to this work

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

Received: 26 July 2021
Accepted: 23 August 2021
Published: 09 September 2021

Citation:
Bai K, Zhao T, Li Y, Li X, Zhang Z, Du Z, Wang Z, Xu Y, Sun B and Bai X (2021) Integrating Genetic and Transcriptomic Data to Reveal Pathogenesis and Prognostic Markers of Pancreatic Adenocarcinoma. Front. Genet. 12:747270. doi: 10.3389/fgene.2021.747270

Pancreatic adenocarcinoma (PAAD) is one of the deadliest malignancies and mortality for PAAD have remained increasing under the conditions of substantial improvements in mortality for other major cancers. Although multiple of studies exists on PAAD, few studies have dissected the oncogenic mechanisms of PAAD based on genomic variation. In this study, we integrated somatic mutation data and gene expression profiles obtained by high-throughput sequencing to characterize the pathogenesis of PAAD. The mutation profile containing 182 samples with 25,470 somatic mutations was obtained from The Cancer Genome Atlas (TCGA). The mutation landscape was generated and somatic mutations in PAAD were found to have preference for mutation location. The combination of mutation matrix and gene expression profiles identified 31 driver genes that were closely associated with tumor cell invasion and apoptosis. Co-expression networks were constructed based on 461 genes significantly associated with driver genes and the hub gene FAM133A in the network was identified to be associated with tumor metastasis. Further, the cascade relationship of somatic mutation-Long non-coding RNA (lncRNA)-microRNA (miRNA) was constructed to reveal a new mechanism for the involvement of mutations in post-transcriptional regulation. We have also identified prognostic markers that are significantly associated with overall survival (OS) of PAAD patients and constructed a risk score model to identify patients' survival risk. In summary, our study revealed the pathogenic mechanisms and prognostic markers of PAAD providing theoretical support for the development of precision medicine.

Keywords: pancreatic cancer, somatic mutation, genomic variation, prognostic marker, complex disease

INTRODUCTION

Pancreatic adenocarcinoma (PAAD) remains one of the deadliest cancer types and has become the leading cause of cancer-related mortality in the United States (Rahib et al., 2014; Ilic and Ilic, 2016). The incidence and mortality rates of PAAD vary widely worldwide and are highest in developed countries (McGuigan et al., 2018). Although studies have shown that smoking, obesity,

hereditary diabetes and irregular diet are risk factors for the development of pancreatic cancer, the pathogenesis was still poorly understood. Several of treatments exist that can improve the prognosis of PAAD patients. For example, nab-paclitaxel plus gemcitabine (Von Hoff et al., 2013) and FOLFIRINOX vs. gemcitabine (Conroy et al., 2011). Although these treatments have improved the survival of some patients, the 5-year survival rate of PAAD still remains severe at 8% (Siegel et al., 2017). Therefore, it is necessary to deeply discover the carcinogenic mechanism and possible therapeutic targets of PAAD.

Genomic variation refers to differences in the structure and composition of DNA between individuals or between populations. With the development of high-throughput sequencing, multiple sources of disease-related genomic variation have been identified such as copy number variation and somatic mutations. Large-scale cancer genome sequencing consortia, such as The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) and ICGC (International Cancer Genome et al., 2010), have provided somatic mutation data from numerous of tumor patients. The role of somatic mutations in the development of specific cancer phenotypes is the main purpose of cancer genomics studies (Vogelstein et al., 2013). Somatic mutations have significant tumor heterogeneity, and each individual has different sets of mutations across many genes. Therefore, exploring the mutation-driven regulation of gene expression can better serve the purpose of precision medicine.

Work from the past decade has given us a whole new perspective on non-coding RNAs. For example, Long non-coding RNA (lncRNA) have been demonstrated to play an important role in chromatin reprogramming, transcription, post-transcriptional modifications and signal transduction (Anastasiadou et al., 2018; Wang et al., 2021). LncRNA could act as a miRNA sponge to participate in competitive endogenous RNA (ceRNA) regulation determined by microRNA (miRNA) response elements (MREs) (Salmena et al., 2011), which is an important way for it to regulate gene expression post-transcriptionally. Somatic mutations in the MRE region of the lncRNA may weaken, enhance or prevent binding to the pro-miRNA, which may cause some imbalance in the ceRNA regulatory network and even alter the expression of related target genes in the regulatory pathway (Thomas et al., 2011; Thomson and Dinger, 2016).

Here, we have collected mutation data, clinical information and transcript expression profile of PAAD from TCGA to conduct a systematic investigation concerning mutation features, pathogenesis and prognostic markers.

MATERIALS AND METHODS

Data Collection

The somatic mutation profiles (182 samples), clinical information (222 samples), and RNA-seq profiles (178 tumor and 4 paracancer samples) of PAAD were collected from TCGA (Tomczak et al., 2015).[1] We collected hallmark gene sets from the molecular feature database (MSigDB v7.4 Liberzon et al., 2015)[2] for enrichment analysis of carcinogenic functions. The human genome annotation data of GRCh38 v29 version including the position and sequence information of lncRNA was collected from GENCODE (Frankish et al., 2019)[3]. The sequence information of 2654 miRNA was obtained from miRbase v22 (Kozomara et al., 2019)[4] database. Further, we downloaded the experimentally validated miRNA-target gene regulatory relationships from miRTarBase v8.0 (Chou et al., 2018)[5] to reconstruct ceRNA regulatory relationships.

Statistical Analysis of Somatic Mutations

The R package maftools (version 2.8.0) (Mayakonda et al., 2018) was used for the statistical and visualization of mutation location, mutation form, mutation frequency and other information. The package enables efficient aggregation, analysis, annotation and visualization of MAF files from TCGA sources or any in-house study. We also used the visualization results of maftools to reveal new discoveries of PAAD.

Driver Gene Identification

We first counted the number of mutations in each gene across samples to generate a mutation matrix. Combined with the gene expression profile of PAAD from TCGA, we retained genes that were mutated in at least two samples. Further, the difference in expression of each gene between mutated and unmutated samples was measured by Student's *t*-test and fold change. We set the cutoff for *p*-value and fold change to 0.05 and 1.5, respectively (He et al., 2021). We define genes that are differentially expressed between mutated and unmutated samples as mutation driver genes.

Construction of Gene Co-expression Networks

For the driver genes affected by mutations, we separately calculated other genes co-expressed with each driver gene, which may interact with each other and play a role in the occurrence and development of PAAD. Pearson's (Bishara and Hittner, 2012) correlation algorithm was used to calculate the correlation between the expression of two genes, which was performed by cor.test function of R. We defined gene pairs with p-value < 0.01 and correlation coefficient $|R| > 0.5$ as those with significantly related expression. For all co-expressed genes, cytoscape (v3.7.0)[6] (Shannon et al., 2003) was used to plot the co-expression network. Further, NetworkAnalyzer was used to calculate the topological properties of the network and to mark the size of the nodes according to their degree.

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

[2] http://software.broadinstitute.org/gsea/msigdb

[3] https://www.gencodegenes.org/

[4] http://www.mirbase.org/ftp.shtml

[5] http://mirtarbase.cuhk.edu.cn/

[6] https://cytoscape.org/

Identification of Putative Mutation-miRNA-LncRNA Regulation Units

Somatic mutations occurring in lncRNA may affect the affinity of the original lncRNA and miRNA binding (Wang P. et al., 2020; Zhang et al., 2021). Based on the lncRNA annotation information collected from GENCODE (v29, GRCH38), we relocated the mutations that occurred in the lncRNA. Considering the requirements of miRNA target prediction tools for predicted sequences, we extracted sequences of 21 approximately nucleotides (nt) upstream and 7 nt downstream of the lncRNA somatic mutation site, which will be used to construct mutation and wild sequences. TargetScan (v.6.0)[7] and miRanda (v2010),[8] which are two miRNA target prediction tools, were used to predict the possible combination of miRNA and mutant/wild sequence. We also set stringent thresholds of score > 160 and energy < -20 for miRanda (Betel et al., 2008) and context score < -0.4 for TargetScan (Friedman et al., 2009), and miRNA-targets that satisfy this threshold are considered to be reliable. We define mutations that affect the affinity of miRNA binding to wild sequences as putative mutations, and the lncRNA in which the putative mutation was located as ceL. Further, the altered binding affinity of miRNA and mutation/wild sequence was divided into four states including gain, up, loss, and down. For these ceRNAs perturbed by somatic mutations, we constructed putative mutation-miRNA-lncRNA (ceL) units.

Next, altered binding affinity of the original lncRNA and miRNA may affect the expression of other downstream mRNAs regulated by this miRNA (Wang et al., 2015; Wang P. et al., 2019; Zhang et al., 2021). We collected miRNA-target mRNA regulatory relationships from the miRTarBase database that were validated by experiments including the luciferase reporter assay, PCR, and western blotting to build the somatic mutation-lncRNA-miRNA-mRNA (ceRNA dysregulation) network.

Functional Enrichment Analysis

For those mutated genes, we sorted the genes with a weight of $-\log10(p\text{-value})$. The sorted genes and hallmark gene set were used for gene set enrichment analysis (GSEA) (Subramanian et al., 2005). Similarly, for those genes co-expressed with the mutation driver genes, we ordered the co-expressed genes for each driver gene using the correlation coefficient as a weight, which was also used for GSEA. The clusterprofiler (v3.18.0) (Yu et al., 2012) R package was used to perform gene ontology (GO) functional enrichment and kyoto encyclopedia of genes and genomes (KEGG) pathway analysis on these mRNA. We set p-value < 0.05 to screen for significantly enriched functions and pathways.

Constructing Survival Prediction Model

We integrated significantly differentially expressed mutant genes (p-value < 0.05 only) and other protein-coding genes perturbed by putative mutations in these genes through the ceRNA mechanism. First, we used univariate COX regression to screen for genes significantly associated with overall survival (OS) in PAAD patients (the cutoff of p-value was 0.05). Considering that univariate cox regression was not sufficiently rigorous, lasso regression (Alhamzawi and Ali, 2018) was used to further screen for prognosis-related genes. Next, we randomly selected 70% of all samples as the training set and the remaining as the testing set. The train set were used to construct a multivariate COX regression model (Fisher and Lin, 1999). The Hazard Ratio hypothesis test was also used in the construction of the regression model. We retained the genes passing the Hazard Ratio hypothesis test to establish survival risk prediction model and nomogram to predict the OS of PAAD. The reliability of this risk prediction model was depicted by the receiver working characteristic curve (ROC), and the area under curve (AUC) also was calculated. The train set and test set was, respectively, divided into high-risk and low-risk groups based on the median risk score calculated by risk score model, and Kaplan-Meier (KM) survival analysis was used to measure the difference in OS between these two groups and bilateral logarithmic rank test was used.

Statistical Analysis

All statistical analyses and graph generation were performed in R (version 4.0.2). The R package resources were obtained from http://www.bioconductor.org/ and https://cran.rstudio.com/bin/windows/Rtools/.

RESULTS

The Landscape of Pancreatic Adenocarcinoma Somatic Mutations

In this study, it is necessary to perform an overall statistical analysis of the somatic mutations in PAAD. First, we evaluated samples in the TCGA database collection for which somatic mutation data were available. The result contained 182 samples with 25,470 somatic mutations. We counted the distribution of somatic mutations on the genome including chromosomal location and transcript type. We found that somatic mutations were significantly enriched on chromosomes 17 and 19 (**Figure 1A**), suggesting the preference of PAAD somatic mutation in the mutation position. Compared with transcripts (mRNA) of protein-coding genes, several somatic mutations occur in lncRNA (**Figure 1A**). Although relatively few mutations occur in the non-coding region, studies have confirmed that mutations within the non-coding genome are a major determinant of human disease (Maurano et al., 2012). Somatic mutations, including missense and nonsense mutations, account for the largest proportion of all somatic mutations, with missense mutations predominating (**Figures 1A,B**). We also found mutations occurring at the transcription start site in only four samples (**Figure 1B**). All these suggest that PAAD patients are more likely to have mutations that alter protein function to disrupt normal physiological mechanisms. Further, we counted the frequency of mutations in each gene and the number of

[7] http://www.targetscan.org/vert_60/

[8] http://www.miranda.org/

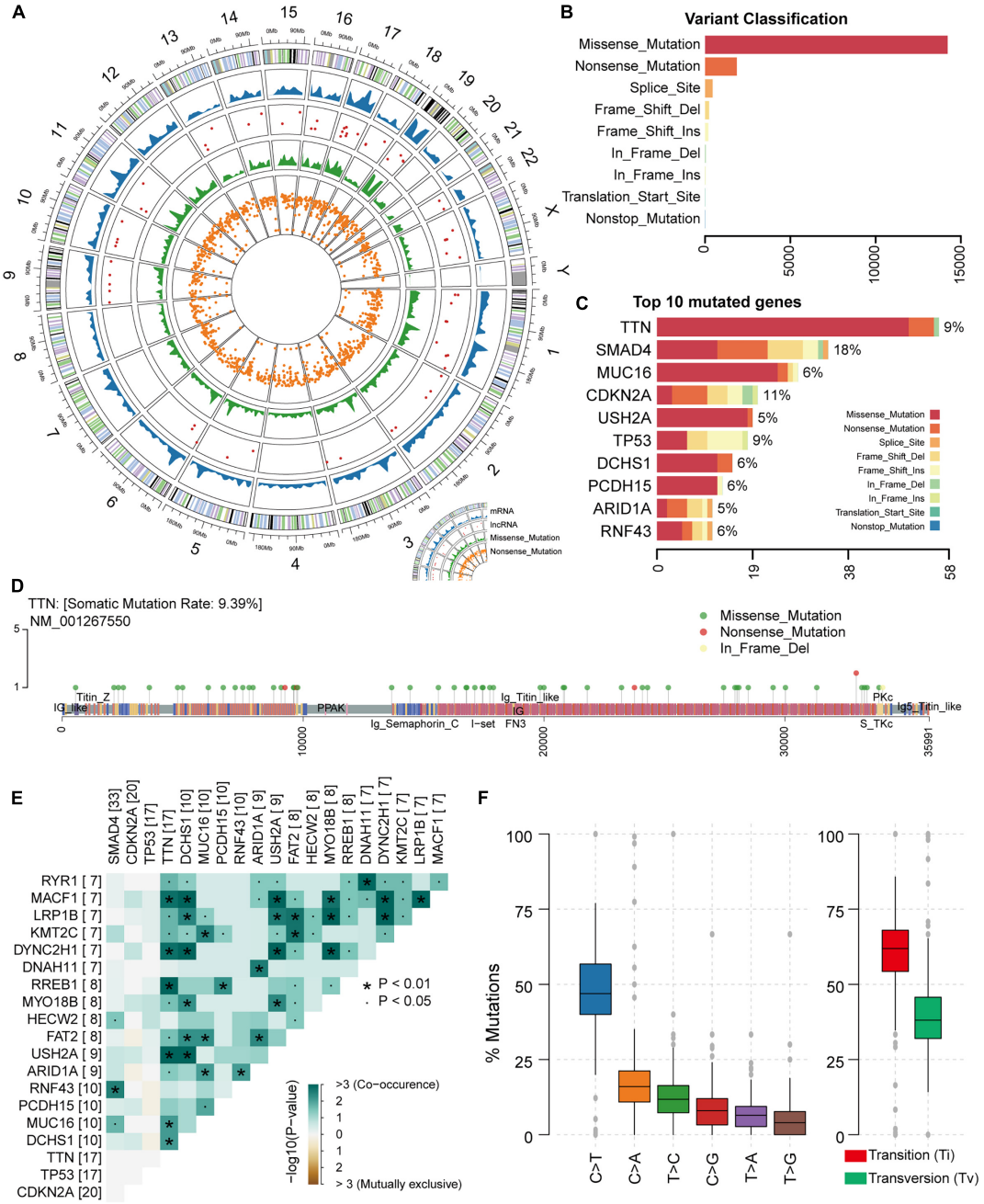


FIGURE 1 | Genomic overview of somatic mutations in PAAD. **(A)** The global view of genomic variations location. **(B)** Somatic mutations were classified into nine clusters according to function and location. The bar plot shows the number of mutations in each cluster. **(C)** The bar plot illustrates the proportion of each cluster of mutation in the top 10 genes in terms of number of mutations. The proportion of samples to which the mutation on each gene belongs was also calculated. **(D)** The location and type of mutations occurring on gene TTN were shown by the lollipop chart. **(E)** The mutation correlation between the top 20 high-frequency mutated genes. **(F)** The frequency of base substitutions (transitions and transversions) in PAAD.

samples with mutations in that gene, and the top mutated genes were illustrated (**Figure 1C** and **Supplementary Figure 1A**). We found that different genes have different preferences in the type of mutation. For example, *TTN*, the gene considered to be most frequently mutated in the pan-cancer cohort (Oh et al., 2020), tended to have missense mutations in PAAD, whereas the *TP53* gene had a high proportion of indel mutations. Studies have shown that the impact of mutations on the prognosis of patients is related to the type and background of the tumor (Hainaut and Pfeifer, 2016). As a mutated gene commonly occurring in PAAD patients, *TTN* has multiple non-sense mutation hot spots (**Figure 1D**), which will have a significant impact on the function

and structure of its encoded protein. We found no significant exclusivity between high-frequency mutated genes in the PAAD samples, and a general correlation between the *TNN* gene and other high-frequency mutated genes (**Figure 1E**), revealing a mutational feature of pancreatic cancer that the coordinated mutation of multiple genes affects the normal physiological mechanism. We found that nearly half of the point mutations (base substitution) in PAAD patients are C > T substitutions (**Figure 1F** and **Supplementary Figure 1B**). Transitions, one of the two types of DNA base conversion, have a high proportion of overall PAAD point mutations, which are capable of being retained by evolution. However, transversions as another type of DNA base conversion account for nearly 30% of overall point mutations, and these mutations may be key factors in the deterioration of pancreatic tissue. Taken together, all these revealed the mutational features of PAAD.

Driver Genes Boost Tumor Invasion

Somatic mutations could indirectly affect biological traits by regulating gene expression. It is thus intriguing to explore genes whose expression changes affected by mutations. We integrated the mutation and gene expression profiles of PAAD, with 173 samples having both mutation and gene expression data. A total of 4,517 genes that were mutated in at least two samples were collected to construct the mutation matrix. By comparing the differential expression of each gene between mutant and non-mutant samples, we identified a total of 31 driver genes that were significantly differentially up/down regulated [p-value < 0.05, $|\log2(\text{fold change})| > \log2(1.5)$] (**Figure 2A**). We next sorted the genes by fold change. The top 10 driver genes were *RP11-97C18.1 (ENSG00000225191)*, *AC024937.4 (ENSG00000231464), DRD1, CD5L, PCDH8, GK2, MAGEB6, SORCS3, TRIM51*, and *PRDM9* (**Figure 2B**). The top driver gene *RP11-97C18.1* is a pseudogene of Adaptor-Related Protein Complex 2, Beta 1 Subunit (*AP2B1*), which is an essential adaptor of the clathrin-mediated endocytosis pathway (Diling et al., 2019; Wang G. et al., 2020). The driver gene *AC024937.4* is also a pseudogene of ADP-ribosylation factor-like 8B (*ARL8B*), which is involved in cellular endocytosis, autophagy and the movement of phagocytic vesicles on microtubule tracks to fuse with lysosomes (Marwaha et al., 2017). All these suggest non-coding genes are essential in the development and progression of PAAD. Further, consensus clustering tools were used to cluster PAAD samples based on driver gene expression profiles. These samples were divided into six clusters (**Supplementary Figure 2**). We found that *PCDH8*, which acts as a tumor-suppressor gene in multiple types of cancer and inhibits tumor cell proliferation, invasion and migration (Yu et al., 2020), was downregulated in clusters 3 and 4 (**Figures 2C,D**), suggesting that tumor cells may be more aggressive in the two clusters with lower *PDCH8* expression. Patients in stage I were mainly concentrated in clusters 5 and 6 (**Figure 2C**). It is intriguing that there is no significant difference in the number of sample mutations in each cluster (**Figure 2E**), revealing that differences in gene expression of samples among clusters are not simply determined by the number of mutations. Taken together, all these suggest that driver genes affected by mutations play an essential role in the proliferation and invasion of PAAD.

Interaction of Essential Factors With Driver Genes Regulates Oncogenic Pathways

For those genes that were mutated, they may play an essential role in the proliferation and invasion of tumors. In order to explore the role of these genes in carcinogenic pathways, we performed GSEA to identify hallmark pathways enriched in mutant genes explaining somatic mutations in the genome of PAAD patients (see section "Materials and Methods"). We found that IL2-STAT5 signaling, glycolysis, apoptosis and allograft rejection pathways are significantly enriched in genes whose expression is affected by somatic mutations (**Figure 3A**). Studies have shown that interleukin-2 (IL-2) and the downstream transcription factor STAT5 are essential for maintaining regulatory T (Treg) cell homeostasis and function (Cheng et al., 2018), suggesting that the immune microenvironment in tumor tissue of PAAD patients affected by somatic mutations may be disrupted. The altered glycolytic machinery in PAAD was designed to adapt to the tumor microenvironment, which is consistent with previous studies showing that cancer cells are preferentially dependent on glycolysis (Ganapathy-Kanniappan and Geschwind, 2013). The allograft rejection pathway affected by mutations may become the key point of PAAD immunotherapy (Land et al., 2016).

Global reprogramming of the transcriptome occurs in order to support tumorigenesis and progression. In addition to the direct effect of mutations on gene expression, there are other regulatory mechanisms such as transcriptional regulation, ceRNA mechanisms, epigenetic. Genes co-expressed with driver genes may have a potential role in tumor development. We performed the Pearson correlation algorithm to identify genes that may be influenced by other regulatory mechanisms co-expressed with driver genes. We identified 495 genes (491 positive and 4 negative) significantly associated with 19 driver genes (p-value < 0.01, $|R| > 0.5$). These significantly related genes were used to construct gene co-expression networks using cytoscape (**Figure 3B**). We also counted the topological properties of the network using the NetworkAnalyzer tool and found that the gene *FAM133A* had the top degree (**Supplementary Table 1**). *FAM133A* has been confirmed in previous studies to be related to the invasion and metastasis of glioma (Huang et al., 2018). Next, we performed a functional enrichment analysis of all genes in the co-expression network using the R package clusterprofiler. We found that these genes were significantly enriched in immune-related functions and apoptotic pathways, such as complement activation, immunoglobulin mediated immune response, B cell mediated immunity, and apoptosis—multiple species (**Figure 3C** and **Supplementary Figure 3**). For the 19 driver genes identified as having co-expressed genes, we used GSEA to analyze the functional features of the driver genes. Hallmark gene sets and genes ordered by correlation coefficients were available for GSEA. We found that the oncogenic pathway was significantly enriched only in genes co-expressed with the driver genes *FAM133A* and *SORCS3*, suggesting that most driver genes are required

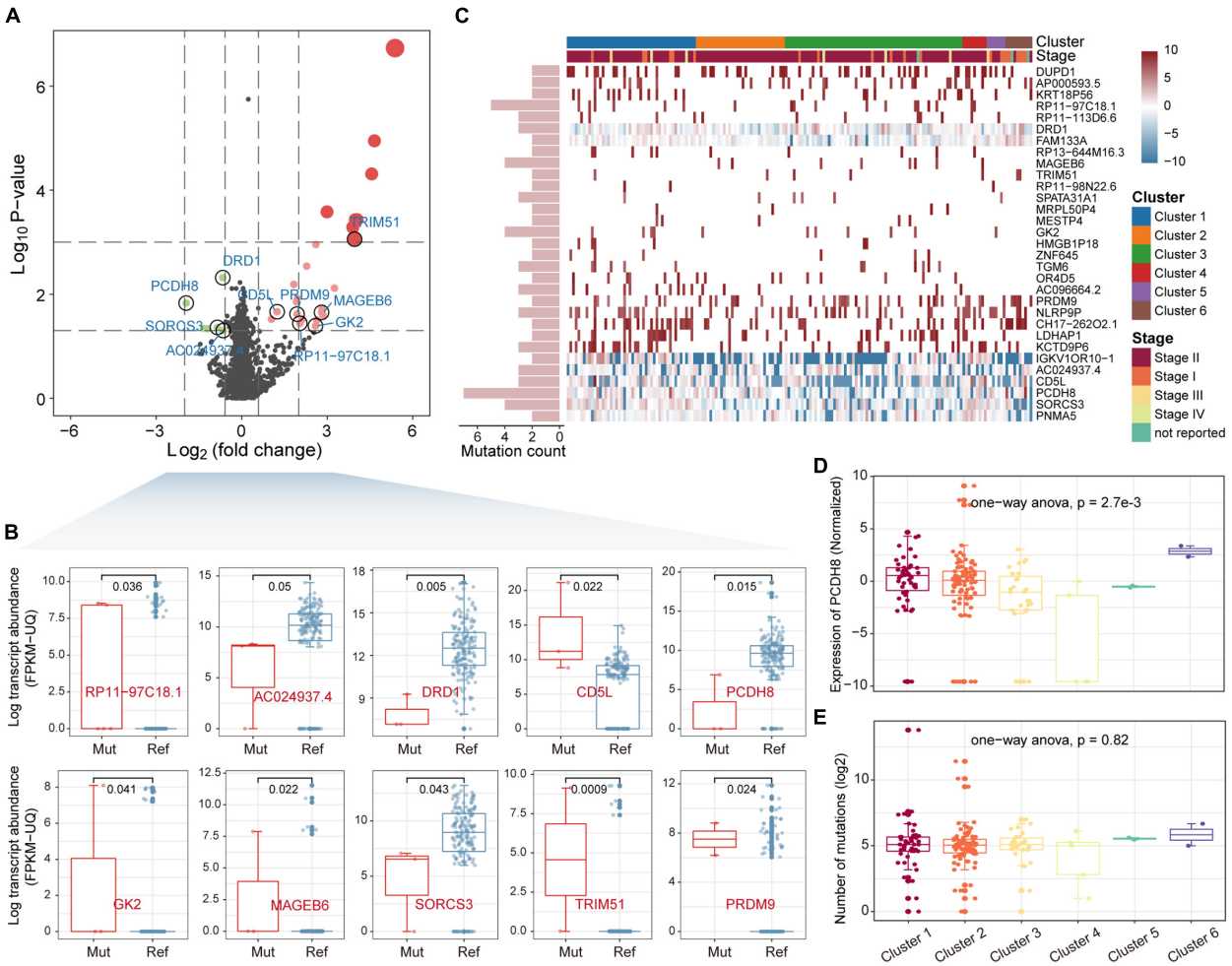


FIGURE 2 | Identification of driver genes. **(A)** The differential expression analysis comparing mutated and unmutated patients. Differentially expressed genes were shown as red or green dots. **(B)** Expression level of the top 10 driver genes between mutated and unmutated samples was shown by boxplot. The rank sum test was used to check the significance. **(C)** Distribution of driver gene expression in the samples. Column labels indicate the cluster and stage to which the sample belongs. Row labels indicate the number of mutations in each driver gene. **(D,E)** Expression and mutation levels of the driver gene PCDH8 between the 6 clusters of samples were shown by boxplot.

to synergistically regulate oncogenic mechanisms. In contrast to the driver gene *FAM133A*, the driver gene *SORCS3*, in combination with its co-expressed genes, plays an important role in tumor metastasis, hypoxia and apoptosis (**Figure 3D**). Taken together, all these indicate that the synergistic interaction network of multiple driver genes may contribute to the complex pathogenesis of PAAD.

LncRNA Mutations-ceRNA Indicates Novel Mechanisms of Mutation Regulation

LncRNA have been confirmed that genes are essential in pre- and post-transcriptional regulation. The lncRNA with (miRNA) response element (MRE) can be used as a miRNA sponge to participate in the ceRNA regulatory mechanism. To explore the impact of somatic mutations occurring on lncRNA MREs on ceRNA regulatory mechanisms, we constructed mutant/wild sequences to identify mutations that alter the affinity of lncRNA-miRNA binding. Based on lncRNA annotation data collected from GENCODE, we identified 497 somatic mutations occurring on lncRNA compared to 24,604 somatic mutations occurring on the genome. Affected by mutations, lncRNA may enhance, reduce and lose their binding affinity to existing miRNAs, or even gain binding affinity to new miRNAs (**Figure 4A**). Next, we examined the influence of lncRNA mutations on miRNA binding sites according to the TargetScan and miRanda. In total, we identified 277 somatic mutations for PAAD in 235 putative miRNA target genes (putative lncRNAs). These mutation sites showed different binding affinities to 447 miRNAs between the mutation and wild sequences (**Figure 4B**). All these constituted 552 mutation-miRNA-lncRNA regulation units. We further constructed ceRNA dysregulation networks based on the identification of mutation-miRNA-lncRNA regulation units (**Figure 4C**). We found that

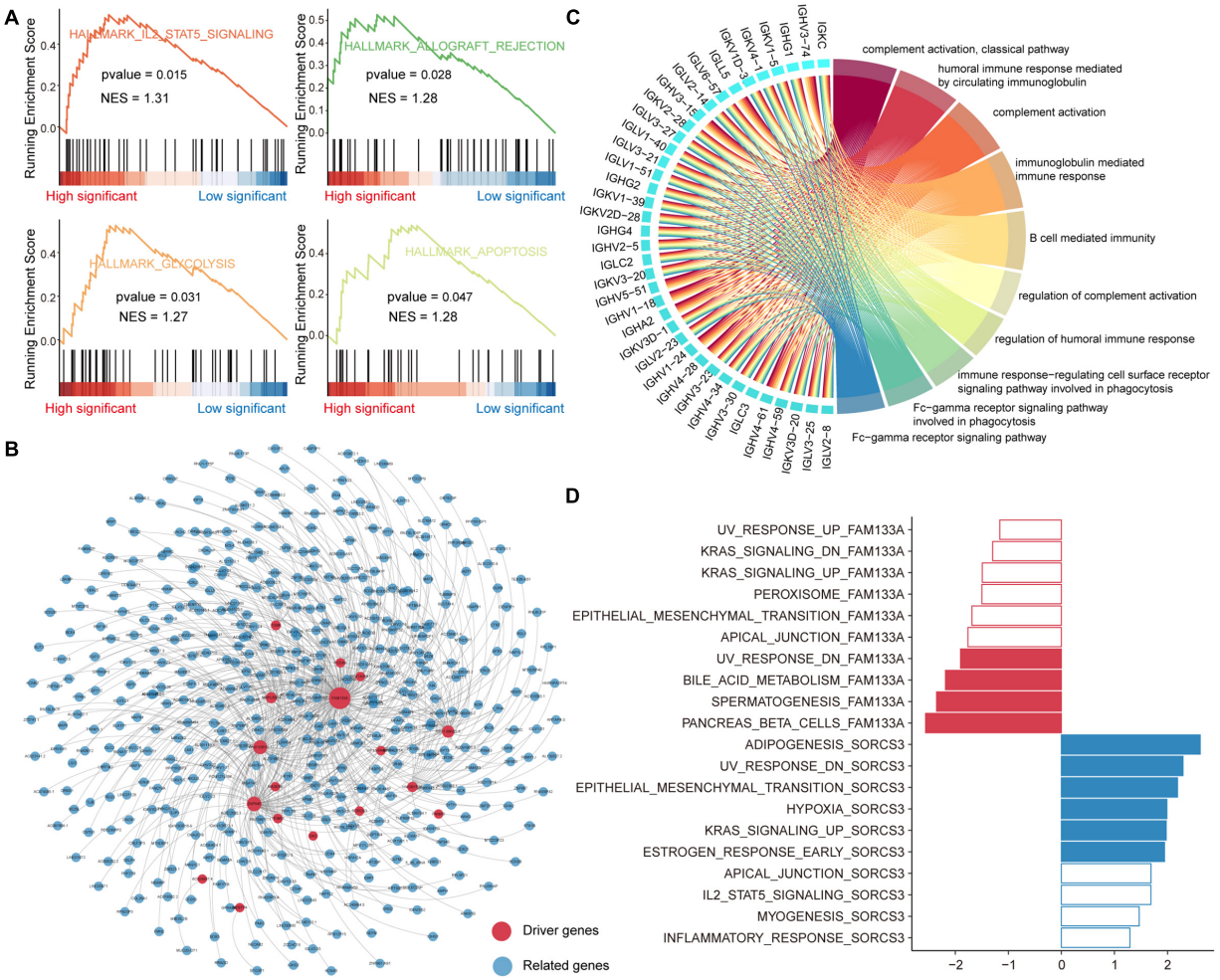


FIGURE 3 | Functional enrichment analysis of driving related genes. **(A)** The GSEA analysis of mutant genes, which explain somatic mutations in the genome of PAAD patients. **(B)** Co-expression network of driver genes and their significantly related genes. The size of the node is related to the degree. **(C)** The top 10 GO items that are enriched by genes significantly related to driver genes. This network shows the interaction between genes and GO items. **(D)** Top 10 hallmark pathways enriched in genes related to driver gene FAM133A or SORCS3. significantly enriched pathways with GSEA *p*-values < 0.05 are highlighted in red (FAM133A) or blue (SORCS3).

TTN-AS1 has top degree in the ceRNA dysregulation networks and that five somatic mutations occurring on it affect the affinity of binding to 11 miRNAs (8 Up/gain and 3 Down/loss, **Figure 4D**). Combining 31 driver genes, we found that only driver lncRNA *AC090099.1 (ENSG00000255470)* has mutations involved in ceRNA regulation imbalance, which suggesting that the mechanisms underlying changes in driver gene expression are complex. We found two mutations in *AC090099.1* that affected binding affinity to four miRNAs (3 Up/gain and 1 Down/loss, **Figure 4E**). In order to verify our prediction results at the transcriptome level, we performed one-sample *t*-test to identify the difference between the gene expression level of the non-mutated sample and the mutant sample. We found significant differences in the expression of *AC090099.1* and the target gene *CEBPB* and *LHFPL3* regulated by miRNA *hsa-miR-663a* between mutated and unmutated samples (**Figures 4F–H**). Taken together, all these results suggest that ceRNA dysregulation due to lncRNA mutations is an essential factor in variations of target gene expression.

Identifying Prognostic Markers for PAAD

Genes affected by mutations played an important role in the mechanism of carcinogenesis. It is meaningful to identify the markers associated with prognosis of PAAD patients from genes that are significantly differentially expressed between mutated and unmutated samples (p-value < 0.05). In total, we obtained 171 genes that were significantly differentially expressed by mutation-driven. We performed univariate cox regression to identify genes associated with overall survival (OS) in PAAD patients, and 53 genes were selected by controlling for p-value < 0.05. We further rigorously screened for these 53 genes using lasso regression and 8 genes including *SLC30A1, RBM10, PNPLA6, DSG2, CHML, DLGAP5, TTLL6,* and *PDE4DIPP5* were identified as significantly associated with patient OS

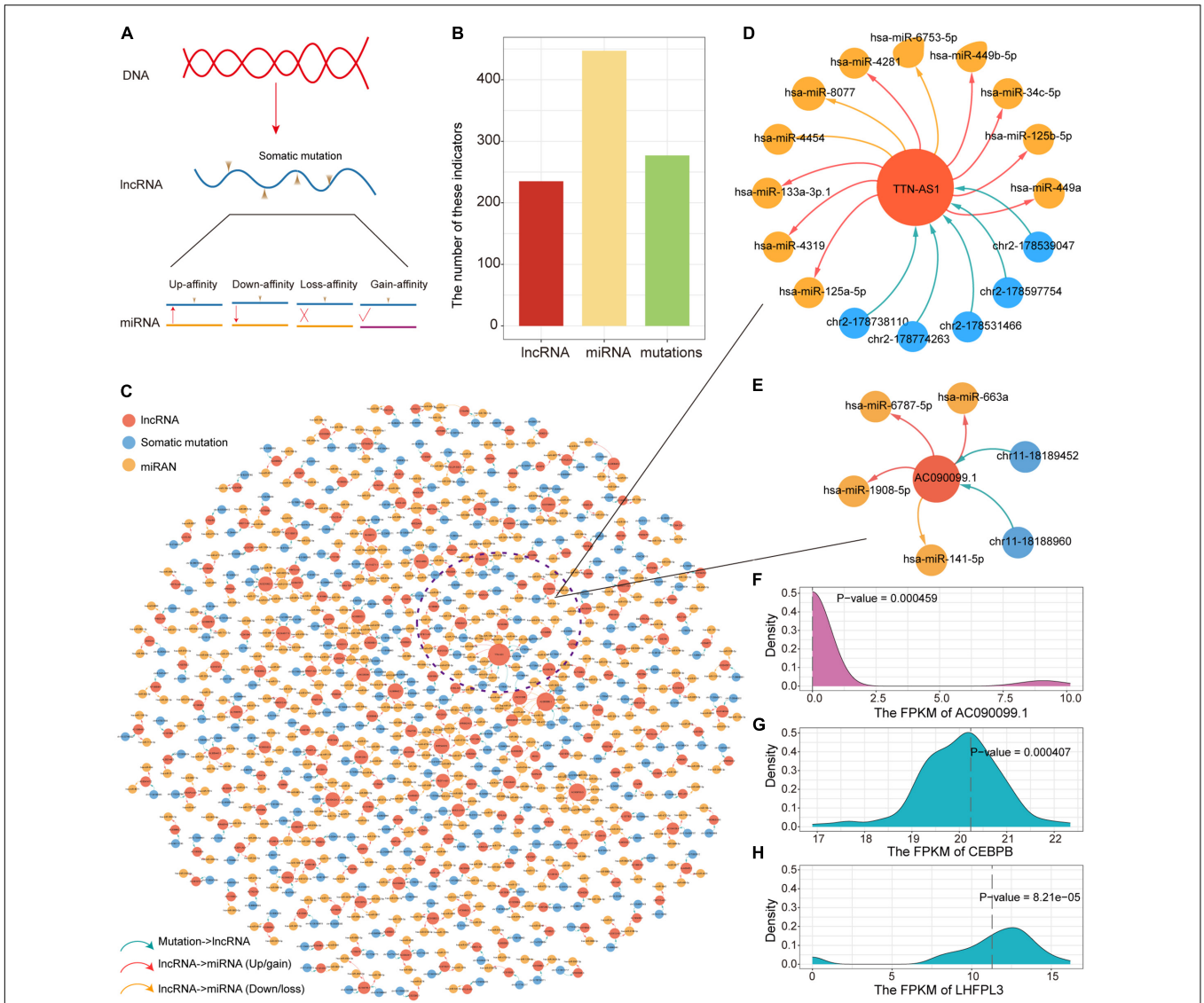


FIGURE 4 | Construction of the ceRNA dysregulated network. **(A)** Mutations in lncRNA affect binding intimacy with miRNA. **(B)** The number of lncRNA, miRNA and somatic mutations in the ceRNA dysregulated network. **(C)** The ceRNA dysregulated network constructed by lncRNA, miRNA, and somatic mutations. The red, blue and yellow nodes represent lncRNA, somatic mutation and miRNA, respectively. Up-regulated or newly gain affinity between miRNA and lncRNA used red line. Down-regulated or loss affinity between miRNA and lncRNA used yellow line. **(D)** The ceRNA dysregulated sub-network of lncRNA TTN-AS1. **(E)** The ceRNA dysregulated sub-network of lncRNA AC090099.1. **(F–H)** The distribution of the expression for lncRNA AC090099.1, CEBPB, and LHFPL3 was shown by density curve. The expression value of these genes in the mutant sample were marked with a red line. One-sample *t*-test was used to calculate statistical significance.

(**Supplementary Figure 4A**). The multivariate Cox regression were performed to construct survival risk prediction model using these eight feature genes and train set, three of which, *RBM10, SLC30A1,* and *DLGAP5*, were major genes that associated with the risk of death in patients (**Figure 5A**). Nomograms were used to illustrate the probability of survival risk at 6, 12, and 18 months (**Figure 5B**). The calibration curve was also used to validate the stability of the risk prediction model (**Supplementary Figure 4B**). In order to identify the best predictive time point for the risk prediction model, we divided the 6–18 months period into six time periods and evaluated the prediction results using ROC curve. We found that the risk prediction result reached the maximum area under curve (AUC) value of 0.84 in the 474.5 days (**Figure 5C**). Further, we used multivariate Cox regression coefficients of eight genes identified by lasso regression to construct risk score models as follows: risk score = 0.65* *SLC30A1*—0.84* *RBM10*—0.27* *PNPLA6* + 0.36* *DSG2*—0.21* *CHML* + 0.54* *DLGAP5*—0.02* *TTLL6*—0.08* *PDE4DIPP5*, and calculated the risk score for each PAAD sample. The samples of train and test set were, respectively, divided into two categories (high-risk and low-risk) based on the median risk score, and we found that high-risk samples in both the training and test sets exhibited an association with poorer PAAD OS (**Figures 5D,E**). By combining clinical information from the PAAD sample with

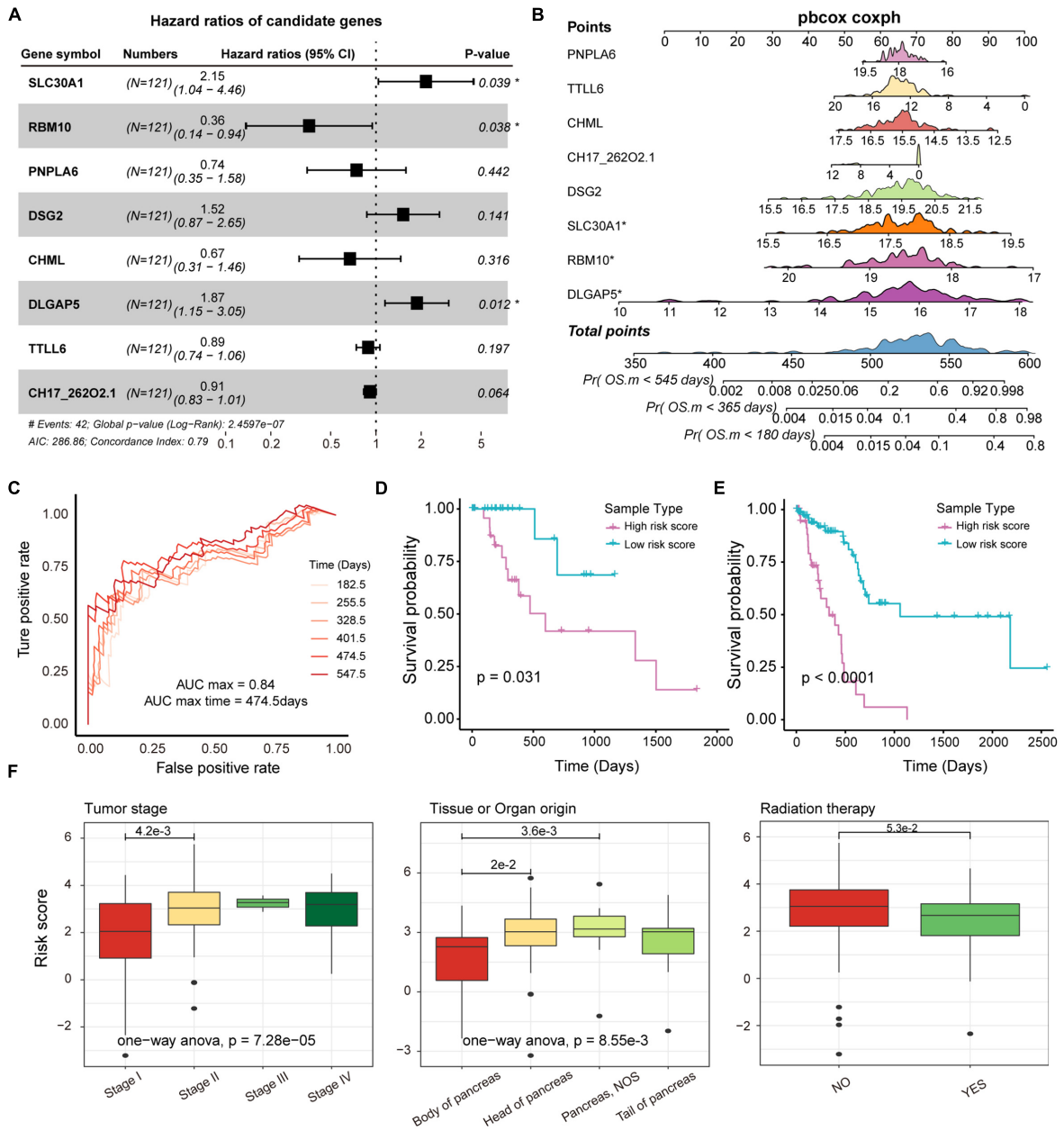


FIGURE 5 | Survival analysis of potential markers in PAAD. **(A)** Forest plots for multivariate Cox risk regression models. **(B)** Nomogram for survival risk prediction of 180, 365, and 545 days. The model contains eight features. **(C)** The ROC curve validation of the risk regression model at 6 time points. The different colored curves represent specific time-points. **(D,E)** KM plot of train and test dataset in which high- and low-risk groups were show as different lines. Log-rank test was used to calculate statistical significance. **(F)** Box plot of risk scores for samples of different tumor stages, tissue origin, and radiation therapy. The rank sum test and ANOVA were used to measure differences between groups.

the risk score, we found that patients in stage II, III, and IV had a significantly higher risk score compared to stage I (**Figure 5F**), and found that the origin of the tumor was significantly related to the patient's survival risk (**Figure 5F**), and found that patients treated with radiation have a significantly lower risk of survival than those who are not treated with radiation (**Figure 5F**). All these may provide support for the treatment of PAAD.

DISCUSSION

In this study, we have used mutational and transcriptomic data to reveal mutational features, driver genes and prognostic markers in PAAD. Statistical analysis of the mutational profile of PAAD revealed that relatively lower number of mutations occurred in non-coding regions of the genome, with most mutations occurring in coding regions affecting the structure and

function of the protein. We identified 31 driver genes based on statistical test that are strongly associated with apoptosis, energy metabolism and invasion of tumor cells. Next, we constructed a co-expression network determined by driver genes, revealing the oncogene interaction mechanism and oncogenic pathways of PPAD. We further constructed a ceRNA dysregulation network using TargetScan and miRanda tools to reveal that somatic mutations on lncRNA regulate the expression of target genes at the post-transcriptional level. Using a dual screen of univariate cox regression and lasso regression, we identified eight genes that were strongly associated with the prognosis of PAAD patients despite the existence of public databases for studying the prognosis of pan-cancers (Qi et al., 2021). We also constructed a risk score model to specify the risk of survival for each patient, showing that higher risk scores have a poorer probability of survival.

Pancreatic cancer is one of the deadliest malignancies (Vincent et al., 2011). Multiple of studies have tried to reveal the pathogenesis of pancreatic cancer and discover effective treatments. For example, exploring the role of the microbiome in the occurrence, development and treatment of PAAD (Wang Y. et al., 2019), and discover the carcinogenic mechanism and possible treatments of PAAD from the perspective of genetics (Bhosale et al., 2018). The development of PAAD is influenced by multiple factors, the most critical is the occurrence of malignant mutations in the chromosomes. Malignant mutations in chromosomes, which hold the genetic material of an organism, will affect the physiological mechanisms of normal cells. Although there are numerous of research results to support the conquering of PAAD, few studies have focused on somatic mutations in the genome (Chang et al., 2014). We integrated mutagenomic and transcriptomic data to discover the oncogenic mechanisms and potential prognostic markers of PAAD, which is the rational application of multi-omics data in the era of big data. In revealing the carcinogenic mechanism, multi-omics research has more advantages than previous single-omics research.

CeRNAs are transcript that regulate each other by competing shared miRNAs. The proposal of the ceRNA competition mechanism provides a new direction for the post-transcriptional regulation of genes. Considering the important role of non-coding RNA in PAAD, we explored the impact of lncRNA mutations on the ceRNA competition network. We have identified 552 mutation-miRNA-lncRNA regulation units and constructed a ceRNA dysregulated network. Although there is not enough gene expression data (massive absence of miRNA expression data) to support our prediction results, it contributes to the exploration of the post-transcriptional regulatory mechanism of PAAD.

In conclusion, this study provided the mutational landscape of PAAD and discovered driver genes. The IL2-STAT5 signaling pathway and allograft rejection affected by mutations provide a new direction for the treatment of PAAD. Marker genes associated with patient prognosis were identified through univariate cox regression and lasso regression. We also provide a survival risk prognostic model for PAAD patients. All these findings in this study may provide theoretical guidance for the diagnosis and treatment of PAAD.

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

KB, TZ, YL, BS, and XB designed the experiments and wrote the manuscript. XL, ZZ, and ZD collected and analyzed the data. ZW and YX validated the method and data. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Postdoctoral Foundation of Heilongjiang Province (LBH-Q20041).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.747270/full#supplementary-material

REFERENCES

Alhamzawi, R., and Ali, H. T. M. (2018). The Bayesian adaptive lasso regression. *Math. Biosci.* 303, 75–82. doi: 10.1016/j.mbs.2018.06.004

Anastasiadou, E., Jacob, L. S., and Slack, F. J. (2018). Non-coding RNA networks in cancer. *Nat. Rev. Cancer* 18, 5–18.

Betel, D., Wilson, M., Gabow, A., Marks, D. S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res.* 36, D149–D153.

Bhosale, P., Cox, V., Faria, S., Javadi, S., Viswanathan, C., Koay, E., et al. (2018). Genetics of pancreatic cancer and implications for therapy. *Abdom. Radiol. (N.Y.)* 43, 404–414. doi: 10.1007/s00261-017-1394-y

Bishara, A. J., and Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol. Methods* 17, 399–417. doi: 10.1037/a0028087

Chang, D. K., Grimmond, S. M., and Biankin, A. V. (2014). Pancreatic cancer genomics. *Curr. Opin. Genet. Dev.* 24, 74–81.

Cheng, Y., Wang, Z. M., Tan, W., Wang, X., Li, Y., Bai, B., et al. (2018). Partial loss of psychiatric risk gene Mir137 in mice causes repetitive behavior and impairs sociability and learning via increased Pde10a. *Nat. Neurosci.* 21, 1689–1703. doi: 10.1038/s41593-018-0261-7

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302.

Conroy, T., Desseigne, F., Ychou, M., Bouche, O., Guimbaud, R., Becouarn, Y., et al. (2011). FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N. Engl. J. Med.* 364, 1817–1825.

Diling, G., Yinrui, Q. Longkai, Xiaocui, T., Yadi, L., Xin, Y., Guoyan, H., et al. (2019). Circular RNA NF1-419 enhances autophagy to ameliorate senile dementia by binding Dynamin-1 and adaptor protein 2 B1 in AD-like mice. *Aging (Albany N.Y.)* 11, 12002–12031. doi: 10.18632/aging.102529

Fisher, L. D., and Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* 20, 145–157. doi: 10.1146/annurev.publhealth.20.1.145

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 92–105. doi: 10.1101/gr.082701.108

Ganapathy-Kanniappan, S., and Geschwind, J. F. (2013). Tumor glycolysis as a target for cancer therapy: progress and prospects. *Mol. Cancer* 12:152.

Hainaut, P., and Pfeifer, G. P. (2016). Somatic TP53 mutations in the era of genome sequencing. *Cold Spring Harb. Perspect. Med.* 6:a026179.

He, L., Liu, L., Li, T., Zhuang, D., Dai, J., Wang, B., et al. (2021). Exploring the imbalance of periodontitis immune system from the cellular to molecular level. *Front. Genet.* 12:653209. doi: 10.3389/fgene.2021.653209

Huang, G. H., Du, L., Li, N., Zhang, Y., Xiang, Y., Tang, J. H., et al. (2018). Methylation-mediated miR-155-FAM133A axis contributes to the attenuated invasion and migration of IDH mutant gliomas. *Cancer Lett.* 432, 93–102. doi: 10.1016/j.canlet.2018.06.007

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

International Cancer Genome, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162.

Land, W. G., Agostinis, P., Gasser, S., Garg, A. D., and Linkermann, A. (2016). DAMP-induced allograft and tumor rejection: the circle is closing. *Am. J. Transplant.* 16, 3322–3337. doi: 10.1111/ajt.14012

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Marwaha, R., Arya, S. B., Jagga, D., Kaur, H., Tuli, A., and Sharma, M. (2017). The Rab7 effector PLEKHM1 binds Arl8b to promote cargo traffic to lysosomes. *J. Cell Biol.* 216, 1051–1070. doi: 10.1083/jcb.201607085

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28, 1747–1756. doi: 10.1101/gr.239244.118

McGuigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., and McCain, R. S. (2018). Pancreatic cancer: a review of clinical diagnosis, epidemiology, treatment and outcomes. *World J. Gastroenterol.* 24, 4846–4861. doi: 10.3748/wjg.v24.i43.4846

Oh, J. H., Jang, S. J., Kim, J., Sohn, I., Lee, J. Y., Cho, E. J., et al. (2020). Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *NPJ Genom. Med.* 5:33.

Qi, Y., Xin, M., Zhang, Y., Hao, Y., Liu, Q., Wang, P., et al. (2021). TTSurv: exploring the multi-gene prognosis in thousands of tumors. *Front. Oncol.* 11:691310. doi: 10.3389/fonc.2021.691310

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

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

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

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

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

Thomas, L. F., Saito, T., and Saetrom, P. (2011). Inferring causative variants in microRNA target sites. *Nucleic Acids Res.* 39:e109. doi: 10.1093/nar/gkr414

Thomson, D. W., and Dinger, M. E. (2016). Endogenous microRNA sponges: evidence and controversy. *Nat. Rev. Genet.* 17, 272–283. doi: 10.1038/nrg.2016.20

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn.)* 19, A68–A77.

Vincent, A., Herman, J., Schulick, R., Hruban, R. H., and Goggins, M. (2011). Pancreatic cancer. *Lancet* 378, 607–620.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, Jr, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Von Hoff, D. D., Ervin, T., Arena, F. P., Chiorean, E. G., Infante, J., Moore, M., et al. (2013). Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N. Engl. J. Med.* 369, 1691–1703.

Wang, G., Jiang, L., Wang, J., Zhang, J., Kong, F., Li, Q., et al. (2020). The G protein-coupled receptor FFAR2 promotes internalization during influenza a virus entry. *J. Virol.* 94, e01707–e01719.

Wang, P., Guo, Q., Hao, Y., Liu, Q., Gao, Y., Zhi, H., et al. (2021). LnCeCell: a comprehensive database of predicted lncRNA-associated ceRNA networks at single-cell resolution. *Nucleic Acids Res.* 49, D125–D133.

Wang, P., Li, X., Gao, Y., Guo, Q., Ning, S., Zhang, Y., et al. (2020). LnCeVar: a comprehensive database of genomic variations that disturb ceRNA network regulation. *Nucleic Acids Res.* 48, D111–D117.

Wang, P., Li, X., Gao, Y., Guo, Q., Wang, Y., Fang, Y., et al. (2019). LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. *Nucleic Acids Res.* 47, D121–D127.

Wang, P., Ning, S., Zhang, Y., Li, R., Ye, J., Zhao, Z., et al. (2015). Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res.* 43, 3478–3489. doi: 10.1093/nar/gkv233

Wang, Y., Yang, G., You, L., Yang, J., Feng, M., Qiu, J., et al. (2019). Role of the microbiome in occurrence, development and treatment of pancreatic cancer. *Mol. Cancer* 18:173.

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

Yu, H., Jiang, X., Jiang, L., Zhou, H., Bao, J., Zhu, X., et al. (2020). Protocadherin 8 (PCDH8) inhibits proliferation, migration, invasion, and angiogenesis in esophageal squamous cell carcinoma. *Med. Sci. Monit.* 26:e920665.

Zhang, Y., Han, P., Guo, Q., Hao, Y., Qi, Y., Xin, M., et al. (2021). Oncogenic landscape of somatic mutations perturbing pan-cancer lncRNA-ceRNA regulation. *Front. Cell Dev. Biol.* 9:658346. doi: 10.3389/fcell.2021.658346

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fgene.2021.741608

Identification of Key Regulators of Hepatitis C Virus-Induced Hepatocellular Carcinoma by Integrating Whole-Genome and Transcriptome Sequencing Data

Guolin Chen, Wei Zhang and Yiran Ben*

Department of Infectious Diseases, The First Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Xinyi Liu,
University of Illinois at Chicago, United States

Reviewed by:
Bo Han,
Capital Medical University, China
Qiong Wu,
The Chinese University of Hong Kong, China
Guangyi Fan,
Beijing Genomics Institute (BGI), China

***Correspondence:**
Guolin Chen
guolinchen139@126.com

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

Received: *15 July 2021*
Accepted: *12 August 2021*
Published: *09 September 2021*

Citation:
Chen G, Zhang W and Ben Y (2021) Identification of Key Regulators of Hepatitis C Virus-Induced Hepatocellular Carcinoma by Integrating Whole-Genome and Transcriptome Sequencing Data. Front. Genet. 12:741608. doi: 10.3389/fgene.2021.741608

Background: Hepatitis C virus (HCV) infection is a major cause of cirrhosis and hepatocellular carcinoma (HCC). Despite recent advances in the understanding of the biological basis of HCC development, the molecular mechanisms underlying HCV-induced HCC (HCC-HCV) remain unclear. The carcinogenic potential of HCV varies according to the genotype and mutation in its viral sequence. Moreover, regulatory pathways play important roles in many pathogenic processes. Therefore, identifying the pathways by which HCV induces HCC may enable improved HCC diagnosis and treatment.

Methods: We employed a systematic approach to identify an important regulatory module in the process of HCV-HCC development to find the important regulators. First, an HCV-related HCC subnetwork was constructed based on the gene expression in HCC-HCV patients and HCC patients. A priority algorithm was then used to extract the module from the subnetworks, and all the regulatory relationships of the core genes of the network were extracted. Integrating the significantly highly mutated genes involved in the HCC-HCV patients, core regulatory modules and key regulators related to disease prognosis and progression were identified.

Result: The key regulatory genes including *EXO1*, *VCAN*, *KIT*, and *hsa-miR-200c-5p* were found to play vital roles in HCV-HCC development. Based on the statistics analysis, *EXO1*, *VCAN*, and *KIT* mutations are potential biomarkers for HCV–HCC prognosis at the genomic level, whereas has-miR-200c-5P is a potential biomarker for HCV–HCC prognosis at the expression level.

Conclusion: We identified three significantly mutated genes and one differentially expressed miRNA, all related to HCC prognosis. As potential pathogenic factors of HCC, these genes and the miRNA could be new biomarkers for HCV-HCC diagnosis.

Keywords: genetic mutation, transcriptome, miRNA, hepatitis C. virus, hepatocellular carcinoma

INTRODUCTION

Hepatocellular carcinoma (HCC)—the second leading cause of cancer−related deaths worldwide (Merte, 1989)—is often diagnosed at an advanced stage and progresses rapidly. Therefore, in HCC patients, early diagnosis is very important to improve their prognosis. Currently, early clinical screening methods for HCC involve serum alpha fetoprotein (AFP) detection and liver ultrasound examination (Sato et al., 1993). However, the sensitivity and specificity of markers such as AFP are marginal; moreover, ultrasound examination considerably relies on the subjective judgment of the operator, and conventional ultrasound results are often not useful for the conclusive identification of liver lesions. Therefore, a more effective, accurate method for screening liver cancer is needed urgently. As the understanding of cancer biology improves, liquid biopsy will become an increasingly useful tool for early diagnosis. Risk factors for HCC include cirrhosis, aflatoxin B intake, alcohol consumption, and hepatitis B virus (HBV) and hepatitis C virus (HCV) infection. Of these, HBV and HCV infections are the most notorious; in general, HBV- or HCV-positive patients have a 15–20-fold higher lifetime relative risk of HCC than HBV- and HCV-negative patients (El-Serag, 2012). To date, few studies have been focused on the factors leading to liver cancer in HCV patients. At present, HCV RNA, cirrhosis, and HCV genotype are thought to affect the occurrence of HCV-related liver cancer, but the involvement of these factors has not been conclusively proven. At present, the number of people affected by chronic HCV infection is 180 million—linked to > 350,000 deaths annually (Li and Lo, 2015). Epidemiological studies have also shown that HCV is a risk factor for various diseases, including oral manifestations, glomerulopathies, type 2 diabetes mellitus, and insulin resistance (Montenegro et al., 2013; Carrozzo and Scally, 2014; Ozkok and Yildiz, 2014).

In total, 55–85% of people with HCV infection will develop chronic hepatitis C, and 20–30% of people with chronic liver disease will develop liver failure or cirrhosis (Mahale et al., 2017). Over the course of 30 years, 1–3% of patients with HCV without cirrhosis will develop HCC eventually (Huang et al., 2011; El-Serag, 2012). Moreover, one-third of HCC cases have been reported to be caused by hepatitis C (Parkin, 2006). At present, there are three major known mechanisms for HCV-induced HCC (HCV-HCC): direct pathways involving HCV core proteins, indirect pathways caused by oxidative stress and steatosis, and microRNA (miRNA)-related pathways (Tholey and Ahn, 2015). While biological signaling systems are complex, the analysis of linear pathways may still provide valuable insights (Weng et al., 1999). In the study of HCV, core genes have been found to be closely related to the carcinogenicity of chronic HCV infection. The expression of core genes has been experimentally shown to immortalize primary liver cells and induce cell transformation and carcinogenesis (Li et al., 2010). In addition, the genome sequencing analysis has demonstrated significant differences in the characteristics of liver cancer patients with or without HCV (Fishman et al., 2009). Taken together, these results indicate that core HCV gene mutations are closely associated with increased liver cancer risks.

In this study, the correlation between the key regulators and prognosis was investigated by integrating whole-genome and transcriptome sequencing data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. We identified differentially expressed and mutated genes between HCV-HCC and HCC groups and performed functional enrichment analysis for genes in the module. Then, we explored the association of the key regulators with patient prognoses. The module and the key regulators may be potential biomarkers for predicting HCV-HCC.

MATERIALS AND METHODS

The Cancer Genome Atlas and Gene Expression Omnibus Data Acquisition

Gene mutation and mRNA and miRNA expression data as well as clinical information were downloaded from TCGA [1] (Deng et al., 2016). In TCGA, liver hepatocellular carcinoma (LIHC) samples are divided into two groups: the first group contains HCV RNA or genotype or hepatitis C antibody in the patient's clinical information, and the other group does not; here, we named the two groups HCC-HCV and HCC. The gene/miRNA microarray as verifying cohorts GSE154211 (Wang et al., 2021) and GSE119159 (Umezu et al., 2020) were downloaded from GEO database. The data were normalized, and R and its packages were employed in all analysis steps.

Differential Analysis

MuTect2 Somatic Mutation data, analyzed using MuTect2, were download from TCGA. TCGA provides somatic mutation data in the MAF format. Therefore, we visualized somatic mutations using the R package "maftools" (Mayakonda et al., 2018). In total, 96 HCC-HCV samples and 269 HCC samples were present in the dataset. We calculated the mutational status of genes using the algorithm in maftools, and the genes with $p < 0.05$, OR > 2, and number of mutations > 5 were selected as the significantly and differentially mutated genes.

According to the groupings, we performed normalization and differential gene expression analysis using the R package "edgeR." False discovery rate (FDR) < 0.01 and $|\log_2$ fold change (FC)$|$ > 1 were used as cutoffs for identify differentially expressed genes (DEGs) for further analysis. Two R packages "pheatmap" and "ggplot2" were used for visualizing the heatmaps and volcano maps, respectively.

In total, 139 HCV patients were enrolled in the GSE119159. They included 99 patients who had not developed HCC and 40 who had developed HCC. A total of 10 samples (tumor and non-tumor regions) from two HCV-related HCC patients and three HCC patients were used to find the gene candidates in HCV-related HCC in the GSE154211. For differential expression analysis, we used the R package "limma." | log2FC| > 1 and logFDR < 0.01 were used as cutoffs to identify DEGs for further analysis.

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

Construction of the Transcription Factor–miRNA–mRNA Regulatory Network

The human transcription factor (TF) and miRNA regulatory networks were constructed by integrating miRTarBase, TRANSFAC, and TransmiR (Vlachos et al., 2015; Chou et al., 2018; Tong et al., 2019). The three databases include curated interactions among human TFs, miRNAs, and target genes. We uniformly named the genes and miRNAs within the regulatory networks according to the National Center for Biotechnology Information (NCBI) and miRbase databases. Moreover, all regulatory relationships within the regulatory

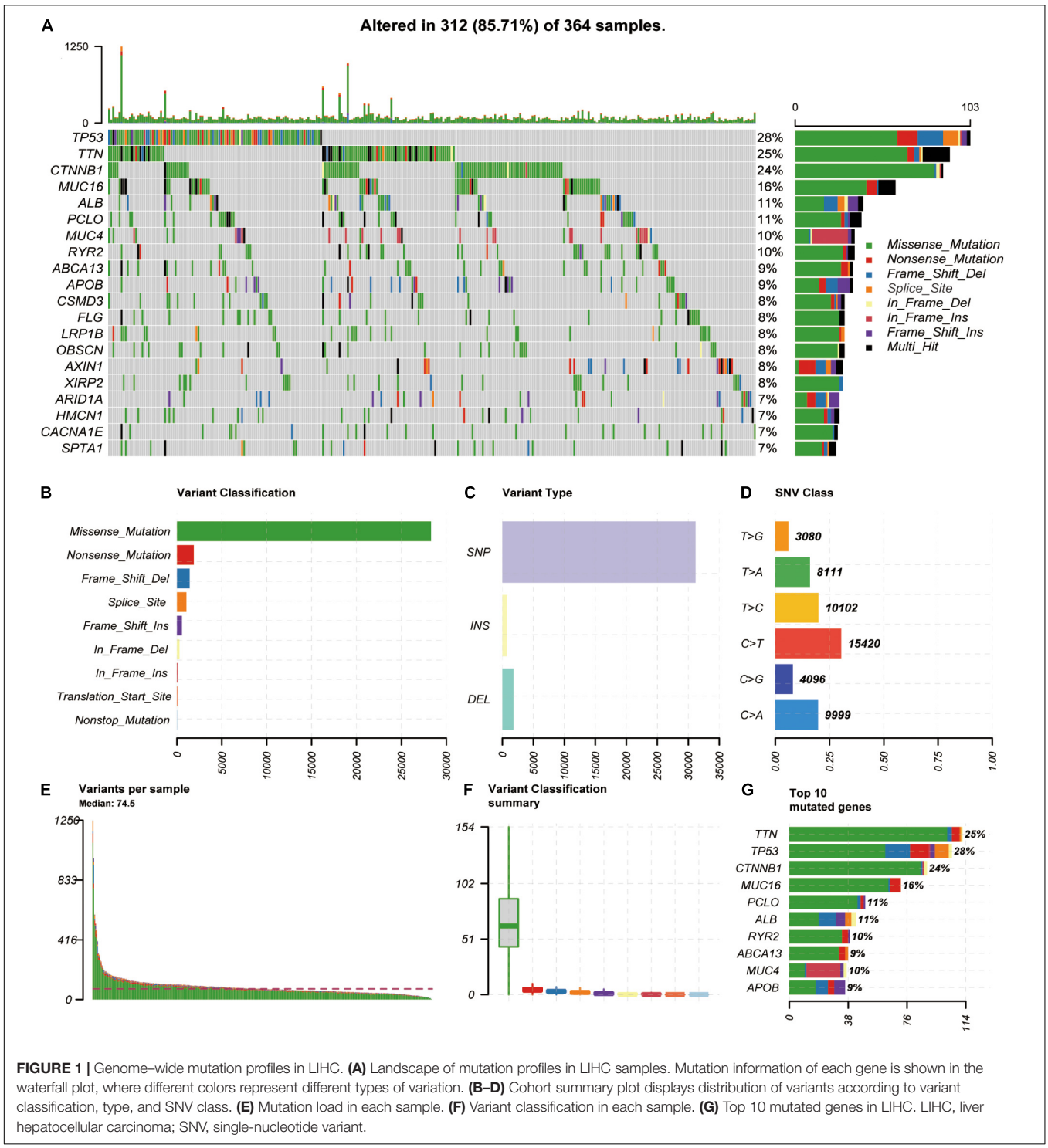


FIGURE 1 | Genome–wide mutation profiles in LIHC. **(A)** Landscape of mutation profiles in LIHC samples. Mutation information of each gene is shown in the waterfall plot, where different colors represent different types of variation. **(B–D)** Cohort summary plot displays distribution of variants according to variant classification, type, and SNV class. **(E)** Mutation load in each sample. **(F)** Variant classification in each sample. **(G)** Top 10 mutated genes in LIHC. LIHC, liver hepatocellular carcinoma; SNV, single-nucleotide variant.

network were supported experimentally. In total, 888 TFs, 1,072 miRNAs, 3,150 target genes, and 18,056 edges were discovered in the regulatory network.

Functional Enrichment Analysis

The key regulatory gene symbols were converted to Entrez ID using the R package "org.Hs.eg.db." To identify the biological pathways involved in HCV-HCC occurrence and development, we employed Gene Ontology–biological process (GO-BP) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and visualized the results using the R packages "clusterProfiler" and "ggplot2."

Survival Analysis

We constructed an HCV-HCC-related subnetwork and identified key regulators from the subnetworks. Next, we investigated whether the key regulators could distinguish HCC patients with good or poor outcomes. From these data, we obtained TCGA HCC dataset with mRNA/miRNA expression and clinical information. Then, we used the key regulator expression values and mutation information to cluster all patients into two groups. The differential survival of the two study groups was finally assessed using the log-rank test.

RESULTS

Mutation Analysis

We downloaded and analyzed the somatic mutation data of 392 TCGA-LIHC samples. The mutation information of all genes in the samples is displayed as a waterfall diagram, with different colors representing different mutation types

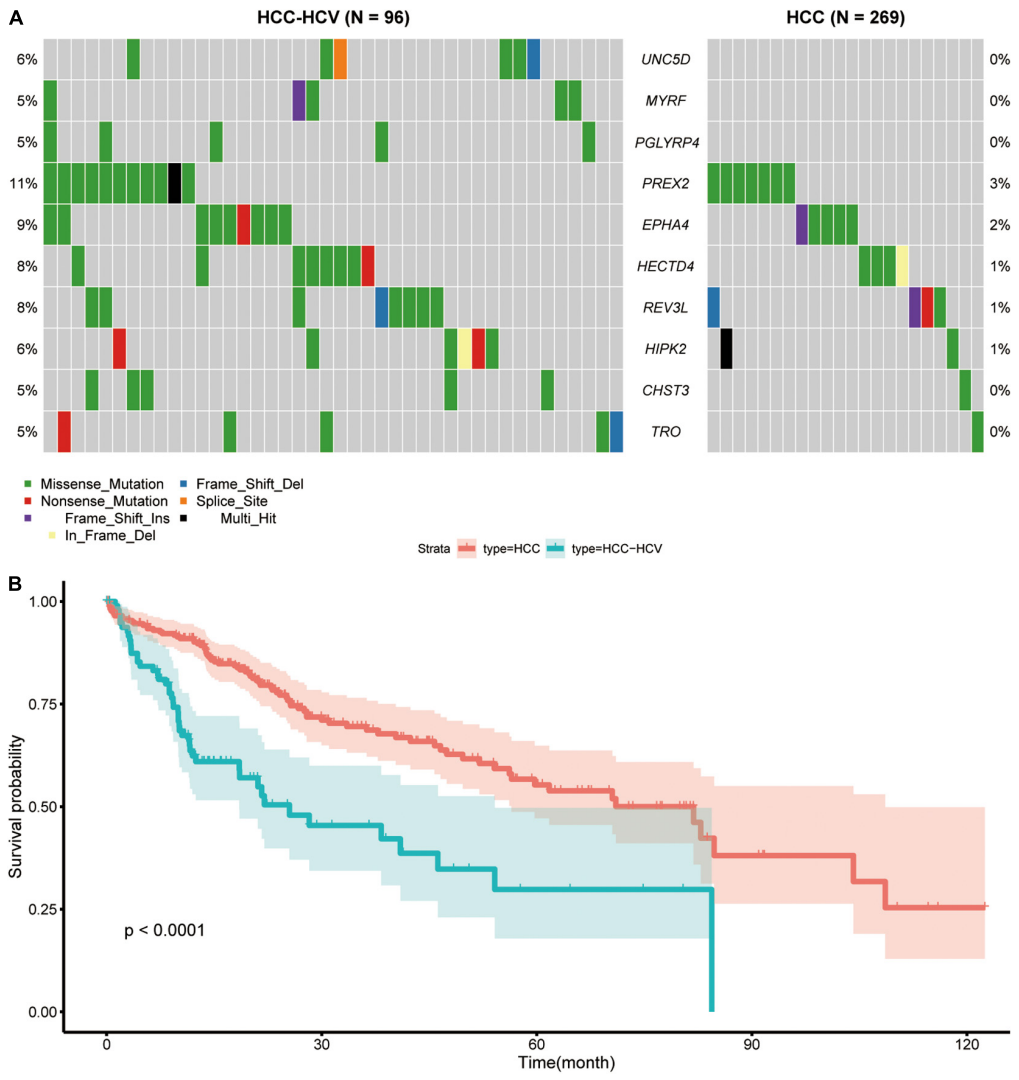


FIGURE 2 | Analyses of different somatic mutations and survival time in HCC-HCV and HCC samples. **(A)** Waterfall plot of detailed information of top 10 differentially mutated genes in each group. **(B)** K-M curves of patients in the HCC group and HCC-HCV group. HCC, hepatocellular carcinoma; HCV, hepatitis C virus; K-M, Kaplan–Meier.

(**Figure 1A**). Further analysis showed that missense mutation, single-nucleotide polymorphisms (SNPs), and C > T accounted for the highest proportion of the variations (**Figures 1B–D**). The median number of variations in all HCC samples was 74.5, and the maximum number of variations in a single sample was 1,250 (**Figure 1E**). The number of variations in different classifications in all samples is shown in a box diagram (**Figure 1F**). The top 10 mutated genes in the 392 samples were TTN (25%), TP53 (28%), CTNNB1 (24%), MUC16 (16%), PCLO (11%), ALB (11%), RYR2 (10%), ABCA13 (9%), MUC4 (10%), and APOB (9%; **Figure 1G**). In total, 96 HCC-HCV samples and 269 HCC samples were present in TCGA dataset; the survival analysis indicated that HCC patients without HCV lived significantly longer than HCC-HCV patients (**Figure 2B**). With the use of the maftools algorithm, 41 differentially mutated genes were identified (**Supplementary Table 1**). The top 10 differential mutated genes between the two groups of patients were UNC5D (6–0), MYRF (5–0), PGLYRP4 (5–0), PREX2 (11–7), EPHA4 (9–5), HECTD4 (8–4), REV3L (8–4), HIPK2 (6–2), CHST3 (5–1), and TRO (5–1; **Figure 2A**). Moreover, HCC-HCV patients with mutations in some genes had a poor prognosis (**Supplementary Figure 1**).

Transcriptome Analysis

Differentially expressed mRNAs and miRNAs were identified from two raw datasets: one containing GSE154211 and GSE119159 downloaded from the GEO database and another dataset from TCGA database. In total, 530 mRNA and 30 miRNA transcripts were observed to be expressed differentially in the HCC-HCV samples compared with HCC samples in TCGA dataset—including, respectively, 412 and 25 upregulated and 118 and five downregulated transcripts. Hierarchical clustering showed systematic variations in mRNA and miRNA expression in the HCC-HCV and HCC samples (**Supplementary Figure 2**). To identify the genes related to HCC-HCV in GSE154211, we first divided the expression data into four groups to identify DEGs between (A) HCC vs. HCC-HCV-adjacent, suggesting related to HCV-related carcinogenesis; (B) HCC-HCV vs. HCC, suggesting related HCV-related hepatocarcinogenesis; (C) HCC-HCV-adjacent vs. HCC-adjacent, suggesting related to HCV-related non-oncogenic effects; and (D) HCC vs. HCC-adjacent, suggesting related to non-HCV-related carcinogenesis. Four groups of data were then analyzed. Consequently, we identified 1,494 DEGs belonging to group A or B, but not group C or D, as genes with strong potential to be relevant to

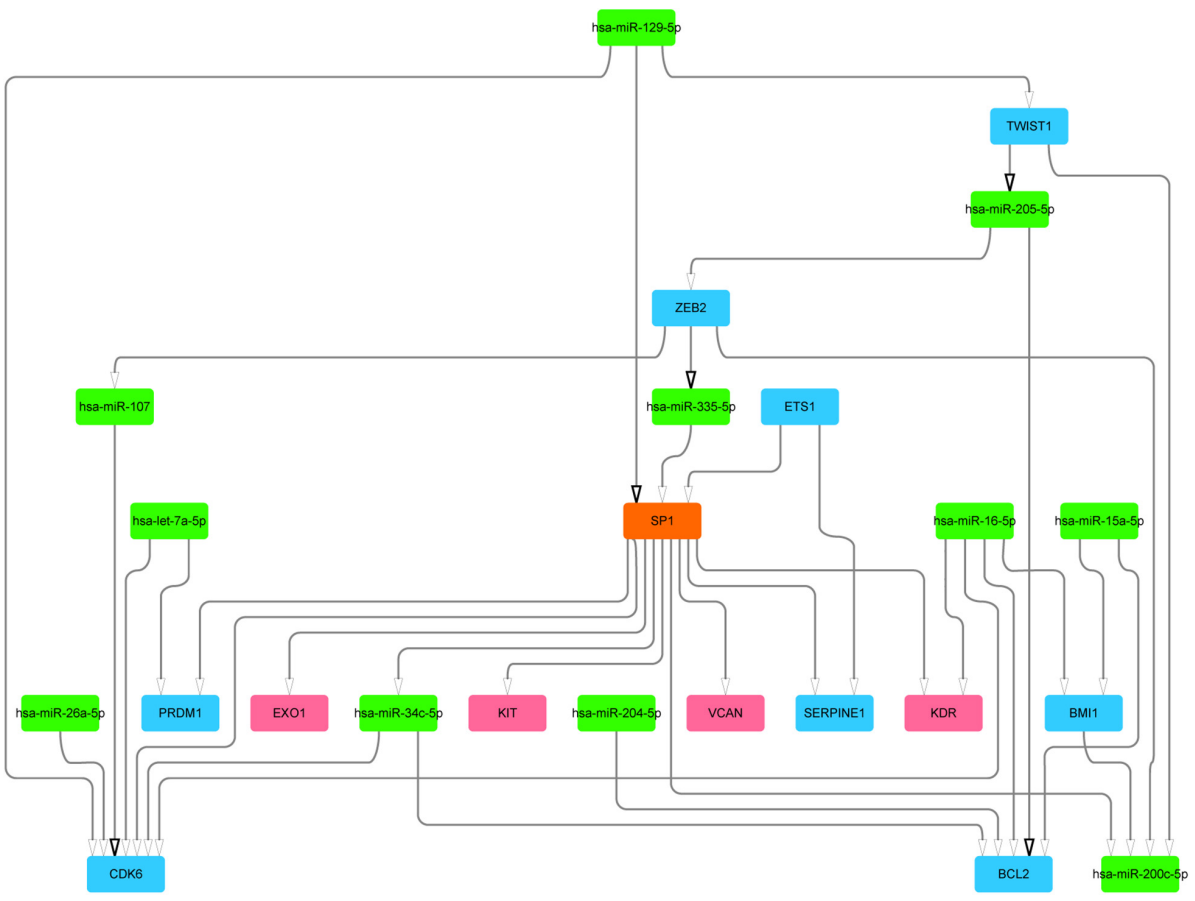


FIGURE 3 | The core regulatory module and key regulators. Blue color represents DEGs, green color represents DE-miRs, and red color represents different mutation genes. DEGs, differentially expressed genes; DE-miRs, differentially expressed microRNAs.

HCC-HCV (**Supplementary Figure 3**). In addition, 21 miRNA transcripts were observed to be differentially expressed in the developed HCC samples compared with the non-developed HCC samples in GSE119159, including nine upregulated and 12 downregulated transcripts.

The Core Regulatory Module and Key Regulators

To mine HCV-HCC-related regulatory relationships, we first constructed a TF–miRNA–mRNA regulatory network as a background network. Then, the HCV-induced HCC-related subnetwork was constructed by mapping DEGs into the background network. The nodes in the subnetwork contained DEGs and genes directly connected to the DEGs. In total, 359 TFs, 395 miRNAs, 739 target genes, and 2626 edges were present in the subnetwork.

We next mined the core regulatory module from the subnetwork by extracting the top 20 nodes ranked by closeness centrality and the edges among them. Notably, the regulatory relationships between these 20 nodes and differential mutated genes were added into the core regulatory module (**Figure 3** and **Supplementary Table 2**). Finally, the core regulatory module contained 24 nodes and 36 edges.

To analyze the function of genes in the module, we conducted enrichment analysis of GO and KEGG, with $FDR < 0.05$ used as the cutoff to identify statistically significant GO terms and KEGG pathways. We found that many GO terms and KEGG pathways were implicated in the HCV-HCC processes in previous studies. As shown in **Figure 4A** and **Supplementary Table 3**, in the biological process and molecular function categories, the significantly enriched genes were for vasculature development regulation (Vescovo et al., 2016), ameboidal-type and epithelial cell migration (Khera et al., 2017), cell aging (Naggie, 2017), cell-matrix adhesion (Ninio et al., 2019), and melanocyte differentiation and angiogenesis involved in wound healing (Mohsen et al., 2014). Furthermore, KEGG pathway analysis showed that the significantly enriched genes were for small cell lung cancer, miRNAs in cancer, PI3K–Akt signaling pathway (Cheng et al., 2015), Ras and p53 signaling pathways (Vescovo et al., 2016), cellular senescence (Shiu et al., 2017), endocrine resistance, and advanced glycation end products (AGE)–receptor for AGE (RAGE) signaling pathway in diabetic complications (Hyogo and Yamagishi, 2008; **Figure 4B** and **Supplementary Table 4**).

We further analyzed the genes in the core regulatory module and found that expression of *EXO1*, *VCAN*, *has-miR-200c-5p*, *BMI1*, *has-miR-204-5p*, and *KIT* was significantly correlated with HCC prognosis in all patients; and thus, these genes were considered key regulators (**Figure 5A**). In particular, we found that the patients with low *EXO1*, *VCAN*, or *KIT* expression had an adverse outcome ($HR < 1$; **Figure 5A**). The HCC-HCV patients with mutations in these three genes have possibly also poor prognoses (**Figures 5B–D**). They may be potential biomarkers to predict the prognosis of patients at the genomic level. Moreover, we found that *has-miR-200c-5P* was significantly overexpressed in HCC-HCV samples (**Figure 6A**). The survival time of patients with high *has-miR-200c-5P* expression was significantly lower than that of patients with low expression (**Figure 6B**), suggesting that *has-miR-200c-5P* may be a potential biomarker to predict the prognosis of patients at the expression level.

Mutations in specific locations in *EXO1* have been reported to inactivate proteins that increase cancer susceptibility (Welchew et al., 2002). *KDR* was also a significantly differential mutated gene in the module. KDR is the principal receptor that promotes the proangiogenic action of vascular endothelial growth factor

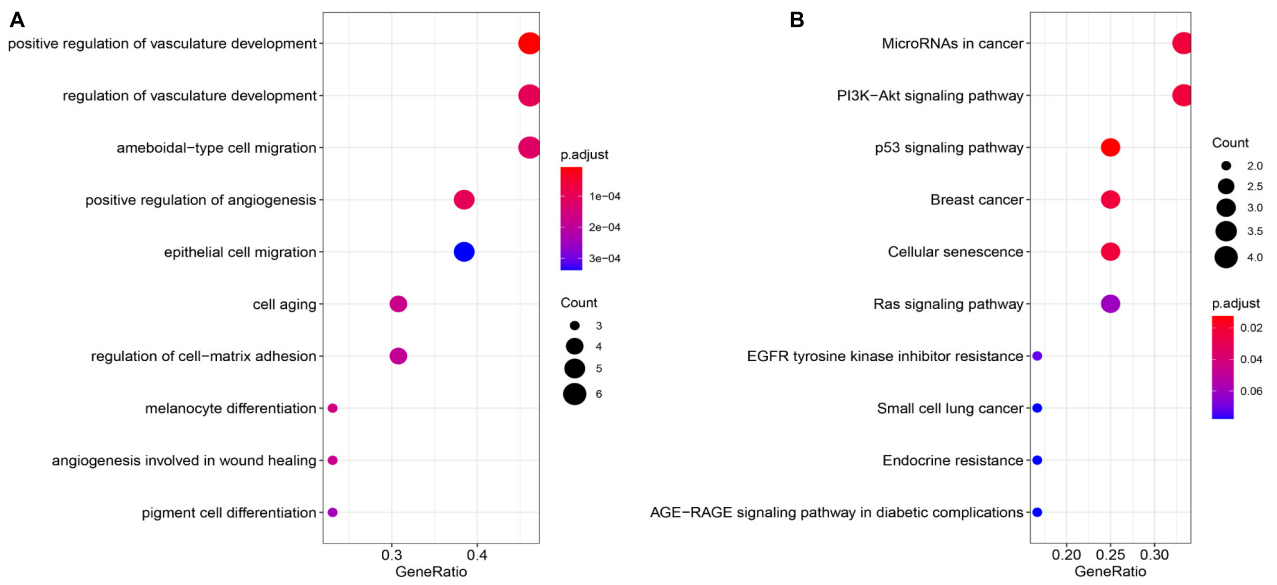


FIGURE 4 | GO and KEGG pathway enrichment analyses. **(A)** GO enrichment analysis of the module genes. **(B)** KEGG pathway enrichment analysis of the module genes. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

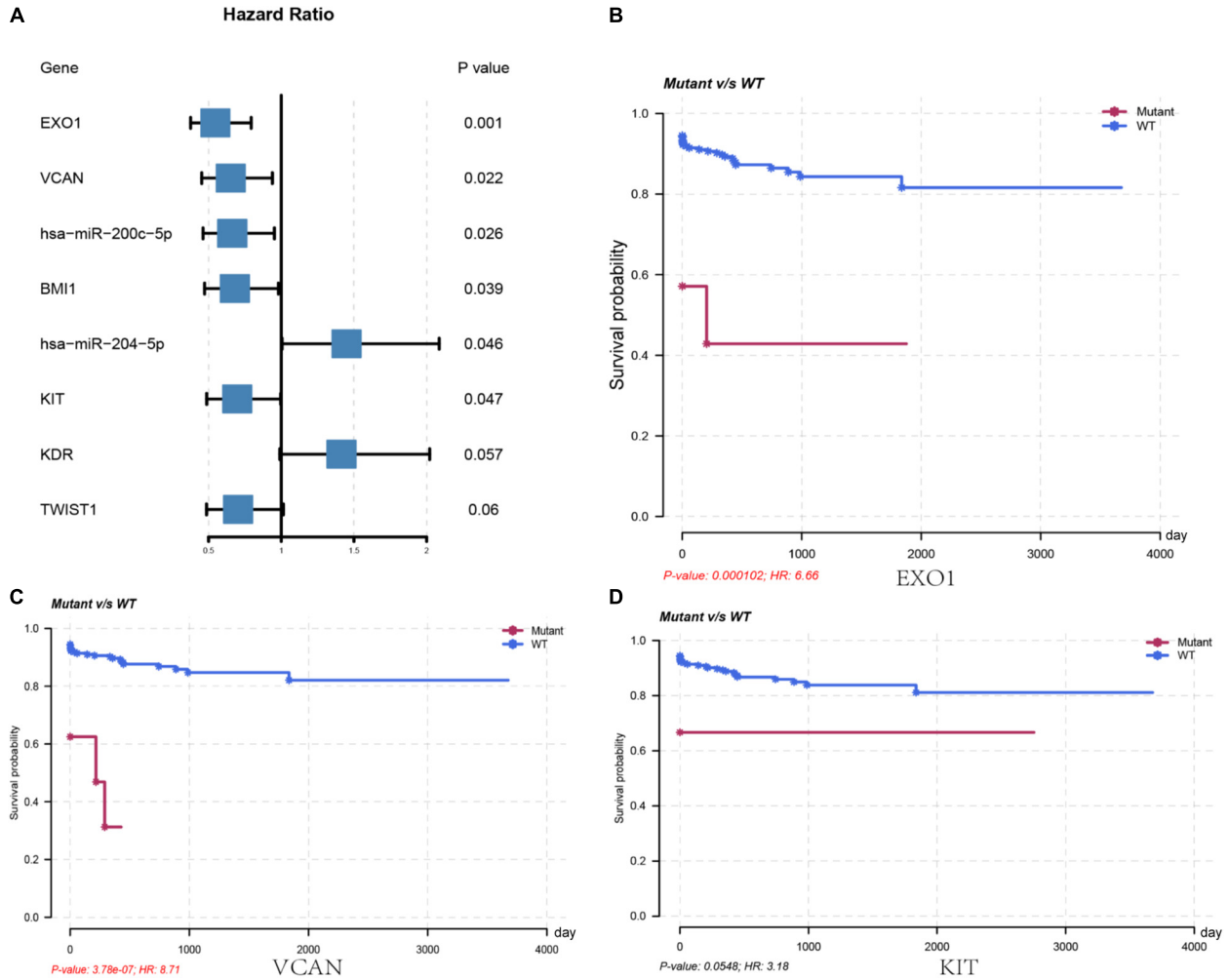


FIGURE 5 | Key regulators are related to worse survival rate. **(A)** Forest plot of hazard ratios showing the prognostic values of genes. **(B–D)** Survival curves of key regulators in the LIHC patients from TCGA dataset. LIHC, liver hepatocellular carcinoma; TCGA, The Cancer Genome Atlas.

and is involved in the tumorigenesis and progression of many malignancies, including HCC (Zheng et al., 2014). Moreover, *BCL2* was the downstream gene in the core regulatory module. *BCL2* can be functionally divided into antiapoptotic and proapoptotic groups. The balance between these two groups may determine the fate of a tumor cell. In HCC, this balance is often tilted toward the antiapoptotic members, leading to resistance to death and rapid proliferation in cancer cells (Alenzi et al., 2010). *BCL2* expression in the HCC-HCV samples was lower than that in the HCC samples, but the difference was non-significant in TCGA data—which may be a reason for the worse prognosis of HCC-HCV.

DISCUSSION

HCC is responsible for the second highest global mortality rate, and HCV infection is a leading HCC risk factor. However, the mechanisms of HCC initiation, development, and metastasis are too complicated and thus unclear (Kanda et al., 2019). Currently, several factors are believed to influence the evolution of HCC from HCV infection. However, due to the lack of appropriate models or data, determining the specific role of HCV in the malignant transformation of liver cells is difficult. To identify and characterize these mechanisms, researchers have conducted genomic, transcriptomic, and epigenomic studies (Khatun and Ray, 2019).

Driver mutations in cancer-associated genes alter downstream signaling and transcription patterns, which are critical in cancer progression (Lai and Yang, 2013; Zhang et al., 2014, 2015; Huh et al., 2019). These studies have revealed that downstream gene mutations and gene expression changes are critical in hepatitis-induced liver cancer development. In this study, we found that mutations in a single gene can have a significant impact on disease prognosis in patients, whereas a combination of mutations in multiple genes is not an effective predictor of prognosis. This may be due to the low probability of simultaneous mutations of multiple genes; this

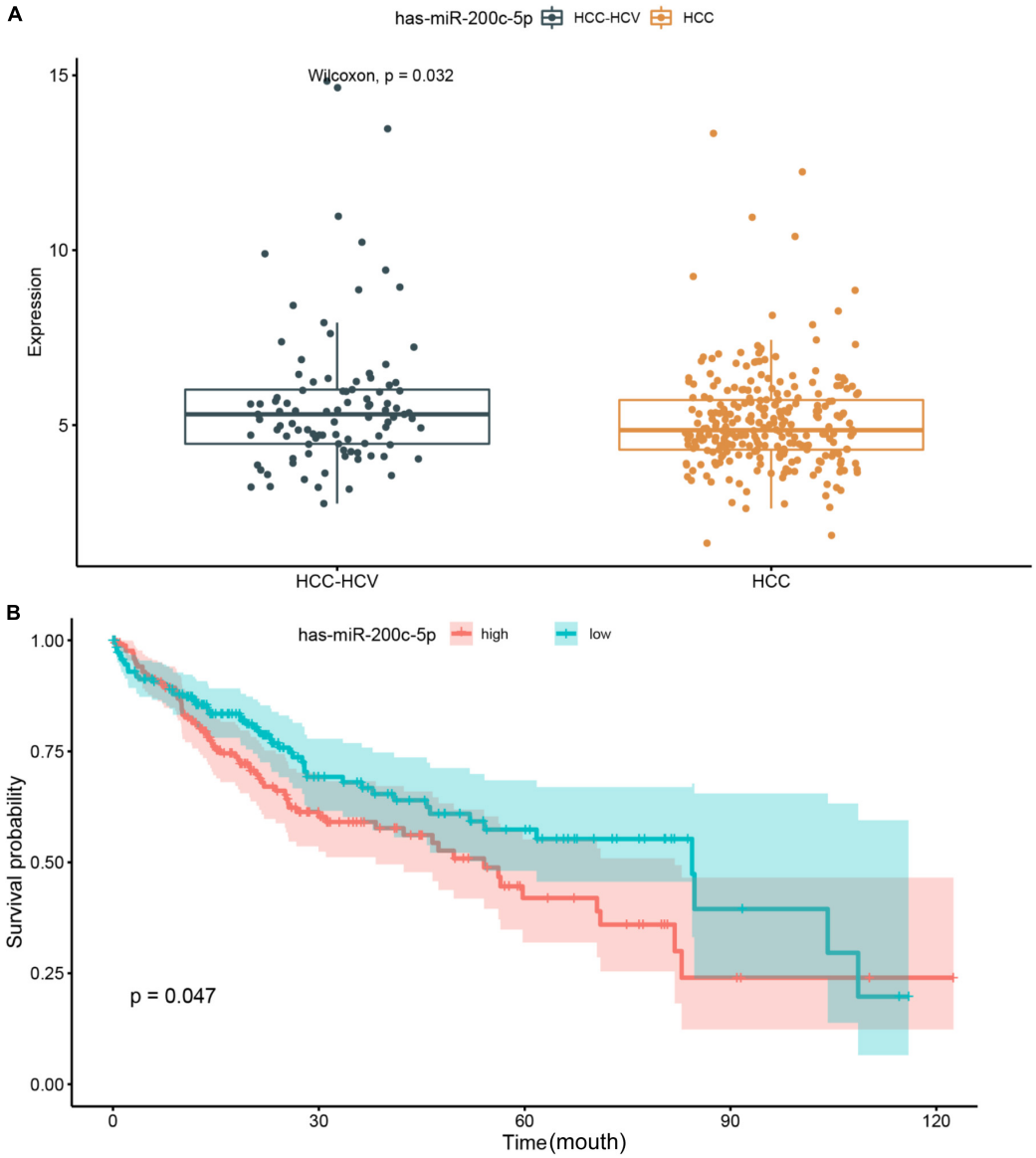


FIGURE 6 | Relationship between miR-200c expression level and patient prognosis. **(A)** Significant difference was found in the miR-200c expression between HCC-HCV and HCC patients. **(B)** HCC patients with high miR-200c expression have a relatively poor prognosis. HCC, hepatocellular carcinoma; HCV, hepatitis C virus.

will be studied further in our future work. Genomic research has found that long-term interactions between hepatitis virus and immune system causes significant stress and damage to the liver cells, making them undergo pathological adaptation—even after elimination of the virus. Non-coding RNA (ncRNA)-related analysis has indicated that miRNAs play a crucial role in the posttranscriptional regulation of gene expression (Wong et al., 2018). Deregulation of certain miRNAs leads to the inactivation of tumor-suppressor genes and activation of HCC-related oncogenes. In this study, we incorporated whole-genome and transcriptomic sequence data to identify key regulators of HCV-HCC and found that abnormal expression of certain genes and miRNAs predict whether a patient with HCV infection will develop HCC. These genes may be potential biomarkers, which could enable HCC detection at significantly earlier stages.

In the functional enrichment analysis, we found that genes in the module were significantly enriched in the PI3K–Akt signaling pathway that promotes survival and growth in response to extracellular signals. KIT is an important receptor tyrosine kinase (RTK) that can stimulate the PI3K–Akt signaling pathway (Zhou et al., 2011). In addition, recent studies have shown that KIT exon 9 had a mutation resistant to TGFβ, which can promote HCC development in HCV patients (El-Houseini et al., 2019). The miR-200 family—the most common family of miRNAs—demonstrates low expression in various cancers

and is closely associated with tumorigenesis and outcome, particularly in HCC (Mao et al., 2020). *has-miR-200c-5P* is significantly overexpressed in HCV patients and promotes hepatic fibrosis (Ramachandran et al., 2013)—consistent with our results. Moreover, the survival time of patients with high *has-miR-200c-5P* expression was significantly lower than that of patients with low expression in the current study. In general, *has-miR-200c-5P* overexpression in esophageal cancer increases resistance to chemotherapeutic drugs by dysregulating PI3K–Akt signaling pathway (Karakatsanis et al., 2013). Therefore, we speculate that *has-miR-200c-5P* and *KIT* may jointly regulate the PI3K–Akt signaling pathway and affect drug response and prognosis in HCV-HCC patients.

Although we identified some important regulatory genes and miRNAs, the specific underlying mechanisms could not be elaborated. Furthermore, HCC is complicated and multifactorial, and taking all factors into consideration was difficult. Therefore, additional studies determining whether genes correlated with HCV-induced cancer are also correlated with liver cancer caused by other factors are warranted.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

GC: study design, manuscript writing, and data analysis. WZ: data analysis, data collection, and manuscript writing. YB: data analysis and data collection. All authors have read, edited and approved of the final version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.741608/full#supplementary-material

REFERENCES

Alenzi, F. Q., El-Nashar, E. M., Al-Ghamdi, S. S., Abbas, M. Y., Hamad, A. M., El-Saeed, O. M., et al. (2010). Original article: investigation of Bcl-2 and PCNA in hepatocellular carcinoma: relation to chronic HCV. *J. Egypt. Natl. Canc. Inst.* 22, 87–94.

Carrozzo, M., and Scally, K. (2014). Oral manifestations of hepatitis C virus infection. *World J. Gastroenterol.* 20, 7534–7543. doi: 10.3748/wjg.v20.i24.7534

Cheng, D., Zhang, L., Yang, G., Zhao, L., Peng, F., Tian, Y., et al. (2015). Hepatitis C virus NS5A drives a PTEN-PI3K/Akt feedback loop to support cell survival. *Liver Int.* 35, 1682–1691. doi: 10.1111/liv.12733

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Deng, M., Bragelmann, J., Schultze, J. L., and Perner, S. (2016). Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. *BMC Bioinformatics* 17:72. doi: 10.1186/s12859-016-0917-9

El-Houseini, M. E., Ismail, A., Abdelaal, A. A., El-Habashy, A. H., Abdallah, Z. F., Mohamed, M. Z., et al. (2019). Role of TGF-beta1 and C-Kit mutations in the development of hepatocellular carcinoma in hepatitis C virus-infected patients: in vitro study. *Biochemistry* 84, 941–953. doi: 10.1134/S0006297919080108

El-Serag, H. B. (2012). Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology* 142, 1264–1273.e1. doi: 10.1053/j.gastro.2011.12.061

Fishman, S. L., Factor, S. H., Balestrieri, C., Fan, X., Dibisceglie, A. M., Desai, S. M., et al. (2009). Mutations in the hepatitis C virus core gene are associated with advanced liver disease and hepatocellular carcinoma. *Clin. Cancer Res.* 15, 3205–3213. doi: 10.1158/1078-0432.CCR-08-2418

Huang, Y. T., Jen, C. L., Yang, H. I., Lee, M. H., Su, J., Lu, S. N., et al. (2011). Lifetime risk and sex difference of hepatocellular carcinoma among patients with chronic hepatitis B and C. *J. Clin. Oncol.* 29, 3643–3650. doi: 10.1200/JCO.2011.36.2335

Huh, H. D., Kim, D. H., Jeong, H. S., and Park, H. W. (2019). Regulation of TEAD transcription factors in cancer biology. *Cells* 8:600. doi: 10.3390/cells8060600

Hyogo, H., and Yamagishi, S. (2008). Advanced glycation end products (AGEs) and their involvement in liver disease. *Curr. Pharm. Des.* 14, 969–972. doi: 10.2174/138161208784139701

Kanda, T., Goto, T., Hirotsu, Y., Moriyama, M., and Omata, M. (2019). Molecular mechanisms driving progression of liver cirrhosis towards hepatocellular carcinoma in chronic hepatitis B and C infections: a review. *Int. J. Mol. Sci.* 20:1358. doi: 10.3390/ijms20061358

Karakatsanis, A., Papaconstantinou, I., Gazouli, M., Lyberopoulou, A., Polymeneas, G., and Voros, D. (2013). Expression of microRNAs, miR-21, miR-31, miR-122, miR-145, miR-146a, miR-200c, miR-221, miR-222, and miR-223 in patients with hepatocellular carcinoma or intrahepatic cholangiocarcinoma and its prognostic significance. *Mol. Carcinog.* 52, 297–303. doi: 10.1002/mc.21864

Khatun, M., and Ray, R. B. (2019). Mechanisms underlying hepatitis C virus-associated hepatic fibrosis. *Cells* 8:1249. doi: 10.3390/cells8101249

Khera, L., Paul, C., and Kaul, R. (2017). Hepatitis C Virus E1 protein promotes cell migration and invasion by modulating cellular metastasis suppressor Nm23-H1. *Virology* 506, 110–120. doi: 10.1016/j.virol.2017.03.014

Lai, D., and Yang, X. (2013). BMP4 is a novel transcriptional target and mediator of mammary cell migration downstream of the Hippo pathway component TAZ. *Cell. Signal.* 25, 1720–1728. doi: 10.1016/j.cellsig.2013.05.002

Li, H. C., and Lo, S. Y. (2015). Hepatitis C virus: virology, diagnosis and treatment. *World J. Hepatol.* 7, 1377–1389. doi: 10.4254/wjh.v7.i10.1377

Li, Z. H., Tang, Q. B., Wang, J., Zhou, L., Huang, W. L., Liu, R. Y., et al. (2010). Hepatitis C virus core protein induces malignant transformation of biliary epithelial cells by activating nuclear factor-kappaB pathway. *J. Gastroenterol. Hepatol.* 25, 1315–1320. doi: 10.1111/j.1440-1746.2009.06201.x

Mahale, P., Torres, H. A., Kramer, J. R., Hwang, L. Y., Li, R., Brown, E. L., et al. (2017). Hepatitis C virus infection and the risk of cancer among elderly US adults: a registry-based case-control study. *Cancer* 123, 1202–1211. doi: 10.1002/cncr.30559

Mao, Y., Chen, W., Wu, H., Liu, C., Zhang, J., and Chen, S. (2020). Mechanisms and functions of MiR-200 family in hepatocellular carcinoma. *Onco Targets Ther.* 13, 13479–13490. doi: 10.2147/OTT.S288791

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28, 1747–1756. doi: 10.1101/gr.239244.118

Merte, B. (1989). [The first general mathematical description of optical lenses without aberrations]. *Klin. Monbl. Augenheilkd.* 194, 59–61.

Mohsen, M. A. A., Hussein, N. A., Ghazal, A. A., El-Ghandour, M. K., Farouk, M., El-Wahab, A., et al. (2014). Angiogenic output in viral hepatitis, C and B, and HCV-associated hepatocellular carcinoma. *Alexandria J. Med.* 50, 235–240.

Montenegro, L., De Michina, A., Misciagna, G., Guerra, V., and Di Leo, A. (2013). Virus C hepatitis and type 2 diabetes: a cohort study in southern Italy. *Am. J. Gastroenterol.* 108, 1108–1111. doi: 10.1038/ajg.2013.90

Naggie, S. (2017). Hepatitis C virus, inflammation, and cellular aging: turning back time. *Top. Antivir. Med.* 25, 3–6.

Ninio, L., Nissani, A., Meirson, T., Domovitz, T., Genna, A., Twafra, S., et al. (2019). Hepatitis C virus enhances the invasiveness of hepatocellular carcinoma

via EGFR-mediated invadopodia formation and activation. *Cells* 8:1395. doi: 10.3390/cells8111395

Ozkok, A., and Yildiz, A. (2014). Hepatitis C virus associated glomerulopathies. *World J. Gastroenterol.* 20, 7544–7554. doi: 10.3748/wjg.v20.i24.7544

Parkin, D. M. (2006). The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* 118, 3030–3044. doi: 10.1002/ijc.21731

Ramachandran, S., Ilias Basha, H., Sarma, N. J., Lin, Y., Crippin, J. S., Chapman, W. C., et al. (2013). Hepatitis C virus induced miR200c down modulates FAP-1, a negative regulator of Src signaling and promotes hepatic fibrosis. *PLoS One* 8:e70744. doi: 10.1371/journal.pone.0070744

Sato, Y., Nakata, K., Kato, Y., Shima, M., Ishii, N., Koji, T., et al. (1993). Early recognition of hepatocellular carcinoma based on altered profiles of alpha-fetoprotein. *N. Engl. J. Med.* 328, 1802–1806. doi: 10.1056/NEJM199306243282502

Shiu, T. Y., Shih, Y. L., Feng, A. C., Lin, H. H., Huang, S. M., Huang, T. Y., et al. (2017). HCV core inhibits hepatocellular carcinoma cell replicative senescence through downregulating microRNA-138 expression. *J. Mol. Med.* 95, 629–639. doi: 10.1007/s00109-017-1518-4

Tholey, D. M., and Ahn, J. (2015). Impact of hepatitis C virus infection on hepatocellular carcinoma. *Gastroenterol. Clin. North Am.* 44, 761–773. doi: 10.1016/j.gtc.2015.07.005

Tong, Z., Cui, Q., Wang, J., and Zhou, Y. (2019). TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res.* 47, D253–D258. doi: 10.1093/nar/gky1023

Umezu, T., Tsuneyama, K., Kanekura, K., Hayakawa, M., Tanahashi, T., Kawano, M., et al. (2020). Comprehensive analysis of liver and blood miRNA in precancerous conditions. *Sci. Rep.* 10:21766. doi: 10.1038/s41598-020-78500-1

Vescovo, T., Refolo, G., Vitagliano, G., Fimia, G. M., and Piacentini, M. (2016). Molecular mechanisms of hepatitis C virus-induced hepatocellular carcinoma. *Clin. Microbiol. Infect.* 22, 853–861. doi: 10.1016/j.cmi.2016.07.019

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., et al. (2015). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.* 43, D153–D159. doi: 10.1093/nar/gku1215

Wang, S. C., Li, C. Y., Chang, W. T., Cheng, W. C., Yen, C. H., Tu, W. Y., et al. (2021). Exosome-derived differentiation antagonizing non-protein coding RNA with risk of hepatitis C virus-related hepatocellular carcinoma recurrence. *Liver Int.* 41, 956–968. doi: 10.1111/liv.14772

Welchew, D. E., Honey, G. D., Sharma, T., Robbins, T. W., and Bullmore, E. T. (2002). Multidimensional scaling of integrated neurocognitive function and schizophrenia as a disconnexion disorder. *Neuroimage* 17, 1227–1239. doi: 10.1006/nimg.2002.1246

Weng, G., Bhalla, U. S., and Iyengar, R. (1999). Complexity in biological signaling systems. *Science* 284, 92–96. doi: 10.1126/science.284.5411.92

Wong, C. M., Tsang, F. H., and Ng, I. O. (2018). Non-coding RNAs in hepatocellular carcinoma: molecular functions and pathological implications. *Nat. Rev. Gastroenterol. Hepatol.* 15, 137–151. doi: 10.1038/nrgastro.2017.169

Zhang, W., Gao, Y., Li, F., Tong, X., Ren, Y., Han, X., et al. (2015). YAP promotes malignant progression of Lkb1-deficient lung adenocarcinoma through downstream regulation of survivin. *Cancer Res.* 75, 4450–4457. doi: 10.1158/0008-5472.CAN-14-3396

Zhang, W., Nandakumar, N., Shi, Y., Manzano, M., Smith, A., Graham, G., et al. (2014). Downstream of mutant KRAS, the transcription regulator YAP is essential for neoplastic progression to pancreatic ductal adenocarcinoma. *Sci. Signal.* 7:ra42. doi: 10.1126/scisignal.2005049

Zheng, Y. B., Huang, J. W., Zhan, M. X., Zhao, W., Liu, B., He, X., et al. (2014). Genetic variants in the KDR gene is associated with the prognosis of transarterial chemoembolization treated hepatocellular carcinoma. *Tumour Biol.* 35, 11473–11481. doi: 10.1007/s13277-014-2478-8

Zhou, Q., Lui, V. W., and Yeo, W. (2011). Targeting the PI3K/Akt/mTOR pathway in hepatocellular carcinoma. *Future Oncol.* 7, 1149–1167. doi: 10.2217/fon.11.95

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 September 2021
doi: 10.3389/fgene.2021.737965

Identification of Mutation Landscape and Immune Cell Component for Liver Hepatocellular Carcinoma Highlights Potential Therapeutic Targets and Prognostic Markers

*Hengzhen Wang, Wenjing Jiang, Haijun Wang, Zheng Wei, Hali Li, Haichao Yan and Peng Han**

Department of General Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China

Liver hepatocellular carcinoma (LIHC) is a primary malignancy, and there is a lack of effective treatment for advanced patients. Although numerous studies exist to reveal the carcinogenic mechanism of LIHC, few studies have integrated multi-omics data to systematically analyze pathogenesis and reveal potential therapeutic targets. Here, we integrated genomic variation data and RNA-seq profiles obtained by high-throughput sequencing to define high- and low-genomic instability samples. The mutational landscape was reported, and the advanced patients of LIHC were characterized by high-genomic instability. We found that the tumor microenvironment underwent metabolic reprograming driven by mutations accumulate to satisfy tumor proliferation and invasion. Further, the co-expression network identifies three mutant long non-coding RNAs as potential therapeutic targets, which can promote tumor progression by participating in specific carcinogenic mechanisms. Then, five potential prognostic markers (*RP11-502I4.3, SPINK5, CHRM3, SLC5A12,* and *RP11-467L13.7*) were identified by examining the association of genes and patient survival. By characterizing the immune landscape of LIHC, loss of immunogenicity was revealed as a key factor of immune checkpoint suppression. Macrophages were found to be significantly associated with patient risk scores, and high levels of macrophages accelerated patient mortality. In summary, the mutation-driven mechanism and immune landscape of LIHC revealed by this study will serve precision medicine.

Keywords: liver hepatocellular carcinoma, somatic mutation, RNA-sequencing, genome variation, precision medicine

OPEN ACCESS

Edited by:
Xinyi Liu,
University of Illinois at Chicago, United States

Reviewed by:
Yuanyuan Zhang,
Peking-Tsinghua Center for Life Sciences, China
Fulong Yu,
Broad Institute, United States

***Correspondence:**
Peng Han
hanpeng88@126.com

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

Received: *08 July 2021*
Accepted: *12 August 2021*
Published: *16 September 2021*

Citation:
Wang H, Jiang W, Wang H, Wei Z, Li H, Yan H and Han P (2021) Identification of Mutation Landscape and Immune Cell Component for Liver Hepatocellular Carcinoma Highlights Potential Therapeutic Targets and Prognostic Markers. Front. Genet. 12:737965. doi: 10.3389/fgene.2021.737965

INTRODUCTION

Liver hepatocellular carcinoma (LIHC) is the most common primary malignancy of the liver and the third leading cause of cancer-related death worldwide (Bosch et al., 1999; Bray et al., 2018). Of these, liver cancer is the second leading cause of cancer-related death in LIHC, accounting for approximately 90% of all primary liver cancer cases

(Llovet et al., 2016). Studies have found that fat accumulation of liver can lead to non-alcoholic steatohepatitis, cirrhosis, liver failure, and LIHC (Kim et al., 2021). Treatments for LIHC include hepatectomy, liver transplant, chemotherapy, and molecular targeted therapy. However, clinical treatment results show that these treatments are not effective for LIHC patients (Heimbach et al., 2018). Therefore, there is an urgent need for the identification of new therapeutic targets for the development of new drugs.

Somatic variations, including copy number variations (CNVs) and point mutations, are considered to be the driving event for the occurrence and development of cancer. In recent years, researchers mainly focused on key mutated genes and their mutational characteristics (Zhang et al., 2021). However, the integration of mutagenomics with other omics data is more powerful in revealing the pathogenesis of patients and potential therapeutic targets (Fujimoto et al., 2016). With the development of next-generation sequencing, multiple somatic variations have been discovered. Especially, accumulated studies have demonstrated that somatic variations, such as single-nucleotide variations and CNVs, could contribute to tumorigenesis (Wang et al., 2020) and used to infer individual medications based on the RNA interaction network (Zhang et al., 2018). Based on the notion that the instability of the genome is related to age (Chatsirisupachai et al., 2021), it is crucial to investigate the relationship between the stability of the genome and the physiological mechanism of the patient. More recently, large-scale biomedical data, including multidimensional molecular profiles of tumor samples of LIHC generated by The Cancer Genome Atlas (TCGA; Tomczak et al., 2015) project, provide opportunities to uncover mutation-driven potential therapeutic targets and potential prognostic markers for liver cancer.

Over the past decade, the immune microenvironment has been a popular area of cancer biology research in relation to therapeutic targets. The immune microenvironment is composed of a variety of lymphocytes, such as T cells, B cells, and macrophages. Previous studies have shown that the composition of immune cells is closely related to tumor proliferation and metastasis. For example, CD8+ T cells show strong cytotoxic activity on tumor cells and have a strong inhibitory effect on tumor progression (Seo et al., 2018). Macrophage polarization plays a key role in subverting adaptive immunity and promoting tumor progression (Mantovani et al., 2002). The development of the immune cell fraction algorithm (Newman et al., 2015) for bulk RNA-seq data provides convenience for investigating the relationship between specific immune cell content and tumor progression.

In the current study, we integrated and analyzed the somatic mutations, CNVs data, and RNA-seq of LIHC collected from the TCGA database. The mutation landscape of LIHC and the metabolic features driven by mutations were revealed. Our work highlights potential therapeutic targets, potential prognostic markers, and the role of macrophages in tumor progression. These results promote the understanding of pathogenesis and provide a basis for the treatment of LIHC.

MATERIALS AND METHODS

Data Collection

The CNV data, somatic mutation data, clinical information, and RNA-seq profiles of LIHC collected by TCGA (Tomczak et al., 2015) were downloaded from UCSC Xena browser.[1] Metabolic pathway and hallmark gene sets that will be used for metabolic feature analysis and enrichment analysis of carcinogenic functions for LIHC were collected from the Molecular Signatures Database (Liberzon et al., 2015).[2] Moreover, the annotation data of GRCh38 v29 for long-noncoding RNA (lncRNA) were collected from GENCODE (Frankish et al., 2019).[3] The signature matrix of 22 immune cell types was collected from the previous studies (Newman et al., 2015) for the analysis of immune cell invasion of tumor samples.

Processing of Mutation Data

We first counted the distribution of mutation sites on the human genome, including mRNA, lncRNA, and transcription start site, as well as the distribution of various types of mutation, including missense and nonsense mutation on the chromosome. Further, the R package maftools (version 2.8.0; Mayakonda et al., 2018) was used for the statistical and visualization of mutation form, mutation frequency, and mutational correlation between genes, which provides great convenience for the research of mutation data and the reveal of characteristics. The number of mutations in each tumor sample was calculated and used to link the CNV data. We downloaded the GDC GISTIC copy number dataset from the UCSC Xena browser, which is derived from focal copy number estimates, and the positions of the variant sequence corresponding to the genes. Both gene amplification and deletion events are thought to increase genome instability. By integrating the mutation information and gene copy number information of patient cohort, we defined the top 20% of patients with copy number amplitude and mutation load as high-genomic instability group, the bottom 20% of patients with copy number amplitude and mutation load as low-genomic instability group, and the remaining patients as median/unknown-genomic instability group.

Gene Set Enrichment Analysis

Considering that there were multiple zero values in the gene expression matrix, we control the number of genes by requiring effective genes to be expressed in at least 10% of tumor samples. Based on the previously defined high/low-genomic instability samples, the rank sum test was used to identify genes that are significantly differentially expressed in the high/low-genomic instability samples. The cutoff of value of p is set to 0.01. For these significantly differentially expressed genes (DEGs), the genes were sorted using the logarithmic fold change as the weight and combined with the hallmark gene set to be used for gene set enrichment analysis (GSEA; Subramanian et al., 2005)

[1]https://xenabrowser.net/
[2]http://software.broadinstitute.org/gsea/msigdb
[3]https://www.gencodegenes.org/

by R package fgsea (version 1.1.0). We set the value of p to <0.05 to screen out carcinogenic functions that are significantly enriched on DEGs.

Calculation of Metabolic Pathway Activity

Gene set variation analysis (GSVA; Hanzelmann et al., 2013), which is an unsupervised manner to estimate changes in pathway activity over a sample population, was used to calculate the metabolic activity of each tumor sample by R package GSVA (version 1.32.0). We set the number of genes in the gene set used for functional enrichment to be at least 10 and not more than 500. The rank sum test and fold change algorithm were also used to calculate the variation of metabolic pathway activity between high and low-genomic instability samples. Metabolic pathways with a value of $p < 0.01$ were considered to be affected by mutations, and reprogramming has occurred.

Construction of Co-expression Network Mediated by Mutant lncRNA

We extracted lncRNA from DEGs which differentially expressed between high- and low-genomic instability samples based on lncRNA annotation data obtained from GENCODE. By combining somatic mutation and CNV data, we identified lncRNAs that were mutated in tumor samples and differentially expressed in the high-genomic instability group, defined as mutation-driven lncRNA (Md-lncRNA). Next, the Pearson correlation algorithm (Bishara and Hittner, 2012) is used to calculate the correlation between Md-lncRNAs and other DEGs, which was performed by cor.test function of R. We have defined that gene pairs with value of $p < 0.01$ and $|\ R\ | > 0.3$ have significant correlation in expression and are co-expressed with each other (van Dam et al., 2018). For these co-expressed genes, cytoscape (Shannon et al., 2003) was used to plot the co-expression network, and Network Analyzer tool was used to calculate the topological properties of the network.

Identification of Potential Prognostic Markers

The genes in the co-expression network mediated by Md-lncRNAs were used as candidate markers. We first used univariate COX regression and lasso regression (Alhamzawi and Ali, 2018) to screen genes that significantly associated with overall survival (OS) of LIHC patients (the cutoff of value of p was 0.05). Next, we randomly selected 60% of all samples as the training set and the remaining as the test set. The training set was used to construct a multivariate COX regression model (Fisher and Lin, 1999). We retained the genes passing the test of multivariate COX regression as potential prognostic markers and establish nomogram to predict the OS of LIHC. The reliability of the prediction model was validated by the receiver operating working characteristic curve (ROC), and the area under curve (AUC) also was calculated. The calibration curve was used to evaluate the predictive power of nomograph for survival risk.

Survival Analysis

The risk score for each patient was calculated according to the linear combination of expression values weighted by the coefficient from the multivariate Cox regression analysis:

$$Risk\ score(i) = \sum_{k=1}^{n} \beta_k * e_{ki}$$

where n denotes the number of prognosis markers ($n = 5$), β was the coefficient of multivariate Cox regression analysis, and e_{ki} was the expression level of kth prognosis-related gene expression of patient i. Further, the samples of train set and test set were, respectively, divided into high- and low-risk categories based on the median risk score calculated by risk score model, and Kaplan–Meier algorithm (Ranstam and Cook, 2017) was used to compare whether the survival data of the two categories are different and bilateral log-rank test was used to validate the significance of the difference.

Calculation of Immune Cell Fraction

Based on the feature matrix of 22 immune cells obtained from previous studies, the CIBERSORTx tool[4] (Newman et al., 2015, 2019) was used to analyze tumor-infiltrating immune cells. CIBERSORTx is a method to characterize the cell composition of complex tissues from the gene expression profile. The parameter perms that the number of permutations when calculating the value of p was set to 1,000, and QN was set to TRUE to perform quantile normalization. In order to see more group differences in other cell types other than plasma cells, we further transformed the original cell components into a log ratio of log (the fraction of plasma-cell +1e-3)/log (the fraction of immune-cell +1e-3) (He et al., 2021).

Statistical Analysis

All statistical analyses and graph generation were performed in R (version 4.0.2) and GEPIA (version 2.0).[5]

RESULT

A Global View of Mutations for Liver Hepatocellular Carcinoma

Malignant mutations in the genome are the underlying cause of tumor development and progression. The identification of mutation characteristics is essential for the exploration of pathogenesis. We have first used maftool to evaluate mutation profiles of LIHC in the TCGA database collection for which somatic mutation data were available. A total of 44,847 somatic mutation sites in 375 samples were included in this study. We counted the distribution of somatic mutations on the genome and found that somatic mutations are significantly enriched in specific regions of chromosomes 1, 11, 17, and 19 (**Figure 1A**), indicating that the global mutations of LIHC have preference for location. Compared with transcripts

[4]https://cibersortx.stanford.edu/
[5]http://gepia2.cancer-pku.cn/#index

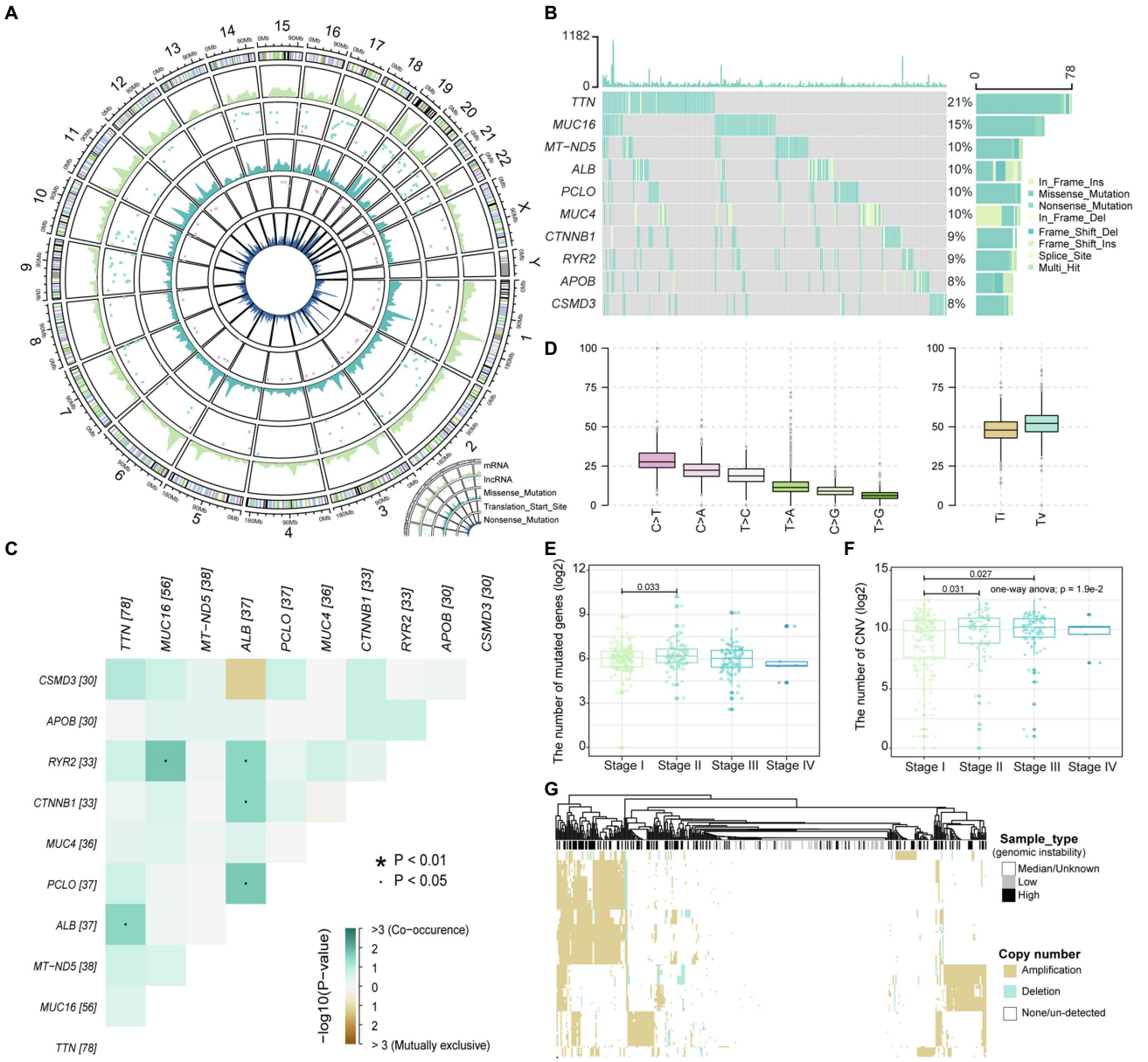


FIGURE 1 | The landscape of liver hepatocellular carcinoma (LIHC) somatic variations. **(A)** The density distribution of somatic mutations on chromosomes. The four-layer circle plot shows the density distribution of nonsense mutations, mutations of transcription start site, missense mutations, mutations of long-noncoding RNA (lncRNA), and mutations of mRNA on chromosomes from inside to outside. **(B)** The waterfall plot shows the top 10 genes in terms of mutation frequency by sample. The mutation type of each gene in each sample is marked. **(C)** Mutation correlation heatmap of the top 10 high-frequency mutated genes. Locations with significant correlations are marked by stars. **(D)** Boxplot shows the frequency of base substitutions including transversion and transition. **(E,F)** The relationship between the number of mutated and copy number variation (CNV) genes in each sample and the stage are displayed with boxplot. The number of mutated and CNV genes is logarithmized, and the rank sum test is used to assess differences between groups. **(G)** The copy number amplitudes of tumor samples are presented in heat map. Column labels show sample types, including high/low-genomic instability and median.

(mRNA) of protein-coding genes, fewer somatic mutations occurred in lncRNAs; **Figure 1A**), indicating that somatic mutations were more likely to directly affect the expression of protein-coding genes and the structure of proteins. However, few mutations in non-coding genes were still the main determinants of human diseases (Maurano et al., 2012). Mutations in the transcription start site will regulate gene expression levels before transcription, which rarely occur on autosomes 4 and 13 in LIHC. Point mutations, including missense and nonsense mutations, are an important part of somatic variations, and LIHC shows the dominant position of missense mutations (**Figure 1A** and **Supplementary Figures S1A,B**). Further, we counted the frequency of mutations in each gene, and the top 10 mutated

genes were identified (**Figure 1B**). *TTN*, the gene considered to be most frequently mutated in the pan-cancer cohort (Oh et al., 2020), tended to have missense mutations in LIHC. The content of albumin encoded by *ALB* has been confirmed to be closely related to tumor development and patient prognosis (Li et al., 2018). We found that there was a significant mutational correlation between the genes *TNN* and *ALB* (**Figure 1C**), which indicates that *TNN* and *ALB* may play a synergistic role in LIHC. We found that almost a quarter of point mutations in LIHC patients were C > T substitutions (**Figure 1D; Supplementary Figure S1C**). Transitions and transversions, as the two types of DNA base transformations, account for similar proportions in the entire LIHC point mutation (**Figure 1D**). Mutations of transversions, which account for a relatively high proportion, may be a key factor in liver tissue degradation. By combining the mutation with the patient's clinical information, we found that patients of stage II have a higher number of mutated genes compared to stage I (**Figure 1E**), which indicates that the accumulation of mutations appears as the stage increases. We introduced copy number data of LIHC patients, further confirming that advanced patients have a higher accumulation of variation and genomic instability (**Figure 1F**). Next, we defined high and low-genomic instability samples by integrating somatic mutation and copy number data. We found that the high-genomic instability samples in LIHC have overall gene amplification (**Figure 1G**). Taken together, all these revealed the mutational features of LIHC.

Metabolic Reprogramming Affected by Accumulation of Mutations

Genome variation can indirectly affect the metabolic efficiency of organisms by regulating gene expression. The rank sum test was used to identify genes that are significantly DEGs between high and low-genomic instability samples. We identified 6,438 DEGs (value of $p < 0.01$), including 2,981 upregulated genes and 3,457 downregulated genes (**Figure 2A**). After GSEA, we identified four carcinogenic functional pathways that are significantly enriched in DEGs (value of $p < 0.05$). We found that the E2F pathway, which forms with CDK-RB driving cell cycle progression (Kent and Leone, 2019), is significantly enriched in upregulated DEGs (**Figure 2B**), indicating that the cell cycle is severely affected by the accumulation of mutations. The G2/M checkpoint can effectively detect the genome and prevent cells from entering mitosis (Anand et al., 2020), which dysfunction may be a key factor in the accumulation of mutations in high-genomic instability samples. We found that the inflammatory response was significantly enriched in the downregulated DEGs (**Figure 2B**), which may be due to the accumulation of mutations that caused the weakening or loss of tumor tissue immunogenicity (Capietto et al., 2020). All these indicate that the resistance of some patients with advanced liver cancer to immune targeted therapy (Zongyi and Xiaowu, 2020) may be due to the loss of immunogenicity caused by the excessive accumulation of mutations.

Metabolic reprograming affected by mutations was the basis for satisfying tumor proliferation and invasion. Gene set variation analysis (GSVA) was used to evaluate the metabolic activity of each tumor sample. By clustering the metabolic pathway activity score matrix, we found that there are obvious differences in metabolic function between the high and low-genomic instability samples (**Figure 2C**). Compared with low-genomic instability samples, high-genomic instability samples had higher pyrimidine synthesis activity (**Figures 2D,E**). Previous studies have shown that inhibiting the metabolic activity of pyrimidine synthesis can effectively reduce the carcinogenic ability of tumors (Wang et al., 2019), which indicates that pyrimidine driver mutations that trigger pyrimidine anabolic remodeling can be used as therapeutic targets for patients with advanced liver cancer. We found that the activity of glycosylphosphatidylinositol (GPI)-anchor biosynthesis pathway is also upregulated in high-genomic instability samples (**Figure 2F**). The enhancement of GPI-anchor biosynthesis pathway activity could recruit macrophages to tumor tissues to generate TAM polarization (Dangaj et al., 2011), suggesting that the high tumor invasion and metastasis ability shown by high-genomic instability samples may be caused by the upregulation of GPI-anchored protein. All these indicate that the reprogramming of metabolic pathways provides the necessary preparations for tumor proliferation and invasion and is also the basis for tumor heterogeneity.

Mutated LncRNA Stimulates Tumor Progression

LncRNA has become an important participant in almost every level of gene function and regulation (Qian et al., 2019; Wang et al., 2021). It is intriguing to identify the driver mutation lncRNA between high- and low-genomic instability samples. We extracted lncRNAs that were significantly differentially expressed between high- and low-genomic instability samples based on lncRNA annotation data, and combined CNV and somatic mutation data to identify three Md-lncRNAs (**Figure 3A**). We found that samples with Md-lncRNA *AL589743.1* copy number amplification clustered in highly mutant samples. Next, the Pearson correlation algorithm was used to identify DEGs that are significantly related to these three Md-lncRNAs at the gene expression level. We found that 412 DEGs (value of $p < 0.01$ and correlation coefficient $|R| > 0.3$) are involved in the regulatory network co-expressed with these three Md-lncRNAs (**Figure 3B**). To identify the role of these three mutation-driven lncRNAs in the carcinogenic mechanism of LIHC, gene ontology (GO) was used to perform functional enrichment analysis on DEGs that are significantly related to these three mutation-driven lncRNAs. We found that DEGs co-expressed with Md-lncRNA *AC037459.4* are mainly involved in the fat metabolism process of liver tissue (**Figure 3C**). The abnormal fat metabolism was the key cause of fatty liver, liver cirrhosis, and even liver cancer (Alves-Bezerra and Cohen, 2017). DEGs significantly related to lncRNA *AL589743.1* were enriched in protein processing and modification functional nodes (**Figure 3D**), suggesting that *AL589743.1* is involved in

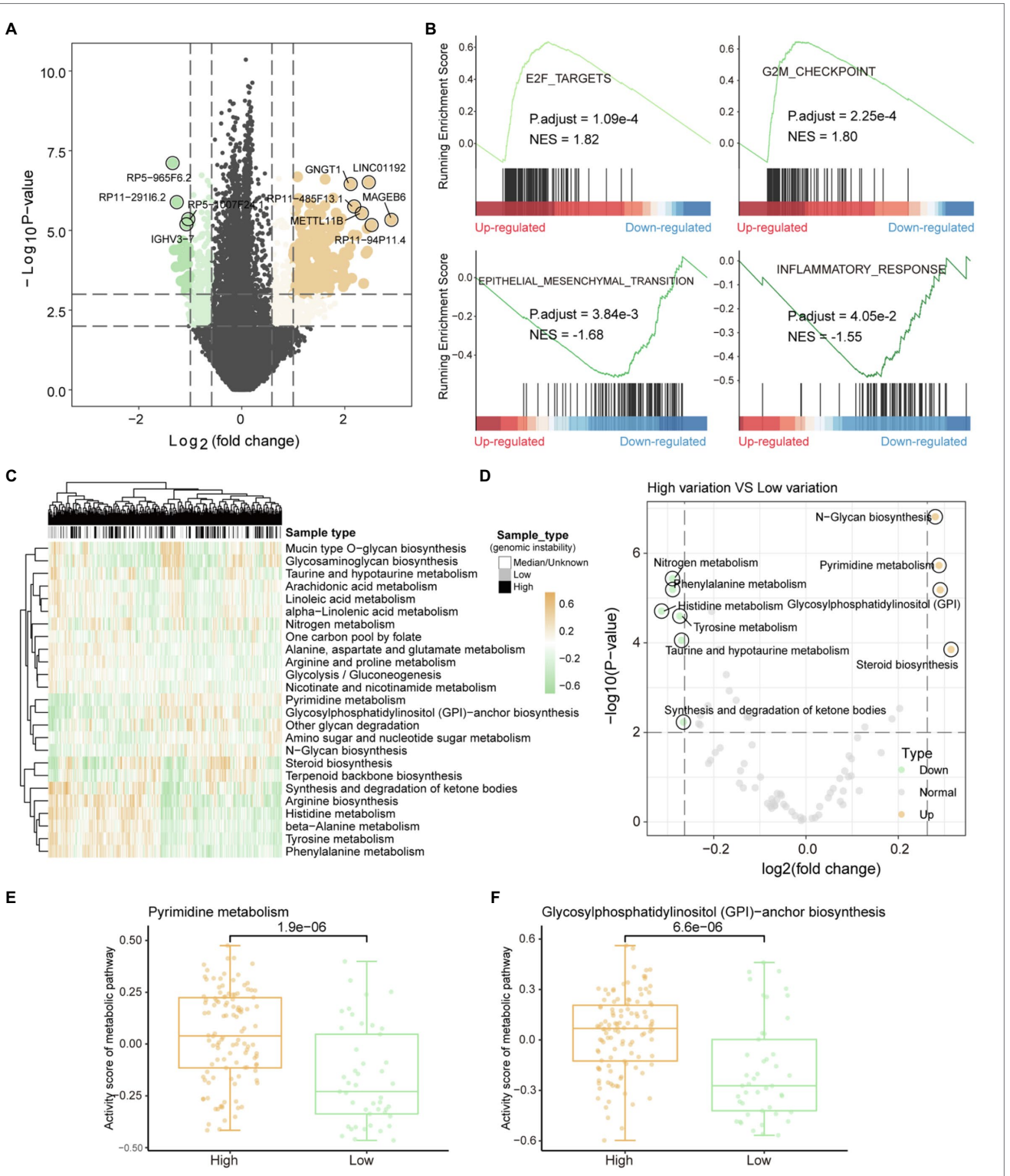


FIGURE 2 | Metabolic remodeling based on genome instability. **(A)** The results of the differential gene expression analysis between high-genomic instability and low-genomic instability samples are shown by the volcano graph. Grey dots represent non-differentially expressed genes, yellow dots indicate genes upregulated in high-genomic instability samples, and the green dots mean the opposite. **(B)** gene set enrichment analysis results of differentially expressed genes. Normalized enrichment score and corrected value of *p* are calculated. **(C)** The enrichment scores of tumor samples in each metabolic pathway calculated by GSVA are displayed by heat map. Column labels show sample types, including high-/low-genomic instability and median. **(D)** Analysis of the difference of metabolic pathway activity scores between high-genomic instability and low-genomic instability samples. **(E,F)** Comparison of pyrimidine synthesis and glycosylphosphatidylinositol-anchor biosynthesis pathway activity among high-/low-genomic instability samples. The rank sum test is used to calculate the significance.

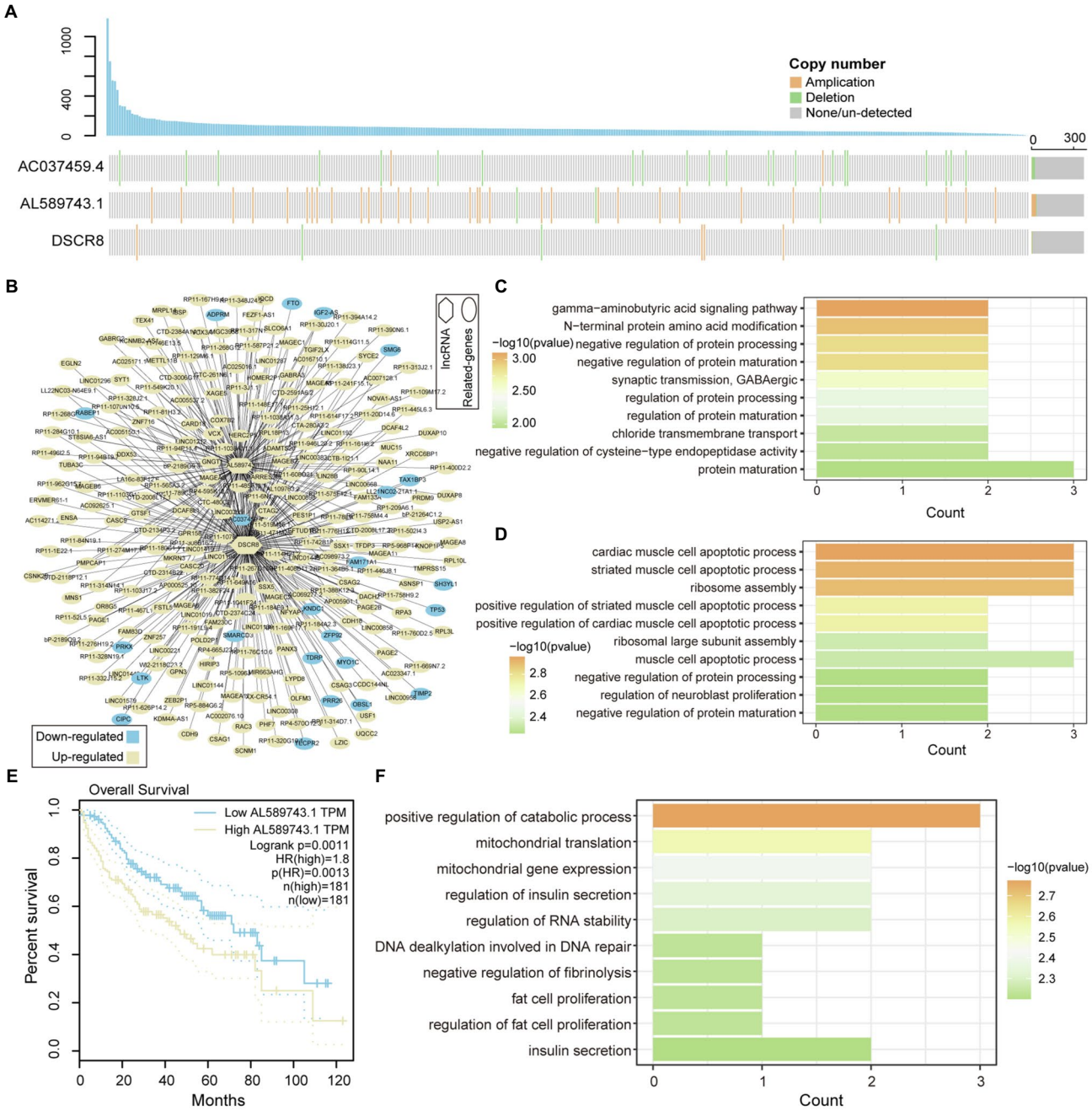


FIGURE 3 | Functional identification of mutation-driven lncRNA. **(A)** Waterfall plot illustrates variation types of copy number (amplification, deletion, and none/un-detected) for each sample on lncRNA, including AC037459.4, AL589743.1, and DSCR8. The up-panel shows the number of mutated genes in each sample. **(B)** The relationship between mutation-driven lncRNA and co-expressed genes is shown by network. Circles represent co-expressed genes with lncRNAs; squares represented mutation-driven lncRNAs. Upregulated genes are marked with yellow; downregulated genes are marked with blue. **(C,D)** The bar graphs show the GO function enrichment results for genes co-expressed with lncRNA AC037459.4 and AL589743.1. **(E,F)** Survival difference between the two groups of samples with high and low expression of AL589743.1 and DSCR8. Univariate cox regression algorithm and log-rank test are used to evaluate the relationship between the expression of Md-lncRNA and patient survival.

carcinogenic mechanisms by regulating the structure and function of proteins. We also found that the high expression of *AL589743.1* was significantly associated with poor patient's prognosis (**Figure 3E**), indicating that *AL589743.1* can be used as an important target for the treatment of patients with advanced liver cancer. Further, DEGs co-expressed with Md-lncRNA *DSCR8* are mainly involved in protein processing and muscle cell apoptosis (**Figure 3F**). In previous studies, it has been

confirmed that *DSCR8* can act as a miRNA sponge to activate the Wnt/β-catenin signaling pathway and promote the progress of LIHC (Wang et al., 2018). Taken together, all these results reveal that three Md-lncRNAs to promote tumor progression by participating in specific carcinogenic mechanisms.

Prognostic Markers Correlated to LIHC

LncRNA and transcripts co-expressed with it play an important role in the carcinogenic mechanism, which can be used as candidate prognostic markers. To identify prognostic markers of LICH, we first performed univariate cox regression and lasso regression algorithm to identify genes associated with OS in LIHC patients (see method). Then, 20 genes were identified and significantly correlated with the patient's OS of LIHC (**Figure 4A**). Through the multivariate Cox regression constructed by the 20 genes and training set, five of which, *RP11-502I4.3, SPINK5, CHRM3, SLC5A12,* and *RP11-467L13.7*, were identified as prognostic markers for LIHC (**Figure 4B**; **Supplementary Figure S2**). To evaluate the predictive performance of the model, we showed the prediction results using ROC for five time points. We found that the risk prediction result reached the maximum AUC value of 0.72 (**Figure 4C**). Further, the nomograms algorithm was used to build a survival risk prediction model for LIHC (**Supplementary Figure S3**). The calibration curve was also used to validate the stability of the risk prediction model (**Figure 4D**). Moreover, the risk scoring model was constructed as follows: risk score = −0.37**RP11-502I4.3*–0.11**SPINK5*–0.16**CHRM3* + 0.06**SLC5A12* + 0.42**RP11-467L13.7*. The samples of train set and test set were, respectively, divided into high- and low-risk groups based on the median risk score. We found that high-risk samples in train set are associated with poor prognosis of LIHC patients (**Figure 4E**). The test set also showed the same prediction results as the train set (**Figure 4F**), indicating the reliability of the risk score model in predicting the prognostic risk of patients. Taken together, we have identified five potential prognostic markers in LIHC, which can be used for clinical diagnosis.

Tumor Progression Regulated by the Immune Microenvironment

The tumor immune microenvironment plays an important role in the occurrence and development of tumors (Lei et al., 2020). The remodeling of the immune microenvironment is conducive to the progress of the tumor (Hinshaw and Shevde, 2019). Therefore, we used the CIBERSORTx tool to calculate the immune cell abundance of each LIHC sample and paracancerous tissue sample through the deconvolution algorithm that is a special kind of forward convolution, where the size of the input image is first enlarged by complementing the 0 at a certain scale, followed by rotating the convolution kernel and then forward convolution. For the 22 immune cell fraction matrices obtained, the consensus clustering algorithm was used to identify the immune subtypes of LIHC. We have defined four reliable tumor immune subtypes (**Figure 5A** and **Supplementary Figure S4**), which have specific immune cell composition. We found that the normal samples are mainly clustered in the third cluster, which has a relatively low content of CD8+ T cell and CD4+ T cell (**Figure 5B**). Multiple tumor samples have similar immune cell composition to normal samples in the third cluster, indicating that these samples are in immunosuppressed state. Different from other clusters, the fourth cluster of tumor samples has a higher content of CD8+ T cells (**Figure 5B**), suggesting that this type of LIHC patients is more suitable for immuno-targeted therapy. In order to explore the formation mechanism of tumor immunosuppressive microenvironment, we calculated the content of major histocompatibility complex (MHC). We found that genes involved in the synthesis of *MHC-I* have lower expression levels in the third cluster and significantly higher expression in the fourth cluster (**Figure 5C**), indicating that the immunosuppression of the third cluster may be caused by the loss of tumor immunogenicity. The *MHC-II* molecule, which is the CD4+ T-cell binding partner (Marty Pyke et al., 2018), also had lower expression level in the third cluster (**Figure 5D**). Next, by linking the immune cell fraction and risk score of each sample, we found that the fraction of Macrophages M0 is significantly related to the patient's prognostic risk (**Figure 5E**). Tumor samples were divided into two categories (high/low fraction) based on the median of macrophages M0 fraction; we found that high-fraction samples are associated with poor patient's prognosis (**Figure 5F**), suggesting that macrophages cells can promote tumor progression in the tumor microenvironment. Taken together, all these indicate that the loss of immunogenicity is a key factor for the formation of immunosuppressive microenvironment in multiple patients of LIHC.

DISCUSSION

In this study, we have integrated multi-omics data to reveal mutation-driven pathogenesis and immune landscape of LIHC. Through the statistics of the mutation location and type, we found the mutation characteristics of LIHC and defined two types of samples (high/low-genomic instability). We found that the inflammatory response was significantly enriched in the downregulated genes of the high-genomic instability samples by GSEA. Metabolic pathway activity analysis has shown that pyrimidine synthesis and GPI-anchor biosynthesis pathway are closely related to tumor progression and have low activity scores in high-genomic instability samples. We identified three mutations driving lncRNA and defined the molecular functions of these three mutations driving lncRNA in LIHC by constructing a co-expression network. Further, based on the genes involved in the co-expression network, we identified four prognostic markers, including *RP11-502I4.3, SPINK5, CHRM3, SLC5A12,* and *RP11-467L13.7*, through univariate cox regression and lasso algorithm screening. We also built risk score model to assess the prognostic risk of LIHC patients. Through the analysis of the immune cell fraction of tumor

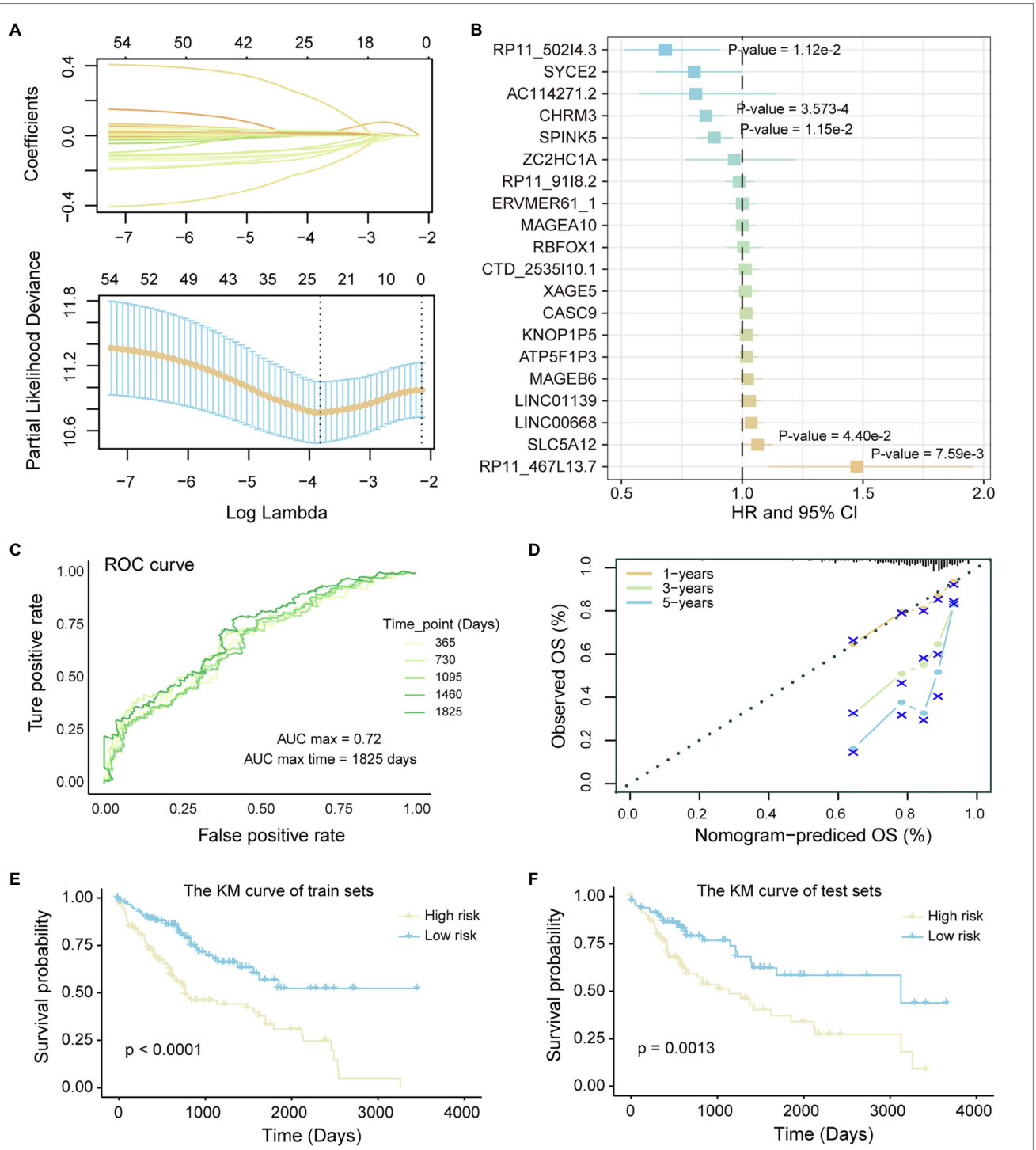


FIGURE 4 | Identification of potential prognostic markers of LIHC. **(A)** Lasso regression model screen genes related to overall survival (OS) of LIHC patients. Variation curve of regression coefficient and λ value is shown. **(B)** COX risk regression to assess the association between the expression level of genes and patient survival. Genes that are significantly related to patient survival are added value of *p*. **(C)** The ROC curve reflects the predictive power of the risk regression model at five time points from 1 to 5 years. The different colored curves represent specific time points. **(D)** Calibration curve of nomogram. **(E,F)** Kaplan–Meier (KM) curves for survival of train set and test set in high- and low-risk groups. Log-rank test was used to calculate statistical significance.

and paracancerous tissue samples, we defined four immune subtypes and found that the samples of immunosuppressive subtypes have low immunogenicity.

LIHC is a primary malignancy of the liver (Huang et al., 2016). Numerous of studies have tried to reveal the pathogenesis of LIHC and find effective treatments. For example, studies have

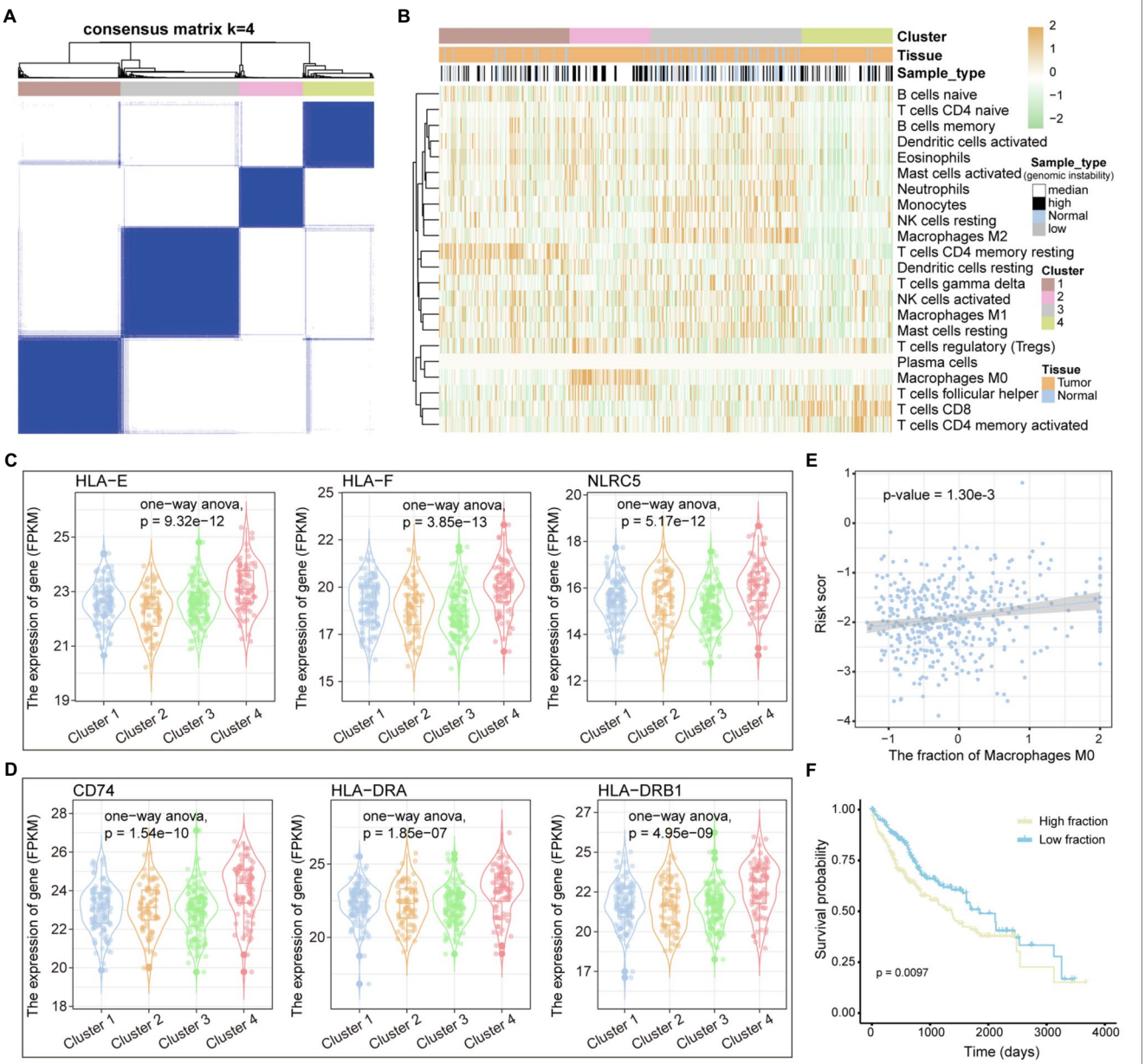


FIGURE 5 | Immune cell components of LIHC patients. **(A)** The samples are divided into several clusters based on the immune cell components of each sample. The consistency matrix is drawn as a heat map, and the column labels show the clusters. **(B)** The 22 immune cell fractions of each sample were displayed by heat map. Column labels, including clusters, tissue origin, and type of variation of samples, are displayed. **(C,D)** The relationship between genes encoding MHC I and MHC II molecules and clusters defined by immune cell components is shown by boxplot. ANOVA is used to calculate statistical significance. **(E)** The correlation between the fraction of macrophages M0 and the risk score is shown. **(F)** KM curves for survival in high and low fraction groups of macrophages M0. Log-rank test is used to calculate statistical significance.

shown that fibrosis of liver cells plays a vital role in the pathogenesis of liver cirrhosis and hepatocellular carcinoma (Liu et al., 2020). *TXNIP* activates the expression of oncogenes to inhibit the proliferation of hepatocellular carcinoma cells and induces apoptosis (Liu et al., 2017). In the last decade, the immune microenvironment of tumor has been a popular area of cancer biology research in relation to therapeutic targets for drug discovery. Although checkpoint inhibitors have been successfully used in cancer treatment, they are only effective in 10–40% of cases (Hamid et al., 2013; Callahan et al., 2014). Previous study has shown that checkpoint inhibitors do not trigger cancer-specific T-cell responses in some patients (Sharma and Allison, 2015). Therefore, it is necessary to reveal the relationship between the immune microenvironment of LIHC and tumor progression and the relationship between immune cells, which can be used to guide the combination medication of liver cancer patients.

Recent reports from developed countries indicate that metabolic disorders caused by diabetes, obesity, and fatty liver are risk factors for LIHC (Makarova-Rusher et al., 2016). Besides, the experimentally confirmed carcinogenic and regulatory mechanisms of lncRNA have been widely revealed (Wang et al., 2019). Genes related to lncRNA *AC037459.4* were identified involved in the fat metabolism process of liver tissue, suggesting that *AC037459.4* may mediate dysregulation of fat metabolism pathways in patients. Based on previous research on the identification of cancer prognostic markers (Yu et al., 2019), we identified five potential prognostic markers by multivariate Cox regression analysis, which can be used in the clinical diagnosis of patients and guiding their treatment. The subtype of LIHC with strong immunogenicity suggests that immune checkpoint inhibitor may have a better effect on these patients. The fraction of macrophages in tumor tissue was found to be significantly associated with the risk of death in patients, consistent with previous studies demonstrating the involvement of macrophages in tumor invasion and metastasis (Chen et al., 2020).

In conclusion, this study provided a mutation-driven metabolic landscape and immune landscape of LIHC. Three mutated lncRNAs were identified to drive transcriptional perturbed oncogenic pathways and affect patient prognosis. Five gene signatures associated with patient prognosis were identified through Cox regression and lasso regression. We also identified four immune subtypes for LIHC. In conclusion, all these findings provide theoretical guidance for the optimization of LIHC treatment strategies.

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

HeW, WJ, HaW, and PH conceived and designed the experiments. ZW, HL, and HY analyzed the data. HeW and WJ collected the data. HeW, WJ, and HaW validated the method and data. HeW and PH wrote this manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [81803010].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.737965/full#supplementary-material

REFERENCES

Alhamzawi, R., and Ali, H. T. M. (2018). The Bayesian adaptive lasso regression. *Math. Biosci.* 303, 75–82. doi: 10.1016/j.mbs.2018.06.004

Alves-Bezerra, M., and Cohen, D. E. (2017). Triglyceride metabolism in the liver. *Compr. Physiol.* 8, 1–8. doi: 10.1002/cphy.c170012

Anand, S. K., Sharma, A., Singh, N., and Kakkar, P. (2020). Entrenching role of cell cycle checkpoints and autophagy for maintenance of genomic integrity. *DNA Repair (Amst)* 86:102748. doi: 10.1016/j.dnarep.2019.102748

Bishara, A. J., and Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, spearman, transformation, and resampling approaches. *Psychol. Methods* 17, 399–417. doi: 10.1037/a0028087

Bosch, F. X., Ribes, J., and Borras, J. (1999). Epidemiology of primary liver cancer. *Semin. Liver Dis.* 19, 271–285. doi: 10.1055/s-2007-1007117

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

Callahan, M. K., Postow, M. A., and Wolchok, J. D. (2014). CTLA-4 and PD-1 pathway blockade: combinations in the clinic. *Front. Oncol.* 4:385. doi: 10.3389/fonc.2014.00385

Capietto, A. H., Jhunjhunwala, S., Pollock, S. B., Lupardus, P., Wong, J., Hansch, L., et al. (2020). Mutation position is an important determinant for predicting cancer neoantigens. *J. Exp. Med.* 217:e20190179. doi: 10.1084/jem.20190179

Chatsirisupachai, K., Lesluyes, T., Paraoan, L., Van Loo, P., and de Magalhaes, J. P. (2021). An integrative analysis of the age-associated multi-omic landscape across cancers. *Nat. Commun.* 12:2345. doi: 10.1038/s41467-021-22560-y

Chen, Z., Zhou, L., Liu, L., Hou, Y., Xiong, M., Yang, Y., et al. (2020). Single-cell RNA sequencing highlights the role of inflammatory cancer-associated fibroblasts in bladder urothelial carcinoma. *Nat. Commun.* 11:5077. doi: 10.1038/s41467-020-18916-5

Dangaj, D., Abbott, K. L., Mookerjee, A., Zhao, A., Kirby, P. S., Sandaltzopoulos, R., et al. (2011). Mannose receptor (MR) engagement by mesothelin GPI anchor polarizes tumor-associated macrophages and is blocked by anti-MR human recombinant antibody. *PLoS One* 6:e28386. doi: 10.1371/journal.pone.0028386

Fisher, L. D., and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annu. Rev. Public Health* 20, 145–157. doi: 10.1146/annurev.publhealth.20.1.145

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773. doi: 10.1093/nar/gky955

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., et al. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* 48, 500–509. doi: 10.1038/ng.3547

Hamid, O., Robert, C., Daud, A., Hodi, F. S., Hwu, W. J., Kefford, R., et al. (2013). Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* 369, 134–144. doi: 10.1056/NEJMoa1305133

Hanzelmann, 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

He, L., Liu, L., Li, T., Zhuang, D., Dai, J., Wang, B., et al. (2021). Exploring the imbalance of periodontitis immune system From the cellular to molecular level. *Front. Genet.* 12:653209. doi: 10.3389/fgene.2021.653209

Heimbach, J. K., Kulik, L. M., Finn, R. S., Sirlin, C. B., Abecassis, M. M., Roberts, L. R., et al. (2018). AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology* 67, 358–380. doi: 10.1002/hep.29086

Hinshaw, D. C., and Shevde, L. A. (2019). The tumor microenvironment innately modulates cancer progression. *Cancer Res.* 79, 4557–4566. doi: 10.1158/0008-5472.CAN-18-3962

Huang, Y. K., Fan, X. G., and Qiu, F. (2016). TM4SF1 promotes proliferation, invasion, and metastasis in human liver cancer cells. *Int. J. Mol. Sci.* 17:661. doi: 10.3390/ijms17050661

Kent, L. N., and Leone, G. (2019). The broken cycle: E2F dysfunction in cancer. *Nat. Rev. Cancer* 19, 326–338. doi: 10.1038/s41568-019-0143-7

Kim, J. H., Matsubara, T., Lee, J., Fenollar-Ferrer, C., Han, K., Kim, D., et al. (2021). Lysosomal SLC46A3 modulates hepatic cytosolic copper homeostasis. *Nat. Commun.* 12:290. doi: 10.1038/s41467-020-20461-0

Lei, X., Lei, Y., Li, J. K., Du, W. X., Li, R. G., Yang, J., et al. (2020). Immune cells within the tumor microenvironment: biological functions and roles in cancer immunotherapy. *Cancer Lett.* 470, 126–133. doi: 10.1016/j.canlet.2019.11.009

Li, S. Q., Jiang, Y. H., Lin, J., Zhang, J., Sun, F., Gao, Q. F., et al. (2018). Albumin-to-fibrinogen ratio as a promising biomarker to predict clinical outcome of non-small cell lung cancer individuals. *Cancer Med.* 7, 1221–1231. doi: 10.1002/cam4.1428

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Liu, X., Huang, K., Zhang, R. J., Mei, D., and Zhang, B. (2020). Isochlorogenic acid A attenuates the progression of liver fibrosis through regulating HMGB1/TLR4/NF-kappaB signaling pathway. *Front. Pharmacol.* 11:582. doi: 10.3389/fphar.2020.00582

Liu, Y., Lou, G., Norton, J. T., Wang, C., Kandela, I., Tang, S., et al. (2017). 6-Methoxyethylamino-numonafide inhibits hepatocellular carcinoma xenograft growth as a single agent and in combination with sorafenib. *FASEB J.* 31, 5453–5465. doi: 10.1096/fj.201700306RR

Llovet, J. M., Zucman-Rossi, J., Pikarsky, E., Sangro, B., Schwartz, M., Sherman, M., et al. (2016). Hepatocellular carcinoma. *Nat. Rev. Dis. Primers.* 2:16018. doi: 10.1038/nrdp.2016.18

Makarova-Rusher, O. V., Altekruse, S. F., McNeel, T. S., Ulahannan, S., Duffy, A. G., Graubard, B. I., et al. (2016). Population attributable fractions of risk factors for hepatocellular carcinoma in the United States. *Cancer* 122, 1757–1765. doi: 10.1002/cncr.29971

Mantovani, A., Sozzani, S., Locati, M., Allavena, P., and Sica, A. (2002). Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes. *Trends Immunol.* 23, 549–555. doi: 10.1016/S1471-4906(02)02302-5

Marty Pyke, R., Thompson, W. K., Salem, R. M., Font-Burgada, J., Zanetti, M., and Carter, H. (2018). Evolutionary pressure against MHC class II binding cancer mutations. *Cell* 175, 416.e13–428.e13. doi: 10.1016/j.cell.2018.08.048

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28, 1747–1756. doi: 10.1101/gr.239244.118

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

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37, 773–782. doi: 10.1038/s41587-019-0114-2

Oh, J. H., Jang, S. J., Kim, J., Sohn, I., Lee, J. Y., Cho, E. J., et al. (2020). Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *NPJ Genom. Med.* 5:33. doi: 10.1038/s41525-019-0107-6

Qian, X., Zhao, J., Yeung, P. Y., Zhang, Q. C., and Kwok, C. K. (2019). Revealing lncRNA structures and interactions by sequencing-based approaches. *Trends Biochem. Sci.* 44, 33–52. doi: 10.1016/j.tibs.2018.09.012

Ranstam, J., and Cook, J. A. (2017). Kaplan-Meier curve. *Br. J. Surg.* 104:442. doi: 10.1002/bjs.10238

Seo, N., Akiyoshi, K., and Shiku, H. (2018). Exosome-mediated regulation of tumor immunology. *Cancer Sci.* 109, 2998–3004. doi: 10.1111/cas.13735

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

Sharma, P., and Allison, J. P. (2015). The future of immune checkpoint therapy. *Science* 348, 56–61. doi: 10.1126/science.aaa8172

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

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinf.* 19, 575–592. doi: 10.1093/bib/bbw139

Wang, P., Guo, Q., Hao, Y., Liu, Q., Gao, Y., Zhi, H., et al. (2021). LnCeCell: a comprehensive database of predicted lncRNA-associated ceRNA networks at single-cell resolution. *Nucleic Acids Res.* 49, D125–D133. doi: 10.1093/nar/gkaa1017

Wang, P., Li, X., Gao, Y., Guo, Q., Ning, S., Zhang, Y., et al. (2020). LnCeVar: a comprehensive database of genomic variations that disturb ceRNA network regulation. *Nucleic Acids Res.* 48, D111–D117. doi: 10.1093/nar/gkz887

Wang, P., Li, X., Gao, Y., Guo, Q., Wang, Y., Fang, Y., et al. (2019). LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. *Nucleic Acids Res.* 47, D121–D127. doi: 10.1093/nar/gky1144

Wang, Y., Sun, L., Wang, L., Liu, Z., Li, Q., Yao, B., et al. (2018). Long non-coding RNA DSCR8 acts as a molecular sponge for miR-485-5p to activate Wnt/beta-catenin signal pathway in hepatocellular carcinoma. *Cell Death Dis.* 9:851. doi: 10.1038/s41419-018-0937-7

Wang, X., Yang, K., Wu, Q., Kim, L. J. Y., Morton, A. R., Gimple, R. C., et al. (2019). Targeting pyrimidine synthesis accentuates molecular therapy response in glioblastoma stem cells. *Sci. Transl. Med.* 11:aau4972. doi: 10.1126/scitranslmed.aau4972

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

Zhang, Y., Han, P., Guo, Q., Hao, Y., Qi, Y., Xin, M., et al. (2021). Oncogenic landscape of somatic mutations perturbing pan-cancer lncRNA-ceRNA regulation. *Front. Cell Dev. Biol.* 9:658346. doi: 10.3389/fcell.2021.658346

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

Zongyi, Y., and Xiaowu, L. (2020). Immunotherapy for hepatocellular carcinoma. *Cancer Lett.* 470, 8–17. doi: 10.1016/j.canlet.2019.12.002

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

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: 21 October 2021
doi: 10.3389/fgene.2021.758926

Recognition of DNA Methylation Molecular Features for Diagnosis and Prognosis in Gastric Cancer

Donghui Liu[1,2†], Long Li[3,4†], Liru Wang[1,2], Chao Wang[5], Xiaowei Hu[6], Qingxin Jiang[7], Xuyao Wang[8], Guiqin Xue[9], Yu Liu[10] and Dongbo Xue[3,4]**

[1]Department of Oncology, Heilongjiang Provincial Hospital, Harbin, China, [2]Harbin Institute of Technology, School of Life Science and Technology, Harbin, China, [3]Department of General Surgery, First Affiliated Hospital of Harbin Medical University, Harbin, China, [4]Key Laboratory of Hepatosplenic Surgery, Ministry of Education, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [5]Department of Cardiology, Second Affiliated Hospital of Harbin Medical University, Harbin, China, [6]Department of Head and Neck and Genito-Urinary Oncology, Harbin Medical University Cancer Hospital, Harbin, China, [7]Department of General Surgery, Harbin 242 Hospital of Genertec Medical, Harbin, China, [8]Department of Pharmacy, Harbin Second Hospital, Harbin, China, [9]Department of General Surgery, Daqing Fifth Hospital, Daqing, China, [10]Department of Endocrine, Heilongjiang Provincial Hospital, Harbin, China

OPEN ACCESS

Edited by:
Xinyi Liu,
University of Illinois at Chicago, United States

Reviewed by:
Yanhui Gao,
Harbin Medical University, China
Yan Qiu,
University of Bristol, United Kingdom
Gang Li,
Peking University Third Hospital, China

***Correspondence:**
Yu Liu
ldhknight1@126.com
Dongbo Xue
xuedongbo@hrbmu.edu.cn

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

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

Received: 15 August 2021
Accepted: 04 October 2021
Published: 21 October 2021

Citation:
Liu D, Li L, Wang L, Wang C, Hu X, Jiang Q, Wang X, Xue G, Liu Y and Xue D (2021) Recognition of DNA Methylation Molecular Features for Diagnosis and Prognosis in Gastric Cancer. Front. Genet. 12:758926. doi: 10.3389/fgene.2021.758926

Background: The management of gastric cancer (GC) still lacks tumor markers with high specificity and sensitivity. The goal of current research is to find effective diagnostic and prognostic markers and to clarify their related mechanisms.

Methods: In this study, we integrated GC DNA methylation data from publicly available datasets obtained from TCGA and GEO databases, and applied random forest and LASSO analysis methods to screen reliable differential methylation sites (DMSs) for GC diagnosis. We constructed a diagnostic model of GC by logistic analysis and conducted verification and clinical correlation analysis. We screened credible prognostic DMSs through univariate Cox and LASSO analyses and verified a prognostic model of GC by multivariate Cox analysis. Independent prognostic and biological function analyses were performed for the prognostic risk score. We performed TP53 correlation analysis, mutation and prognosis analysis on eleven-DNA methylation driver gene (DMG), and constructed a multifactor regulatory network of key genes.

Results: The five-DMS diagnostic model distinguished GC from normal samples, and diagnostic risk value was significantly correlated with grade and tumor location. The prediction accuracy of the eleven-DMS prognostic model was verified in both the training and validation datasets, indicating its certain potential for GC survival prediction. The survival rate of the high-risk group was significantly lower than that of the low-risk group. The prognostic risk score was an independent risk factor for the prognosis of GC, which was significantly correlated with N stage and tumor location, positively correlated with the VIM gene, and negatively correlated with the CDH1 gene. The expression of CHRNB2 decreased significantly in the TP53 mutation group of gastric cancer patients, and there were significant differences in CCDC69, RASSF2, CHRNB2, ARMC9, and RPN1 between the TP53 mutation group and the TP53 non-mutation group of gastric cancer patients. In addition, CEP290, UBXN8, KDM4A, RPN1 had high frequency mutations and the function

of eleven-DMG mutation related genes in GC patients is widely enriched in multiple pathways.

Conclusion: Combined, the five-DMS diagnostic and eleven-DMS prognostic GC models are important tools for accurate and individualized treatment. The study provides direction for exploring potential markers of GC.

Keywords: gastric cancer, tumor marker, diagnosis, prognosis, DNA methylation, mutation

INTRODUCTION

According to the statistics released by the World Health Organization in 2018, the incidence and mortality rate of gastric cancer (GC) ranked fifth and third, respectively, among cancers worldwide. GC is a characteristic cancer in East Asia with an incidence rate of 32.1/100,000 and a mortality rate of 13.2/100,000 (1). Among Eastern Asian countries, Japan, South Korea, and China have the highest GC morbidity and mortality rates in the world (Bray et al., 2018). Therefore, the prevention and treatment of GC are essential for improving patient outcomes. Although advances in surgery, radiotherapy, chemotherapy, molecular targeting, and immunotherapy have improved overall prognosis, diagnosis of GC is often delayed, resulting in unsatisfactory outcomes (Bang et al., 2017; Cats et al., 2018; Sundar et al., 2019). It is, thus, urgent to explore effective biomarkers for early diagnosis and prognosis prediction of GC.

Epigenetic markers have been widely recognized in recent years, particularly promoter hypermethylation. Compared with a wide range of mutational variations in a specific gene, promoter hypermethylation occurs in the same defined region of a gene in all forms of cancer (Fu, 2015). Therefore, diagnosis and prognosis prediction of patients with GC can be reliably obtained at the epigenetic level *via* differential expression of common DNA methylation (DNAm). DNAm is a major epigenetic modification that participates in many important life activities, such as cell proliferation, differentiation, development, apoptosis, tumor development, and occurrence of other diseases, and it is also one of the earliest discovered DNA modifications. DNAm can cause changes in chromatin structure and DNA stability, thereby regulating gene expression (Neri et al., 2017). Abnormal DNAm located in the promoter region usually leads to silencing of tumor suppressor genes or high expression of proto-oncogenes, thereby promoting tumor progression (Das and Singal, 2004). Among them, hypermethylation of tumor suppressor genes is the most common and can be used as an early tumor marker. Some specific DNAm sites are closely related to GC, such as cell cycle-related genes P16 and MDGA2 (Hibi et al., 2003; Wang et al., 2016), tumor suppressor genes, apoptosis-related genes PCDH10 and BCL6B (Yu et al., 2009; Xu et al., 2012), signal transduction-related genes FOXF2 and RUNX3 (Sakakura et al., 2005; Higashimori et al., 2018), and proto-oncogenes RAS and c-myc (Nishigaki et al., 2005; Licchesi et al., 2010). The discovery of these DNAm sites has broad application value in the early diagnosis, prognosis, and even treatment of GC. However, only a small number of DNAm sites have been approved for use as basic tumor markers (NDRG4, BMP3, and SEPTIN9) (Imperiale et al., 2014; FDA). There are many reasons for this, such as small numbers of test samples, patient selection bias, lagging research design and data analysis methods, lack of substantial clinical value, and other factors have prevented thorough evaluation of the clinical value of GC biomarkers. With the development of bioinformatics, enabling the establishment of GC diagnostic and prognostic models based on big data, the above problems can be resolved.

Few studies have described the application of a differential methylation site (DMS) scoring system to construct individualized GC diagnostic and prognostic models. In this study, we integrated publicly available GC DNAm datasets obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases to construct a diagnostic model and verify its ability to distinguish GC from normal tissues. The DMSs were then matched with overall survival (OS) data and a prognostic model was constructed. Finally, the prognostic model was analyzed to explore its clinical application and potential molecular mechanisms in patients with GC. The correlations between clinical correlation analysis of the diagnostic and analysis of independent prognostic factors will help achieve accurate and individualized treatment in a clinical setting.

MATERIALS AND METHODS

Obtaining DNAm Data of Gastric Cancer

We downloaded TCGA GC DNAm profiles (Illumina Human Methylation 450 BeadChip, Illumina Human Methylation 27 BeadChip), expression profiles, and corresponding clinical data through the UCSC Xena database (https://xena.ucsc.edu/) (Wang et al., 2019). The Illumina Human Methylation 450 BeadChip DNAm dataset contained two normal samples and 395 GC samples, while the Illumina Human Methylation 27 BeadChip DNAm dataset contained 25 normal samples and 48 GC samples. The expression profile dataset contained 32 normal samples and 372 GC samples. **Table 1** lists the clinicopathological characteristics of the patients with GC. We downloaded the GC DNAm profile dataset GSE30601 from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) (Kurashige et al., 2016). The GSE30601 dataset was based on the GPL8490 platform (Illumina Human Methylation 27 BeadChip), containing 94 normal samples and 203 GC samples. The data from TCGA GC DNAm profiles were sorted and merged as the training dataset; the GEO GC DNAm profile dataset was used as the validation dataset. Because of the availability of public data in

TABLE 1 | The clinicopathological characteristics of GS patients.

	Alive (*n* = 216)	Dead (*n* = 107)	Total (*n* = 323)
Gender			
FEMALE	89 (41.2%)	34 (31.8%)	123 (38.1%)
MALE	127 (58.8%)	73 (68.2%)	200 (61.9%)
Age			
Mean (SD)	63.9 (10.7)	65.8 (10.3)	64.5 (10.6)
Median [MIN, MAX]	65 [30,90]	67 [41,90]	66 [30,90]
Grade			
G1	5 (2.3%)	2 (1.9%)	7 (2.2%)
G2	74 (34.3%)	34 (31.8%)	108 (33.4%)
G3	137 (63.4%)	71 (66.3%)	208 (64.4%)
Stage			
Stage I	30 (13.9%)	8 (7.5%)	38 (11.8%)
Stage II	84 (38.9%)	26 (24.3%)	110 (34.0%)
Stage III	94 (43.5%)	62 (57.9%)	156 (48.3%)
Stage IV	8 (3.7%)	11 (10.3%)	19 (5.9%)
T			
T1	13 (6.0%)	1 (0.9%)	14 (4.3%)
T2	41 (19.0%)	16 (15.0%)	57 (17.6%)
T3	106 (49.1%)	56 (52.3%)	162 (50.2%)
T4	56 (25.9%)	34 (31.8%)	90 (27.9%)
M			
M1	209 (96.8%)	99 (92.5%)	308 (95.4%)
M2	7 (3.2%)	8 (7.5%)	15 (4.6%)
N			
N0	84 (38.9%)	24 (22.4%)	108 (33.5%)
N1	52 (24.1%)	30 (28.0%)	82 (25.4%)
N2	42 (19.4%)	25 (23.4%)	67 (20.7%)
N3	38 (17.6%)	28 (26.2%)	66 (20.4%)
Race			
ASIAN	63 (29.2%)	21 (19.6%)	84 (26%)
BLACK	3 (1.4%)	6 (5.6%)	9 (2.8%)
WHITE	150 (69.4%)	80 (74.8%)	230 (71.2%)
Position			
Body of stomach	54 (25%)	18 (16.8%)	72 (22.3%)
Cardia, NOS	49 (22.7%)	29 (27.1%)	78 (24.1%)
Fundus of stomach	33 (15.3%)	14 (13.1%)	47 (14.6%)
Gastric antrum	77 (35.6%)	40 (37.4%)	117 (36.2%)
Stomach, NOS	3 (1.4%)	6 (5.6%)	9 (2.8%)

TCGA and GEO databases, this study did not require ethical approval or informed consent.

Identification of Differential Methylated Sites

We performed background correction and normalization on the DNAm data in the training set (Zhang et al., 2019). Using normal samples as controls, we screened the DMSs in GC samples using the Wilcoxon test (Xu et al., 2017), with |log2 fold change (FC)| > 1 and false discovery rate (FDR) < 0.01 set as the threshold considered to have biological significance. The "pheatmap" package in R software was used to draw a DNAm heatmap of DMSs in GC.

Screening of Diagnostic DNAm Markers

We used the random forest method in R software to predict key DNAm sites in GC. The DNAm sites were sorted from high to low according to their calculated "Mean Decrease Accuracy" value, and 10-fold cross validation was performed five times to screen representative DNAm markers in GC. We also used the "glmnet" package in R software to predict key DNAm sites in GC through LASSO regression analysis. DMSs that could distinguish tumors from normal samples were defined as representative DNAm markers in GC. Finally, shared DNAm markers predicted by both methods were selected as reliable DNAm markers for GC diagnosis (Zhou et al., 2019a).

Construction of DNAm Diagnostic Model

The "glm" package in R software was used to construct a diagnosis prediction model with five reliable DNAm markers through multivariate logistic regression analysis. The constructed GC diagnosis prediction model was applied to distinguish GC from normal samples in the training and validation datasets, and the model's accuracy was evaluated. Unsupervised hierarchical clustering was used to show the DNAm status of five credible diagnostic DNAm markers in the training set and validation set.

Correlation Analysis of DNAm Diagnostic Model With Clinical Indicators

To evaluate the clinical application of the DNAm diagnostic model in GC, we calculated the scores of patients with GC in TCGA dataset using the constructed DNAm diagnostic model. Samples with missing clinical characteristics were removed, and correlations between diagnostic score and clinical characteristics of patients were analyzed. The t-test was used for comparisons between two groups, and the Kruskal–Wallis test was used for comparisons between two or more groups. $p < 0.05$ was considered statistically significant.

Construction of Prognostic Model Based on Differential Methylated Sites

The "survival" package in R software was used to determine DNAm sites of differential methylation associated with survival of patients with GC through univariate Cox regression analysis, and the random forest map was plotted for the top 20 DNAm sites with the most significant differences ($p < 0.01$). Based on the selected prognosis-related DNAm sites, the "glmnet" package in R software was used to perform 10,000 simulations through LASSO regression analysis, and key DNAm sites were obtained after removing overlap through cross validation.

We used multivariate Cox regression analysis to construct the following risk score formula for each patient (cg07990939 Methylation levels*(−8.908))+(cg08317263 Methylation levels*(−1.739))+(cg10301990 Methylation levels *(−4.088))+(cg10968649 Methylation levels *(−20.267))+(cg13801416 Methylation levels *(−1.009))+(cg19614321 Methylation levels*(−1.779))+(cg20074795 Methylation levels *(12.778))+(cg21052164 Methylation levels *(−0.941))+(cg26069252 Methylation levels *(7.734))+(cg26089280 Methylation levels *(−8.569))+(cg27662379 Methylation levels *(−7.672)). Patients were divided into low-risk and high-risk groups according to the risk score formula using the median risk as the cut-off point. We assessed survival differences between the two groups using the Kaplan–Meier method, and compared these survival differences using log-rank statistics. Receiver operating characteristic (ROC)

curve analysis was used to determine the accuracy of model predictions (Xu et al., 2017).

Analysis of Independent Prognostic Factors and Prognostic Risk Model

To evaluate the prognostic model and the effect of different clinical characteristics of patients with GC on prognosis and survival, we obtained phenotypic information of all samples from the clinical data in TCGA dataset and extracted the risk model samples separately, as well as the corresponding age, gender, and other phenotypic and clinical information. We combined the information in the risk model with the survival status of patients, then used the "survival" package in R software to perform univariate and multivariate independent prognostic analyses to test the ability of the prognostic risk model and the clinical characteristics of patients with GC to predict the prognosis (Vasiljević et al., 2014).

Functional Analysis of Prognostic Risk Score

To evaluate the clinical application and important functions of the DNAm prognostic model in GC, we first calculated the risk scores of patients with GC in TCGA dataset using the constructed DNAm prognostic model and combined the risk scores with their clinical data. Samples with missing clinical traits were removed, and the correlation between risk scores and clinical characteristics was analyzed. We used the t-test to compare two groups and the Kruskal–Wallis test to compare two or more groups. $p < 0.05$ was considered statistically significant. We then extracted the expression levels of regulatory, cytotoxic, and epithelial–mesenchymal transition (EMT) factors of known immune checkpoint sites from the GC samples in TCGA dataset and correlated these levels with the risk scores of these samples to investigate whether the risk scores played an important regulatory role in GC by influencing the above factors. Finally, patients were divided into low-risk and high-risk groups according to the prognostic risk model using the median risk as the cut-off point. The low-risk group was used as the control. We used the Wilcoxon test to screen significant differentially expressed genes in the high-risk group, using the standard threshold $|\log2FC| > 0$ and FDR <0.05. The "clusterProfiler" package in R language was used to perform gene set enrichment analysis (GSEA) for the potential mechanism of c2 (c2.cp.kegg.v7.1.entrez.gmt, c2.cp.biocarta.v7.1.entrez.gmt) and c5 (c5.bp.v7.1.entrez.gmt) in the molecular signature database (MSigDB). The number of random sample arrangements was set to 1,000, and the significance threshold was set to $p < 0.05$ (Zhou et al., 2019a).

Analysis of the Correlation Between Eleven Prognostic-Related DMG and TP53 Mutations

UALCAN (http://ualcan.path.uab.edu/analysis.html) is a comprehensive, user-friendly and interactive online data analysis website based on relevant cancer data found in TCGA database. We used the UALCAN database to evaluate the expression levels of eleven prognostic-related DMG in gastric cancer and normal gastric tissues (Chandrashekar et al., 2017). Considering the unequal variances, the significance of differences in the transcriptional levels was evaluated using the Student's t-test, and a p value of <0.05 was considered statistically significant.

Mutation and Prognostic Analysis of Eleven Prognostic-Related DMG

The cBioPortal (http://www.cbioportal.org) integrates data from large-scale cancer research projects, such as TCGA and the International Cancer Genome Consortium (ICGC), whose gene data types cover somatic mutations, DNA copy number changes, mRNA and microRNA expression, DNA methylation, protein and phosphorus protein abundance, and provides visual and multidimensional cancer genomic data (Cerami et al., 2012; Gao et al., 2013). This study based on TCGA database, gene expression data of 412 GC patients were included. We obtained the relevant module information about 11-DMG mutation from the cBioPortal. Set the parameters: "Enter a z-score threshold±1.8", then enter DMG to generate a mutation frequency visualization chart, and then select the top 10 genes significantly related to each gene mutations in "Co-expression" module, delete duplicates and import them into Metascape. Metascape (https://metascape.org/gp/index.html#/main/step1) is a gene list analysis tool. It integrates data from over 40 types of biological information databases for gene annotation and analysis, and provides a unique protein–protein interaction (PPI) network analysis function. We used the "Multiple Gene list" module of the Metascape tool to perform gene annotation and enrichment analyses on the genes obtained from the cBioPortal that were highly related to DMG mutations(27), and set the parameters: "enrichment factor Min overlap = 3," "p-value cut-off value <0.01," "Min enrichment >1.5" is considered statistically significant, then select Gene Ontology (GO) enriching "Biological Processes," "Cellular Components" and "Molecular Functions" and "KEGG pathways" classification. To further capture the relationships between the terms, a subset of enriched terms was selected and rendered as a network plot, where terms with a similarity >0.3 were connected by edges. We selected the terms with the best p-values from each of the 20 clusters, with the constraint that there were no more than 15 terms per cluster and no more than 250 terms in total. The network was visualized using Cytoscape (Shannon et al., 2003), where each node represented an enriched term and was colored first by its cluster ID, and then by its p-value. For each given gene list, PPI enrichment analysis was carried out using the following databases: STRING (Szklarczyk et al., 2019), BioGrid (Oughtred et al., 2019), OmniPath (Li et al., 2017), and InWeb_IM (Li et al., 2017). Only physical interactions in STRING (physical score >0.132) and BioGrid were used (details). The molecular complex detection (MCODE) algorithm (Bader and Hogue, 2003) was applied to identify densely connected network components.

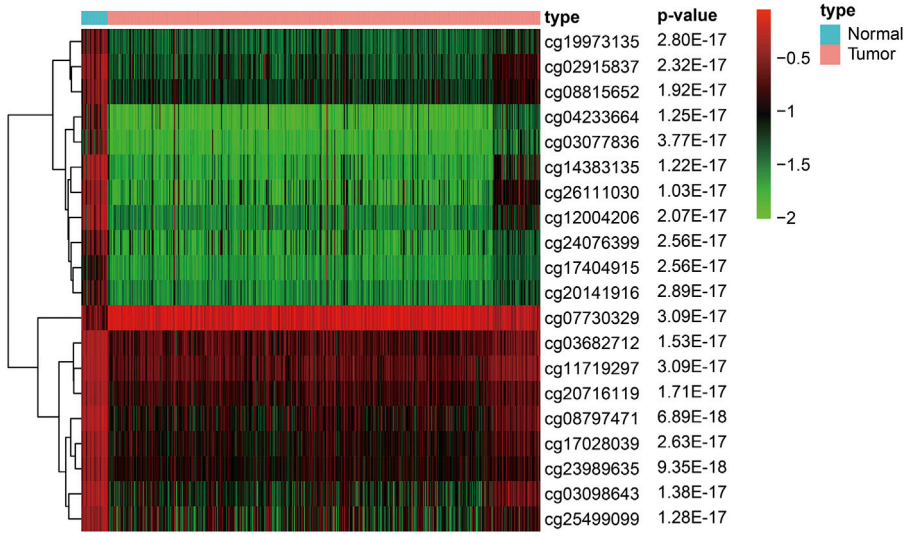


FIGURE 1 | Heat map of the top 20 significanly different methylation sites in gastric cancer (Arranged in *p*-value order).

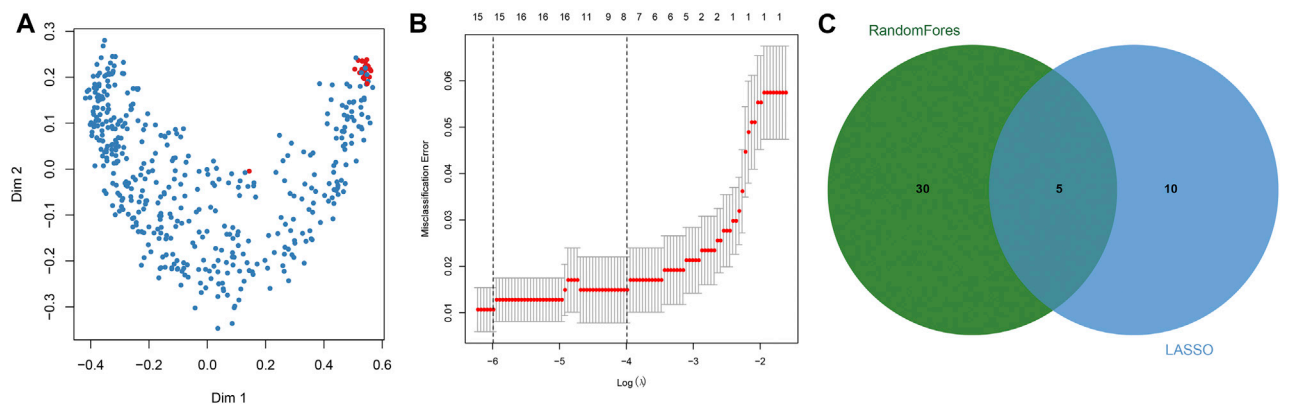


FIGURE 2 | Screening of diagnostic DNA methylation (DNAm) markers in gastric cancer. **(A)** Multi-dimensional scaling plot of the proximity matrix generated from random forest analysis in the training dataset. Red dots represent normal samples and blue dots indicate tumor samples. **(B)** Misclassification error for different numbers of variables revealed by the LASSO regression model. Red dots represent the value of misclassification error, grey lines represent the standard error (SE), the two vertical dotted lines on the left and right represent optimal values by the minimum and 1-SE criteria, respectively. "Lambda" is the tuning parameter. **(C)** Screening of DNAm markers for reliable diagnosis. The green circle represents representative DNAm markers selected by random forest analysis, and the blue circle indicates representative DNAm markers screened by LASSO regression analysis.

Construction of Multi-Factor Regulatory Network of Key Genes

In order to predict the regulatory factors of key genes related to the prognostic model constructed in gastric cancer, we predicted the upstream regulated miRNAs of key genes through Starbase (http://starbase.sysu.edu.cn/index.php) and TargetScan (http://www.targetscan.org/vert_71/), and intersected the prediction results to obtain reliable miRNAs. After that, we further predicted the lncRNA upstream regulated by the trusted miRNA through the Starbase database, and predicted the transcription factors (TF) that can regulate key genes through the TRRUST (https://www.grnpedia.org/trrust) database. Finally, the regulatory network among mRNA, miRNA, lncRNA and TF was constructed by Cytoscape v3.6.1 software.

RESULTS

Identification of Differential Methylated Sites in Gastric Cancer

To construct the diagnostic and prognostic GC models, we performed background correction and normalization on the DNAm data from 27 normal samples and 443 GC samples in the training dataset. Among them, 1842 hypermethylated and 899

TABLE 2 | Characteristics of five methylation markers and their coefficients in GC diagnosis.

Markers	Ref.Gene	Coefficients	SE	z.value	P.value
		12.209	3.242	3.766	<0.001
cg14383135	NPAS2	−2.609	7.309	−0.357	0.721
cg08797471	DAPK1	−19.390	5.950	−3.259	0.001
cg26619317	CNN3	−2.969	7.454	−0.398	0.690
cg17028039	FGFR2	−6.982	9.783	−0.714	0.475
cg25764464	PLEKHA5	−2.097	6.914	−0.303	0.762

SE: standard errors of coe-cients; z value: Wald z-statistic value.

hypomethylated sites were screened out in the GC samples. We used R software package pheatmap to draw the methylation heat map of the top 20 significanly different methylation sites in gastric cancer, arranged in *p*-value order (**Figure 1**) (**Supplementary Table S1**).

Screening of Diagnostic DNAm Markers

Key DNAm sites in GC were predicted through random forest analysis, combined with five repeated ten-fold cross validations, resulting in 35 representative DNAm markers (**Figure 2A**). At the same time, we also predicted 15 key DNAm sites in GC by LASSO regression analysis (**Figure 2B**). The intersection of the representative DNAm markers predicted by both methods yielded five reliable diagnostic DNAm markers in GC (**Figure 2C**).

Construction of a DNAm Diagnostic Model

Using multivariate logistic regression analysis, we established a GC diagnosis prediction model with the five selected DNAm markers (**Table 2**). Applying the model to the training dataset yielded a sensitivity of 99.1% and specificity of 81.5% samples (**Figure 3A**) and a sensitivity of 87.2% and specificity of 63.8% in the validation dataset (**Figure 3B**). We also demonstrated this model could differentiate GC from normal samples both in the training dataset (AUC = 0.994) and the validation dataset (AUC = 0.829) (**Figures 3C,D**). Unsupervised hierarchical clustering of these five markers distinguished GC from normal samples with high specificity and sensitivity (**Figures 3E,F**). These results indicated that the DNAm diagnostic model could be a significant tool for distinguishing GC from normal samples.

Correlation Between DNAm Diagnostic Model and Clinical Indicators

After excluding samples with missing clinical data, we analyzed correlations between the diagnostic risk score and the clinical characteristics of 323 patients obtained from TCGA dataset. The results indicated that diagnostic risk score was significantly correlated with grade and tumor location in patients with GC, but not with age, gender, stage, extent of the tumor (T), presence of metastasis (M), extent of spread to the lymph nodes (N), or race of the patient (**Figures 4A–I**).

Prognostic Model Based on Differential Methylated Sites

We combined the DNAm values of the DMSs in GC samples with the survival data of the corresponding patients, using $p < 0.01$ as the threshold standard to perform univariate Cox proportional hazard regression analysis. We found that 137 DMSs significantly affected the survival of patients with GC, among which the top 20

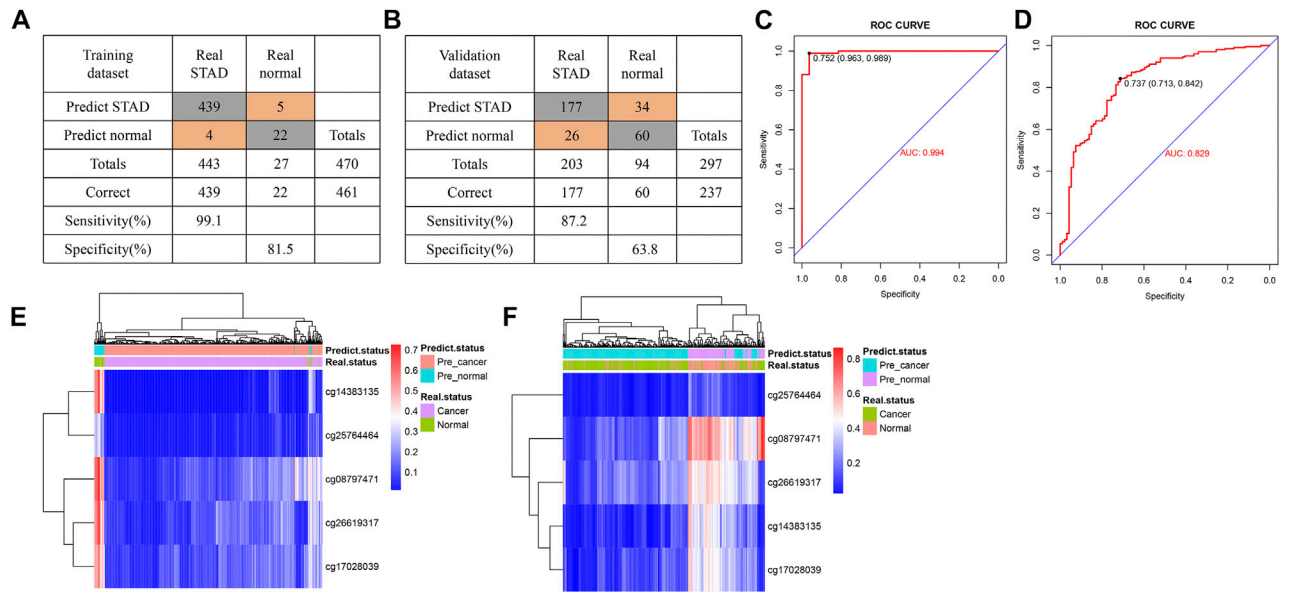


FIGURE 3 | Construction of a diagnostic model of DNA methylation (DNAm) in gastric cancer. **(A,B)** Confusion tables of binary results of the diagnostic prediction model in the training **(A)** and validation datasets **(B)**. **(C,D)** Receiving operating characteristics curve analysis of the diagnostic prediction model with DNAm markers in the training **(C)** and validation datasets **(D)**. **(E,F)** Unsupervised hierarchical clustering of five DNAm markers selected for use in the diagnostic prediction model in the training **(E)** and validation data sets **(F)**.

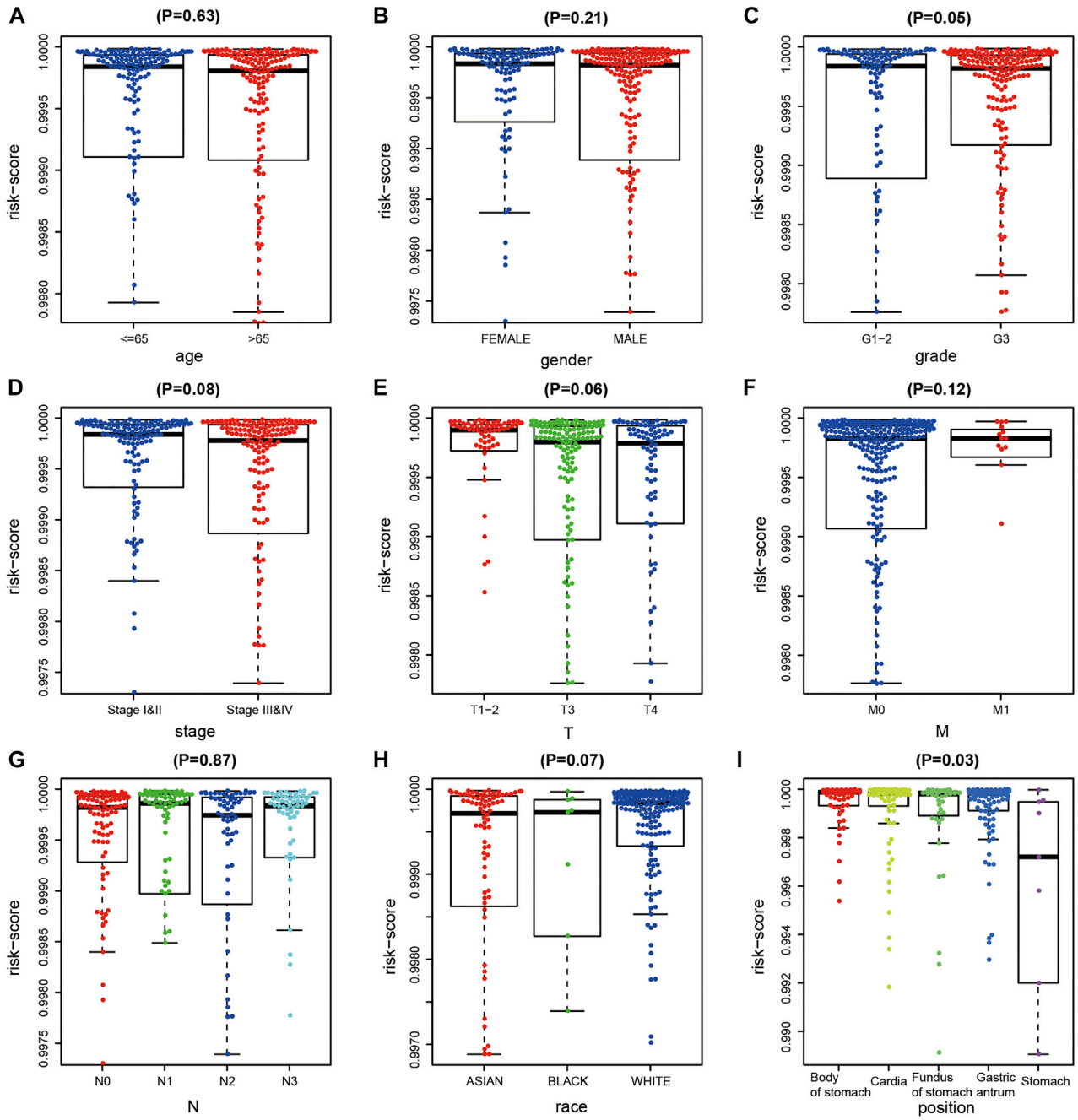


FIGURE 4 | Correlation analysis of DNA methylation (DNAm) diagnostic model and clinical indicators in gastric cancer (GC). **(A–I)** Correlation analysis between diagnostic risk score and age, gener, tumor grade, T, M, N stage, race, and tumor site of gastric cancer.

DNAm sites with the most significant differences are shown (**Figure 5A**). We used LASSO regression analysis to remove redundant DNAm sites, performed 10,000 simulations, removed overlaps through cross validation, and finally obtained 25 prognostic-related DMSs (**Figures 5B,C**). We constructed a prognostic risk score formula for each patient based on these 25 prognosis-related DMSs (**Table 3**). The DNAm heatmap demonstrated the DMSs in the low-risk and high-risk groups based on the prognostic (**Figure 5D**). The corresponding ROC curve analysis demonstrated that the area under the curve (AUC) value of the constructed prognostic model was 0.747, which indicated the predictive power of the prognostic model based on the expression of DMSs in GC (**Figure 5E**). Further, the Kaplan–Meier curves suggested that the survival rate of patients in the high-risk group was significantly lower than that in the low-risk group (**Figure 5F**).

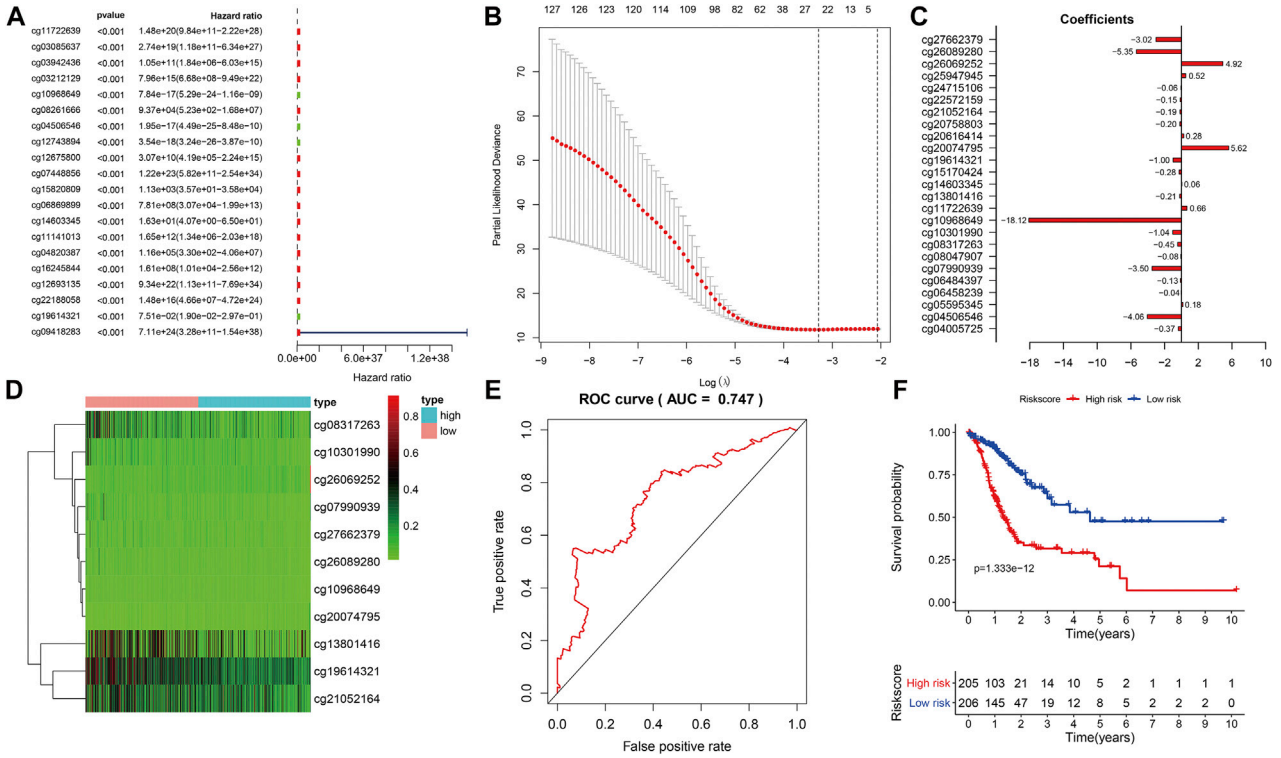


FIGURE 5 | Prognostic model based on differential methylation sites (DMSs). **(A)** Random forest plot of the top DMS with the most significant differences through univariate Cox proportional hazard regression analysis. **(B)** Results of LASSO regression analysis and 10,000 simulations. **(C)** Corresponding coefficient values of each DMS in LASSO regression analysis. **(D)** DNAm heatmap of DMSs in the low-risk and high-risk groups with increasing prognostic risk score value. **(E)** Receiver operating characteristics curve analysis of the prognostic model. **(F)** Kaplan–Meier survival analysis of DMSs related to prognostic model, abscissa for survival time, ordinate for survival rate, blue curve for low-risk patients, red curve for high-risk patients. The number of high-risk and low-risk patients at each time point are located on the bottom axis of the graph.

TABLE 3 | Characteristics of eleven methylation markers and their coefficients in GC prognosis prediction.

Markers	Ref.Gene	Coefficients	HR	CI	SE	z.value	P.value
cg07990939	CEP290	−8.908	1.35E-04	2.05e-07–8.94e-02	3.313	−2.689	7.17E-03
cg08317263	CCDC69	−1.739	1.76E-01	3.38e-02–9.13e-01	0.841	−2.068	3.86E-02
cg10301990	UBXN8	−4.088	1.68E-02	1.99e-04–1.41e+00	2.263	−1.807	7.08E-02
cg10968649	KDM4A	−20.267	1.58E-09	1.73e-17–1.44e-01	9.352	−2.167	3.02E-02
cg13801416	AKR1B1	−1.009	3.65E-01	1.65e-01–8.05e-01	0.404	−2.496	1.26E-02
cg19614321	RASSF2	−1.779	1.69E-01	3.55e-02–8.02e-01	0.795	−2.237	2.53E-02
cg20074795	KDELR3	12.778	3.54E+05	1.54e+01–8.15e+09	5.124	2.494	1.26E-02
cg21052164	CHRNB2	−0.941	3.90E-01	1.22e-01–1.24e+00	0.592	−1.59	1.12E-01
cg26069252	EGR1	7.734	2.29E+03	1.69e+02–3.08e+04	1.327	5.826	5.67E-09
cg26089280	ARMC9	−8.569	1.90E-04	1.60e-09–2.25e+01	5.96	−1.438	1.51E-01
cg27662379	RPN1	−7.672	4.66E-04	1.65e-07–1.31e+00	4.053	−1.893	5.84E-02

HR: Hazard Ratio; CI: 95.0% confidence interval; SE: standard errors of coe-cients; z value: Wald z-statistic value.

Analysis of Independent Prognostic Factors in the Prognostic Risk Model

To further evaluate the prognostic model and the impact of different clinical characteristics of patients with GC on prognosis and survival, we obtained the corresponding age, gender, phenotype, and clinical information for 315 patients with GC from TCGA dataset. We performed univariate and multivariate independent prognostic analyses (**Figures 6A,B**), revealing that the prognostic risk score value and tumor site were significant high-risk factors and were significantly correlated with the survival status of patients with GC ($p < 0.05$). The corresponding ROC curve analysis demonstrated that the constructed prognostic model had the largest AUC value of 0.782, which also indicated the predictive power of the prognostic model based on DMSs in GC (**Figure 6C**).

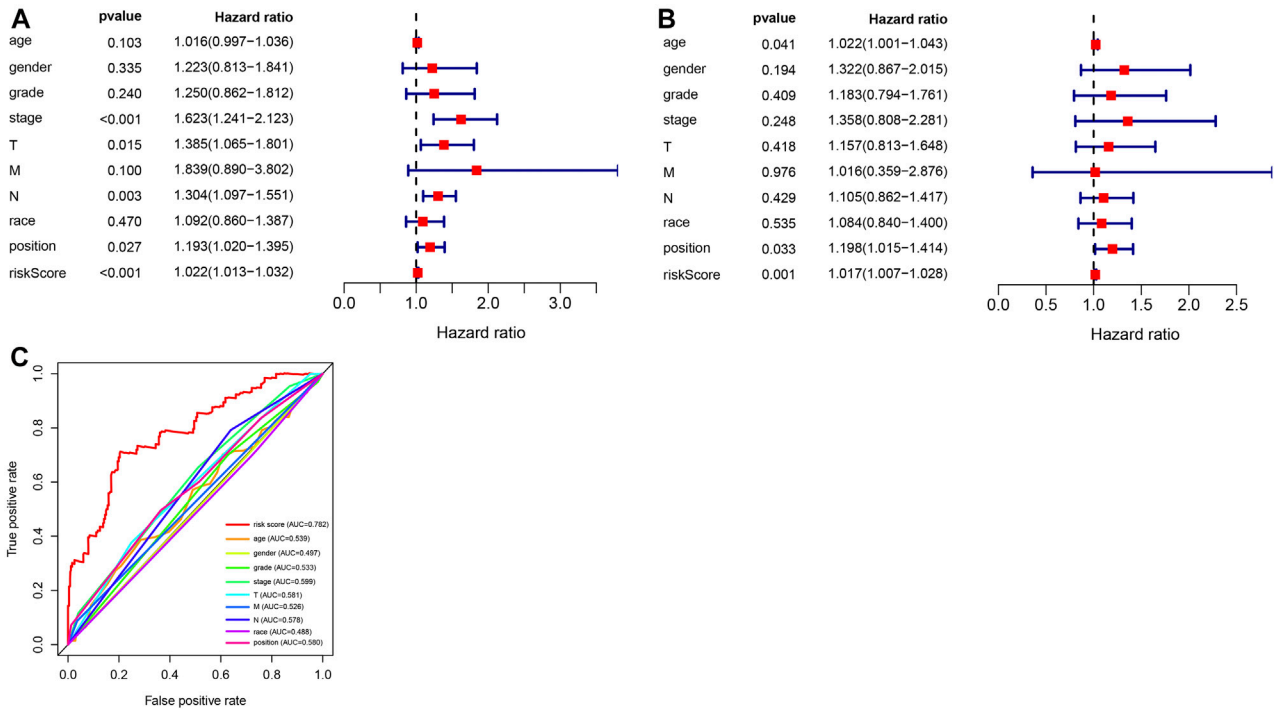


FIGURE 6 | Analysis of independent prognostic factors in the prognostic risk model. **(A)** Random forest plot of univariate independent prognostic analysis; the left side indicates clinical characteristics of gastric cancer (GC), the middle is the *p*-value. The hazard ratio indicates the risk rate with hazard ratio >1 indicating high-risk clinical features, and hazard ratio <1 indicating low-risk clinical features. **(B)** Random forest plot of multivariate independent prognostic analysis; the left side represents clinical characteristics of GC, the middle is the *p*-value. The hazard ratio represents the risk rate with hazard ratio >1 indicating high-risk clinical features, and hazard ratio <1 indicating low-risk clinical features. **(C)** Receiver operating characteristics curve analysis of the prognostic model constructed with eleven differential methylation sites (DMSs) in GC.

Functional Analysis of Prognostic Risk Score

To evaluate the clinical application and important functions of the DNAm prognostic model in GC, we calculated the prognostic risk score of patients with GC from TCGA dataset and then analyzed correlations with patient clinical characteristics. The prognostic risk score was significantly correlated with extent of spread to the lymph nodes (N) and tumor site in patients with GC but not significantly correlated with other clinical features (**Figure 7A**). We also analyzed correlations between prognostic risk score and expression levels of regulatory, cytotoxic, and EMT factors of immune checkpoint sites. The results indicated that prognostic risk score was significantly positively correlated with VIM, which was significantly positively correlated with PDCD1, CTLA4, LAG3, TIGIT, GZMB, and TNF and significantly negatively correlated with CDH1 (**Figure 7B**). We screened 6,172 significant differentially expressed genes in the high-risk group samples. GSEA on the potential mechanism of c2 (c2.cp.kegg.v7.1.entrez.gmt, c2.cp.biocarta.v7.1.entrez.gmt) and c5 (c5.bp.v7.1.entrez.gmt) in the MSigDB (**Figures 7C–E**) revealed that highly expressed genes in the high-risk group were significantly enriched in multiple biological processes, such as the "calcium signaling pathway," "cytokine receptor interaction," "focal adhesion," "neuroactive ligand receptor interaction," and "regulation of actin cytoskeleton," indicating that these pathways may play important roles in the development of GC.

Analysis of the Correlation Between Eleven Prognostic-Related DMG and TP53 Mutations

We further analyzed the relationship between DMG mRNA expression levels and TP53 mutation status in patients with gastric cancer using the UALCAN data mining website. In the correlation analysis of TP53 mutation status, it is worth noting that the expression of CHRNB2 decreased significantly only in the TP53 mutation group of gastric cancer patients. CCDC69, RASSF2, CHRNB2, ARMC9, and RPN1 were significantly different in the TP53 mutation group and TP53 non-mutation group of gastric cancer patients (**Figure 8**).

Mutation and Prognostic Analysis of Eleven Prognostic-Related DMG

We analyzed eleven prognostic-related DMG mutations and their relationship with OS and PFS in gastric cancer patients using the cBioportal website. Among 412 patients with gastric cancer, 242

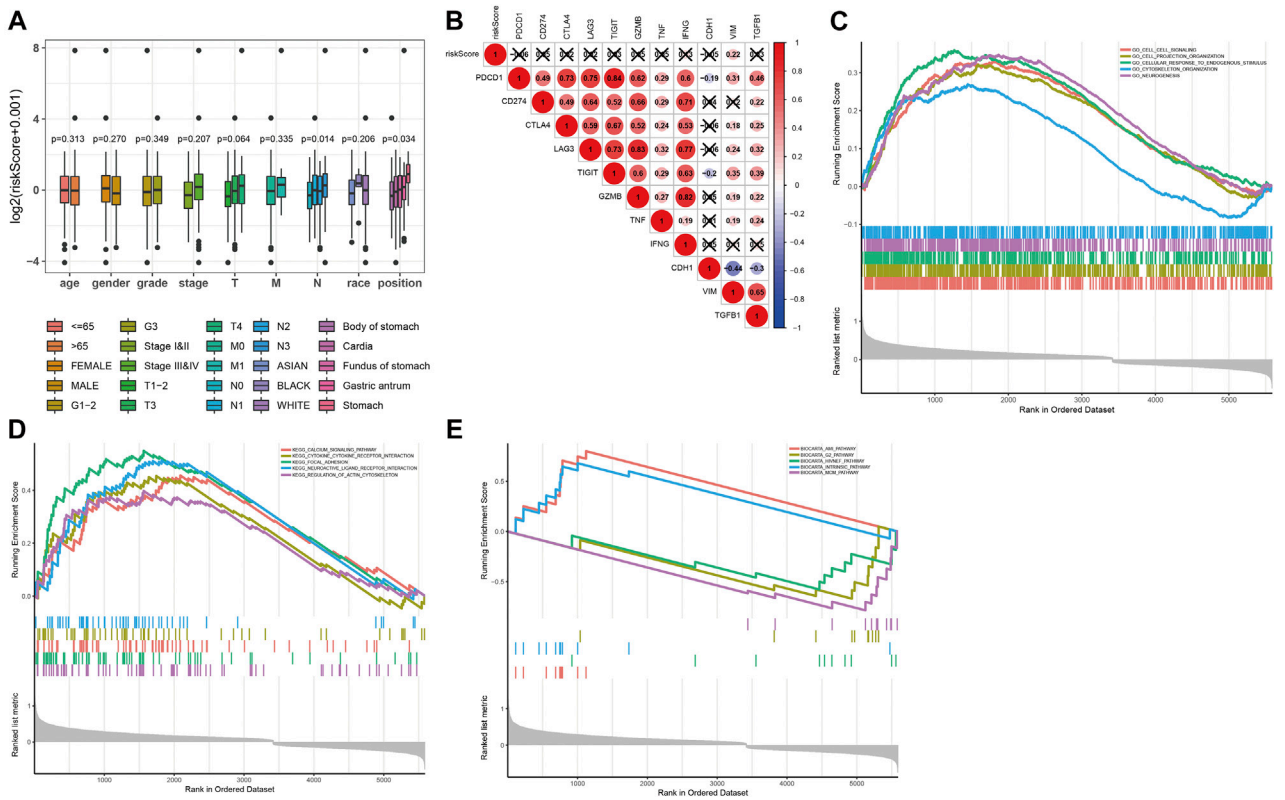


FIGURE 7 | Functional analysis of prognostic risk score. **(A)** Correlation analysis between prognostic risk score and age, gender, tumor grade, N stage, T stage, race, and tumor location of the patient in the prognostic risk model. **(B)** Correlation analysis between prognostic risk score and expression levels of regulatory, cytotoxic, and epithelial–mesenchymal transition (EMT) factors of immune checkpoint sites. **(C–E)** The results of gene set enrichment analysis on the potential mechanism of c5 (c5.bp.v7.1.entrez.gmt) and c2 (c2.cp.kegg.v7.1.entrez.gmt, c2.cp.biocarta.v7.1.entrez.gmt) in the molecular signatures database.

had gene mutations, with a mutation rate of 59%. The mutation rates of CEP290, CCDC69, UBXN8, KDM4A, AKR1B1, RASSF2, KDELR3, CHRNB2, EGR1, ARMC9, RPN1 were 10, 5, 12, 11, 8, 2.9, 6, 7, 6, 7, and 13%, respectively. We observed that the mutation rates of CEP290, UBXN8, KDM4A, and RPN1 were more than 10% (10, 12, 11, 13%) (**Figure 9A**). In addition, high mRNA expression was an important factor leading to high mutation frequency in gastric cancer (**Figure 9B**). However, Kaplan-Meier plotter and log-rank test analysis showed that SMYD family mutations had no significant correlation with OS and PFS in patients with gastric cancer (OS: p value = 0.887, PFS: p value = 0.548) (**Figure 9C**). Next, we used the cBioportal to search for genes that were significantly related to gastric cancer and DMG mutations (the top 10, respectively). After deduplication, a total of 108 genes were obtained, ZDHHC17, ARID4A, ATRX, ARID4B, UPF2, ZNF37BP, CEP162, MDM4, CCDC66, PHIP, ASB2, PRKCB, GYPC, SLC9A9, RASGRP2, JAM2, FNBP1, MAP3K3, PLEKHO, GTF2E2, MAK16, CNOT7, PPP2CB, CCDC25, DCTN6, INTS10, PPP2R2A, LEPROTL1, ELP3, AGO1, PTPRF, COMMD6, NCOA2, COPS9, MRPL53, POLR3A, UHMK1, CSNK1G1, AIDA, ADAP2, NRROS, HVCN1, LY86, TM6SF1, TRPV2, MAP7, CSF1R, CHST11, TNFAIP8L2, FLI1, ARHGEF6, ZEB2, RCSD1, MEF2C, FMNL3, ARHGAP31, CYRIA, SYNE1, GIMAP8, CREB3L1, ARF4, AGR2, KCNK1, SEC13, BACE2, CD55, KDELR2, S100P, BSN, RUNDC3A, CHGB, SCG3, AP3B2, SYP, CACNA2D2, SEZ6, CELF3, GNG4, FOS, FOSB, ZFP36, DUSP1, CSRNP1, NR4A1, JUNB, EGR3, CCN1, ATF3, COL8A1, MAP1A, PKD2, EDNRA, AEBP1, TIMP2, SYDE1, KANK2, SCARF2, DDR2, SEC61A1, COPG1, SRPRB, TFG, P4HB, COPB2, UMPS, TMEM39A, RUVBL1 and PDIA5, respectively. The 108 genes significantly related to 11-DMG mutation obtained from the cBioportal were used through the Meatascape website to perform GO and KEGG enrichment analysis (**Figures 10A–C**). GO enrichment was divided into three functional groups: biological processes (15 items), molecular functions (1 item), and cellular components (2 items), and KEGG functional group (2 items). We found that these genes were mainly involved in cellular response to calcium, skeletal muscle cell differentiation, blood vessel development, cellular response to growth factor stimulus, endoplasmic reticulum to Golgi vesicle-mediated transport, peptidyl-serine dephosphorylation, myeloid cell differentiation, transmembrane receptor protein tyrosine kinase signaling pathway, MAPK cascade, placenta blood vessel development, maintenance of protein location, positive regulation of cell-substrate adhesion, positive regulation of phospholipase activity, multicellular organismal movement, positive regulation of cell motility. The

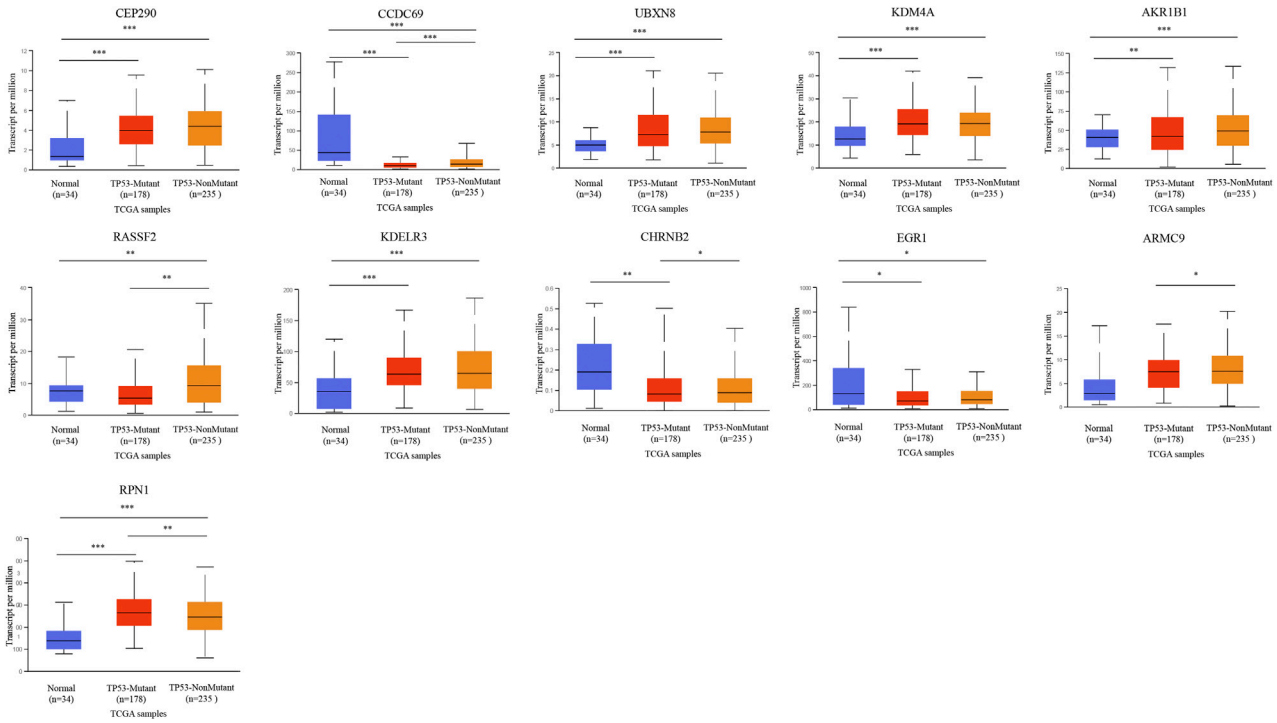


FIGURE 8 | The relationship between 11-DMG mRNA expression levels and TP53 mutation in gastric cancer (GC) (mutation: red, non-mutation: orange, and normal gastric tissues: blue) (UALCAN) (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

molecular function of these genes mainly played a role in the activity of calcium channels. The cellular components involved in these genes were cytoplasmic ribonucleoprotein granules and cytoplasmic regions (**Table 4**). In addition, in order to better understand the relationship between DMG mutation-related genes and GC, we conducted protein interaction network analysis. After pathway and process enrichment analysis for each MCODE component, it was found that the main component of the cell involved was the endoplasmic reticulum lumen, and the biological function was mainly related to COPI-coated vesicle membrane, endoplasmic reticulum to Golgi vesicle-mediated transport, COPI-coated vesicle, P-body, nuclear-transcribed mRNA catabolic process, mRNA catabolic process (**Figures 10D–E**).

Construction of Multi-Factor Regulatory Network of Key Genes

Using databases such as Starbase, TargetScan and other databases to predict the miRNAs upstream regulated of 11 key genes, and intersect the prediction results, a total of 90 reliable miRNAs capable of regulating 11 mRNAs were obtained. By predicting the upstream of reliable miRNA regulated lncRNAs through the Starbase database to, a total of 2,469 lncRNAs were obtained, and the most reliable first three lncRNAs were selected for each miRNA, and finally 270 credible lncRNAs were obtained. The TRRUST database predicted transcription factors that can regulate 11 key genes, and 13 TFs were obtained. Finally, the regulatory network between mRNA, miRNA, lncRNA and TF was constructed (**Figure 11**).

DISCUSSION

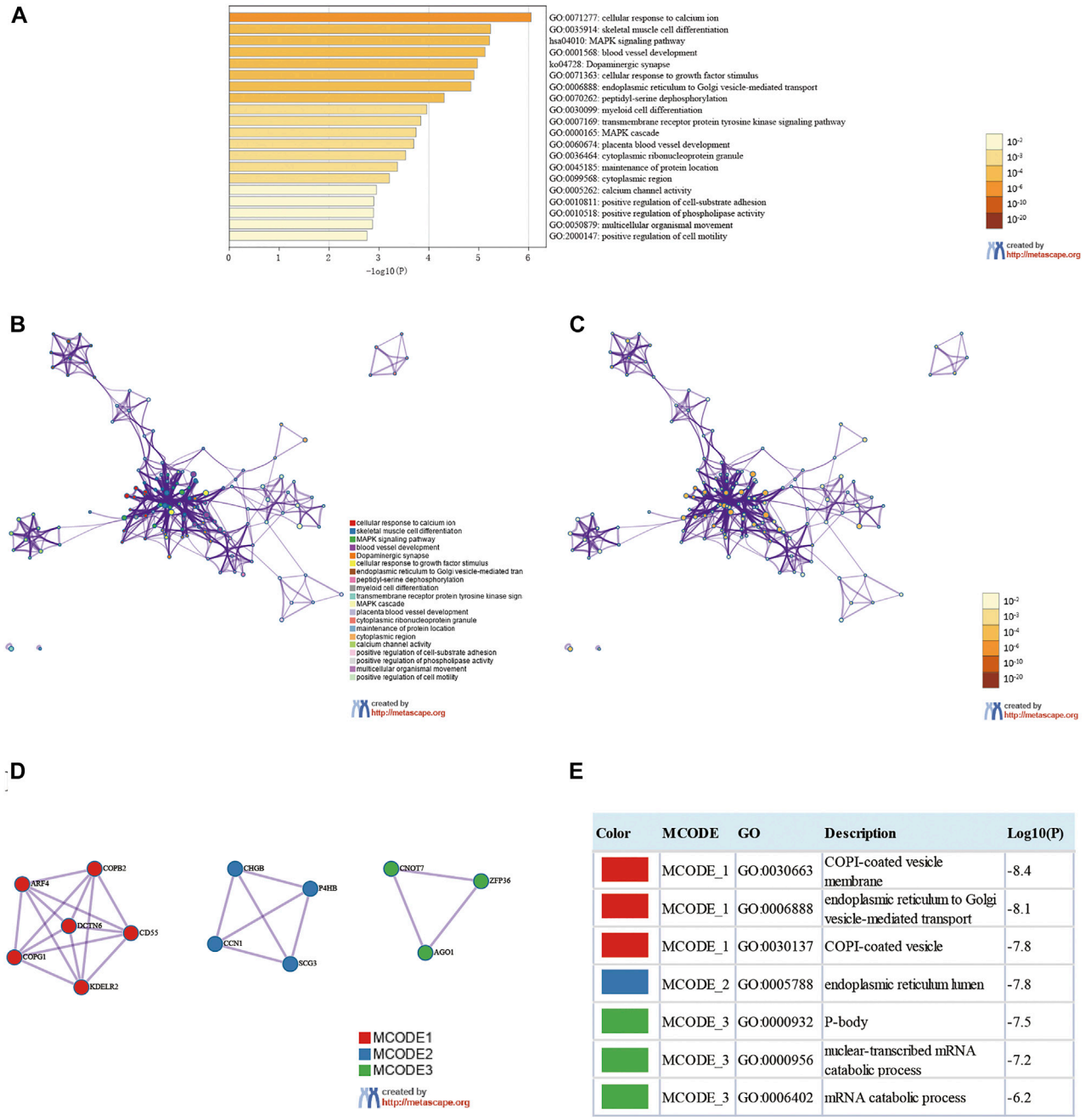
Although tumor markers for different types of cancers have been rapidly discovered in recent years, there remains a lack of specific and sensitive tumor markers for the management of GC. With the development and deeper understanding of epigenetics, abnormal DNAm has become the most extensively studied epigenetic mechanism in GC research, and the relationship between DNAm and tumors has become a research hotspot. The mechanism whereby DNAm promotes cancer may be related to activation or inhibition of certain signaling pathways, and DNAm is thus recognized as a potential tumor marker (Rashid and Issa, 2004). However, the performance of a single DNAm site in predicting the prognosis of GC is unreliable. A large prospective trial with 7,941 patients with colorectal cancer was conducted to evaluate the accuracy of screening circulating DNAm by detecting the methylation level of SEPT9. The results revealed a specificity of 91.5% but a sensitivity of only 48.2% (Church et al., 2014). Some studies have shown that the prediction accuracy of GC models is improved by combining multiple tumor markers (Li et al., 2020a; Bai et al., 2020). This is because multiple markers can take advantage of the complementary effects of genetic information and effectively eliminate redundant genes through machine learning

FIGURE 9 | Mutation of 11-DMG in gastric cancer (GC) patients (cBioportal). **(A)** A visual summary of 11-DMG mutation frequency. **(B)** Summary of mutation frequency of 11-DMG in gastric cancer patients. **(C)** Kaplan-Meier plotter was used to compare the relationship between gene mutation (red) and gene non-mutation (blue) of 11-DMG mutation with OS and PFS ($p < 0.05$ statistical significance).

algorithms. As a result, we developed a GC diagnostic model with a 5-DMS signature and a GC prognostic model with an 11-DMS signature. Through clinical correlation analysis of the diagnostic models, independent prognostic factors analysis of prognostic models and enrichment analysis of the high-risk prognostic risk score group, our study provides potential targets and related mechanisms for clinical diagnosis and treatment of GC.

The accuracy of a DNAm diagnostic model has been confirmed for liver cancer (Luo et al., 2020). In the current study, we developed a 5-DMS (NPAS2, DAPK1, CNN3, FGFR2, PLEKHA5) signature diagnostic model and calculated GC diagnostic risk scores to accurately distinguish GC from normal tissues. The predicted results were highly consistent with the actual results, indicating the model's potential for wide application. In addition, unsupervised hierarchical clustering analysis demonstrated high specificity and sensitivity. In subsequent analysis, the diagnostic risk score was significantly correlated with grade and tumor site in patients with GC. Since the disease state of gastric cancer patients is often manifested in clinical characteristics, the correlation analysis between the risk score calculated by this diagnostic model and the clinical characteristics can further understand the quality of our model and assess the clinical status of GC patients, which is of great significance. In clinical practice, the gold standard for GC diagnosis is pathological results, but the diagnostic model still has high clinical value. At the same time, this model and pathology are used for diagnosis. If the two diagnostic results are consistent, it is more convincing. Generally, pathological diagnosis is the main method, and model diagnosis is the auxiliary method. In addition, the model can assist in the diagnosis and classification of patients with difficult pathological diagnosis, and can also be used for the detection of tumor residual, recurrence and metastasis for subsequent accurate and personalized treatment.

The prognostic model constructed in the current study employed an 11-DMS (CEP290, CCDC69, UBXN8, KDM4A,



Color	MCODE	GO	Description	Log10(P)
	MCODE_1	GO:0030663	COPI-coated vesicle membrane	-8.4
	MCODE_1	GO:0006888	endoplasmic reticulum to Golgi vesicle-mediated transport	-8.1
	MCODE_1	GO:0030137	COPI-coated vesicle	-7.8
	MCODE_2	GO:0005788	endoplasmic reticulum lumen	-7.8
	MCODE_3	GO:0000932	P-body	-7.5
	MCODE_3	GO:0000956	nuclear-transcribed mRNA catabolic process	-7.2
	MCODE_3	GO:0006402	mRNA catabolic process	-6.2

FIGURE 10 | Enrichment analysis of genes related to 11-DMG mutation in gastric cancer (GC) (Metascape). **(A)** Heat maps of Go and KEGG enrichment analysis of 108 adjacent genes related to 11-DMG mutation were stained with *p*-value. **(B)** Term-enriched network: colored by cluster ID, where nodes sharing the same cluster ID are usually close to each other, **(C)** colored by *p*-value, terms containing more genes tend to have more significant *p*-values. **(D)** For the MCODE components identified in the protein-protein interaction network, **(E)** the three best score items divided by *p*-value are used as the functional description of the corresponding components, which are represented by the grid diagram.

AKR1B1, RASSF2, KDELR3, CHRNB2, EGR1, ARMC9, and RPN1) signature. In this model, prognostic risk score effectively distinguished patients with GC into high-risk and low-risk groups. Kaplan–Meier curves also confirmed that the survival rate of patients in the high-risk group was significantly lower than that in the low-risk group. By univariate and multivariate Cox analyses, prognostic risk score was proven to be an independent prognostic risk factor for GC. Compared with other clinical factors (age, gender, tumor grade, clinical stage, T, N, and M stage, race, tumor location), prognostic risk score had higher predictive potential, which indicated the reliability of the model for predicting the prognosis of patients with GC. Although

TABLE 4 | The GO and KEGG function enrichment analysis of genes related to 11-DMG mutation in GC.

GO	Category	Description	Count	%	$Log_{10}(P)$	$Log_{10}(q)$
GO:0071277	GO Biological Processes	cellular response to calcium ion	6	5.61	-6.05	-1.71
GO:0035914	GO Biological Processes	skeletal muscle cell differentiation	5	4.67	-5.24	-1.69
hsa04010	KEGG Pathway	MAPK signaling pathway	8	7.48	-5.21	-1.69
GO:0001568	GO Biological Processes	blood vessel development	13	12.15	-5.13	-1.69
ko04728	KEGG Pathway	Dopaminergic synapse	6	5.61	-4.97	-1.69
GO:0071363	GO Biological Processes	cellular response to growth factor stimulus	12	11.21	-4.90	-1.69
GO:0006888	GO Biological Processes	endoplasmic reticulum to Golgi vesicle-mediated transport	6	5.61	-4.84	-1.68
GO:0070262	GO Biological Processes	peptidyl-serine dephosphorylation	3	2.80	-4.31	-1.24
GO:0030099	GO Biological Processes	myeloid cell differentiation	8	7.48	-3.96	-1.01
GO:0007169	GO Biological Processes	transmembrane receptor protein tyrosine kinase signaling pathway	10	9.35	-3.84	-0.92
GO:0000165	GO Biological Processes	MAPK cascade	11	10.28	-3.74	-0.89
GO:0060674	GO Biological Processes	placenta blood vessel development	3	2.80	-3.70	-0.86
GO:0036464	GO Cellular Components	cytoplasmic ribonucleoprotein granule	6	5.61	-3.54	-0.71
GO:0045185	GO Biological Processes	maintenance of protein location	4	3.74	-3.37	-0.62
GO:0099568	GO Cellular Components	cytoplasmic region	6	5.61	-3.21	-0.53
GO:0005262	GO Molecular Functions	calcium channel activity	4	3.74	-2.95	-0.38
GO:0010811	GO Biological Processes	positive regulation of cell-substrate adhesion	4	3.74	-2.90	-0.36
GO:0010518	GO Biological Processes	positive regulation of phospholipase activity	3	2.80	-2.89	-0.36
GO:0050879	GO Biological Processes	multicellular organismal movement	3	2.80	-2.87	-0.36
GO:2000147	GO Biological Processes	positive regulation of cell motility	8	7.48	-2.76	-0.29

It includes the first 20 clusters and their representative enrichment terms (one for each cluster). "Count" is the number of genes in the provided list that have membership in the given ontology term. "%" is the percentage of all genes provided found in a given ontology term (only input genes with at least one ontology term annotation are included in the calculation). "Log10(P)" is the p *value based on Log10. "Log10(q)" is a multi-test adjusted* p *value based on Log10.*

TNM stage is still the gold standard for the classification and prognosis of GC patients, from the perspective of data analysis, this prognostic model can better reflect the prognosis of gastric cancer patients than TNM stage. With the continuous expansion of subsequent data, the constructed prognostic model will with higher stability and accuracy, it is not impossible to replace TNM stage. In clinical practice, we often encounter GC patients with the same TNM stage and other clinical characteristics, but their prognosis is quite different, and the subsequent treatment plans given are not completely the same. For this situation, we can apply this prognostic model to classify and predict the prognosis, so that doctors can summarize the treatment plans of patients in the high-risk group and the low-risk group, and provide corresponding treatment plans. Therefore, this prognostic model has great potential value in the prognosis judgment and treatment of GC patients, which is helpful for accurate and personalized treatment in the clinical environment.

Among the eleven DMGs in the prognostic model, five DMGs (KDM4A, AKR1B1, RASSF2, CHRNB2, and EGR1) are known to be closely related to the occurrence and development of GC. The protein encoded by the KDM4A gene acts as a trimethylation-specific demethylase, which can specifically demethylate the "Lys-9" and "Lys-36" residues of histone H3, thereby playing a central role in coding for histones (Bavetsias et al., 2016). This protein can also control the growth and invasion of GC cells by inhibiting the KDM4A/YAP1 pathway (Chen et al., 2019). The AKR1B1 gene encodes a member of the aldose/keto reductase superfamily, which is composed of more than 40 known enzymes and proteins. The related pathways include acetone degradation I (conversion to methylglyoxal) and glycerolipid metabolism (Sivenius et al., 2004; Wolford et al., 2006). AKR1B1 plays an important role in the occurrence and development of GC, which had a certain reference value for the prognosis of patients with GC (Li et al., 2020b). The protein encoded by the RASSF2 gene has been found to be a potential tumor suppressor and can act as a KRAS-specific effector protein. It may promote apoptosis and cell cycle arrest, stabilizing STK3/MST2 by protecting it from proteasome degradation (Cooper et al., 2009). Meta-analysis has shown that RASSF2 is significantly more methylated in GC, which can predict the risk of GC (Zhou et al., 2019c). Neuronal acetylcholine receptors are homo- or heteropentameric complexes composed of homologous α and β subunits, of which the CHRNB2 gene encodes one of several β subunits. The related pathways include nicotine addiction and chemical synaptic transmission (Chen et al., 2009). CHRNB2 and TP53 may also play a role in *Helicobacter* pylori-associated GC, but the specific mechanism is unknown (Hu et al., 2018). The protein encoded by the EGR1 gene belongs to the EGR family of C2H2-type zinc-finger proteins and is a transcriptional regulator (Hu et al., 2010). Its functions are diverse and can regulate the transcription of many target genes, thus, playing an important role in regulating the response to growth factors, DNA damage, and ischemia. Its role in regulating cell survival, proliferation, and cell death cannot be ignored. EGR1 protein can directly bind to the HNF1A-AS1 promoter region and activate its transcription to promote the GC cell cycle (Liu et al., 2018). The relationship between the remaining six DMGs and GC is unknown. Further exploration of the potential functions and mechanisms of these DMGs may deepen our understanding of GC development and provide potential tumor markers.

Regulatory, cytotoxic, and EMT factors are significantly associated with the occurrence, development, and immunity of tumor (Zhou et al., 2019b), and their analysis can further explore potentially important biological phenotypes. Correlation analysis with these three factors revealed that prognostic risk score was significantly positively correlated with VIM. This gene encodes a type III intermediate filament protein responsible for maintaining

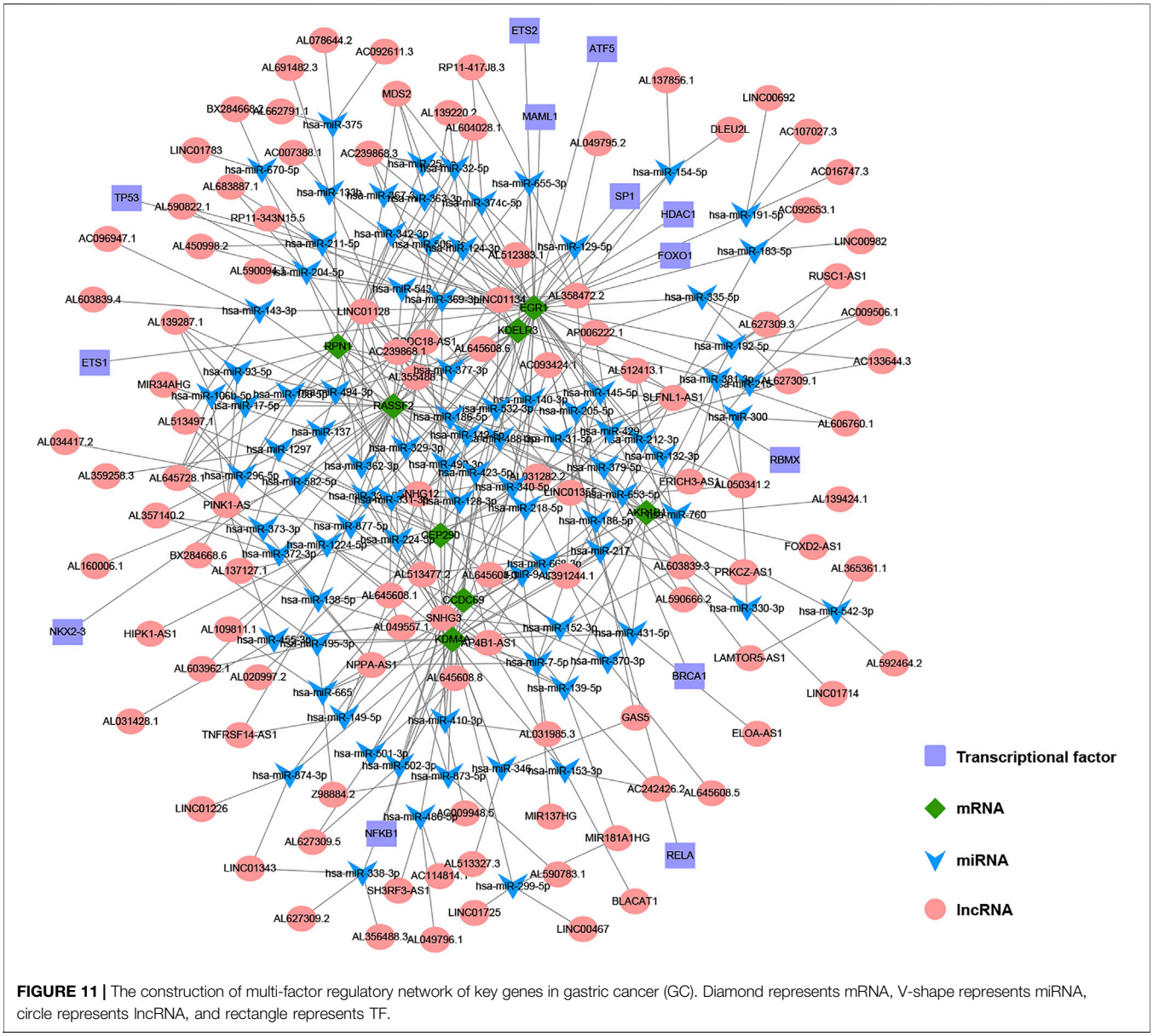


FIGURE 11 | The construction of multi-factor regulatory network of key genes in gastric cancer (GC). Diamond represents mRNA, V-shape represents miRNA, circle represents lncRNA, and rectangle represents TF.

cell shape and cytoplasm integrity and stabilizing cytoskeletal interactions. VIM protein is involved in neurogenesis, cholesterol transport, and functions as an organizer of a number of other critical proteins involved in cell attachment, migration, and signaling. EMT is widespread in malignant tumor cells, of which VIM is a marker gene. The higher the risk score, the more likely EMT will occur. We performed GSEA to clarify the potential mechanisms involved in GC that were identified in the high-risk score group. The differentially expressed genes were mainly distributed in five pathways: "calculation signaling pathway," "cytokine receptor interaction," "focal assignment," "neural ligand receptor interaction," and "regulation of actin cytoskeleton." This indicates that the above pathways may be related to the origin of GC, which concurs with the results of previously published research (Liu et al., 2016; Zhu et al., 2017; Xu et al., 2019; Zhou et al., 2020).

In order to understand the correlation between 11-DMG and TP53 mutation, we analyzed their correlation on the data website through UALCAN. In the analysis, we found for the first time that the expression of CHRNB2 was significantly reduced only in the TP53 mutation group of gastric cancer patients, and the mutation of tumor suppressor gene TP53 may be involved in the regulation of mRNA expression in CCDC69, RASSF2, CHRNB2, ARMC9, and RPN1(Sartorio and Morabito, 1988; Hu et al., 2018; Wang et al., 2020). In the analysis of 11-DMG mutation and prognosis, we found that CEP290, UBXN8, KDM4A, RPN1 had high frequency mutations. The genes related to their mutations are mainly related to pathways such as COPI-coated vesicle membrane, endoplasmic reticulum to Golgi vesicle-mediated transport, COPI-coated vesicle, P-body, nuclear-transcribed mRNA catabolic process, mRNA catabolic process.

To the best of our knowledge, the 5-DMS diagnostic and 11-DMS prognostic models of GC have not been previously reported. The models were verified by external datasets and demonstrated good generalization ability, which can facilitate clinical treatment decision-making. The DMSs selected in this study are relatively novel, and subsequent research on these DMSs will be of great significance. However, this study also has some shortcomings. The small normal sample size may lead to some bias in the results. Other omics fields, such as genome, transcriptome, proteome, and metabolome, have shown respective advantages in GC diagnostic and prognostic models (Li et al., 2010; Chan et al., 2016; Deng et al., 2018; Zhang et al., 2018; Shen et al., 2019); therefore, it is too early to assert that our model is optimal. The models should be validated in a real-world cohort. We hope to address these concerns in our future work.

In conclusion, the GC diagnostic and prognostic models established in the current study are low cost, highly sensitive, specific, and may facilitate accurate and individualized treatment for patients with GC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://xena.ucsc.edu/. https://www.ncbi.nlm.nih.gov/geo/. http://ualcan.path.uab.edu/analysis.html.

AUTHOR CONTRIBUTIONS

Conceptualization, DL; Methodology, LW and GX; Formal Analysis, XH and CW; Investigation, QJ and XW; Writing–Original Draft Preparation, DX; Writing–Review and Editing, LL; Supervision, YL; Project Administration, DX; Funding Acquisition, YL. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 81770634) and Heilongjiang Province General Undergraduate Colleges and Universities Young Innovative Talents Training Plan (UNPYSCT-2018073).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.758926/full#supplementary-material

Supplementary TableS1 | The methylation status of DMS: 1842 hypermethylation sites and 899 hypomethylation sites were screened from 27 normal samples and 443 GC samples.

REFERENCES

Bader, G. D., and Hogue, C. W. (2003). An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics* 4, 2. doi:10.1186/1471-2105-4-2

Bai, Y., Wei, C., Zhong, Y., Zhang, Y., Long, J., Huang, S., et al. (2020). Development and Validation of a Prognostic Nomogram for Gastric Cancer Based on DNA Methylation-Driven Differentially Expressed Genes. *Int. J. Biol. Sci.* 16 (7), 1153–1165. doi:10.7150/ijbs.41587

Bang, Y.-J., Xu, R.-H., Chin, K., Lee, K.-W., Park, S. H., Rha, S. Y., et al. (2017). Olaparib in Combination with Paclitaxel in Patients with Advanced Gastric Cancer Who Have Progressed Following First-Line Therapy (GOLD): a Double-Blind, Randomised, Placebo-Controlled, Phase 3 Trial. *Lancet Oncol.* 18 (12), 1637–1651. doi:10.1016/S1470-2045(17)30682-4

Bavetsias, V., Lanigan, R. M., Ruda, G. F., Atrash, B., McLaughlin, M. G., Tumber, A., et al. (2016). 8-Substituted Pyrido[3,4-D]pyrimidin-4(3h)-One Derivatives as Potent, Cell Permeable, KDM4 (JMJD2) and KDM5 (JARID1) Histone Lysine Demethylase Inhibitors. *J. Med. Chem.* 59 (4), 1388–1409. doi:10.1021/acs.jmedchem.5b01635

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer J. clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Cats, A., Jansen, E. P. M., van Grieken, N. C. T., Sikorska, K., Lind, P., Nordsmark, M., et al. (2018). Chemotherapy versus Chemoradiotherapy after Surgery and Preoperative Chemotherapy for Resectable Gastric Cancer (CRITICS): an International, Open-Label, Randomised Phase 3 Trial. *Lancet Oncol.* 19 (5), 616–628. doi:10.1016/S1470-2045(18)30132-3

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.CD-12-0095

Chan, A. W., Mercier, P., Schiller, D., Bailey, R., Robbins, S., Eurich, D. T., et al. (2016). 1H-NMR Urinary Metabolomic Profiling for Diagnosis of Gastric Cancer. *Br. J. Cancer* 114 (1), 59–62. doi:10.1038/bjc.2015.414

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* 19 (8), 649–658. doi:10.1016/j.neo.2017.05.002

Chen, L.-h., Wang, L.-p., and Ma, X.-q. (2019). Circ_SPECC1 Enhances the Inhibition of miR-526b on Downstream KDM4A/YAP1 Pathway to Regulate the Growth and Invasion of Gastric Cancer Cells. *Biochem. Biophysical Res. Commun.* 517 (2), 253–259. doi:10.1016/j.bbrc.2019.07.065

Chen, Y., Wu, L., Fang, Y., He, Z., Peng, B., Shen, Y., et al. (2009). A Novel Mutation of the Nicotinic Acetylcholine Receptor Gene CHRNA4 in Sporadic Nocturnal Frontal Lobe Epilepsy. *Epilepsy Res.* 83 (2-3), 152–156. doi:10.1016/j.eplepsyres.2008.10.009

Church, T. R., Wandell, M., Lofton-Day, C., Mongin, S. J., Burger, M., Payne, S. R., et al. (2014). Prospective Evaluation of methylatedSEPT9in Plasma for Detection of Asymptomatic Colorectal Cancer. *Gut* 63 (2), 317–325. doi:10.1136/gutjnl-2012-304149

Cooper, W. N., Hesson, L. B., Matallanas, D., Dallol, A., von Kriegsheim, A., Ward, R., et al. (2009). RASSF2 Associates with and Stabilizes the Proapoptotic Kinase MST2. *Oncogene* 28 (33), 2988–2998. doi:10.1038/onc.2009.152

Das, P. M., and Singal, R. (2004). DNA Methylation and Cancer. *Jco* 22 (22), 4632–4642. doi:10.1200/JCO.2004.07.151

Deng, X., Xiao, Q., Liu, F., and Zheng, C. (2018). A Gene Expression-Based Risk Model Reveals Prognosis of Gastric Cancer. *PeerJ* 6, e4204. doi:10.7717/peerj.4204

FDA (2016). Premarket Approval (PMA) for Epi proColon. US Food and Drug Administration. Available at: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma.cfm?id=P130001 (Accessed April 21, 2016).

Fu, D.-G. (2015). Epigenetic Alterations in Gastric Cancer (Review). *Mol. Med. Rep.* 12 (3), 3223–3230. doi:10.3892/mmr.2015.3816

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

Hibi, K., Koike, M., Nakayama, H., Fujitake, S., Kasai, Y., Ito, K., et al. (2003). A Cancer-Prone Case with a Background of Methylation of P16 Tumor Suppressor Gene. *Clin. Cancer Res.* 9 (3), 1053–1056.

Higashimori, A., Dong, Y., Zhang, Y., Kang, W., Nakatsu, G., Ng, S. S. M., et al. (2018). Forkhead Box F2 Suppresses Gastric Cancer through a Novel FOXF2-Irf2bpl-β-Catenin Signaling Axis. *Cancer Res.* 78 (7), 1643–1656. doi:10.1158/0008-5472.CAN-17-2403

Hu, C.-T., Chang, T.-Y., Cheng, C.-C., Liu, C.-S., Wu, J.-R., Li, M.-C., et al. (2010). Snail Associates with EGR-1 and SP-1 to Upregulate Transcriptional Activation of p15INK4b. *FEBS J.* 277 (5), 1202–1218. doi:10.1111/j.1742-4658.2009.07553.x

Hu, Y., He, C., Liu, J. P., Li, N. S., Peng, C., Yang-Ou, Y. B., et al. (2018). Analysis of Key Genes and Signaling Pathways Involved inHelicobacter Pylori-associated Gastric Cancer Based on the Cancer Genome Atlas Database andRNAsequencing Data. *Helicobacter* 23 (5), e12530. doi:10.1111/hel.12530

Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Levin, T. R., Lavin, P., Lidgard, G. P., et al. (2014). Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *N. Engl. J. Med.* 370 (14), 1287–1297. doi:10.1056/NEJMoa1311194

Kurashige, J., Hasegawa, T., Niida, A., Sugimachi, K., Deng, N., Mima, K., et al. (2016). Integrated Molecular Profiling of Human Gastric Cancer Identifies DDR2 as a Potential Regulator of Peritoneal Dissemination. *Sci. Rep.* 6, 22371. doi:10.1038/srep22371

Li, C., Zheng, Y., Pu, K., Zhao, D., Wang, Y., Guan, Q., et al. (2020). A Four-DNA Methylation Signature as a Novel Prognostic Biomarker for Survival of Patients with Gastric Cancer. *Cancer Cel Int.* 20, 88. doi:10.1186/s12935-020-1156-8

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2017). A Scored Human Protein-Protein Interaction Network to Catalyze Genomic Interpretation. *Nat. Methods* 14 (1), 61–64. doi:10.1038/nmeth.4083

Li, X., Yang, J., Gu, X., Xu, J., Li, H., Qian, J., et al. (2020). The Expression and Clinical Significance of Aldo-Keto Reductase 1 Member B1 in Gastric Carcinoma. *DNA Cel Biol.* 39 (7), 1322–1327. doi:10.1089/dna.2020.5550

Li, X., Zhang, Y., Zhang, Y., Ding, J., Wu, K., and Fan, D. (2010). Survival Prediction of Gastric Cancer by a Seven-microRNA Signature. *Gut* 59 (5), 579–585. doi:10.1136/gut.2008.175497

Licchesi, J. D. F., Van Neste, L., Tiwari, V. K., Cope, L., Lin, X., Baylin, S. B., et al. (2010). Transcriptional Regulation of Wnt Inhibitory Factor-1 by Miz-1/c-Myc. *Oncogene* 29 (44), 5923–5934. doi:10.1038/onc.2010.322

Liu, H.-T., Liu, S., Liu, L., Ma, R.-R., and Gao, P. (2018). EGR1-mediated Transcription of lncRNA-HNF1A-AS1 Promotes Cell Cycle Progression in Gastric Cancer. *Cancer Res.* 78 (20), 5877. doi:10.1158/0008-5472.CAN-18-1011

Liu, J.-j., Liu, J.-y., Chen, J., Wu, Y.-x., Yan, P., Ji, C.-d., et al. (2016). Scinderin Promotes the Invasion and Metastasis of Gastric Cancer Cells and Predicts the Outcome of Patients. *Cancer Lett.* 376 (1), 110–117. doi:10.1016/j.canlet.2016.03.035

Luo, H., Zhao, Q., Wei, W., Zheng, L., Yi, S., Li, G., et al. (2020). Circulating Tumor DNA Methylation Profiles Enable Early Diagnosis, Prognosis Prediction, and Screening for Colorectal Cancer. *Sci. Transl. Med.* 12 (524), eaax7533. doi:10.1126/scitranslmed.aax7533

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., et al. (2017). Intragenic DNA Methylation Prevents Spurious Transcription Initiation. *Nature* 543 (7643), 72–77. doi:10.1038/nature21373

Nishigaki, M., Aoyagi, K., Danjoh, I., Fukaya, M., Yanagihara, K., Sakamoto, H., et al. (2005). Discovery of Aberrant Expression of R-RAS by Cancer-Linked DNA Hypomethylation in Gastric Cancer Using Microarrays. *Cancer Res.* 65 (6), 2115–2124. doi:10.1158/0008-5472.CAN-04-3340

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., et al. (2019). The BioGRID Interaction Database: 2019 Update. *Nucleic Acids Res.* 47 (D1), D529–D541. doi:10.1093/nar/gky1079

Rashid, A., and Issa, J. P. J. (2004). CpG Island Methylation in Gastroenterologic Neoplasia: a Maturing Field. *Gastroenterology* 127 (5), 1578–1588. doi:10.1053/j.gastro.2004.09.007

Sakakura, C., Hasegawa, K., Miyagawa, K., Nakashima, S., Yoshikawa, T., Kin, S., et al. (2005). Possible Involvement of RUNX3 Silencing in the Peritoneal Metastases of Gastric Cancers. *Clin. Cancer Res.* 11 (18), 6479–6488. doi:10.1158/1078-0432.CCR-05-0729

Sartorio, A., and Morabito, F. (1988). The Disability of Short Stature. *Arch. Dis. Child.* 63 (2), 222. doi:10.1136/adc.63.2.222-a

Shannon, P., Markiel, A., and Ozier, O. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shen, Q., Polom, K., Williams, C., de Oliveira, F. M. S., Guergova-Kuras, M., Lisacek, F., et al. (2019). A Targeted Proteomics Approach Reveals a Serum Protein Signature as Diagnostic Biomarker for Resectable Gastric Cancer. *EBioMedicine* 44, 322–333. doi:10.1016/j.ebiom.2019.05.044

Sivenius, K., Niskanen, L., Voutilainen-Kaunisto, R., Laakso, M., and Uusitupa, M. (2004). Aldose Reductase Gene Polymorphisms and Susceptibility to Microvascular Complications in Type 2 Diabetes. *Diabet Med.* 21 (12), 1325–1333. doi:10.1111/j.1464-5491.2004.01345.x

Sundar, R., Huang, K. K., Qamra, A., Kim, K.-M., Kim, S. T., Kang, W. K., et al. (2019). Epigenomic Promoter Alterations Predict for Benefit from Immune Checkpoint Inhibition in Metastatic Gastric Cancer. *Ann. Oncol.* 30 (3), 424–430. doi:10.1093/annonc/mdy550

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

Vasiljević, N., Ahmad, A. S., Thorat, M. A., Fisher, G., Berney, D. M., Møller, H., et al. (2014). DNA Methylation Gene-Based Models Indicating Independent Poor Outcome in Prostate Cancer. *BMC cancer* 14, 655. doi:10.1186/1471-2407-14-655

Wang, K., Li, L., Fu, L., Yuan, Y., Dai, H., Zhu, T., et al. (2019). Integrated Bioinformatics Analysis the Function of RNA Binding Proteins (RBPs) and Their Prognostic Value in Breast Cancer. *Front. Pharmacol.* 10, 140. doi:10.3389/fphar.2019.00140

Wang, K., Liang, Q., Li, X., Tsoi, H., Zhang, J., Wang, H., et al. (2016). MDGA2 Is a Novel Tumour Suppressor Cooperating with DMAP1 in Gastric Cancer and Is Associated with Disease Outcome. *Gut* 65 (10), 1619–1631. doi:10.1136/gutjnl-2015-309276

Wang, X., Duanmu, J., Fu, X., Li, T., and Jiang, Q. (2020). Analyzing and Validating the Prognostic Value and Mechanism of colon Cancer Immune Microenvironment. *J. Transl Med.* 18 (1), 324. doi:10.1186/s12967-020-02491-w

Wolford, J. K., Yeatts, K. A., Eagle, A. R. R., Nelson, R. G., Knowler, W. C., and Hanson, R. L. (2006). Variants in the Gene Encoding Aldose Reductase (AKR1B1) and Diabetic Nephropathy in American Indians. *Diabet Med.* 23 (4), 367–376. doi:10.1111/j.1464-5491.2006.01834.x

Xu, L., Li, X., Chu, E. S. H., Zhao, G., Go, M. Y. Y., Tao, Q., et al. (2012). Epigenetic Inactivation ofBCL6B, a Novel Functional Tumour Suppressor for Gastric Cancer, Is Associated with Poor Survival. *Gut* 61 (7), 977–985. doi:10.1136/gutjnl-2011-300411

Xu, R.-h., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., et al. (2017). Circulating Tumour DNA Methylation Markers for Diagnosis and Prognosis of Hepatocellular Carcinoma. *Nat. Mater* 16 (11), 1155–1161. doi:10.1038/nmat4997

Xu, Z., Li, Z., Wang, W., Xia, Y., He, Z., Li, B., et al. (2019). MIR-1265 Regulates Cellular Proliferation and Apoptosis by Targeting Calcium Binding Protein 39 in Gastric Cancer and, Thereby, Impairing Oncogenic Autophagy. *Cancer Lett.* 449, 226–236. doi:10.1016/j.canlet.2019.02.026

Yu, J., Cheng, Y. Y., Tao, Q., Cheung, K. F., Lam, C. N. Y., Geng, H., et al. (2009). Methylation of Protocadherin 10, a Novel Tumor Suppressor, Is Associated with Poor Prognosis in Patients with Gastric Cancer. *Gastroenterology* 136 (2), 640–651. doi:10.1053/j.gastro.2008.10.050

Zhang, C., Zhang, B., Meng, D., and Ge, C. (2019). Comprehensive Analysis of DNA Methylation and Gene Expression Profiles in Cholangiocarcinoma. *Cancer Cel Int.* 19, 352. doi:10.1186/s12935-019-1080-y

Zhang, Y., Li, H., Zhang, W., Che, Y., Bai, W., and Huang, G. (2018). LASSO-based Cox-PH M-odel I-dentifies an 11-lncRNA S-ignature for P-rognosis P-rediction in G-astric C-ancer. *Mol. Med. Rep.* 18 (6), 5579–5593. doi:10.3892/mmr.2018.9567

Zhou, K., Cai, C., He, Y., Zhou, C., Zhao, S., Ding, X., et al. (2019). Association between RASSF2 Methylation and Gastric Cancer: A PRISMA-Compliant Systematic Review and Meta-Analysis. *DNA Cel Biol.* 38 (10), 1147–1154. doi:10.1089/dna.2019.4922

Zhou, Q., Wu, X., Wang, X., Yu, Z., Pan, T., Li, Z., et al. (2020). The Reciprocal Interaction between Tumor Cells and Activated Fibroblasts Mediated by TNF-

α/IL-33/ST2L Signaling Promotes Gastric Cancer Metastasis. *Oncogene* 39 (7), 1414–1428. doi:10.1038/s41388-019-1078-x

Zhou, R., Zhang, J., Zeng, D., Sun, H., Rong, X., Shi, M., et al. (2019). Immune Cell Infiltration as a Biomarker for the Diagnosis and Prognosis of Stage I-III colon Cancer. *Cancer Immunol. Immunother.* 68 (3), 433–442. doi:10.1007/s00262-018-2289-7

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10 (1), 1523. doi:10.1038/s41467-019-09234-6

Zhu, M., Wang, H., Cui, J., Li, W., An, G., Pan, Y., et al. (2017). Calcium-binding Protein S100A14 Induces Differentiation and Suppresses Metastasis in Gastric Cancer. *Cell Death Dis.* 8 (7), e2938. doi:10.1038/cddis.2017.297

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

The reviewer YG declared a shared affiliation, with the authors LL, DX, LW, CW, XW, and GX to the handling editor at the time of the review

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

GLOSSARY

GC gastric cancer
DNAm DNA methylation
DMS DNA methylation sites
DMG DNA methylation driver gene
TCGA the cancer genome atlas
GEO gene expression omnibus
MsigDB molecular signatures database
LASSO least absolute shrinkage and selection operator
FDR false discovery rate
GSEA gene set enrichment analysis
ROC receiver operating characteristic
NPAS2 neuronal PAS domain protein 2
DAPK1 death associated protein kinase 1
CNN3 calponin 3
FGFR2 fibroblast growth factor receptor 2
PLEKHA5 pleckstrin homology domain containing A5
CEP290 centrosomalprotein290
CCDC69 coiled-coil domain containing 69
UBXN8 UBX domain protein 8
KDM4A lysine demethylase 4A
AKR1B aldo-keto reductase family 1 member B
RASSF2 ras association domain family member 2
KDELR3 KDEL endoplasmic reticulum protein retention receptor 3
CHRNB2 cholinergic receptor nicotinic beta 2 subunit
EGR1 early growth response 1
ARMC9 armadillo repeat containing 9
RPN1 ribophorin I
PDCD1 programmed cell death 1
CTLA4 cytotoxic T-lymphocyte associated protein 4
LAG3 lymphocyte activating 3
TIGIT T cell immunoreceptor with Ig and ITIM domains
GZMB granzyme B
TNF tumor necrosis factor
EMT epithelial-mesenchymal transition
CDH1 cadherin 1
TF transcription factors.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 November 2021
doi: 10.3389/fgene.2021.762160

USH2A Mutation is Associated With Tumor Mutation Burden and Antitumor Immunity in Patients With Colon Adenocarcinoma

Yuanyuan Sun†, Long Li†, Wenchao Yao, Xuxu Liu, Yang Yang, Biao Ma and Dongbo Xue**

Laboratory of Hepatosplenic Surgery, Department of General Surgery, Ministry of Education, The First Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Xinyi Liu,
University of Illinois at Chicago, United States

Reviewed by:
Xinting Pan,
The Affiliated Hospital of Qingdao University, China
Dechao Bu,
Institute of Computing Technology (CAS), China

***Correspondence:**
Biao Ma
mabiaohero@126.com
Dongbo Xue
xuedongbo@hrbmu.edu.cn

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

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

Received: *21 August 2021*
Accepted: *14 October 2021*
Published: *02 November 2021*

Citation:
Sun Y, Li L, Yao W, Liu X, Yang Y, Ma B and Xue D (2021) USH2A *Mutation is Associated With Tumor Mutation Burden and Antitumor Immunity in Patients With Colon Adenocarcinoma. Front. Genet. 12:762160. doi: 10.3389/fgene.2021.762160*

Colon adenocarcinoma (COAD) is one of the diseases with the highest morbidity and mortality in the world. At present, immunotherapy has become a valuable method for the treatment of COAD. Tumor mutational burden (TMB) is considered to be the most common biomarker for predicting immunotherapy. According to reports, the mutation rate of COAD ranks third. However, whether these gene mutations are related to TMB and immune response is still unknown. Here, COAD somatic mutation data were downloaded from The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) databases. Bioinformatics methods were used to study the relationships among gene mutations, COAD survival prognosis, and tumor immune response. A total of 22 of the top 40 mutations in TCGA and ICGC databases were the same. Among them, the *USH2A* mutation was associated with high TMB and poor clinical prognosis. According to Gene Set Enrichment Analysis (GSEA) and the CIBERSORT algorithm, we determined that the *USH2A* mutation upregulates signaling pathways involved in the immune system and the antitumor immune response. In cases with a *USH2A* mutation, the immune score and MSI score of TCGA samples increased, the expression of immune checkpoint genes decreased significantly, and the TIDE score decreased significantly. Dependent on the presence or absence of a *USH2A* mutation, TCGA COAD samples were analyzed for differentially expressed genes, 522 of which were identified. Using a univariate Cox analysis and LASSO COX analysis of these differential genes, a prediction model was established, which established significant differences in the infiltration of immune cells, immune checkpoint gene expression, immune score, MSI score, TMB, and TIDE in patients in high- and low-risk groups. In conclusion, mutation of *USH2A* is frequent in COAD and is related to an increase in TMB and the antitumor immunity. The differential genes screened by *USH2A* mutation allowed the construction of a risk model for predicting the survival and prognosis of cancer patients, in addition to providing new ideas for COAD immunotherapy.

Keywords: colon adenocarcinoma, *USH2A*, tumor mutation burden, immunotherapy response, bioinformatics analysis

1 INTRODUCTION

Colon cancer is the third leading cause of cancer deaths, with more than one million new cases diagnosed each year (Labianca et al., 2010). COAD is the main pathological type of colon cancer. The incidence of COAD is mainly related to age and eating habits, and partly related to genetic diseases (Cunningham et al., 2010; Watson and Collins, 2011). COAD is heterogeneous, and there are significant differences in mutation patterns across different patients (Punt et al., 2017). Increasing evidence has shown that COAD is a molecular heterogeneous disease that contains a series of genetic changes (Choi et al., 2015). Mutations in key genes can affect tumor cell proliferation, differentiation, apoptosis, viability, and distant metastasis (The Cancer Genome Atlas Network, 2012). Surgery combined with postoperative chemotherapy is currently the main treatment for COAD. Although current treatment methods including chemotherapy and surgery have improved the survival rate of COAD patients, the prognosis of COAD patients is still poor (Neri et al., 2010; Roncucci and Mariani, 2015). The use of reliable biomarkers and the timely diagnosis of treatment targets can significantly improve the mortality of COAD patients and reduce the incidence of COAD (Herzig and Tsikitis, 2015; Tsimberidou, 2015). The immune system plays an important role in the occurrence and development of cancer (Patel and Minn, 2018). The 2020 ESMO clinical practice guidelines for colon cancer recommend the use of immune scores to improve the prognosis of colon cancer (Argilés et al., 2020). Therefore, it is necessary to study the relationship between specific genetic variants and immune events, as well as alternative methods of treating patients with different genetic characteristics. The accumulation of somatic mutations is one of the main causes of tumors and contributes to the expression of neoantigens (Gubin et al., 2015). Studies have shown that TMB is correlated with immunotherapy response (Goodman et al., 2017). It was reported that a high TMB can predict the prognosis of non-small-cell lung cancer and melanoma (Chen et al., 2019a; Chen et al., 2019b). Furthermore, TMB is considered to be a predictive biomarker of tumor behavior and immune response (Goodman et al., 2017).

Immune checkpoint blocking therapy (ICB), which targets programmed cell death ligand 1 (*PDL1*) and cytotoxic T lymphocyte antigen 4 (*CTLA4*) pathways, has become a treatment strategy for various types of cancer (Long et al., 2017; Zhang et al., 2021). TMB is an indicator that is independent of the expression level of *PDL1* and can better indicate the response to ICB treatment (Hodges et al., 2017; Rizvi et al., 2018). A comprehensive analysis of 27 cancer types reported that TMB is associated with better ICB treatment effects (Yarchoan et al., 2017). At present, the proportion of patients benefiting from ICB treatment in clinical practice is still very low, and new biomarkers that predict the ICB response rate of patients need to be developed (Ancevski Hunter et al., 2018; Janjigian et al., 2018). The Tumor Immune Dysfunction and Exclusion (TIDE) algorithm is a calculation method that uses gene expression profiles to predict the ICB response in non-small-cell lung cancer and melanoma (Jiang et al., 2018). TIDE uses a set of gene expression markers to estimate two different mechanisms of tumor immune evasion, including tumor-infiltrating cytotoxic T lymphocyte (CTL) dysfunction and immunosuppressive factor rejection of CTL. A higher TIDE score denotes a higher chance of antitumor immune escape and a lower response rate of ICB therapy (Jiang et al., 2018). TIDE score is more accurate than *PDL1* expression level and TMB in predicting the survival and prognosis of cancer patients treated with ICB (Jiang et al., 2018; Kaderbhaï et al., 2019; Keenan et al., 2019; Wang et al., 2019b). Several recent studies have reported its use in predicting or evaluating the effects of ICB treatment (Bretz et al., 2019; George et al., 2019; Liu et al., 2019; Pallocca et al., 2019; Wang et al., 2019b). At present, whether gene mutations are related to the COAD immune response and ICB treatment response remains unclear.

In this study, we used The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) databases to identify somatic mutations in COAD patients in the United States and China. Then, we identified common mutant genes in both cohorts, which were found to be related to TMB and prognosis, thus confirming that gene mutations are related to

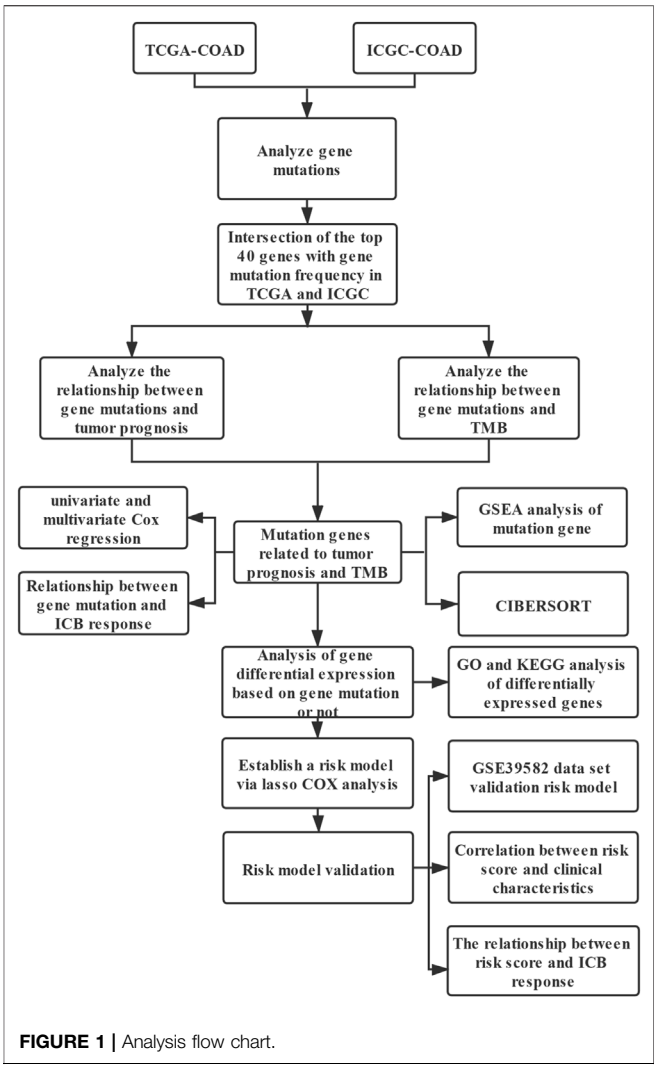


FIGURE 1 | Analysis flow chart.

immune response and ICB treatment response. On the basis of their differential expression caused by mutations, we constructed a prognostic model composed of two genes with a predictive effect on tumor prognosis and ICB treatment response. These findings reveal that a gene mutation can be used as a biomarker for predicting immune response and for evaluating the response to ICB treatment in patients with COAD.

2 MATERIALS AND METHODS

2.1 Data Collection

We used a method similar to that of Gongmin Zhu et al (2020). As shown in the flowchart (**Figure 1**), we downloaded transcriptome data (n = 444), clinical data (n = 336), and somatic gene mutation data (n = 398) from TCGA database (http://portal.gdc.cancer.gov/projects) (data updated on 29 October 2020). For clinical data, patients with COAD were included only when their clinical information was complete, and patients without survival time, survival status, age, gender, grade, or TNM classification data were not included. Next, the somatic gene mutation data of Chinese COAD patients (n = 305) was downloaded from the ICGC database (http://dcc.icgc.org/releases/current/Projects) (data updated on 27 November 2019), and the COAD dataset GSE39582 was downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/).

2.2 Bioinformatics Analysis

PERL software (version 5.32) was used to extract and sort TCGA transcription data, somatic mutation data, clinical data, ICGC mutation data, and GEO transcription data for subsequent analysis. R software (version 4.0.3) package GenVisR was used to analyze and visualize the MAF file of Varscn, the somatic mutation data of colon cancer in the TCGA database. R software package GenVisR was used to analyze and visualize the colon cancer somatic mutation data TSV file of the ICGC database based on the hg19 genome reference information. The venn package in R software was used to take the intersection of the top 40 genes in the TCGA and ICGC datasets with mutation evaluation rates, followed by obtaining the intersection genes with the top mutation frequencies in both databases. Next, R software package ggpubr was used to analyze the relationship between gene mutations and TMB. The Kaplan–Meier (KM) method was used to analyze the relationship between gene mutation and survival prognosis. Univariate and multivariate Cox methods were used to analyze the relationships among patient clinical information (age, gender, tumor stage, and TNM classification), TMB, gene mutations, and tumor survival prognosis. In all comparisons, a p-value < 0.05 was considered statistically significant. Software GSEA (version 4.1.0) was used for gene enrichment analysis. According to the gene mutations, TCGA expression data were divided into two groups: mutation and wild-type. The arrangement was set to 1,000, and the standardized enrichment score (NES) was applied using an FDR q-value < 0.05 as the significance threshold for enrichment (Subramanian et al., 2005). The edge R package was used to analyze the differentially expressed genes between the gene mutant group and the unmutated group (wild-type group). In the analysis process, genes were considered significantly differentially expressed for a p-value < 0.05 and a fold-change (FC) difference >2 (i.e., absolute value of $\log_2$ FC > 1). The enrichment analysis tool DAVID (Da et al., 2009) was used to analyze the Gene Ontology (GO) (Ashburner et al., 2000) functions and KEGG (Minoru and Susumu, 2000) pathways involved in upregulated and downregulated genes (number of parameter-enriched genes ≥2, p-value of hypergeometric test <0.05). R software was used to perform KM survival analysis and univariate Cox analysis of the differential genes using a p-value < 0.05 as the filter value. R software package glmnet was used to perform LASSO COX regression analysis and construct a prognosis-related risk model. To evaluate the risk model, R software package survival ROC was used to analyze the prediction accuracy of the model, and univariate and multivariate Cox analyses were used to evaluate whether the risk score of the tumor patient model could be used as an independent prognostic factor.

2.3 Tumor Mutation Burden and Evaluation of Microsatellite Instability

Tumor mutation burden (TMB) refers to the total number of gene coding errors, base substitutions, and gene insertion or deletion errors per megabit (Mb) of tumor tissue. All base substitutions and insertions in the coding region of the target gene are counted, whereas silent mutations that cannot cause amino-acid changes are not counted. The total number of mutations counted was divided by the exome size (the estimated value of the exome size was 38 Mb) to calculate the TMB score for each sample (Chalmers et al., 2017). A microsatellite is defined as a region of 10–60 base pairs containing 1–5 repeated base-pair motifs (Shia, 2015). Nucleotides in repetitive DNA fragments are spontaneously lost or repeated to form microsatellite instabilities (MSIs) (de la Chapelle and Hampel, 2010). According to the methods of Bonneville et al., the MSI score of COAD samples was characterized (Bonneville et al., 2017).

2.4 Tumor Immune Cell Infiltration Analysis

CIBERSORT is a deconvolution algorithm that can evaluate the proportion of 22 tumor-infiltrating lymphocyte subsets in a large number of tumor samples (Newman et al., 2015). This algorithm is used to evaluate the relative abundance of immune cell infiltration in tumor tissues. The number of permutations was set to 1,000, and a p-value < 0.05 was used as the basis for the successful calculation of the sample.

2.5 Prediction of ICB Treatment Response

R package estimate was used to calculate the immune score of the tumor sample. The Tumor Immune Dysfunction and Exclusion (TIDE) algorithm is a calculation method that uses gene expression profiles to predict immune checkpoint blockade (ICB) responses in non-small-cell lung cancer and melanoma (Jiang et al., 2018). Accordingly, it was used to predict the potential ICB response (Jiang et al., 2018).

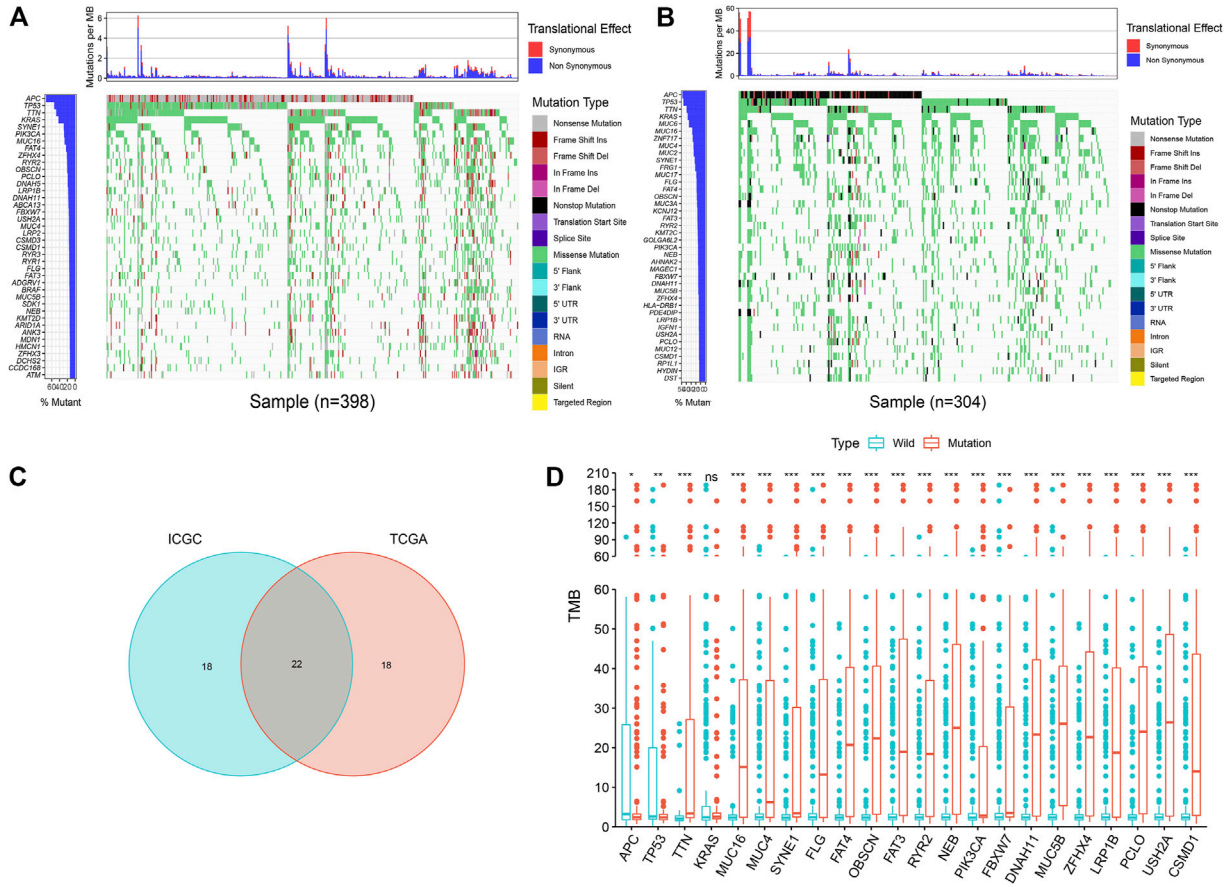


FIGURE 2 | Overview of frequently mutated genes and TMB in COAD. **(A)** Waterfall plot shows the frequently mutated genes in COAD from TCGA database. The left panel shows mutation frequency, and genes are ordered by their mutation frequencies. The right panel presents different mutation types. **(B)** Waterfall plot displaying the frequently mutated genes in COAD from the ICGC cohort. The left panel shows the genes ordered by their mutation frequencies. The right panel presents different mutation types. **(C)** Venn diagram shows 22 frequently mutated genes covered by both the TCGA and ICGC cohorts. **(D)** 21 genes with high mutation frequency are associated with a higher TMB. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns has no statistical difference.

2.6 Data Analysis

R software (version 4.0.3) was used for statistical analysis and graphing. The logrank test was used for KM survival analysis, and the Mann–Whitney U test was used for analysis of the relationship between gene mutation and TMB. In all comparisons, a p-value < 0.05 was considered statistically significant.

3 RESULTS

3.1 COAD Somatic Mutations

The analysis found that, in the mutation data of TCGA samples, the top five most frequently mutated genes were *APC*, *TP53*, *TTN*, *KRAS*, and *SYNE1* (**Figure 2A**). In the mutation data of ICGC samples, the top five most frequently mutated genes were *APC*, *TP53*, *TTN*, *KRAS*, and *MUC6* (**Figure 2B**); thus, we identified some genes with high mutation frequencies in both databases. Therefore, we selected the top 40 genes in both databases, whereby we found an overlap of 22 genes, as depicted in a Venn diagram (**Figure 2C**).

3.2 *USH2A* Mutation is Associated With Tumor Mutation Burden and Survival Prognosis

The mutation burden of COAD ranges from 0.05 to 188.31/Mb, with a median of 2.45/Mb. Among the 22 genes screened by the Venn diagram, the mutation of 21 genes was statistically related to the tumor mutation burden in the sample (**Figure 2D**). In order to study the relationship between these gene mutations related to tumor mutation burden and the prognosis of COAD, we further performed Kaplan–Meier analysis. The calculation results of KM analysis (**Table 2**) showed that the *USH2A* mutation and *MUC4* mutation were related to tumor survival and prognosis (**Figure 3**). Next, the mutation gene and the tumor patient's age, gender, tumor stage, and tumor mutation burden were analyzed by univariate and multivariate Cox regression. The

TABLE 1 | Univariate and multivariate COX overall survival analysis of patients with COAD.

Factors	Univariate		Multivariate	
	HR(95% CI)	*p*-value	HR(95% CI)	*p*-value
Age (year) (≤65, >65)	1.829 (1.087–4.246)	0.023	2.475 (1.442–4.246)	0.001
Gender (male, female)	1.345 (1.087–4.246)	0.227		
Stage (I and II, III and IV)	2.831 (1.723–4.652)	<0.001	3.380 (2.022–5.647)	<0.001
TMB (low, high)	1.002 (0.991–1.012)	0.780		
MUC4	2.232 (1.301–3.829)	0.004	2.054 (1.196–3.528)	0.009
USH2A	1.909 (1.088–3.351)	0.024	2.067 (1.169–3.655)	0.012

results (**Table 1**) showed that the *USH2A* mutation (HR = 1.909; 95% CI = 1.088–3.351; $p = 0.024$) and *MUC4* mutation (HR = 2.232; 95% CI = 1.301–3.829; $p = 0.004$) were associated with a poor prognosis of COAD; thus, they could be considered independent risk factors. We further studied the relationship between the location of the USH2A mutation site in the COAD sample of the TCGA database and the survival of COAD. We searched the UCSC database (http://genome.ucsc.edu/, hg38) and found that the mutation sites provided are distributed in the exon region of the USH2A gene (**Supplementary Table S2**). We analyzed the mutation regions with a sample size greater than 2 (exons 17, 61, 63, 64, 70) and found that the mutations located in exon 17 and exon 63 of USH2A are related to the survival of COAD (**Supplementary Figure S2**).

3.3 Gene Set Enrichment Analysis

Since TMB has been reported as a biomarker for immunotherapy, and since *USH2A* and *MUC4* mutations are associated with increased TMB, we further studied the relationship between *USH2A*/*MUC4* mutations and immune response using TCGA data for GSEA. The *MUC4* mutation revealed no pathway with an FDR q-value < 0.05 (**Figures 4D–F**), whereas the *USH2A* mutation featured the following significantly upregulated pathways (**Figures 4A–C**): antigen processing and presentation pathways, thyroid autoimmune disease pathways, and NK cell-mediated cytotoxic pathways. These results indicate that the *USH2A* mutation affects the signaling pathways of the immune system.

3.4 *USH2A* Mutation in COAD is Associated With Tumor-Infiltrating Immune Cells

GSEA results showed that the *USH2A* mutation affects the signaling pathways of the immune system. Therefore, we used the CIBERSORT algorithm to evaluate the relationship between the *USH2A* mutation and tumor-infiltrating immune cells in the colon cancer microenvironment. The results showed that the composition of 22 immune cells in each sample was significantly different (**Figure 5A**), and the immune score of *USH2A* mutation samples was significantly increased (**Figure 5D**). We also found that activated NK cells, follicular helper T cells (TFH cells), and $\gamma\delta$T cells were enriched in *USH2A* mutation samples (**Figure 5C**). In addition, the immune cell correlation matrix indicated activated NK cells, TFH cells, and $\gamma\delta$T, which were positively correlated with each other (**Figure 5B**).

TABLE 2 | The clinical prognostic calculation results of gene mutations related to TMB.

Gene	*p*-value	Gene	*p*-value
MUC4	0.002	PIK3CA	0.466
USH2A	0.010	MUC5B	0.469
TTN	0.077	TP53	0.527
RYR2	0.151	FBXW7	0.616
NEB	0.154	MUC16	0.675
SYNE1	0.167	FAT3	0.772
LRP1B	0.222	APC	0.819
FLG	0.325	PCLO	0.863
ZFHX4	0.395	OBSCN	0.876
CSMD1	0.447		

3.5 *USH2A* Mutation Affects Immunotherapy

In cases with a *USH2A* mutation, among the common immune checkpoint genes (*PDL1*, *CTLA4*, *LAG3*, *SIGLEC15*, *HAVCR2*, *PDCD1LG2*, *PD1*, and *TIGIT*) (Wang et al., 2019a; Zeng et al., 2019), the expression levels of *PDL1*, *CTLA4*, *LAG3*, *HAVCR2*, *PD1*, *LG2*, and *TIGIT* were significantly increased (**Figure 5E**). We compared the MSI scores of the *USH2A* mutant group and the wild-type group, which showed that the MSI scores of the former were significantly increased (**Figure 5F**). TIDE uses a set of gene expression markers to estimate two different mechanisms of tumor immune evasion: tumor-infiltrating cytotoxic T lymphocyte (CTL) dysfunction and immunosuppressive factor rejection of CTL. A higher TIDE score denotes a higher chance of antitumor immune escape and a lower response rate of ICB therapy (Jiang et al., 2018). We compared the TIDE scores of the *USH2A* gene mutation group and the wild-type group, which showed that the TIDE score of the former was significantly reduced (**Figure 5G**). These results indicate that the *USH2A* mutation affects the tumor immune response and may lead to a better ICB treatment response.

3.6 Analysis of Differential Genes in Tumor Samples After *USH2A* Mutation and Constructing a Tumor Prognostic Risk Model Based on Differential Genes

In order to further study the differential expression of tumor tissue genes after *USH2A* mutation, we divided TCGA COAD

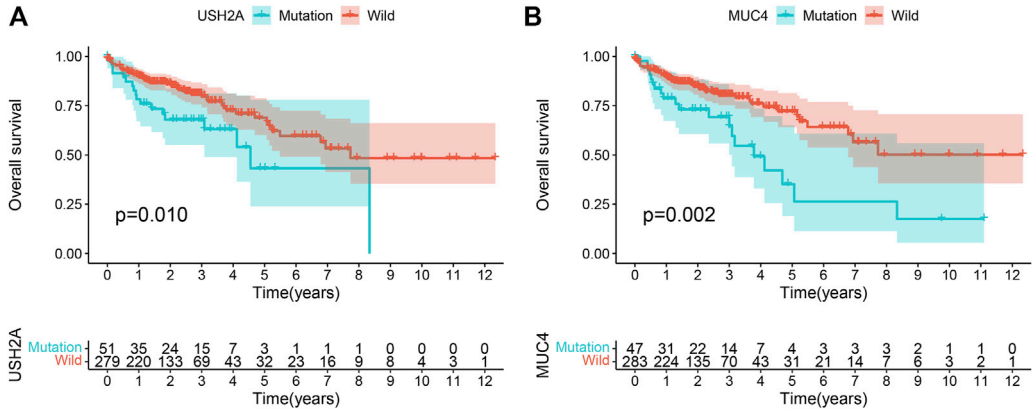


FIGURE 3 | Gene mutation is associated with clinical prognosis. Kaplan-Meier survival analysis was used to determine survival curves that reflect the association between gene mutations and prognosis. The *p*-value is shown each plot. **(A)** USH2A mutation is associated with the prognosis of COAD. **(B)** MUC4 mutation is associated with the prognosis of COAD.

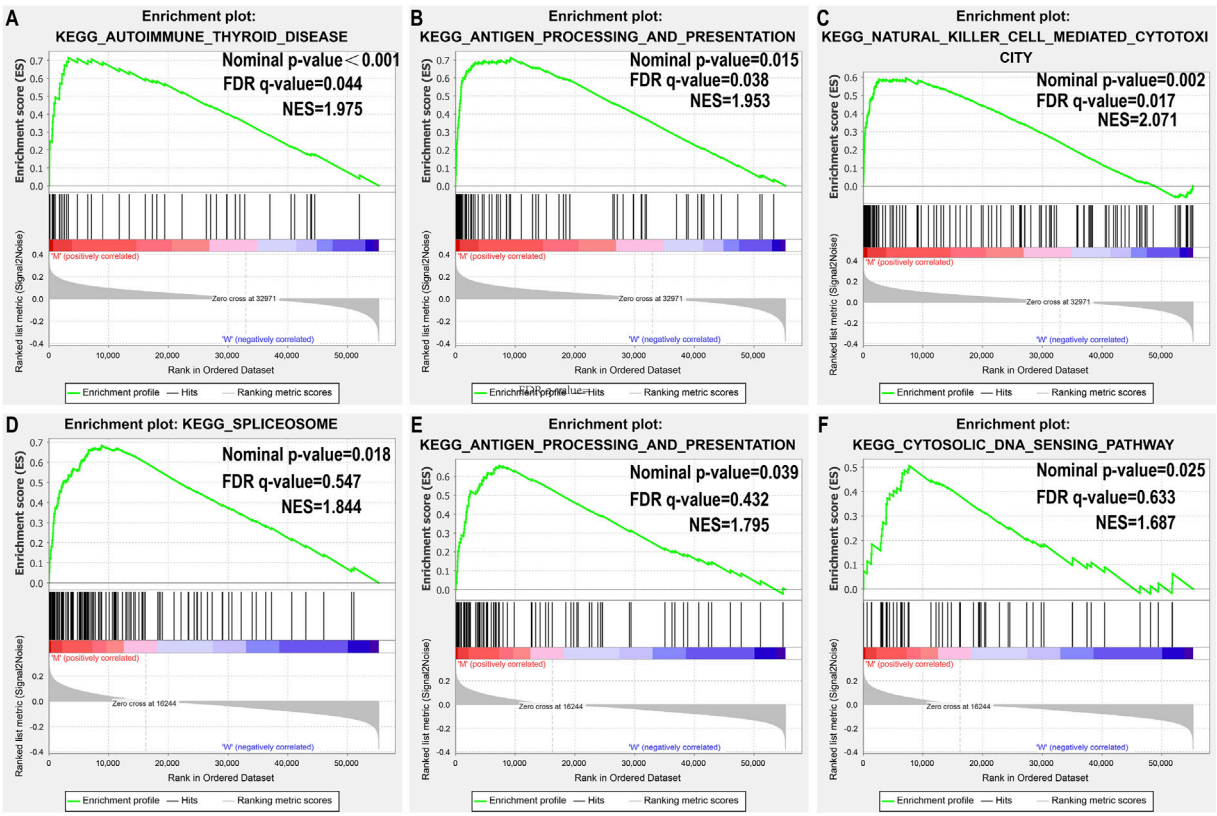


FIGURE 4 | USH2A mutation is associated with immune-related pathways. Gene set enrichment analysis was performed with the TCGA. **(A–C)** Gene enrichment plots display that a series of immune-related gene sets are enriched in the USH2A-mutant group; **(D–F)** Gene enrichment plots display enrichment pathways in the MUC4-mutant group. The nominal *p*-value and FDR *q*-value is shown in each plot.

samples into a *USH2A* mutant group and a wild-type group, and we used the edgeR package to analyze the differential expression of genes. A total of 522 differentially expressed genes (DEGs) were obtained, among which 440 DEGs were upregulated and 82 DEGs were downregulated (**Figure 6A**). Univariate Cox analysis and LASSO COX analysis were performed on the above mentioned DEGs, and a prognostic risk model based on the expression of genes *TNNT1* and *ERFE* was established

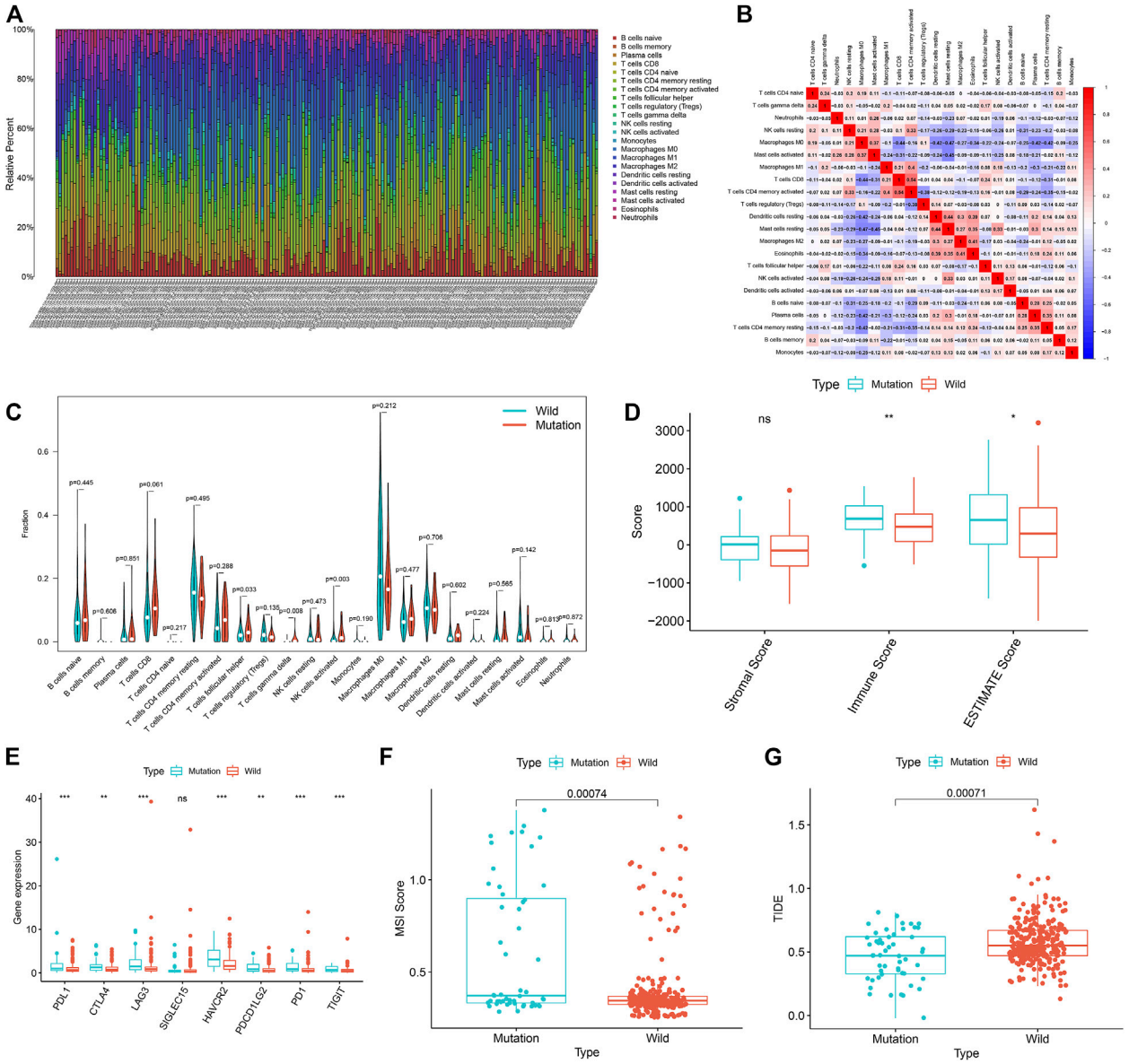


FIGURE 5 | USH2A mutation is associated with tumor immune cell infiltration and antitumor immunity. **(A)** The stacked bar chart shows the distribution of 22 immune cells in each sample. **(B)** Correlation matrix of immune cell proportions. The red color represents a positive correlation, and the blue color represents a negative correlation. **(C)** Violin plot displaying the differentially infiltrated immune cells between the USH2A-mutant groups and the wild-type USH2A group. **(D)** Immune score of USH2A mutant group and USH2A wild group samples. **(E)** Gene expressions of eight common immune checkpoints in USH2A mutant group and USH2A wild group samples. **(F)** MSI scores of samples from the USH2A mutant group and USH2A wild group. **(G)** TIDE scores of USH2A mutant group and USH2A wild group samples. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns has no statistical difference. Green represents the USH2A wild group, and red represents the USH2A mutant group.

(lambda.min = 0.0021, RiskScore = (0.1141) × *TNNT1* + (0.2032 × *ERFE*) (**Figures 6B–E,G**). The ROC curve was drawn using R software package survival ROC (**Figure 6H**). We used the GEO database COAD dataset GSE39582 to validate the risk model (**Figure 6F**) and draw the ROC curve (**Figure 6I**). In the GSE39582 dataset, the survival of patients with high and low risk scores is significantly different (**Figure 6F**), indicating that the model has the ability to predict risk. Using the clinical data of TCGA COAD to test the correlation between the risk score and clinical characteristics, it was found that the age, survival status, and tumor T stage of patients were significantly different in the high- and low-risk groups (**Figure 7C**). Univariate and multivariate Cox analysis found that the risk score is an independent risk factor for the survival and prognosis of cancer patients in TCGA cohort (**Figures 7A,B**) and GSE39582 dataset (**Supplementary Figure S1**). Comparing the immune checkpoint gene expression, immune scores, MSI scores, TMB, and TIDE of patients in the high- and low-risk groups, we found significant differences (**Figures 8A–E**). We also compared the immune cell infiltration of samples from the

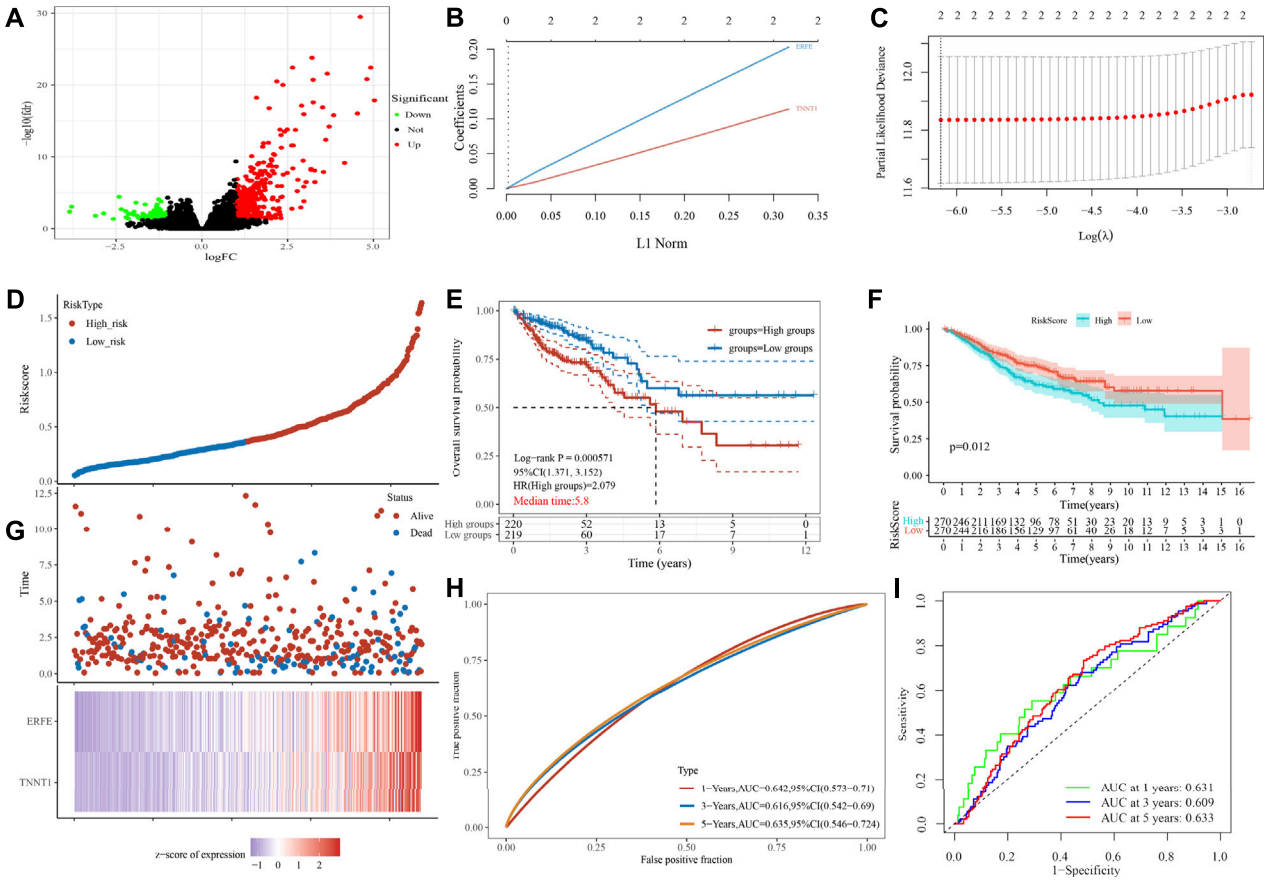


FIGURE 6 | Analysis of gene differential expression after USH2A mutation and construction of a prognostic model based on the differentially expressed genes after USH2A mutation. **(A)** Volcano map of gene differential expression analysis. **(B)** Constructing the lasso coefficient prediction model. **(C)** Selecting variables in lasso regression with minimum criteria by 1,000 times cross-validation. **(D)**, **(G)** Risk plots for patients with higher and lower risk score. **(E)** Kaplan-Meier curve analysis of the high-risk and low-risk groups. **(H)** ROC curves of the risk score. **(F)** Kaplan-Meier curve analysis of the GSE39582 dataset based on risk model. **(I)** ROC curves of the risk score in the GSE39582 dataset based on risk model.

high- and low-risk score groups, which showed that, in the high-risk group, CD8 T cells, TFH cells, and activated NK cells were significantly increased (**Figure 8G**). Our analysis found that the TIDE score was negatively correlated with the sample risk score (**Figure 8F**). These results indicate that patients at high risk with a poor survival prognosis may have a better response to ICB therapy, thereby improving their prognosis. Therefore, the risk model can predict the survival prognosis of cancer patients and guide the clinical treatment decisions of cancer patients.

4 DISCUSSION

In summary, by analyzing the somatic mutation characteristics of 398 USA COAD samples in TCGA database and 305 Chinese COAD samples in the ICGC database, we found that *USH2A* is frequently mutated in both cohorts, and its mutation is associated with high TMB and poor clinical prognosis. We also found that the *USH2A* mutation is positively related to the signaling pathway of the immune system. The results of tumor-infiltrating immune cell analysis showed an enrichment of activated NK cells, TFH cells, and γδT cells in the *USH2A* mutation samples, which is consistent with the results of previous studies (Bindea et al., 2013; Meraviglia et al., 2017; Zhang et al., 2020b). Dependent on the presence or absence of a *USH2A* mutation, we divided the TCGA COAD samples into two groups and analyzed the DEGs. According to the GO (**Supplementary Figure S3A**) and KEGG (**Supplementary Figure S3B**) enrichment analysis, we found that the DEGs mainly involved processes linked to cytokine activity and antibacterial humoral response and were significantly enriched in the IL-17 signaling pathway, which are all related to immune response. Querying the KEGG database (https://www.kegg.jp/), we found that NK cells and γδT cells are involved in the IL-17 signaling pathway, which confirms that the pathway enrichment of DEGs after *USH2A* mutation is correlated with tumor immune cell infiltration. After *USH2A* mutation, analysis showed that immune checkpoint gene expression and TIDE score decreased significantly, whereas immune score and MSI score increased significantly, thus indicating that *USH2A* mutation affects the antitumor immunity and is conducive to ICB treatment. By performing univariate Cox analysis and LASSO COX analysis of DEGs, we established a prognostic risk model

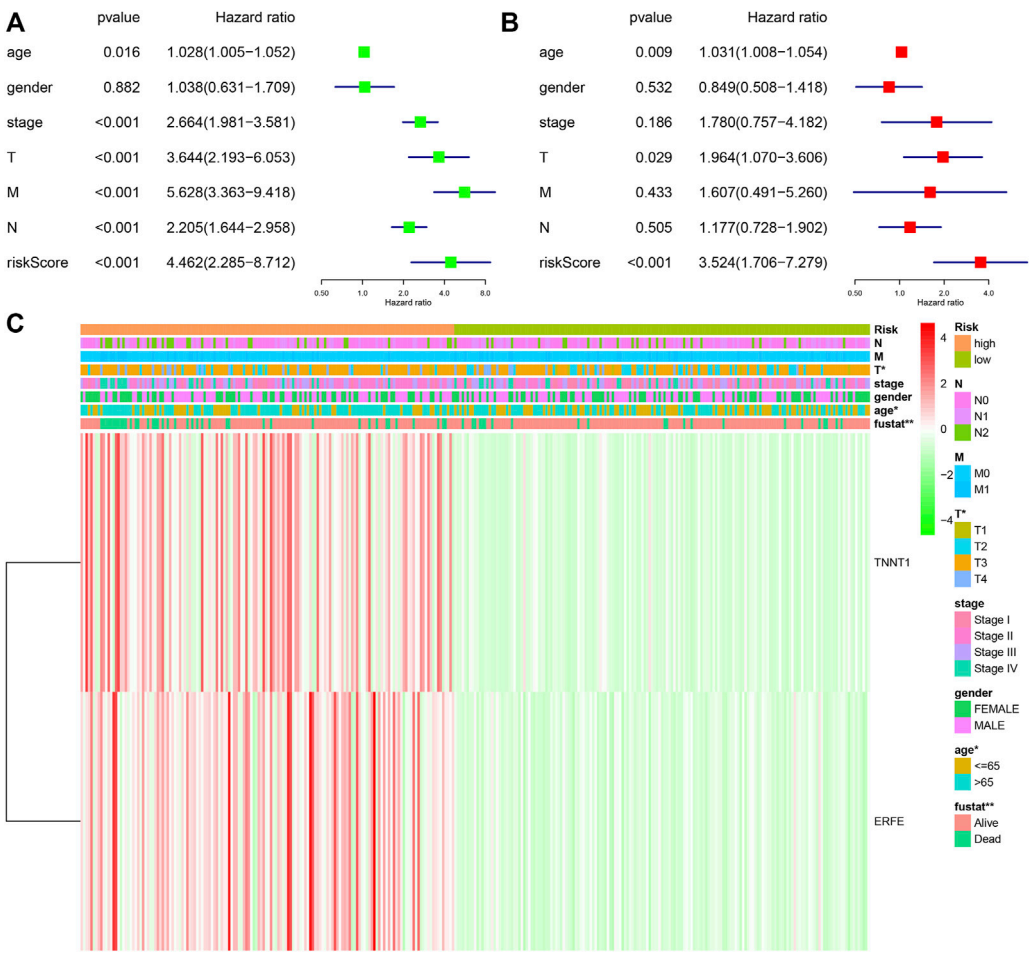


FIGURE 7 | Correlation between risk model and clinical characteristics. **(A)** Univariate COX analysis forest plot based on TCGA samples. **(B)** Multivariate COX analysis forest plot based on TCGA samples. **(C)** Heat map of the correlation between risk scores and clinical characteristics in TCGA samples.

based on the expression of genes *TNNT1* and *ERFE*. Cox analysis showed that risk score is an independent risk factor for tumor survival and prognosis. The verification of the ROC curve using the GSE39582 dataset showed that the model has the ability to predict risk. Comparing the immune checkpoint gene expression, immune score, MSI score, TMB, and TIDE of patients in the high- and low-risk groups, significant differences were found, whereby CD8 T cells, TFH cells, and activated NK cells were all significantly increased in the high-risk group. These results indicate that the risk model can predict the survival prognosis of COAD patients and assess whether the patients will have a good ICB treatment response.

The *USH2A* (also known as Usherin) gene encodes a protein. The protein exists in the basement membrane and may play an important role in the development and homeostasis of the inner ear and retina (Weston et al., 2000). *USH2A* mutations are associated with Usher syndrome type IIa, retinitis pigmentosa (Xing et al., 2020), and tongue squamous cell carcinoma (Zhang et al., 2020). In lung adenocarcinoma, the *USH2A* mutation is one of the most frequently mutated genes for predicting neoantigens (Cai et al., 2018). In our research, we found that *USH2A* mutation is associated with the overexpression of immune checkpoint genes and increased TMB. TMB represents the accumulation of somatic mutations in tumors. A high TMB helps to expose more neoantigens, which may trigger a T-cell-dependent immune response (Mcgranahan et al., 2016). Immune checkpoint blockade (ICB), which targets programmed cell death ligand 1 (*PDL1*) and cytotoxic T lymphocyte antigen 4 (*CTLA4*) pathways, has become a treatment strategy for various types of cancer (Long et al., 2017; Zhang et al., 2021). We used TCGA dataset to analyze the tumor response to immunotherapy after *USH2A* mutation. We found that *USH2A* mutant tumors have stronger immunogenicity, exhibited as higher TMB, increased immune cell infiltration into tumor tissues, and overexpression of immune checkpoint factors such as *PD1*, *PDL1*, and *CTLA4*. This indicates that the *USH2A* mutation can enhance tumor immunogenicity, allowing tumor patients to benefit from antitumor immunotherapy. The expression of *PDL1* and TMB are correlated with the clinical benefit of patients treated with ICB (Long et al., 2017). However, these two biomarkers are continuous variables with no clearly defined cutoff point above which a response is guaranteed. In

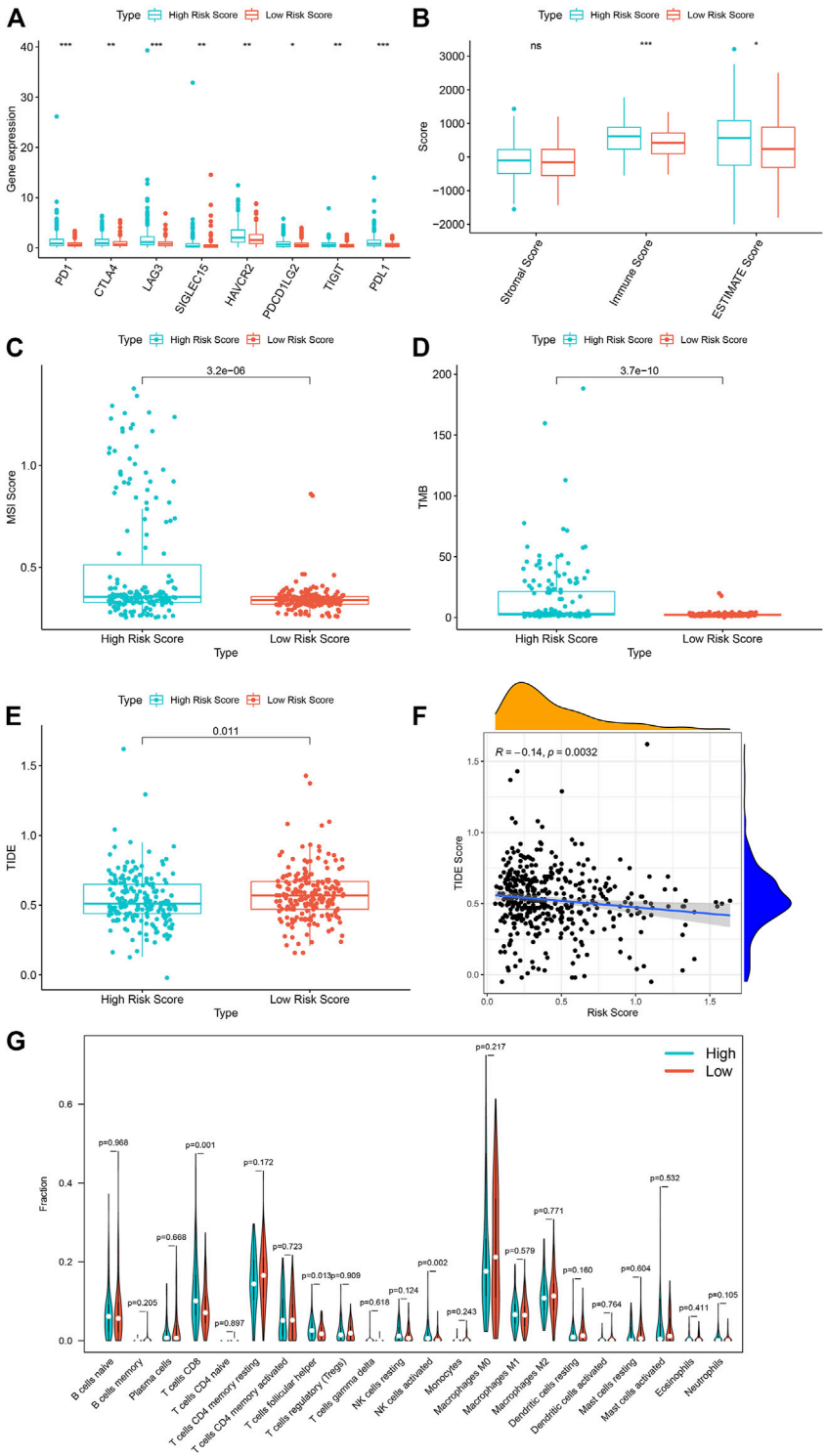


FIGURE 8 | The high risk score is associated with a better tumor immunotherapy response in TCGA sample. **(A)** The expression of immune checkpoint genes in TCGA samples in high and low risk groups. **(B)** The immune scores of the high and low risk groups of TCGA samples. **(C)** The MSI scores of the high and low risk groups of the TCGA sample. **(D)** TMB situation of high and low risk groups of TCGA sample. **(E)** TIDE scores of the high and low risk groups of the TCGA sample. **(F)** Correlation analysis between TCGA sample risk score and TIDE. **(G)** Differences in immune cell infiltration between high and low model scores. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns, has no statistical difference. Green represents the high-scoring group of the model, and red represents the low-scoring group of the model.

addition, the expression of *PDL1* and TMB also vary depending on the detection method and platform (Tsao et al., 2018; Addeo et al., 2019). In contrast, *USH2A* mutations are easily detected by next-generation sequencing, and their presence in this study was closely related to the response to ICB treatment. Therefore, it is worth considering the *USH2A* mutation as a potential biomarker for the sensitization of patients to ICB therapy.

NK cells play a key role in innate and adaptive immune response and tumor immune surveillance by recognizing and killing tumor cells (Mandal and Viswanathan, 2015). Although tumor-related NK cells are not common in tumor immune infiltration, they have been associated with increased survival of colon cancer patients (Melero et al., 2014). Higher NK cell activity is associated with poor prognosis of skin T cell lymphoma (Mundy-Bosse et al., 2018). Explanations for this difference include the impaired recognition of malignant $CD4^+$ T cells mediated by NK cells and the inability of NK cells to form functional immune synapses (Mundy-Bosse et al., 2018). TFH cells are specialized T helper cells, and their most significant role is to promote the formation and maintenance of germinal centers, as well as the maturation of B cells and the acquisition of immune memory (Vinuesa et al., 2016). Currently, it is generally believed that the TFH cell–B cell axis in tumor-associated tertiary lymphoid structures (TLSs) is conducive to the formation of an antitumor immune environment (Galon et al., 2013). TFH cells produce chemokine ligand 13 (CXCL13), which targets B cells and TFH cells themselves *via* chemokine receptor 5 (CXCR5). High numbers of TFH cells and high levels of CXCL13 are associated with increased survival of colon cancer patients (Bindea et al., 2013). $\gamma\delta$ T cells in the colon are the first line of defense against pathogens in the intestinal tissue immune monitoring program (Suzuki et al., 2020). Evidence has shown that human $\gamma\delta$ T cells have antitumor effects in colon cancer, which is related to their ability to kill established colon cancer cells (Suzuki et al., 2020). The knowledge of how $\gamma\delta$ T cells promote colon cancer is still limited, but the $\gamma\delta$ T cells known to promote the progression of colon cancer are mainly concentrated in the $\gamma\delta$ T cell subset that produces IL-17 (Van hede et al., 2017). In breast tumors and gallbladder tumors, an increase in $\gamma\delta$ T cells is associated with poor prognosis (Ma et al., 2012; Patil et al., 2016). There has not been a comprehensive histological analysis of the prognostic ability of $\gamma\delta$ T cells in colon cancer. In a follow-up study of colon cancer patients, the immune score (including tumor-infiltrating $\gamma\delta$ T cells) was used to group patients. The 5 years recurrence rate of patients in the high-immune-score group was only 4.8% (Pages et al., 2009). Considering the relationship between high immune score and good prognosis, it has been speculated that $\gamma\delta$ T cells may be associated with a better colon cancer prognosis (Suzuki et al., 2020). In this study, the survival prognosis of patients with *USH2A* mutations was poor; however, in the *USH2A* mutant tumor samples, there was an enrichment of activated NK cells, TFH cells, and $\gamma\delta$ T cells, indicating a change in the recognition of immune surveillance, as well as an antitumor effect. Therefore, we found that, in COAD, the *USH2A* mutation can induce changes in infiltrating immune cells, thereby enhancing antitumor immunity.

The increased migration and invasion potential of colon cancer cells leads to a significant decrease in the 5 years survival rate of colon cancer patients. Therefore, an accurate prediction of prognosis is essential for individualized treatment of these patients. Today, gene expression profiling has become an adjunct to cancer treatment; for example, Gene-expression prediction models were built using transcriptome to predicte colorectal cancer risk (Guo et al., 2021), the expression characteristics of six lncRNAs were used as indicators to evaluate the prognosis of patients with colorectal cancer (Zhao et al., 2018), and an eleven gene signature was used as prognostic index to predict systemic recurrences in colorectal cancer (Kim et al., 2019). In this study, we identified the expression levels of two mRNAs as reliable prognostic indicators of colon cancer. In this risk model, the *TNNT1* gene encodes a protein of the troponin subunit, which is a regulatory complex located on the sarcomere filaments (Wei and Jin, 2016). Studies have reported that *TNNT1* is significantly upregulated in colon cancer samples and cell lines. The upregulation of *TNNT1* is also related to a variety of clinicopathological characteristics, and its high expression is related to the poor prognosis of patients. Inhibition of *TNNT1* can significantly inhibit cell proliferation, migration, and invasion, while promoting cell apoptosis (Chen et al., 2020). *TNNT1* may promote the progress of COAD and mediate the EMT process (Hao et al., 2020). On the other hand, ERFE is a glycoprotein hormone encoded by *FAM132B*, which is produced upon the stimulation of red blood cells by erythropoietin in the bone marrow and spleen (Ganz, 2019). It has been reported that this gene is mainly related to anemia and metabolic abnormalities (Seldin et al., 2012; Bondu et al., 2019), whereas there are no reports of a tumor connection. *TNNTI* gene expression in our research model was associated with poor tumor prognosis, which is consistent with previous studies (Hao et al., 2020). Patients with a higher TIDE score have a higher chance of antitumor immune escape, thus showing a lower response rate to ICB therapy (Jiang et al., 2018). In our study, a comparison of the TIDE of patients in the high-risk and low-risk groups revealed a lower score in the former, which indicates the potential for high-risk patients to improve their survival prognosis through a better response to ICB therapy, which can facilitate the choice of clinical treatment for cancer patients.

The novelty of this study lies in the discovery that *USH2A* mutations can affect the antitumor immunity of COAD and the responsiveness to ICB therapy. Furthermore, we constructed a prognostic model consisting of two DEGs, which could predict 1, 3, and 5 years survival rates in TCGA dataset and GEO validation dataset GSE39582 with a relatively high AUC. The main limitation of this study is that the ICGC database lacks corresponding clinical data on Chinese COAD; thus, we could not verify the significance of the *USH2A* mutation in the prognosis of Chinese COAD patients and whether it can cause the same immune response. Even though *USH2A* was frequently mutated in Chinese COAD samples, its impact may be somewhat heterogeneous among different races. Therefore, the relationship between *USH2A* mutation and prognosis, including the analysis of infiltrating immune cells and signaling pathways, needs to be further verified in Chinese colon samples. In addition, the differential expression of the two genes used to construct the risk model was identified from TCGA data; although TCGA data are of high quality, further experimental verification of the role of these two differential genes in colon cancer is needed *in vitro* and *in vivo*.

In summary, this study showed that *USH2A* is frequently mutated in COAD, which is associated with a high TMB and poor prognosis. In addition, the *USH2A* mutation upregulates immune signaling pathways and promotes an antitumor immune response. On the basis of two DEGs associated with the *USH2A* mutation, we constructed a model with a predictive effect on the prognosis of tumor survival. These findings reveal a new gene whose mutation can be used as a biomarker for predicting the response to antitumor immunity and ICB treatment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

DX, BM, YS, LL, and WY conceived and designed the study; YS, LL, WY, XL, YY performed the data acquisition and analysis. DX, YS, and LL edited the manuscript. Each author contributed to the writing of manuscript and approved it.

FUNDING

This study was supported by grants from: The National Natural Sciences Foundation of China (81770634). Heilongjiang Province General Undergraduate Colleges and Universities Young Innovative Talents Training Plan (No. UNPYSCT-2018073). National Natural Science Foundation of China Youth Project (No. 81602337). Heilongjiang Province Educational Science Planning Project (No. UNPYSCT-2017062).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.762160/full#supplementary-material

Supplementary Figure 1 | Riskscore is an independent prognostic factor in the validation data set GSE39582. **(A)** Univariate COX analysis forest plot based on GSE39582 dataset. **(B)** Multivariate COX analysis forest plot based on GSE39582 dataset.

Supplementary Figure 2 | The relationship between the location of the USH2A mutation site and the survival of COAD in the COAD sample of the TCGA database. Kaplan-Meier survival analysis was used to determine survival curves that reflect the association between gene mutations in exons 17, 61, 63, 64, 70 and prognosis.

Supplementary Figure 3 | DAVID was used for GO enrichment and KEGG pathway analysis of DEGs. **(A)** Analysis of GO enrichment of differentially expression genes. **(B)** Analysis of GO enrichment of differentially expression genes.

REFERENCES

Addeo, A., Banna, G. L., and Weiss, G. J. (2019). Tumor Mutation Burden-From Hopes to Doubts. *JAMA Oncol.* 5, 934–935. doi:10.1001/jamaoncol.2019.0626

Ancevski Hunter, K., Socinski, M. A., and Villaruz, L. C. (2018). PD-L1 Testing in Guiding Patient Selection for PD-1/pd-L1 Inhibitor Therapy in Lung Cancer. *Mol. Diagn. Ther.* 22, 1–10. doi:10.1007/s40291-017-0308-6

Argilés, G., Tabernero, J., Labianca, R., Hochhauser, D., Salazar, R., Iveson, T., et al. (2020). Localised colon Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* 31, 1291–1305. doi:10.1016/j.annonc.2020.06.022

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

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39, 782–795. doi:10.1016/j.immuni.2013.10.003

Bondu, S., Alary, A.-S., Lefèvre, C., Houy, A., Jung, G., Lefebvre, T., et al. (2019). A Variant Erythroferrone Disrupts Iron Homeostasis inSF3B1-Mutated Myelodysplastic Syndrome. *Sci. Transl. Med.* 11, eaav5467. doi:10.1126/scitranslmed.aav5467

Bonneville, R., Krook, M. A., Kautto, E. A., Miya, J., Wing, M. R., Chen, H.-Z., et al. (2017). Landscape of Microsatellite Instability across 39 Cancer Types. *JCO Precision Oncol.* 2017, 1–15. doi:10.1200/po.17.00073

Bretz, A. C., Parnitzke, U., Kronthaler, K., Dreker, T., Bartz, R., Hermann, F., et al. (2019). Domatinostat Favors the Immunotherapy Response by Modulating the Tumor Immune Microenvironment (TIME). *J. Immunotherapy Cancer* 7, 294. doi:10.1186/s40425-019-0745-3

Cai, W., Zhou, D., Wu, W., Tan, W. L., Wang, J., Zhou, C., et al. (2018). MHC Class II Restricted Neoantigen Peptides Predicted by Clonal Mutation Analysis in Lung Adenocarcinoma Patients: Implications on Prognostic Immunological Biomarker and Vaccine Design. *BMC Genomics* 19, 582. doi:10.1186/s12864-018-4958-5

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., et al. (2017). Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational burden. *Genome Med.* 9, 34. doi:10.1186/s13073-017-0424-2

Chen, H., Chong, W., Wu, Q., Yao, Y., Mao, M., and Wang, X. (2019a). Association of LRP1B Mutation with Tumor Mutation Burden and Outcomes in Melanoma and Non-small Cell Lung Cancer Patients Treated with Immune Check-Point Blockades. *Front. Immunol.* 10, 1113. doi:10.3389/fimmu.2019.01113

Chen, Y., Liu, Q., Chen, Z., Wang, Y., Yang, W., Hu, Y., et al. (2019b). PD-L1 Expression and Tumor Mutational burden Status for Prediction of Response to Chemotherapy and Targeted Therapy in Non-small Cell Lung Cancer. *J. Exp. Clin. Cancer Res.* 38, 193. doi:10.1186/s13046-019-1192-1

Chen, Y., Wang, J., Wang, D., Kang, T., Du, J., Yan, Z., et al. (2020). TNNT1, Negatively Regulated by miR-873, Promotes the Progression of Colorectal Cancer. *J. Gene Med.* 22, e3152. doi:10.1002/jgm.3152

Choi, M. R., Gwak, M., Yoo, N. J., and Lee, S. H. (2015). Regional Bias of Intratumoral Genetic Heterogeneity of Apoptosis-Related Genes BAX, APAF1, and FLASH in Colon Cancers with High Microsatellite Instability. *Dig. Dis. Sci.* 60, 1674–1679. doi:10.1007/s10620-014-3499-2

Cunningham, D., Atkin, W., Lenz, H.-J., Lynch, H. T., Minsky, B., Nordlinger, B., et al. (2010). Colorectal Cancer. *The Lancet* 375, 1030–1047. doi:10.1016/s0140-6736(10)60353-4

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

De La Chapelle, A., and Hampel, H. (2010). Clinical Relevance of Microsatellite Instability in Colorectal Cancer. *Jco* 28, 3380–3387. doi:10.1200/jco.2009.27.0652

Galon, J., Angell, H. K., Bedognetti, D., and Marincola, F. M. (2013). The Continuum of Cancer Immunosurveillance: Prognostic, Predictive, and Mechanistic Signatures. *Immunity* 39, 11–26. doi:10.1016/j.immuni.2013.07.008

Ganz, T. (2019). Erythropoietic Regulators of Iron Metabolism. *Free Radic. Biol. Med.* 133, 69–74. doi:10.1016/j.freeradbiomed.2018.07.003

George, A. P., Kuzel, T. M., Zhang, Y., and Zhang, B. (2019). The Discovery of Biomarkers in Cancer Immunotherapy. *Comput. Struct. Biotechnol. J.* 17, 484–497. doi:10.1016/j.csbj.2019.03.015

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

Gubin, M. M., Artyomov, M. N., Mardis, E. R., and Schreiber, R. D. (2015). Tumor Neoantigens: Building a Framework for Personalized Cancer Immunotherapy. *J. Clin. Invest.* 125, 3413–3421. doi:10.1172/jci80008

Guo, X., Lin, W., Wen, W., Huyghe, J., Bien, S., Cai, Q., et al. (2021). Identifying Novel Susceptibility Genes for Colorectal Cancer Risk from a Transcriptome-wide Association Study of 125,478 Subjects. *Gastroenterology* 160, 1164–1178. e6. doi:10.1053/j.gastro.2020.08.062

Hao, Y.-H., Yu, S.-Y., Tu, R.-S., and Cai, Y.-Q. (2020). TNNT1, a Prognostic Indicator in colon Adenocarcinoma, Regulates Cell Behaviors and Mediates EMT Process. *Biosci. Biotechnol. Biochem.* 84, 111–117. doi:10.1080/09168451.2019.1664891

Herzig, D. O., and Tsikitis, V. L. (2015). Molecular Markers for colon Diagnosis, Prognosis and Targeted Therapy. *J. Surg. Oncol.* 111, 96–102. doi:10.1002/jso.23806

Hodges, T. R., Ott, M., Xiu, J., Gatalica, Z., Swensen, J., Zhou, S., et al. (2017). Mutational burden, Immune Checkpoint Expression, and Mismatch Repair in Glioma: Implications for Immune Checkpoint Immunotherapy. *Neuro Oncol.* 19, 1047–1057. doi:10.1093/neuonc/nox026

Janjigian, Y. Y., Sanchez-Vega, F., Jonsson, P., Chatila, W. K., Hechtman, J. F., Ku, G. Y., et al. (2018). Genetic Predictors of Response to Systemic Therapy in Esophagogastric Cancer. *Cancer Discov.* 8, 49–58. doi:10.1158/2159-8290.Cd-17-0787

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, 1550–1558. doi:10.1038/s41591-018-0136-1

Kaderbhaï, C., Tharin, Z., and Ghiringhelli, F. (2019). The Role of Molecular Profiling to Predict the Response to Immune Checkpoint Inhibitors in Lung Cancer. *Cancers* 11, 201. doi:10.3390/cancers11020201

Keenan, T. E., Burke, K. P., and Van Allen, E. M. (2019). Genomic Correlates of Response to Immune Checkpoint Blockade. *Nat. Med.* 25, 389–402. doi:10.1038/s41591-019-0382-x

Kim, S.-K., Kim, S.-Y., Kim, C. W., Roh, S. A., Ha, Y. J., Lee, J. L., et al. (2019). A Prognostic index Based on an Eleven Gene Signature to Predict Systemic Recurrences in Colorectal Cancer. *Exp. Mol. Med.* 51, 1–12. doi:10.1038/s12276-019-0319-y

Labianca, R., Beretta, G. D., Kildani, B., Milesi, L., Merlin, F., Mosconi, S., et al. (2010). Colon Cancer. *Crit. Rev. oncology/hematology* 74, 106–133. doi:10.1016/j.critrevonc.2010.01.010

Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., et al. (2019). Integrative Molecular and Clinical Modeling of Clinical Outcomes to PD1 Blockade in Patients with Metastatic Melanoma. *Nat. Med.* 25, 1916–1927. doi:10.1038/s41591-019-0654-5

Long, J., Lin, J., Wang, A., Wu, L., Zheng, Y., Yang, X., et al. (2017). PD-1/PD-L Blockade in Gastrointestinal Cancers: Lessons Learned and the Road toward Precision Immunotherapy. *J. Hematol. Oncol.* 10, 146. doi:10.1186/s13045-017-0511-2

Ma, C., Zhang, Q., Ye, J., Wang, F., Zhang, Y., Wevers, E., et al. (2012). Tumor-Infiltrating γδ T Lymphocytes Predict Clinical Outcome in Human Breast Cancer. *J. Immunol.* 189, 5029–5036. doi:10.4049/jimmunol.1201892

Mandal, A., and Viswanathan, C. (2015). Natural Killer Cells: In Health and Disease. *Hematology/oncology Stem Cel. Ther.* 8, 47–55. doi:10.1016/j.hemonc.2014.11.006

Mcgranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., et al. (2016). Clonal Neoantigens Elicit T Cell Immunoreactivity and Sensitivity to Immune Checkpoint Blockade. *Science* 351, 1463–1469. doi:10.1126/science.aaf1490

Melero, I., Rouzaut, A., Motz, G. T., and Coukos, G. (2014). T-cell and NK-Cell Infiltration into Solid Tumors: a Key Limiting Factor for Efficacious Cancer Immunotherapy. *Cancer Discov.* 4, 522–526. doi:10.1158/2159-8290.Cd-13-0985

Meraviglia, S., Lo Presti, E., Tosolini, M., La Mendola, C., Orlando, V., Todaro, M., et al. (2017). Distinctive Features of Tumor-Infiltrating γδ T Lymphocytes in Human Colorectal Cancer. *Oncoimmunology* 6, e1347742. doi:10.1080/2162402x.2017.1347742

Minoru, K., and Susumu, G. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Mundy-Bosse, B., Denlinger, N., Mclaughlin, E., Chakravarti, N., Hwang, S., Chen, L., et al. (2018). Highly Cytotoxic Natural Killer Cells Are Associated with Poor Prognosis in Patients with Cutaneous T-Cell Lymphoma. *Blood Adv.* 2, 1818–1827. doi:10.1182/bloodadvances.2018020388

Neri, E., Faggioni, L., Cini, L., and Bartolozzi, C. (2010). Colonic Polyps: Inheritance, Susceptibility, Risk Evaluation, and Diagnostic Management. *Cancer Manag Res.* 3, 17–24. doi:10.2147/cmr.S1570510.2147/cmar.s15705

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

Pagès, F., Kirilovsky, A., Mlecnik, B., Asslaber, M., Tosolini, M., Bindea, G., et al. (2009). *In Situ* cytotoxic and Memory T Cells Predict Outcome in Patients with Early-Stage Colorectal Cancer. *J. Clin Oncol.* 27, 5944–5951. doi:10.1200/jco.2008.19.6147

Pallocca, M., Angeli, D., Palombo, F., Sperati, F., Milella, M., Goeman, F., et al. (2019). Combinations of Immuno-Checkpoint Inhibitors Predictive Biomarkers Only Marginally Improve Their Individual Accuracy. *J. Transl Med.* 17, 131. doi:10.1186/s12967-019-1865-8

Patel, S. A., and Minn, A. J. (2018). Combination Cancer Therapy with Immune Checkpoint Blockade: Mechanisms and Strategies. *Immunity* 48, 417–433. doi:10.1016/j.immuni.2018.03.007

Patil, R. S., Shah, S. U., Shrikhande, S. V., Goel, M., Dikshit, R. P., Chiplunkar, S. V., et al. (2016). IL17 Producing γδT Cells Induce Angiogenesis and Are Associated with Poor Survival in Gallbladder Cancer Patients. *J. Int. Du Cancer.* 138, 869–81. doi:10.1002/ijc.30134

Punt, C. J. A., Koopman, M., and Vermeulen, L. (2017). From Tumour Heterogeneity to Advances in Precision Treatment of Colorectal Cancer. *Nat. Rev. Clin. Oncol.* 14, 235–246. doi:10.1038/nrclinonc.2016.171

Rizvi, H., Sanchez-Vega, F., La, K., Chatila, W., Jonsson, P., Halpenny, D., et al. (2018). Molecular Determinants of Response to Anti-programmed Cell Death (PD)-1 and Anti-programmed Death-Ligand 1 (PD-L1) Blockade in Patients with Non-small-cell Lung Cancer Profiled with Targeted Next-Generation Sequencing. *J. Clin Oncol.* 36, 633–641. doi:10.1200/jco.2017.75.3384

Roncucci, L., and Mariani, F. (2015). Prevention of Colorectal Cancer: How many Tools Do We Have in Our Basket? *Eur. J. Intern. Med.* 26, 752–756. doi:10.1016/j.ejim.2015.08.019

Seldin, M. M., Peterson, J. M., Byerly, M. S., Wei, Z., and Wong, G. W. (2012). Myonectin (CTRP15), a Novel Myokine that Links Skeletal Muscle to Systemic Lipid Homeostasis. *J. Biol. Chem.* 287, 11968–11980. doi:10.1074/jbc.M111.336834

Shia, J. (2015). Evolving Approach and Clinical Significance of Detecting DNA Mismatch Repair Deficiency in Colorectal Carcinoma. *Semin. Diagn. Pathol.* 32, 352–361. doi:10.1053/j.semdp.2015.02.018

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

Suzuki, T., Hayman, L., Kilbey, A., Edwards, J., and Coffelt, S. B. (2020). Gut γδ T Cells as Guardians, Disruptors, and Instigators of Cancer. *Immunol. Rev.* 298, 198–217. doi:10.1111/imr.12916

The Cancer Genome Atlas Network (2012). Comprehensive Molecular Characterization of Human colon and Rectal Cancer. *Nature* 487, 330–337. doi:10.1038/nature11252

Tsao, M. S., Kerr, K. M., Kockx, M., Beasley, M.-B., Borczuk, A. C., Botling, J., et al. (2018). PD-L1 Immunohistochemistry Comparability Study in Real-Life Clinical Samples: Results of Blueprint Phase 2 Project. *J. Thorac. Oncol.* 13, 1302–1311. doi:10.1016/j.jtho.2018.05.013

Tsimberidou, A.-M. (2015). Targeted Therapy in Cancer. *Cancer Chemother. Pharmacol.* 76, 1113–1132. doi:10.1007/s00280-015-2861-1

Van hede, D., Polese, B., Humblet, C., Wilharm, A., Renoux, V., Dortu, E., et al. (2017). Human Papillomavirus Oncoproteins Induce a Reorganization of

Epithelial-Associated γδ T Cells Promoting Tumor Formation. *Proc. Natl. Acad. Sci. USA* 114, E9056–E9065. doi:10.1073/pnas.1712883114

Vinuesa, C. G., Linterman, M. A., Yu, D., and Maclennan, I. C. M. (2016). Follicular Helper T Cells. *Annu. Rev. Immunol.* 34, 335–368. doi:10.1146/annurev-immunol-041015-055605

Wang, J., Sun, J., Liu, L. N., Flies, D. B., Nie, X., Toki, M., et al. (2019a). Siglec-15 as an Immune Suppressor and Potential Target for Normalization Cancer Immunotherapy. *Nat. Med.* 25, 656–666. doi:10.1038/s41591-019-0374-x

Wang, S., He, Z., Wang, X., Li, H., and Liu, X.-S. (2019b). Antigen Presentation and Tumor Immunogenicity in Cancer Immunotherapy Response Prediction. *Elife* 8. e49020. doi:10.7554/eLife.49020

Watson, A. J. M., and Collins, P. D. (2011). Colon Cancer: a Civilization Disorder. *Dig. Dis.* 29, 222–228. doi:10.1159/000323926

Wei, B., and Jin, J.-P. (2016). TNNT1, TNNT2, and TNNT3: Isoform Genes, Regulation, and Structure-Function Relationships. *Gene* 582, 1–13. doi:10.1016/j.gene.2016.01.006

Weston, M. D., Eudy, J. D., Fujita, S., Yao, S.-F., Usami, S., Cremers, C., et al. (2000). Genomic Structure and Identification of Novel Mutations in Usherin, the Gene Responsible for Usher Syndrome Type IIa. *Am. J. Hum. Genet.* 66, 1199–1210. doi:10.1086/302855

Xing, D., Zhou, H., Yu, R., Wang, L., Hu, L., Li, Z., et al. (2020). Targeted Exome Sequencing Identified a Novel USH2A Mutation in a Chinese Usher Syndrome Family: a Case Report. *BMC Ophthalmol.* 20, 485. doi:10.1186/s12886-020-01711-7

Yarchoan, M., Hopkins, A., and Jaffee, E. M. (2017). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N. Engl. J. Med.* 377, 2500–2501. doi:10.1056/NEJMc1713444

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, 737–750. doi:10.1158/2326-6066.Cir-18-0436

Zhang, H., Song, Y., Du, Z., Li, X., Zhang, J., Chen, S., et al. (2020a). Exome Sequencing Identifies New Somatic Alterations and Mutation Patterns of Tongue Squamous Cell Carcinoma in a Chinese Population. *J. Pathol.* 251, 353–364. doi:10.1002/path.5467

Zhang, S., Liu, W., Hu, B., Wang, P., Lv, X., Chen, S., et al. (2020b). Prognostic Significance of Tumor-Infiltrating Natural Killer Cells in Solid Tumors: A Systematic Review and Meta-Analysis. *Front Immunol.* 11, 1242. doi:10.3389/fimmu.2020.01242

Zhang, Z., Wu, H.-X., Lin, W.-H., Wang, Z.-X., Yang, L.-P., Zeng, Z.-L., et al. (2021). EPHA7 Mutation as a Predictive Biomarker for Immune Checkpoint Inhibitors in Multiple Cancers. *BMC Med.* 19, 26. doi:10.1186/s12916-020-01899-x

Zhao, J., Xu, J., Shang, A.-q., and Zhang, R. (2018). A Six-LncRNA Expression Signature Associated with Prognosis of Colorectal Cancer Patients. *Cell Physiol Biochem* 50, 1882–1890. doi:10.1159/000494868

Zhu, G., Pei, L., Li, Y., and Gou, X. (2020). EP300 Mutation Is Associated with Tumor Mutation burden and Promotes Antitumor Immunity in Bladder Cancer Patients. *Aging* 12, 2132–2141. doi:10.18632/aging.102728

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fgene.2021.754445

Composition and Dynamics of H1N1 and H7N9 Influenza A Virus Quasispecies in a Co-infected Patient Analyzed by Single Molecule Sequencing Technology

Peng Lin[1,2], Tao Jin[2,3], Xinfen Yu[4], Lifeng Liang[3], Guang Liu[2], Dragomirka Jovic[3], Zhou Sun[4], Zhe Yu[3], Jingcao Pan[4] and Guangyi Fan[2,3]**

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [2]BGI-Qingdao, BGI-Shenzhen, Qingdao, China, [3]BGI-Shenzhen, Shenzhen, China, [4]Hangzhou Center for Disease Control and Prevention, Hangzhou, China

A human co-infected with H1N1 and H7N9 subtypes influenza A virus (IAV) causes a complex infectious disease. The identification of molecular-level variations in composition and dynamics of IAV quasispecies will help to understand the pathogenesis and provide guidance for precision medicine treatment. In this study, using single-molecule real-time sequencing (SMRT) technology, we successfully acquired full-length IAV genomic sequences and quantified their genotypes abundance in serial samples from an 81-year-old male co-infected with H1N1 and H7N9 subtypes IAV. A total of 26 high diversity nucleotide loci was detected, in which the A-G base transversion was the most abundant substitution type (67 and 64%, in H1N1 and H7N9, respectively). Seven significant amino acid variations were detected, such as NA:H275Y and HA: R222K in H1N1 as well as PB2: E627K and NA: K432E in H7N9, which are related to viral drug-resistance or mammalian adaptation. Furtherly, we retrieved 25 H1N1 and 22 H7N9 genomic segment haplotypes from the eight samples based on combining high-diversity nucleotide loci, which provided a more concise overview of viral quasispecies composition and dynamics. Our approach promotes the popularization of viral quasispecies analysis in a complex infectious disease, which will boost the understanding of viral infections, pathogenesis, evolution, and precision medicine.

Keywords: H1N1 and H7N9, quasispecies, composition and dynamics, SMRT, precision medicine

OPEN ACCESS

Edited by:
Peng Wang,
Harbin Medical University, China

Reviewed by:
Yanqun Wang,
Guangzhou Medical University, China
Min Guo,
University of Macau, China

****Correspondence:***
Guangyi Fan
fanguangyi@genomics.cn
Jingcao Pan
jingcaopan@sina.com

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

Received: *06 August 2021*
Accepted: *10 September 2021*
Published: *03 November 2021*

Citation:
Lin P, Jin T, Yu X, Liang L, Liu G, Jovic D, Sun Z, Yu Z, Pan J and Fan G (2021) Composition and Dynamics of H1N1 and H7N9 Influenza A Virus Quasispecies in a Co-infected Patient Analyzed by Single Molecule Sequencing Technology. Front. Genet. 12:754445. doi: 10.3389/fgene.2021.754445

INTRODUCTION

Influenza A virus (IAV) is a contagious pathogen that constantly infects many hosts, including but not limited to humans, birds, and pigs (Medina and García-Sastre, 2011). Annual influenza virus infections have significant health and economic burdens to mankind and livestock (Gordon and Reingold, 2018). IAV is a member of the *Orthomyxoviridae* family, and its genome contains eight negative-sense single-stranded RNA segments, ranging from 850 to 2,350 bp (Pleschka, 2013; Hutchinson, 2018). IAV can be subtyped as HxNy by viral surface antigens hemagglutinin (HA) and neuraminidase (NA) proteins, which govern the viral lifecycle at cellular entry and release of virions (Dou et al., 2018). So far, eighteen different HA subtypes (H1-H18) and eleven different NA

subtypes (N1-11) have been observed (Boktor and Hafner, 2019). There are two common IAV cellular receptors:α-2,3-Sialic acid (α-2,3-SA) and α-2,6-Sialic acid (α-2,6-SA) in hosts (Nelli et al., 2010; França et al., 2013; Byrd-Leotis et al., 2017; Chen et al., 2018a; Xu et al., 2019). The avian influenza viruses such H7N9 preferentially recognize α-2,3-SA receptors, while human influenza viruses such H1N1 has a priority to α-2,6-SA receptors (Xiong et al., 2013; de Graaf and Fouchier, 2014). The α-2,6-SA receptors are dominant in the upper respiratory tract (URT) of humans, while α-2,3-SA receptors are relatively more abundant than α-2,6-SA receptors found in the lower respiratory tract (LRT) of humans (Walther et al., 2013; Lakdawala et al., 2015; Long et al., 2019).

Influenza A viruses in a host exist as a population including thousands of virions containing closely related (but nonidentical) genomes, also called quasispecies (Lauring and Andino, 2010; Martínez et al., 2012; Watanabe et al., 2018; Bonomo et al., 2019; Domingo and Perales, 2019). These closely related genomes result from error-prone replication and frequent reassortment of influenza virus genomes (Steel and Lowen, 2014; Pauly et al., 2017). The complicated interactions (cooperativity or interference) among genomes and their productions collectively determine the biological or medical implications of a viral population such as fitness, virulence, pathogenesis, immune escape, or drug resistance (Vignuzzi et al., 2006; Sanz-Ramos et al., 2008; Aragri et al., 2016; Schuster, 2016; Perales, 2020). Therefore, it is primary to reveal the composition and dynamic of viral quasispecies to better understand viral infection, adaptation, and evolution at the level of population (Xue et al., 2017; Donohue et al., 2019; Jary et al., 2020).

Achieving thousands of full-length genomes in a viral population is decisive for quasispecies composition. Previous short-read massively parallel sequencing (MPS) projects have collected abundant consensus genomic sequences (CGSs) and single nucleotide variants (SNVs) to explore influenza virus quasispecies (Van den Hoecke et al., 2015; Ali et al., 2016; McGinnis et al., 2016). Nevertheless, the quasispecies composition is still unclear because of degenerated CGSs and scattered SNVs rooted in short reads from MPS (Schadt et al., 2010; Beerenwinkel et al., 2012; Chen et al., 2018b). The long-read single-molecule real-time sequencing (SMRT) provides an access to full-length influenza virus genomes even with a low frequency in a viral population (Ardui et al., 2018; Lui et al., 2019). The circular consensus sequencing (CCS) reads (average length 13.5 kb) produced by SMRT are 5–15 times as long as the genomic RNAs of influenza A virus, avoiding the fragmentation and assembly of genomes before and after sequencing (Wenger et al., 2019a; Van Poelvoorde et al., 2020).

The sample co-infected with two IAV subtypes is a very good opportunity to embody the advantage of SMRT in distinguishing different-subtype IAV genomic sequences and to quantify their abundances. The co-infection in avian hosts is common (29.59% in the live poultry market during 2016–2019 in China) (Bi et al., 2020). However, to our knowledge, there were only two human cases co-infected with two IAV subtypes reported in China since 2013. One case was a 15-year-old male co-infected with H7N9 and H3N2 in Jiangsu Province in April, 2013 and the other was a 58-year-old male with H7N9 and H1N1 in Zhejiang Province in January, 2014 (Zhu et al., 2013; Li et al., 2014). In this study, an 81-year-old male was diagnosed with H1N1 and H7N9 IAV by RT-PCR in Zhejiang Province in January 2016. Furtherly, the composition and dynamics of H1N1 and H7N9 IAV quasispecies in eight serial samples from this patient were revealed by SMRT, which provided a window to observe the viral quasispecies changes during the patient's hospitalization treated with anti-viral drug oseltamivir.

MATERIALS AND METHODS

Patient, Symptoms and Therapies

An 81-year-old male had a slight cough and chest distress on 1/12/2016 at his home in Xihu District, Hangzhou City, Zhejiang Province, China. On the morning of 1/15/2016, the symptoms worsened with a nasty cough, chest distress and fever. That afternoon, the man went to the local community hospital, where the temperature was 38.5 C, and then he was sent to the Hangzhou First People's Hospital for medical treatment. The examination showed that the white blood cell count (WBC) was 13.4×10^9/L, the percentage of neutrophils (N%) was 92.2%, and the C-reactive protein (CRP) was 110 mg/L. The chest radiographs showed an infection of the right lower lung. The mezlocillin sodium and sulbactam sodium were given for intravenous injection as an anti-infective therapy. The patient was sent to the respiratory department for hospital treatment and pneumonia was confirmed on 1/16/2016.

The next day, a PCR result from a throat swab was positive on influenza A virus and the patient was sent to infection ward for further treatment which included oseltamivir (75mg/bid) and meropenem drugs. On 1/19/2016, patient's symptoms worsened further and chest radiographs confirmed that infections spread on both lungs. The patient received endotracheal intubation and then admitted to intensive care unit (ICU)The treatment was continued, the dose of oseltamivir was doubled (150mg/bid) and meropenem was changed to imipenem. On the same day, the patient's RT-PCR test taken from the throat was positive on the M gene, H7 gene, N9 gene and H1 gene of influenza A virus. Next day, the patient was transferred to Hangzhou Xixi Hospital for treatment in isolation where the anti-viral oseltamivir (150mg/bid) continued to be given until 2/20/2016. Although, the symptomatic treatments such as diuresis, analgesia, vasodilation and nutritional support has been given in the Xixi hospital, the patient did not show signs of improvement, and passed away on 2/28/2016.

In the patient's anamnesis it is stated that the patient went to local live poultry market and bought a live duck at a merchant's site about a week before the first symptoms appeared on 1/12/2016. The live duck was slaughtered, depilated, and bellied by the merchant at his site. After returning home, the patient salted the duck. The patient had a history of hypertension and denied the history of diabetes, viral hepatitis, tuberculosis, and other diseases. There is no history of trauma, surgery, or blood transfusion. Denying any history of drug or food allergies.

Samples and RT-PCR

Ten serial samples were collected from this patient from 1/19/2016 to 2/19/2016, including seven throat swabs and three sputa. The swabs and sputa were placed into 1 ml viral transport medium, transported to the laboratory within 24 h at 4°C, and then frozen at −80°C. Viral RNAs were extracted from samples using a RNeasy Mini Kit (QIAGEN, Germany). Identification of influenza A virus was achieved by RT-PCR using specific primers targeting the M, H7, N9, and H1 gene according to the protocol provided by WHO manual (Organization, W.H, 2002). This study was approved by the Institutional Review Board of BGI (NO.BGI-IRB 16008).

Single-Molecule Real-Time Sequencing

Top eight samples were taken to perform single-molecule real-time sequencing (SMRT). The cDNAs were synthesized from viral RNAs by reverse transcription using Uni12 and Uni13 primers (Bi et al., 2016). The PCR was performed using a Phusion High-Fidelity PCR Kit (New England Biolabs) utilizing the barcoded influenza A virus general primers (**Supplementary Table S1**) (Mei et al., 2016). The concentration of PCR product was quantified by the Agilent Technologies 2,100 bioanalyzer. The two corresponding volumes of PCR products (containing equal mass of dsDNA) were mixed into one sample and quantified in the bioanalyzer again. About 2–3 μg mixed sample was used to SMRTbell library construction following the 2 kb template preparation protocol (Roberts et al., 2013b). The sequencing was performed on a PacBio RS II instrument (Pacific Biosciences, USA) with one SMRT Cell used for each library, using P6/C4 chemistry with a 4 h movie (Bull et al., 2016). SMRTbell adapter sequences were removed and circular consensus sequence (CCS) reads were achieved with SMRT Analysis v2.3 (Roberts et al., 2013a).

Sequence Quality Control

The raw CCS reads were filtered by removing low quality reads (length<800bp, passes<5 or estimated accuracy<99.9%). The 800 bp length near the lower limit of influenza A virus genomic RNAs was used to exclude non-full-length genomic sequences. The other two criteria ensured reads with at least 99.9% estimated accuracy and necessary passes (Korlach, 2015). The sequencing error bases (frequency<0.3%) were additionally corrected to improve sequence reliability as follows: First, the remaining sequences after filtration were split into corresponding samples by 100% base match with barcodes. Then, the sequences of one sample were grouped by subtypes and genomic segments according to the sequence annotation result against an influenza virus genomes database downloaded from NCBI (https://ftp.ncbi.nih.gov/genomes/INFLUENZA) using BLASR (v5.1 with options: -bestn 1) (Chaisson and Tesler, 2012). All full-length genomic sequences of one group were aligned end to end using MUSCLE (v3.8.31 with default options) (Edgar, 2004). Following, the number and percentage of base A/C/T/G in the same nucleotide locus of genomic sequences were stated. Finally, the very low frequency base which percentage was less than 0.3% in the number of four type bases of the same nucleotide locus, was replaced with the dominant type of base with the largest proportion in this nucleotide locus.

Diversity Index of Genomic Sequences

The diversity index (Shannon entropy) of one group sequences is calculated by the formula (Crooks and Brenner, 2004):

$$S = -100 * \sum_{i=1}^{n} P_i * \log_2 P_i$$

In which S is the Shannon entropy and P_i is the ration of the number of one type of sequence to the number of total types of sequence in one group.

Nucleotide Loci With High-Diversity Base Composition

In order to screen out nucleotide loci with high-diversity base composition, we stated the number and percentage of base A/C/T/G in one nucleotide locus and screened out the loci in which the percentages of at least two types of bases were more than 10%.

RESULTS

Sequences Quality Control

The clinical symptoms and therapeutic schedule of this patient were recorded in **Table 1**. Ten samples were collected from this patient on different days, including seven throat swabs (S1-4, S7-10) and three sputa (S5-7). The collected date, sample types and Ct value of RT-PCR for H1N1 and H7N9 of each sample were listed in **Table 2**. The top eight samples (S1-8) were performed using SMRT with four SMRT cells (S9 and S10 were RT-PCR negative for influenza A virus). A total of 142,496 CCS reads (221.72 Mb) were generated from four SMRT cells, of which 82,471 high quality reads (≥99.9% estimated accuracy) were selected for further analysis according to strict filtering criteria. The IAV mutation rate was about 0.018–0.025%, which means that each replicated influenza genome (~13 kb) contained an average of 2–3 mutations (Pauly et al., 2017). However, the estimated base sequencing error rate of high-quality CCS reads (~0.1%) in this study, was notable higher than the normal replication mutation rate (0.018–0.025%) of IAV genomes. Therefore, it was necessary to correct sequencing error bases in the prevention of taking them for real mutations. Finally, 69 group sequences were clustered from four SMRT cells according to samples, subtypes, and genomic segments (**Figure 1A**). Among them, 58 group sequences were with a satisfactory abundance (the sequences number ≥20). In case that the base type (A/C/G/T) in one nucleotide locus of one group sequences was less than 0.3%, it was considered as sequencing error base. The base sequencing error rates of 58 group sequences ranged from 0.13 to 0.21%, approximate to the estimated sequencing error rate of 0.1%, which are composed of 78.37% mismatches, 13.78% deletions and 7.16% insertions (**Figures 1B,C** and **Supplementary Table S2**). Thus, we set 0.3% as the cutoff value to screen out sequencing error base types in nucleotide loci of each group sequences and correct them with dominant base types in these loci.

TABLE 1 | The clinical symptoms and therapeutic schedule of this patient.

Date	Symptoms	Therapies
1/12/2016	Cough and chest tightness	NA
1/15/2016	Cough and chest tightness worsen, fever	The patient was sent to hospital
1/17/2016	RT-PCR positive for influenza A virus	Oseltamivir (75mg/bid) + Meropenem
1/19/2016	Lung injury was confirmed by Chest radiographs	Oseltamivir (150mg/bid) + Imipenem, Trachea cannula, ICU
1/20/2016–2/20/2016	There were no signs of improvement	Oseltamivir (150mg/bid), Symptomatic treatment (diuresis, analgesia, vasodilation, nutritional support)
2/28/2016	This patient passed away	NA

NA = Not available.
bid: Drug use frequency, twice daily.
ICU = Intensive Care Unit.

TABLE 2 | The sampling information and RT-PCR screening of H1N1 and H7N9.

Name	Collected date	Sample type	RT-PCR screening			
			M(Ct)	H7(Ct)	N9(Ct)	H1(Ct)
S1	1/19/2016	Throat swab	+ (28)	+ (29)	+ (30)	+ (28)
S2	1/28/2016	Throat swab	+ (26)	+ (38)	+ (38)	+ (27)
S3	2/02/2016	Throat swab	+ (30)	+ (38)	+ (38)	+ (30)
S4	2/03/2016	Throat swab	+ (30)	–	–	+ (31)
S5	2/04/2016	Sputum	+ (28)	+ (36)	+ (37)	+ (29)
S6	2/05/2016	Sputum	+ (29)	–	–	+ (31)
S7	2/06/2016	Sputum	+ (28)	+ (31)	+ (31)	+ (29)
S8	2/12/2016	Throat swab	+ (31)	–	–	+ (31)
S9	2/16/2016	Throat swab	–	–	–	–
S10	2/19/2016	Throat swab	–	–	–	–

Ct: The cycle threshold value of RT-PCR.
+: Positive RT-PCR result (0 < Ct < 40).
-: Negative RT-PCR result.

Monitoring the Composition and Dynamics of H1N1 and H7N9 Sequences

All the influenza virus genomic sequences produced by SMRT were clustered into 69 groups by eight samples, two subtypes, and eight genomic segments (**Figure 1A**). To monitor the composition and dynamic of H1N1 and H7N9 genomic sequences, we calculated the number of sequence reads, the number of sequence types, and the diversity index of sequence types in each group (**Figure 2**). The 16 groups (eight from H1N1 and eight from H7N9) in the first sample S1 confirmed that this patient was coinfected with H1N1 and H7N9 IAV. Interestingly, in sample S1, although the number of H1N1 and H7N9 sequence reads were almost equal, the diversity index of H7N9 sequence types was obviously higher than that of H1N1 (**Figure 2**). The reasons for the diversity difference of sequence types between H1N1 and H7N9 were unclear. One possible explanation was that in the upper respiratory tract (URT) environment with the dominant α-2,6-SA receptors preferentially recognized by H1N1 virus; the H7N9 virus had to generate more various genomic sequences to adapt to the relative hostile environment (Chen et al., 2013).

Furtherly, there was a sharp decrease of H7N9 viral load in the URT samples from S1 to S4. But the H7N9 viral load was still relative abundant in the subsequent LRT samples from S5 to S7 (**Table 2** and **Figure 2B**). This might indicate that the H7N9 virus had transferred to LRT environment from URT (Gao et al., 2013). Meanwhile, there was an obvious increase of H1N1 viral load in the URT samples from S1 to S2 (**Figure 2A**). The reason for this increase might be due to the transfer of H7N9 from URT to LRT that contributes freeing up more cellular resources for H1N1 growth in URT environment.

High Diversity Nucleotide Loci in H1N1 and H7N9 Genomes

Fifteen and eleven high diversity nucleotide loci were detected in H1N1 and H7N9 genomes, respectively, in which the A-G transversion was the most abundant substitution type (67% in H1N1 and 64% in H7N9) (**Table 3**). In H1N1, another two substitutions were C-T and A-C transversion (20 and 13% respectively). IN H7N9, the C-T, A-C and G-T took the same 0.09% proportion respectively (**Table 3**). In terms of amino acid, the percentage of non-synonymous mutation was greater than synonymous mutation, especially the percentage of non-synonymous mutation was up to 91% (10/11) in H7N9, comparing with the non-synonymous mutation of 66% (10/15) in H1N1(**Table 3**). In H1N1, all five synonymous mutations are in the internal genes of IAV, including two in NS gene (R78R and L69L), one in PB2 gene (T25T), one in NP gene (E64E) and one in M gene (R134R). In H7N9, the only one synonymous mutation was R753R in internal PB2 gene (**Table 3**). The H7N9 with a high frequency of non-synonymous mutation might indicate that H7N9 as avian-origin influenza virus was subjected to higher selection pressures in the human host (Wei et al., 2012; Xu et al., 2019). It is noteworthy to mention that among of these high-diversity loci, four mutations in H1N1 were involved in the evolution or viral drug-resistance, such as the R222K in the HA protein and the H275Y, A204T and K207R in NA protein (**Table 3**). In H7N9, three mutations were related to host adaption or viral drug-resistance, including the G685R and E627K in PB2 protein, and K432E in NA protein (**Table 3**).

The Haplotype Analysis of H1N1 and H7N9 Virus Quasispecies

We achieved 25 and 22 haplotypes for H1N1 and H7N9 genomic segments in all eight samples based on combining high diversity nucleotide loci in the same genomic sequences, respectively. For

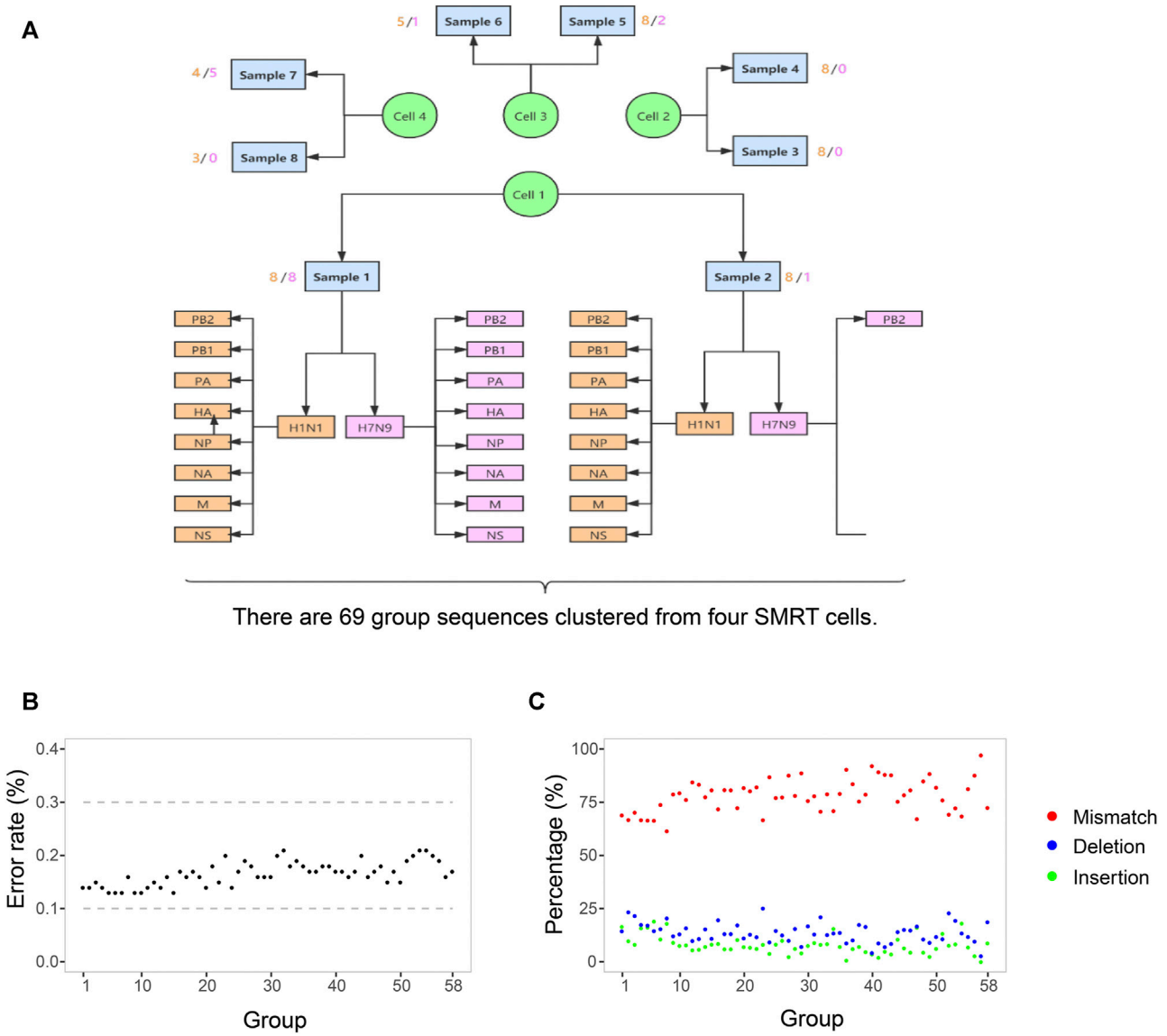


FIGURE 1 | The process of sequences grouping and the SMRT sequencing error rate. **(A)** There are 69 group sequences clustered from four SMRT cells by samples, subtypes, and genomic segments. The sequences in one group indicates that they are with the same sample resource, the same Influenza subtype, and the same genomic RNA segments. The light yellow and pink figures around each sample indicate the number of groups of H1N1 and H7N9. **(B)** The base sequencing error rates of 58 group sequences. In each group, the number of sequences is more than 20. The rate is the ratio of the number of sequencing error bases to the number of total bases in one group sequences. **(C)** The three types and their percentage of sequencing error bases. red indicates mismatch, blue indicates deletion, and green indicates insertion. The percentage is the ratio of the number of one type of sequencing error bases to the number of total sequencing error bases in one group sequences. The detailed values about sequencing error bases in each group are listed in **Supplementary Table S2**.

instance, we obtained seven haplotypes of NA in H1N1 based on four high diversity nucleotide loci (**Table 4**). The bases on high diversity nucleotide loci of each haplotype and its abundance in each sample were listed in **Supplementary Table S4**. Then the composition and dynamics of viral quasispecies in this patient co-infected with H1N1 and H7N9 IAV were displayed along the clinical treatment process (**Figure 3**). In NA gene, the haplotype Hap_2 replaced Hap_1 as the dominant haplotype with more than half proportion (53.11%,1,119/2,107) in sample S2, which contain H275Y drug-resistant mutation. However, Hap_2 in NA failed to be continuously dominant and Hap_1 return the dominant haplotype in following samples (S3 to S6) (**Figure 3A**). Similarly, we found Hap_3 of HA with antigenic drift mutation R222K transiently become dominant in sample S6 (56.19%, 127/226) (**Figure 3A**). Compared with the genomic segment integrity of H1N1 viral quasispecies among almost all samples, The H7N9 quasispecies had a poor integrity except for the first sample S1 (**Figure 3B**). In conclusion, the haplotype provides a more concise overview of viral quasispecies composition and dynamics than whole genomic sequences (**Figure 2** and **Figure 3**).

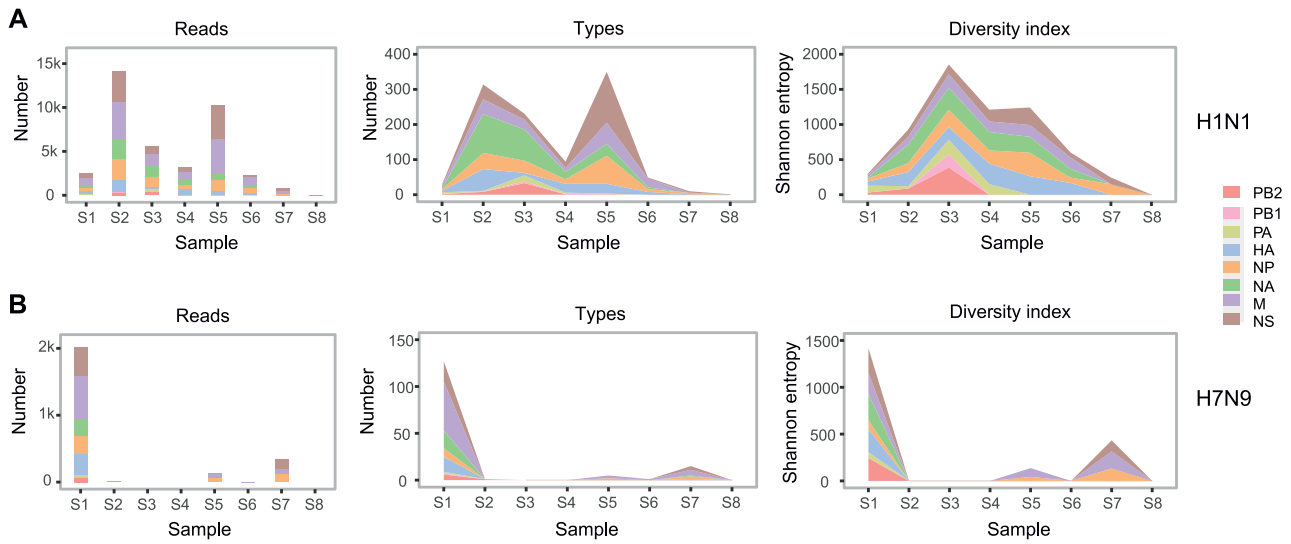


FIGURE 2 | The abundance and diversity of H1N1 and H7N9 genomic sequences. In H1N1 **(A)** and H7N9 **(B)**, from left to right, three subgraphs respectively indicate the number of sequence reads, the number of sequence types and the sequence diversity index in each group. In each subgraph, eight colors represent eight genomic segments of influenza A virus. One sequence type is defined that there are one or more different bases from any other sequence. The sequence diversity index is calculated as the Shannon entropy formula described in the "Materials and Methods". The detailed values about abundance and diversity of genomic sequences are listed in **Supplementary Table S3.**

TABLE 3 | The high-diversity nucleotide loci in H1N1 and H7N9 genomes.

Subtype	Segment	Nucleotide loci	Base composition[a]				Amino acid[b]
			A	C	G	T	
H1N1	PB2	2045	124	0	608	0	G673D
	PB1	99	0	38	0	138	T25T
	PA	182	0	58	0	484	F53S
	HA	86	148	3,071	0	0	D18E
	HA	697	127	0	3,092	0	R222K
	NP	133	155	0	6,391	0	G30R
	NP	237	101	0	6,445	0	E64E
	NA	147	127	5,288	0	0	Q43K
	NA	630	107	0	5,308	0	A204T
	NA	640	5,280	0	135	0	K207R
	NA	843	0	4,217	0	1,198	H275Y
	M	427	663	0	11,809	0	R134R in M1
	M	835	107	0	12,365	0	W41* in M2
	NS	260	9,591	0	114	0	R78R in NS1
	NS	705	100	0	9,605	0	L69L in NEP
H7N9	PB2	1906	35	0	31	0	E627K
	PB2	2080	10	0	56	0	G685R
	PB2	2,286	52	0	14	0	R753R
	HA	722	43	0	269	0	G234D
	HA	1,163	208	0	104	0	Q381R
	NP	661	76	0	0	385	F206I
	NA	339	165	76	0	0	R107S
	NA	448	215	0	26	0	T144A
	NA	1,312	168	0	73	0	K432E
	NS	523	0	101	0	457	L166P in NEP
	NS	707	0	0	497	61	S70I in NEP

[a]*The number of base A/C/G/T in one nucleotide locus.*
[b]*The site and types of amino acid corresponding to base composition.*
Synonymous mutations.
**: Termination codon.*

TABLE 4 | The forming and abundance of NA haplotypes in H1N1.

Haplotype	Loci and bases				Abundance in each sample[a]								
	147	630	640	843	S1	S2	S3	S4	S5	S6	S7	S8	Total[b]
hap_1	C	G	A	C	282 (100%)	980 (46.51%)	1,228 (91.99%)	460 (78.23%)	710 (90.91%)	322 (100%)	0	0	3,982
hap_2	C	G	A	T	0	1,119 (53.11%)	0	0	71 (9.09%)	0	0	0	1,190
hap_3	A	G	G	C	0	0	0	126 (21.43%)	0	0	0	0	126
hap_4	C	A	A	C	0	0	107 (8.01%)	0	0	0	0	0	107
Hap_3	C	G	G	T	0	8 (0.38%)	0	0	0	0	0	0	8
hap_6	C	G	G	C	0	0	0	1 (0.17%)	0	0	0	0	1
hap_7	A	G	A	C	0	0	0	1 (0.17%)	0	0	0	0	1

[a]*The number and percentage of haplotype in each sample.*
[b]*The number of haplotypes in all samples.*

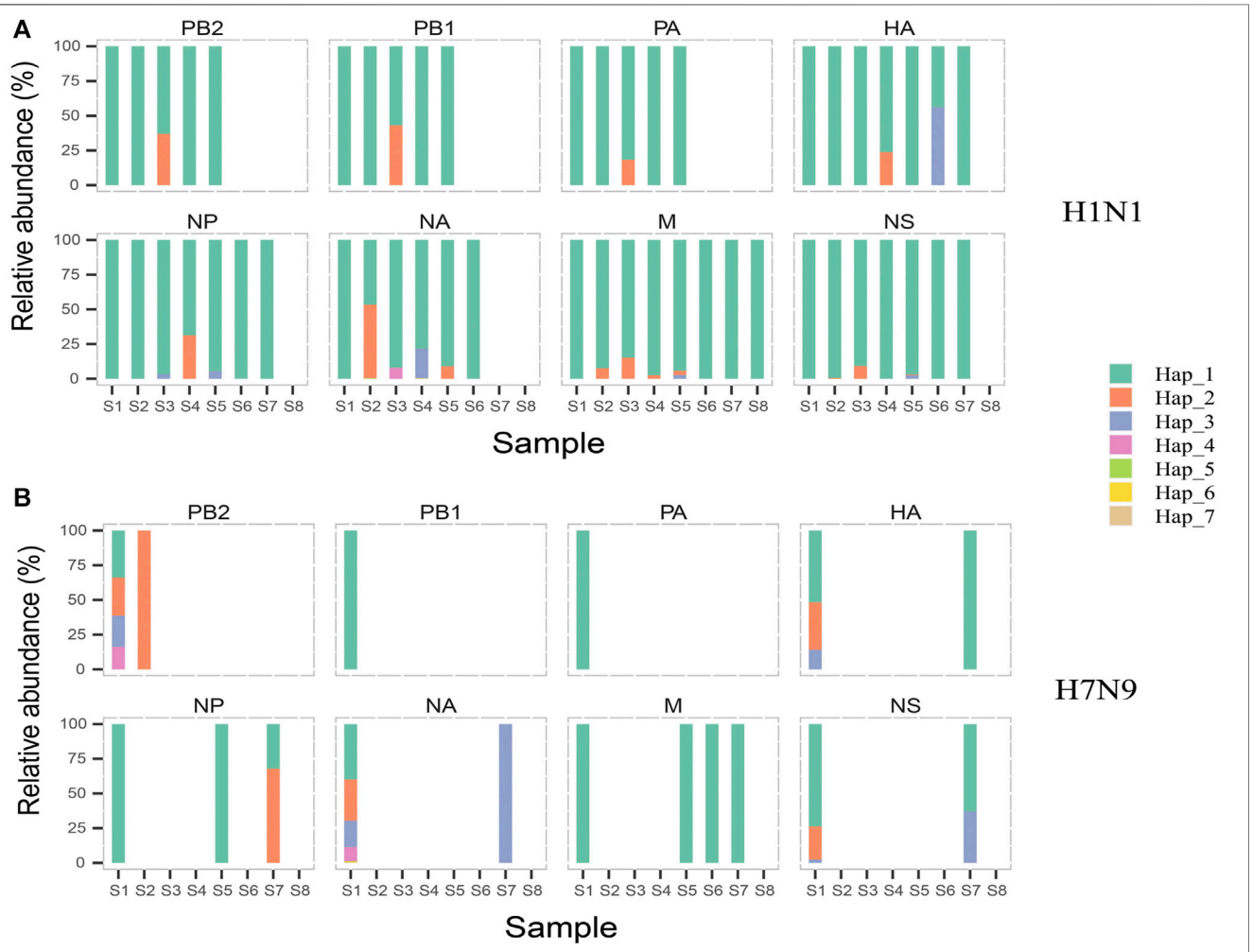


FIGURE 3 | The composition and dynamics of H1N1 and H7N9 virus quasispecies. In H1N1 **(A)** or H7N9 **(B)**, eight subgraphs respectively eight genomic segments of influenza A virus. In each subgraph, different colors indicate different haplotypes. In the same subgraph, the same color represents the same haplotype, while in different subgraphs, same color is irrelevant. The forming and abundance of each haplotype are listed in **Supplementary Table S4**.

DISCUSSION

Four mutations in H1N1 and three in H7N9 were related to viral drug-resistance, host adaption or evolution. The H275Y in NA protein of H1N1 was the widely investigated drug-resistant mutation against oseltamivir as the commonly used first-line drug for the treatment or prophylaxis of influenza (Hurt et al., 2012; Vidaña et al., 2020). Another two mutations A204T and K207R in NA protein of H1N1 were recently reported to have effects on drug-resistance and vaccine efficacy (Nandhini and Sistla, 2020; Skowronski et al., 2020). The antigenic drift mutation R222K in the HA protein was believed to play a role in virus evolution by altering receptor binding specificity (Al Khatib et al., 2019). For H7N9, the mutation E627K on PB2 is a well-characterized host adaption mutation from the avian signature Glu (E) to the mammalian-adapted signature Lys (K), which have been associated with enhanced polymerase activity, high virus replication and pathogenicity in humans (Baccam et al., 2001). Besides, the mutation G685R on PB2 also help to promotes the mammalian adaptation of avian influenza virus (Baccam et al., 2003; Capitán et al., 2011). Interestingly, the mutation K432E on NA, alone or together with mutation H275Y on NA, had a significant impact on the binding pattern and affinity of oseltamivir for neuraminidase, rendering neuraminidase less susceptible (Aguirre and Manrubia, 2008).

The viral quasispecies as a viral population plays a very important role in the process of viral infection, adaption, and evolution through complex cooperative or competitive interactions (Domingo et al., 2012). A population of viruses can be partitioned into subpopulations by the genetic similarities (Baccam et al., 2001; Baccam et al., 2003). The spatial interactions of subpopulations have effects on host cell availability and defense responses (Aguirre and Manrubia, 2008; Capitán et al., 2011). A specific cooperative interaction is that mixed populations of D151 and G151 variants in H3N2 viruses grow better than pure populations of either variant, in which one subpopulation is good at entering new cells, while the other is better at exiting cells to spread the infection (Xue et al., 2016). In our case, the co-existent viral population of H1N1 and H7N9 might help to H7N9 subpopulation migration from URT to LRT and growth in LRT. The H1N1 subpopulation can grow in both URT and LRT of this patient, while H7N9 grow better in LRT than URT, which is related to the different distribution of α-2,3-SA and α-2,6-SA receptors in URT and LRT, as well as the preferential recognition of H1N1 and H7N9 with two receptors (de Graaf and Fouchier, 2014; Byrd-Leotis et al., 2017). The H7N9 subpopulation was easier to transfer to the LRT with the assistance of H1N1 subpopulation in the co-existence of H1N1 and H7N9 than that only in H7N9.

Besides, the competitive interactions among viral quasispecies are also reported (Andino and Domingo, 2015). An example is that in co-infected cells with wild-type polioviruses at a high multiplicity of infection (MOI) and drug-resistant virus at a much lower MOI, the yield of drug-resistant virus was significantly reduced to 3–7% of the output from a single infection due to the interference of chimeric capsid formation (Crowder and Kirkegaard, 2005). We detected the drug-resistant mutation H275Y in NA protein and antigenic drift mutation R222K in HA protein. But both failed to be continuously dominant in the subsequent viral quasispecies composition. A possible explanation is that the forming of the viral particles containing mutations were interfered by normal strains with a similar mechanism illustrated in above poliovirus.

The composition, complexity and dynamic of a viral quasispecies determined its biological or medical implications, such as host range, pathogenesis, and coping with selection pressure (Roedig et al., 2011; Gregori et al., 2016; Barbezange et al., 2018). In this work, the composition and dynamics of viral quasispecies in a patient co-infected with H1N1 and H7N9 influenza A virus were clearly revealed along his treatment process using the single-molecule real-time sequencing (SMRT). Compared with the flaws of consensus genomic sequences (CGSs) and single nucleotide variants (SNVs) by short-read massively parallel sequencing (MPS), the SMRT embody the obvious advantage in investigating the complex haplotype distribution of an IAV population, especially a population coexisting with two subtype IAV. Because of a human co-infected with two subtype IAV is very rare, the single patient in this study restricted the conclusions. Fortunately, the co-existence of two or more subtype IAV in wildfowls is not rare. These works studying the composition and dynamics of viral quasispecies in wildfowls co-infected with multi subtype IAV will be conducted in future. One scarcity of SMRT is the limit of detection (LOD) not good enough to detect the low abundant IAV sequences. For example, when the Ct values of RT-PCR for IAV in samples of this study were less than 30, SMRT was hard to generate IAV genomic sequences. This is also the reason why several types of sequences were not detected in some samples and only 69 rather than 128 groups to be conducted in the study. Besides, the expensive sequencing costs is another limitation. With the optimization and upgrade of SMRT in terms of limit of detection and sequencing cost, using SRMT to reveal the composition and dynamic of influenza A virus quasispecies will become a necessary method for studying viral biological behaviors and medical implications, which will boost the understanding of viral infections, pathogenesis, evolution, and precision medicine (Nakano et al., 2017; Wenger et al., 2019b; Beaulaurier et al., 2020).

Accession Number

The data that support the findings of this study have been deposited into CNGB Sequence Archive (Guo et al., 2020)of CNGBdb (Chen et al., 2020)with accession number CNP0001131.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://db.cngb.org/search/project/CNP0001131/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of BGI. The patients/

participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XY and ZS collected samples and performed experimental detection. PL, TJ, and LL performed single molecule real-time sequencing. PL, LL and GL wrote bioinformatics pipeline. PL, TJ, JP, and GF designed study and analyzed data. PL, GL, ZY, DJ, JP, and GF contributed to writing this study. All authors had full access to the final version of the manuscript and agreed to its submission.

FUNDING

This study was supported by Shenzhen Science and Technology Research and Development Project (No. JCYJ20151029151932602). The funders had no role in the study design, data collection and interpretation, or the decision to submit the study for publication.

ACKNOWLEDGMENTS

We thank staff of Hangzhou First People's Hospital and Hangzhou Xixi Hospital for their clinical care given to the patient and facilitating access to the relevant medical records. We also thank China National GeneBank for storing the data from this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.754445/full#supplementary-material

REFERENCES

Aguirre, J., and Manrubia, S. C. (2008). Effects of Spatial Competition on the Diversity of a Quasispecies. *Phys. Rev. Lett.* 100 (3), 038106. doi:10.1103/PhysRevLett.100.038106

Al Khatib, H. A., Al Thani, A. A., Gallouzi, I., and Yassine, H. M. (2019). Epidemiological and Genetic Characterization of pH1N1 and H3N2 Influenza Viruses Circulated in MENA Region during 2009-2017. *BMC Infect. Dis.* 19 (1), 314. doi:10.1186/s12879-019-3930-6

Ali, R., Blackburn, R. M., and Kozlakidis, Z. (2016). Next-Generation Sequencing and Influenza Virus: A Short Review of the Published Implementation Attempts. *HAYATI J. Biosciences* 23 (4), 155–159. doi:10.1016/j.hjb.2016.12.007

Andino, R., and Domingo, E. (2015). Viral Quasispecies. *Virology* 479-480, 46–51. doi:10.1016/j.virol.2015.03.022

Aragri, M., Alteri, C., Battisti, A., Di Carlo, D., Minichini, C., Sagnelli, C., et al. (2016). Multiple Hepatitis B Virus (HBV) Quasispecies and Immune-Escape Mutations Are Present in HBV Surface Antigen and Reverse Transcriptase of Patients with Acute Hepatitis B. *J. Infect. Dis.* 213 (12), 1897–1905. doi:10.1093/infdis/jiw049

Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. (2018). Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics. *Nucleic Acids Res.* 46 (5), 2159–2168. doi:10.1093/nar/gky066

Baccam, P., Thompson, R. J., Fedrigo, O., Carpenter, S., and Cornette, J. L. (2001). PAQ: Partition Analysis of Quasispecies. *Bioinformatics* 17 (1), 16–22. doi:10.1093/bioinformatics/17.1.16

Baccam, P., Thompson, R. J., Li, Y., Sparks, W. O., Belshan, M., Dorman, K. S., et al. (2003). Subpopulations of Equine Infectious Anemia Virus Rev Coexist *In Vivo* and Differ in Phenotype. *J. Virol.* 77 (22), 12122–12131. doi:10.1128/jvi.77.22.12122-12131.2003

Barbezange, C., Jones, L., Blanc, H., Isakov, O., Celniker, G., Enouf, V., et al. (2018). Seasonal Genetic Drift of Human Influenza A Virus Quasispecies Revealed by Deep Sequencing. *Front. Microbiol.* 9, 2596. doi:10.3389/fmicb.2018.02596

Beaulaurier, J., Luo, E., Eppley, J. M., Uyl, P. D., Dai, X., Burger, A., et al. (2020). Assembly-free Single-Molecule Sequencing Recovers Complete Virus Genomes from Natural Microbial Communities. *Genome Res.* 30 (3), 437–446. doi:10.1101/gr.251686.119

Beerenwinkel, N., Günthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and Opportunities in Estimating Viral Genetic Diversity from Next-Generation Sequencing Data. *Front. Microbio.* 3, 329. doi:10.3389/fmicb.2012.00329

Bi, Y., Chen, Q., Wang, Q., Chen, J., Jin, T., Wong, G., et al. (2016). Genesis, Evolution and Prevalence of H5N6 Avian Influenza Viruses in China. *Cell Host & Microbe* 20 (6), 810–821. doi:10.1016/j.chom.2016.10.022

Bi, Y., Li, J., Li, S., Fu, G., Jin, T., Zhang, C., et al. (2020). Dominant Subtype Switch in Avian Influenza Viruses during 2016-2019 in China. *Nat. Commun.* 11 (1), 5909. doi:10.1038/s41467-020-19671-3

Boktor, S. W., and Hafner, J. W. (2019). Influenza. Available from https://www.ncbi.nlm.nih.gov/books/NBK459363/.

Bonomo, M. E., Kim, R. Y., and Deem, M. W. (2019). Modular Epitope Binding Predicts Influenza Quasispecies Dominance and Vaccine Effectiveness: Application to 2018/19 Season. *Vaccine* 37 (24), 3154–3158. doi:10.1016/j.vaccine.2019.03.068

Bull, R. A., Eltahla, A. A., Rodrigo, C., Koekkoek, S. M., Walker, M., Pirozyan, M. R., et al. (2016). A Method for Near Full-Length Amplification and Sequencing for Six Hepatitis C Virus Genotypes. *BMC Genomics* 17, 247. doi:10.1186/s12864-016-2575-8

Byrd-Leotis, L., Cummings, R. D., and Steinhauer, D. A. (2017). The Interplay between the Host Receptor and Influenza Virus Hemagglutinin and Neuraminidase. *Int. J. Mol. Sci.* 18 (7). doi:10.3390/ijms18071541

Capitán, J. A., Cuesta, J. A., Manrubia, S. C., and Aguirre, J. (2011). Severe Hindrance of Viral Infection Propagation in Spatially Extended Hosts. *PLoS One* 6 (8), e23358. doi:10.1371/journal.pone.0023358

Chaisson, M. J., and Tesler, G. (2012). Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory. *BMC Bioinformatics* 13, 238. doi:10.1186/1471-2105-13-238

Chen, F. Z., You, L. J., Yang, F., Wang, L. N., Guo, X. Q., Gao, F., et al. (2020). CNGBdb: China National GeneBank DataBase. *Yi Chuan* 42 (8), 799–809. doi:10.16288/j.yczz.20-080

Chen, J., Zhao, Y., and Sun, Y. (2018). De Novo haplotype Reconstruction in Viral Quasispecies Using Paired-End Read Guided Path Finding. *Bioinformatics* 34 (17), 2927–2935. doi:10.1093/bioinformatics/bty202

Chen, X., Liu, S., Goraya, M. U., Maarouf, M., Huang, S., and Chen, J.-L. (2018). Host Immune Response to Influenza A Virus Infection. *Front. Immunol.* 9, 320. doi:10.3389/fimmu.2018.00320

Chen, Y., Liang, W., Yang, S., Wu, N., Gao, H., Sheng, J., et al. (2013). Human Infections with the Emerging Avian Influenza A H7N9 Virus from Wet Market Poultry: Clinical Analysis and Characterisation of Viral Genome. *The Lancet* 381 (9881), 1916–1925. doi:10.1016/s0140-6736(13)60903-4

Crooks, G. E., and Brenner, S. E. (2004). Protein Secondary Structure: Entropy, Correlations and Prediction. *Bioinformatics* 20 (10), 1603–1611. doi:10.1093/bioinformatics/bth132

Crowder, S., and Kirkegaard, K. (2005). Trans-dominant Inhibition of RNA Viral Replication Can Slow Growth of Drug-Resistant Viruses. *Nat. Genet.* 37 (7), 701–709. doi:10.1038/ng1583

de Graaf, M., and Fouchier, R. A. M. (2014). Role of Receptor Binding Specificity in Influenza A Virus Transmission and Pathogenesis. *EMBO J.* 33 (8), 823–841. doi:10.1002/embj.201387442

Domingo, E., and Perales, C. (2019). Viral Quasispecies. *Plos Genet.* 15 (10), e1008271. doi:10.1371/journal.pgen.1008271

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral Quasispecies Evolution. *Microbiol. Mol. Biol. Rev.* 76 (2), 159–216. doi:10.1128/mmbr.05023-11

Donohue, R. C., Pfaller, C. K., and Cattaneo, R. (2019). Cyclical Adaptation of Measles Virus Quasispecies to Epithelial and Lymphocytic Cells: To V, or Not to V. *Plos Pathog.* 15 (2), e1007605. doi:10.1371/journal.ppat.1007605

Dou, D., Revol, R., Östbye, H., Wang, H., and Daniels, R. (2018). Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. *Front. Immunol.* 9, 1581. doi:10.3389/fimmu.2018.01581

Edgar, R. C. (2004). MUSCLE: a Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics* 5, 113. doi:10.1186/1471-2105-5-113

França, M., Stallknecht, D. E., and Howerth, E. W. (2013). Expression and Distribution of Sialic Acid Influenza Virus Receptors in Wild Birds. *Avian Pathol.* 42 (1), 60–71. doi:10.1080/03079457.2012.759176

Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., et al. (2013). Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus. *N. Engl. J. Med.* 368 (20), 1888–1897. doi:10.1056/NEJMoa1304459

Gordon, A., and Reingold, A. (2018). The Burden of Influenza: a Complex Problem. *Curr. Epidemiol. Rep.* 5 (1), 1–9. doi:10.1007/s40471-018-0136-1

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J. I., Quer, J., and Domingo, E. (2016). Viral Quasispecies Complexity Measures. *Virology* 493, 227–237. doi:10.1016/j.virol.2016.03.017

Guo, X., Chen, F., Gao, F., Li, L., Liu, K., You, L., et al. (2020). CNSA: A Data Repository for Archiving Omics Data. *Database (oxford).* 2020, baaa055. doi:10.1093/database/baaa055

Hurt, A. C., Hardie, K., Wilson, N. J., Deng, Y. M., Osbourn, M., Leang, S. K., et al. (2012). Characteristics of a Widespread Community Cluster of H275Y Oseltamivir-Resistant A(H1N1)pdm09 Influenza in Australia. *J. Infect. Dis.* 206 (2), 148–157. doi:10.1093/infdis/jis337

Hutchinson, E. C. (2018). Influenza Virus. *Trends Microbiol.* 26 (9), 809–810. doi:10.1016/j.tim.2018.05.013

Jary, A., Leducq, V., Malet, I., Marot, S., Klement-Frutos, E., Teyssou, E., et al. (2020). Evolution of Viral Quasispecies during SARS-CoV-2 Infection. *Clin. Microbiol. Infect.* 26 (11), 1560–e4. doi:10.1016/j.cmi.2020.07.032

Korlach, J. (2015). Understanding Accuracy in SMRT Sequencing. Available from http://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf.

Lakdawala, S. S., Jayaraman, A., Halpin, R. A., Lamirande, E. W., Shih, A. R., Stockwell, T. B., et al. (2015). The Soft Palate Is an Important Site of Adaptation for Transmissible Influenza Viruses. *Nature* 526 (7571), 122–125. doi:10.1038/nature15379

Lauring, A. S., and Andino, R. (2010). Quasispecies Theory and the Behavior of RNA Viruses. *Plos Pathog.* 6 (7), e1001005. doi:10.1371/journal.ppat.1001005

Li, J., Kou, Y., Yu, X., Sun, Y., Zhou, Y., Pu, X., et al. (2014). Human Co-infection with Avian and Seasonal Influenza Viruses, China. *Emerg. Infect. Dis.* 20 (11), 1953–1955. doi:10.3201/eid2011.140897

Long, J. S., Mistry, B., Haslam, S. M., and Barclay, W. S. (2019). Host and Viral Determinants of Influenza A Virus Species Specificity. *Nat. Rev. Microbiol.* 17 (2), 67–81. doi:10.1038/s41579-018-0115-z

Lui, W.-Y., Yuen, C.-K., Li, C., Wong, W. M., Lui, P.-Y., Lin, C.-H., et al. (2019). SMRT Sequencing Revealed the Diversity and Characteristics of Defective Interfering RNAs in Influenza A (H7N9) Virus Infection. *Emerging Microbes & Infections* 8 (1), 662–674. doi:10.1080/22221751.2019.1611346

Martínez, M. A., Martrus, G., Capel, E., Parera, M., Franco, S., and Nevot, M. (2012). Quasispecies Dynamics of RNA Viruses. *Viruses: Essential Agents of Life*, 21–42. doi:10.1007/978-94-007-4899-6_2

McGinnis, J., Laplante, J., Shudt, M., and George, K. S. (2016). Next Generation Sequencing for Whole Genome Analysis and Surveillance of Influenza A Viruses. *J. Clin. Virol.* 79, 44–50. doi:10.1016/j.jcv.2016.03.005

Medina, R. A., and García-Sastre, A. (2011). Influenza A Viruses: New Research Developments. *Nat. Rev. Microbiol.* 9 (8), 590–603. doi:10.1038/nrmicro2613

Mei, K., Liu, G., Chen, Z., Gao, Z., Zhao, L., Jin, T., et al. (2016). Deep Sequencing Reveals the Viral Adaptation Process of Environment-Derived H10N8 in Mice. *Infect. Genet. Evol.* 37, 8–13. doi:10.1016/j.meegid.2015.10.016

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., et al. (2017). Advantages of Genome Sequencing by Long-Read Sequencer Using SMRT Technology in Medical Area. *Hum. Cel* 30 (3), 149–161. doi:10.1007/s13577-017-0168-8

Nandhini, P., and Sistla, S. (2020). Genetic Sequencing of Influenza A (H1N1) Pdm09 Isolates from South India, Collected between 2011 and 2015 to Detect Mutations Affecting Virulence and Resistance to Oseltamivir. *Indian J. Med. Microbiol.* 38 (3), 324–337. doi:10.4103/ijmm.IJMM_20_83

Nelli, R. K., Kuchipudi, S. V., White, G. A., Perez, B. B., Dunham, S. P., and Chang, K.-C. (2010). Comparative Distribution of Human and Avian Type Sialic Acid Influenza Receptors in the Pig. *BMC Vet. Res.* 6, 4. doi:10.1186/1746-6148-6-4

Organization, W.H. (2002). *WHO Manual on Animal Influenza Diagnosis and Surveillance*. Geneva, Switzerland.

Pauly, M. D., Procario, M. C., and Lauring, A. S. (2017). A Novel Twelve Class Fluctuation Test Reveals Higher Than Expected Mutation Rates for Influenza A Viruses. *Elife*. 6, e26437. doi:10.7554/eLife.26437

Perales, C. (2020). Quasispecies Dynamics and Clinical Significance of Hepatitis C Virus (HCV) Antiviral Resistance. *Int. J. Antimicrob. Agents* 56 (1), 105562. doi:10.1016/j.ijantimicag.2018.10.005

Pleschka, S. (2013). Overview of Influenza Viruses. *Curr. Top. Microbiol. Immunol.* 370, 1–20. doi:10.1007/82_2012_272

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013a). The Advantages of SMRT Analysis v2.3 Software Release sequencing. *Genome Biol.*

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013b). The Advantages of SMRT Sequencing. *Genome Biol.* 14, 405. doi:10.1186/gb-2013-14-6-405

Roedig, J. V., Rapp, E., Höper, D., Genzel, Y., and Reichl, U. (2011). Impact of Host Cell Line Adaptation on Quasispecies Composition and Glycosylation of Influenza A Virus Hemagglutinin. *PLoS One* 6 (12), e27989. doi:10.1371/journal.pone.0027989

Sanz-Ramos, M., Di´az-San Segundo, F., Escarmi´s, C., Domingo, E., and Sevilla, N. (2008). Hidden Virulence Determinants in a Viral Quasispecies *In Vivo*. *J. Virol.* 82 (21), 10465–10476. doi:10.1128/jvi.00825-08

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A Window into Third-Generation Sequencing. *Hum. Mol. Genet.* 19 (R2), R227–R240. doi:10.1093/hmg/ddq416

Schuster, P. (2016). Quasispecies on Fitness Landscapes. *Curr. Top. Microbiol. Immunol.* 392, 61–120. doi:10.1007/82_2015_469

Skowronski, D. M., Zou, M., Sabaiduc, S., Murti, M., Olsha, R., Dickinson, J. A., et al. (2020). Interim Estimates of 2019/20 Vaccine Effectiveness during Early-Season Co-circulation of Influenza A and B Viruses, Canada, February 2020. *Euro Surveill.* 25 (7). doi:10.2807/1560-7917.ES.2020.25.7.2000103

Steel, J., and Lowen, A. C. (2014). Influenza A Virus Reassortment. *Curr. Top. Microbiol. Immunol.* 385, 377–401. doi:10.1007/82_2014_395

Van den Hoecke, S., Verhelst, J., Vuylsteke, M., and Saelens, X. (2015). Analysis of the Genetic Diversity of Influenza A Viruses Using Next-Generation DNA Sequencing. *BMC Genomics* 16, 79. doi:10.1186/s12864-015-1284-z

Van Poelvoorde, L. A. E., Saelens, X., Thomas, I., and Roosens, N. H. (2020). Next-Generation Sequencing: An Eye-Opener for the Surveillance of Antiviral Resistance in Influenza. *Trends Biotechnol.* 38 (4), 360–367. doi:10.1016/j.tibtech.2019.09.009

Vidaña, B., Martínez-Orellana, P., Martorell, J. M., Baratelli, M., Martínez, J., Migura-Garcia, L., et al. (2020). Differential Viral-Host Immune Interactions Associated with Oseltamivir-Resistant H275Y and Wild-type H1N1 A(pdm09) Influenza Virus Pathogenicity. *Viruses* 12 (8). doi:10.3390/v12080794

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., and Andino, R. (2006). Quasispecies Diversity Determines Pathogenesis through Cooperative Interactions in a Viral Population. *Nature* 439 (7074), 344–348. doi:10.1038/nature04388

Walther, T., Karamanska, R., Chan, R. W. Y., Chan, M. C. W., Jia, N., Air, G., et al. (2013). Glycomic Analysis of Human Respiratory Tract Tissues and Correlation with Influenza Virus Infection. *Plos Pathog.* 9 (3), e1003223. doi:10.1371/journal.ppat.1003223

Watanabe, Y., Arai, Y., Kawashita, N., Ibrahim, M. S., Elgendy, E. M., Daidoji, T., et al. (2018). Characterization of H5N1 Influenza Virus Quasispecies with Adaptive Hemagglutinin Mutations from Single-Virus Infections of Human Airway Cells. *J. Virol.* 92 (11), e02004. doi:10.1128/JVI.02004-17

Wei, K., Chen, Y., Chen, J., Wu, L., and Xie, D. (2012). Evolution and Adaptation of Hemagglutinin Gene of Human H5N1 Influenza Virus. *Virus Genes* 44 (3), 450–458. doi:10.1007/s11262-012-0717-x

Wenger, A. M., Peluso, P., and Rowell, W. J. (2019a). Highly-accurate Long-Read Sequencing Improves Variantdetection and Assembly of a Human Genome. *Nat. Biotechnol* 37, 1155. doi:10.1038/s41587-019-0217-9

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., et al. (2019b). Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome. *Nat. Biotechnol.* 37 (10), 1155–1162. doi:10.1038/s41587-019-0217-9

Xiong, X., Martin, S. R., Haire, L. F., Wharton, S. A., Daniels, R. S., Bennett, M. S., et al. (2013). Receptor Binding by an H7N9 Influenza Virus from Humans. *Nature* 499 (7459), 496–499. doi:10.1038/nature12372

Xu, Y., Peng, R., Zhang, W., Qi, J., Song, H., Liu, S., et al. (2019). Avian-to-Human Receptor-Binding Adaptation of Avian H7N9 Influenza Virus Hemagglutinin. *Cel Rep.* 29 (8), 2217–2228. doi:10.1016/j.celrep.2019.10.047

Xue, K. S., Hooper, K. A., Ollodart, A. R., Dingens, A. S., and Bloom, J. D. (2016). Cooperation between Distinct Viral Variants Promotes Growth of H3N2 Influenza in Cell Culture. *Elife* 5, e13974. doi:10.7554/eLife.13974

Xue, Y., Wang, M. J., Yang, Z. T., Yu, D. M., Han, Y., Huang, D., et al. (2017). Clinical Features and Viral Quasispecies Characteristics Associated with Infection by the Hepatitis B Virus G145R Immune Escape Mutant. *Emerg. Microbes Infect.* 6 (3), e15. doi:10.1038/emi.2017.2

Zhu, Y., Qi, X., Cui, L., Zhou, M., and Wang, H. (2013). Human Co-infection with Novel Avian Influenza A H7N9 and Influenza A H3N2 Viruses in Jiangsu Province, China. *Lancet* 381 (9883), 2134. doi:10.1016/S0140-6736(13)61135-6

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

The reviewer MG declared a past co-authorship with one of the authors GF to the handling editor.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fgene.2021.775470

Identification of Genetic Predisposition in Noncirrhotic Portal Hypertension Patients With Multiple Renal Cysts by Integrated Analysis of Whole-Genome and Single-Cell RNA Sequencing

Yanjing Wu[1†], Yongle Wu[1†], Kun Liu[2], Hui Liu[3], Shanshan Wang[4], Jian Huang[5] and Huiguo Ding[1*]*

[1]Department of Gastroenterology and Hepatology, Beijing You'an Hospital, Affiliated with Capital Medical University, Beijing, China, [2]Department of General Surgery, Beijing Friendship Hospital, Affiliated with Capital Medical University, Beijing, China, [3]Department of Pathology, Beijing You'an Hospital, Affiliated with Capital Medical University, Beijing, China, [4]Beijing Institute of Hepatology, Beijing You'an Hospital, Affiliated with Capital Medical University, Beijing, China, [5]Experimental Center, Beijing Friendship Hospital, Affiliated with Capital Medical University, Beijing, China

OPEN ACCESS

Edited by:
Peng Wang,
Harbin Medical University, China

Reviewed by:
Chungang Feng,
Nanjing Agricultural University, China
Yi Ding,
Allen Institute for Brain Science, United States

***Correspondence:**
Jian Huang
huangj1966@hotmail.com
Huiguo Ding
dinghuiguo@ccmu.edu.cn

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

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

Received: *14 September 2021*
Accepted: *26 October 2021*
Published: *12 November 2021*

Citation:
Wu Y, Wu Y, Liu K, Liu H, Wang S, Huang J and Ding H (2021) Identification of Genetic Predisposition in Noncirrhotic Portal Hypertension Patients With Multiple Renal Cysts by Integrated Analysis of Whole-Genome and Single-Cell RNA Sequencing. Front. Genet. 12:775470. doi: 10.3389/fgene.2021.775470

Background and Aims: The multiple renal cysts (MRC) occur in some patients with noncirrhotic portal hypertension (NCPH) could be a subset of ciliopathy. However, the potential genetic influencers and/or determinants in NCPH with MRC are largely unknown. The aim of this study was to explore the potential candidate variants/genes associated with those patients.

Methods: 8,295 cirrhotic patients with portal hypertension were enrolled in cohort 1 and 267 patients affected with NCPH were included in cohort 2. MRC was defined as at least two cysts in both kidneys within a patient detected by ultrasonography or computed tomography. Whole-genome sequencing (WGS) was performed in nine patients (four from cohort 1 and five from cohort 2). Then we integrated WGS and publicly available single-cell RNA sequencing (scRNA-seq) to prioritize potential candidate genes. Genes co-expressed with known pathogenic genes within same cell types were likely associated NCPH with MRC.

Results: The prevalence of MRC in NCPH patients (19.5%, 52/267) was significantly higher than cirrhotic patients (6.2%, 513/8,295). Further, the clinical characteristics of NCPH patients with MRC were distinguishable from cirrhotic patients, including late-onset, more prominent portal hypertension however having preserved liver functions. In the nine whole genome sequenced patients, we identified three patients with early onset harboring compound rare putative pathogenic variants in the known disease gene *PKHD1*. For the remaining patients, by assessing cilia genes profile in kidney and liver scRNA-seq data, we identified *CRB3* was the most co-expressed gene with *PKHD1* that highly expressed in ureteric bud cell, kidney stromal cell and hepatoblasts. Moreover, we found a homozygous

variant, *CRB3* p.P114L, that caused conformational changes in the evolutional conserved domain, which may associate with NCPH with MRC.

Conclusion: ScRNA-seq enables unravelling cell heterogeneity with cell specific gene expression across multiple tissues. With the boosting public accessible scRNA-seq data, we believe our proposed analytical strategy would effectively help disease risk gene identification.

Keywords: whole-genome sequencing, single-cell RNA sequencing, gene mutations, multiple renal cysts, noncirrhotic portal hypertension

INTRODUCTION

Cirrhotic portal hypertension complicated esophageal-gastric variceal bleeding (EGVB), ascites, hepatic encephalopathy (HE), acute kidney injury (AKI) or hepatorenal syndrome (HRS) and splenomegaly accompanied by severe liver disfunctions, is almost accounted for 80–85%. In clinical practice, a small number of patients present with portal hypertension, such as splenomegaly and or EGVB, HE, their clinical manifestations are similar to the liver cirrhosis,but in fact, these patients do not have liver cirrhosis, that is non-cirrhotic portal hypertension (NCPH) (Khanna and Sarin, 2019; Gao et al., 2020). Currently, chronic infections, autoimmune disorders, and genetic determinants have been reported to be associated with pathogenesis of NCPH (Vilarinho et al., 2016; Wu et al., 2019). Interestingly, the NCPH is common in cystic fibrosis-associated liver disease (Boëlle et al., 2019). Therefore, we speculate that liver disease in multiple renal cysts (MRC) present as NCPH, except for a subset of patients with ciliopathy affected by hepatorenal fibrocystic diseases (HFDs), such as autosomal recessive polycystic kidney disease (ARPKD) or Caroli syndrome (Abdul Majeed et al., 2020; McConnachie et al., 2021). HFDs are a group of ciliopathies and genetic disorders that involve developmental abnormalities in the portobiliary system in association with fibrocystic degeneration of the kidney (Lasagni et al., 2021). HFDs can cause enlarged kidneys, cyst formation, biliary duct dilation, and congenital hepatic fibrosis (CHF), resulting in portal hypertension (Myram et al., 2021). Therefore, patients with NCPH with MRC may be a non-classical genetic mutations or HFD phenotype.

Currently, single-cell RNA sequencing (scRNA-seq) has revolutionized developmental biology and genomics. ScRNA-seq is a powerful tool that can be used to elucidate the cellular composition in the interest tissue, to define undescribed rare cell subsets, to dissect regulators controlling cell fate transition, to pinpoint cell-type specific responses to stress or stimulation, and to identify mechanisms of cell-cell crosstalk (Han et al., 2020). In this study, we applied a novel analytic approach that integrated whole-genome sequencing (WGS) and scRNA-seq data to survey potential genetic modifiers or candidate disease genes in NCPH patients with MRC.

METHODS

Study Design and Patients

Total 8,295 cirrhotic portal hypertensive patients, diagnosed by medical history, signs and imaging of portal hypertension according to chinese guidelines on the management of liver cirrhosis (Xu et al., 2020), were enrolled in cohort 1. The etiologies of liver cirrhosis included hepatitis B (65.04%), hepatitis C (15.79%), alcoholic cirrhosis (12.35%), and autoimmune liver disease (6.82%). Two hundred sixty-seven patients with NCPH in cohort 2, confirmed by radiologists, hepatologists and pathologists based on enhanced computed tomography (CT) or nuclear magnetic resonance imaging (MRI) and/or pathology according to guidelines on the management of NCPH of EASL (Khanna and Sarin, 2014). The signs and symptoms, laboratory and endoscopy data were obtained from Electronic Medical Record Management System (EMRS). The portal and splenic vein diameter and splenic thickness were measured using abdominal ultrasonography. Model for end-stage liver disease (MELD) and Child-Pugh score were assessed for the severity of cirrhosis. Transient elastography liver stiffness measurement (LSM) was performed using FibroScan™ (Echosens, Paris, France). The MRC were defined as more than two cysts in both kidneys detected by ultrasonography or CT.

The study protocol was performed in compliance with the Declaration of Helsinki and approved by the Ethics Committees of Beijing You'an Hospital Capital Medical University. Signed informed consent was obtained from each participant for using samples, materials and publication.

Whole-Genome DNA Sequencing

The peripheral blood of nine patients with MRC (four from cohort 1 and five from cohort 2 separately) were collected for WGS to explore potential genetic modifiers or candidate disease genes. First, the genomic DNA was extracted from the peripheral blood of those patients. The whole-genome DNA Sequencing libraries were prepared according to the manufacturer's instructions. The raw reads were produced by a BGISEQ-500 sequencer at an average depth of 40×. The rare putatively pathogenic variants were validated by Sanger sequencing.

Genomic DNA Analysis

Sequencing data were quality controlled with adapter and aligned to human reference genome build hg19 (http://www.gencodegenes.org/releases/19.html) with BWA aligner (Li and Durbin, 2009). The GATK best practice Haplotype Caller pipeline was implemented for SNV and indel calling (Li et al., 2009; McKenna et al., 2010). SV was called with lumpy software, and CNV was detected with FreeC software (Layer et al., 2014). All

SNV variants were annotated using ANNOVAR for bioinformatics analysis (Wang et al., 2010). Several genomic databases, including the 1,000 Genomes (1000G), ExAC (Exome Aggregation Consortium), Exome Sequencing Project (ESP), gnom AD (both WES and WGS databases), and CG46, were used to assess the variant frequency in the population. MCAP, SIFT, Polyphen2-HDIV, Polyphen2-HVAR, MutationTaster, MutationAssessor, and Clinvar were implemented to annotate the effect of missense variants. GERPs were used to evaluate the conservation of the variant locus. Rare putative pathogenic variants were filtered as follows (Gao et al., 2020): the allele frequency of the candidate variant should be lower than 0.01 among 1000G, ExAC, ESP 6500, Genome Aggregation (GA) and Complete Genomics 46 (CG46) Databases (Khanna and Sarin, 2019); amino acid changing variants were kept, and GERP scores should be higher than 2.0 (Vilarinho et al., 2016); truncating variants were kept, and for missense variants, MCAP scores higher than 0.6 were automatically kept, whereas for other variants, the effect should be annotated as "Deleterious" or "Highly pathogenic" by at least two software programs for MCAP scores between 0.025 and 0.6. PKHD1 protein domain prediction was obtained using SMART (Letunic et al., 2021). Cilia genes were obtained from Syscilia.org for downstream analysis.

TABLE 1 | Comparison of clinical characteristics between NCPH patients with MRC and Hepatitis B cirrhosis without MRC.

	NCPH with MRC	Hepatitis B cirrhosis without MRC	P
	(*N* = 52)	(*N* = 92)	
Age (y)	60.35 ± 15.47	52.13 ± 10.13	0.011
Sex (female/male)	24/28	31/61	0.076
EGVB	23 (44.23%)	31 (33.70%)	0.027
Ascites	15 (28.85%)	49 (53.26%)	<0.001
Platelet (10^9/L)	108.00 (74.50, 141.50)	56.50 (41.00, 88.00)	<0.001
INR	1.12 (1.04, 1.26)	1.23 (1.12, 1.37)	0.003
ALT (U/L)	23.65 (19.23, 36.38)	22.15 (17.45, 32.95)	0.13
AST (U/L)	37.15 (26.73, 60.45)	30.65 (23.95, 47.25)	0.308
TBIL (μmol/L)	19.85 (14.58, 41.40)	21.60 (15.83, 34.68)	0.087
ALB (g/L)	34.83 ± 6.62	34.23 ± 5.87	0.128
Creatinine (μmol/L)	67.20 (53.45, 87.88)	63.35 (53.75, 72.10)	0.133
Child-pugh score	6.00 (5.00, 8.00)	6.00 (5.00, 7.75)	0.965
MELD	5 (2, 7)	6 (3, 9)	0.559
LSM (kPa)	18.80 (13.25, 30.15)	21.80 (13.95, 30.35)	0.146
PV (mm)	13.05 ± 2.66	14.03 ± 3.15	0.637
SV (mm)	8.15 ± 1.37	9.05 ± 2.31	0.322
Platelet/LSM	5.81 ± 1.16	2.56 ± 0.57	<0.001

MRC, multiple renal cysts; EGVB, esophageal and gastric varices bleeding; ALB, Albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; TBIL, total bilirubin; INR, international normalized ratio; MELD, model for end-stage liver disease; LSM, liver stiffness measurement; PV, portal vein diameter; SV, splenic vein diameter.

Single-Cell RNA Sequencing Analysis

Five sets of scRNA-seq data were used in this study. Summarized gene expression matrices derived from multiple organs from human fetuses were obtained from the GEO database via accession number GSE156793 (Cao et al., 2020). Single-cell expression data from kidneys were generated from mice under accession number GSE140023 (Conway et al., 2020). The adult human kidney and human liver single cell transcriptome was achieved by accession GSE114530 (Hochane et al., 2019), GSE131685 (He et al., 2020) and GSE159929 (Liao et al., 2020), respectively. The single cell RNA sequencing analysis was implemented with Seurat package (Version 3.9.9). Cells were discarded according to the following criteria (Gao et al., 2020): cells that had fewer than 400 genes (UMI >0) (Khanna and Sarin, 2019); cells that had fewer than 600 UMI or over 10,000 UMI; and (Vilarinho et al., 2016) cells that had more than 15% mitochondrial UMI counts. After the above quality control, for mouse kidney, human fetal kidney, human adult kidney and human liver scRNA-seq analysis, we performed log-normalization with the "vst" method and identified 2000, 3,000, 3,000 and 3,000 variable features, respectively. We then scaled by setting the parameter "vars.to.regress" to "percent.mito" and "nCount_RNA". Principal component analysis (PCA) was performed using the "RunPCA" function. The number of PCs was chosen by visualization plot with the "ElbowPlot" function. A shared nearest neighbor (SNN) graph was constructed using the "FindNeighbors" function with the top 40 PCs, then cells were clustered by the "FindClusters" function with the "resolution" parameter set to 0.5. The "RunUMAP" function was used for the visualization plot with the "umap-learn" method, setting "n.neighbors" to 40L, "dims" to 1:40, and "min.dist" to 0.3. Marker genes for each cluster were detected using the "FindAllMarkers" function, setting the parameter "min.pct" to 0.3 and "logfc.threshold" to 0.6. Subsequently, cell clusters were annotated manually to the major cell types according to known markers. Any cluster with multiple markers of two types of cells was manually discarded as a doublet.

Protein Conformation Modeling

The full-length human *CRB3* was not annotated in the PDB Database; thus, the intact wild-type and mutated CRB3 protein sequences were annotated by Phyre2 (Kelley et al., 2015). Then, protein conformational alteration was predicted using Chimera software (Pettersen et al., 2004).

Data Visualization and Statistics

Microsoft R Open (version 3.6.1, https://mran.microsoft.com/) was used. The R packages ggplot 2 (version 3.1.0) and pheatmap (version 1.0.12) were used to generate graphs of the data. Continuous variables were compared using independent T test if data were normally distributed or Mann Whitney U test. The categorical variables were compared using X^2 tests performed with SPSS 22.0 (IBM, United States).

RESULTS

Prevalence and Clinical Characteristics of NCPH Patients With MRC

The prevalence of MRC in NCPH patients accounted for 19.5% (52/267). It was significantly higher than that in cirrhotic patients

TABLE 2 | The clinical data for nine patients with MRC enrolled for whole-genome sequencing.

	P1	P2	P3	P4	P5	P6	P7	P8	P9
Age (at first diagnose)	26 (5)	36 (34)	35 (13)	60 (55)	75 (73)	71 (60)	53 (45)	46 (36)	47 (42)
Complication of PH	EVB	EVB	No	No	No	Ascites	EVB	EVB	No
Child-Pugh Score	5	5	8	6	8	10	8	5	5
LSM (kPa)	14.3	35	15.2	5.8	8	9	12	2.8	3.8
EGV	Severe	Severe	Severe	No	Moderate	Moderate	Severe	Severe	No
Renal function	Normal	Normal	CRF II	Normal	Normal	Normal	Normal	Normal	Normal
ALT/AST	43.2/47.4	Normal	53.2/42.4	49.7/40.2	53.2/82.7	Normal	Normal	Normal	Normal
Hypersplenism	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No
No. renal cysts	3	3	4	3	2	3	2	3	2
Maximum renal cyst (mm)	38	13	24	24	13	15	4	30	25
No. hepatic cysts	0	0	3	1	5	0	3	3	2
Maximum liver cyst (mm)	NA	NA	23	4	17	NA	8	7	14
Dilatation of the intrahepatic bile duct	Yes	Yes	Yes	No	No	No	No	No	No
PV (mm)	9	10	12	14	13	13	14	11	12

PH, portal hypertension; EGVs, esophageal and gastric varices; EVB, esophageal varices bleeding; LSM, liver stiffness measurement; PV, portal vein diameter; CRF: chronic renal failure; ALT, alanine aminotransferase; AST, aspartate aminotransferase. The units for ALT, and AST, was U/L. NA, no data.

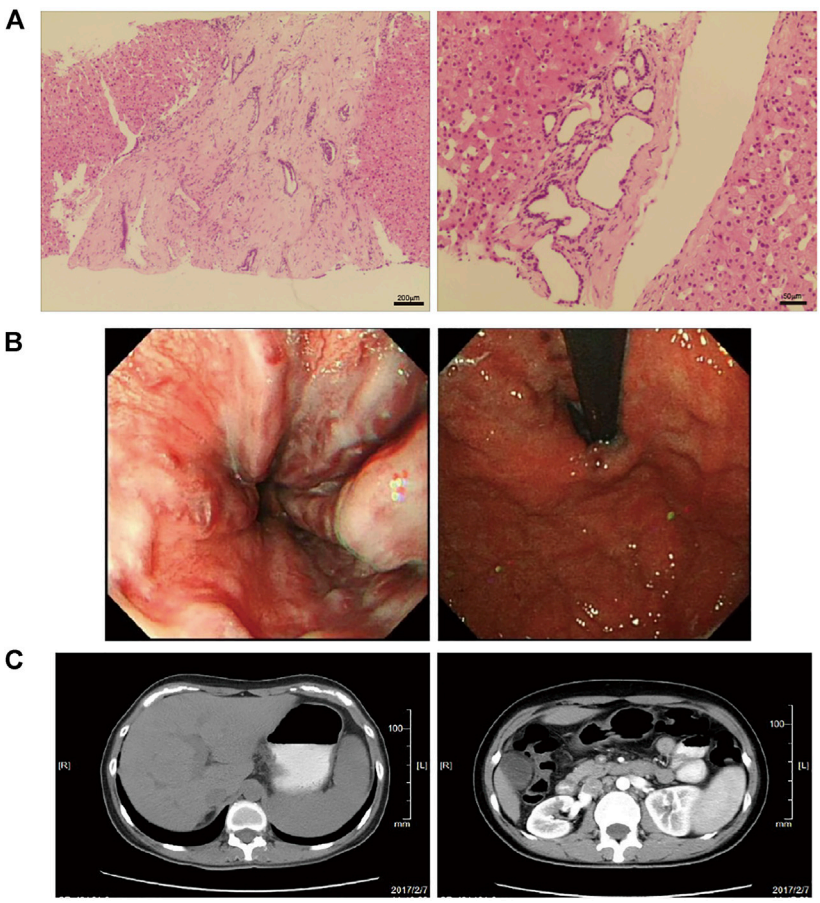


FIGURE 1 | The pathological and clinical presentation of one CHF patient with NCPH. **(A)** The left panel shows diffuse fibrosis in the liver. The right panel shows a small bile duct hamartoma. **(B)** Endoscopy showed esophageal varices (left) and gastric varices (right). **(C)** The abdomen CT images show slight dilatation of the intrahepatic bile duct and splenomegaly (left), and multiple renal cysts in the bilateral kidneys (right).

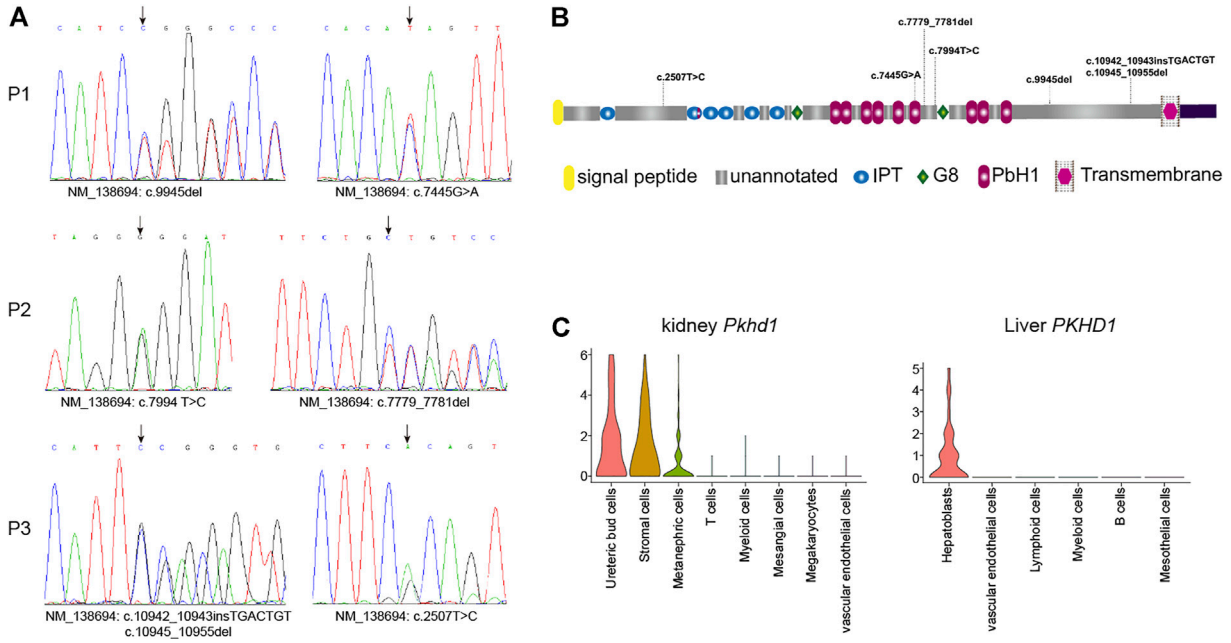


FIGURE 2 | Gene mutations analysis of *PKHD1*. **(A)** The rare putative pathogenic variants in *PKHD1* were validated by Sanger sequencing. **(B)** Schematic representation of variant distributions in the *PKHD1* domain. **(C)** Violin plot of *PKHD1* expression in mouse kidney and human liver tissues by scRNA-seq analysis. The Y axis was normalized expression level for PKHD1 in different cells.

with portal hypertension (6.2%, 513/8,295), $p < 0.05$. The clinical characteristics of NCPH patients with MRC were compared with 92 hepatitis B virous (HBV) related cirrhosis without MRC, which randomly selected (1:2) in cohort 1 (**Table 1**). The NCPH patients with MRC had a relatively older onset complications of portal hypertension ($p < 0.05$). In terms of the manifestations of portal hypertension, the proportion of EGVB was prominent ($p < 0.05$), while ascites was less ($p < 0.001$). The platelet counts were higher than that of hepatitis B cirrhosis ($p < 0.001$). Although there was no significant difference in LSM between the two groups, the ratio of platelet counts/LSM in NCPH patients with MRC was significantly higher than it in hepatitis B cirrhosis without MRC patients ($p < 0.001$). The clinical characteristics of nine patients with MRC enrolled for WGS were summarized in **Table 2**. Three patients (P1-3) showed CHF by liver biopsy (**Figure 1A**), in which two patients exhibit early-onset in childhood or adolescence and underwent splenectomy procedure before their visit to our hospital. The endoscopy and CT imaging confirmed EGV, dilated bile ducts and multiple renal cysts in those patients (**Figures 1B,C**).

The others NCPH patients (P4 and P5) showed late-onset of complication, their clinical manifestations, including large hepatorenal cysts and normal LSM, implied underlying HFDs although there is no pathological evidence. Moreover, four HBV-positive patients (P6-P9) were diagnosed at middle age with normal liver function may also pinpoint to the need of dissecting genetic factors for HFD phenotypic expression.

Early Onset Harbored Compound Rare Pathogenic Variants in *PKHD1*

The copy number variations or structural variants spanning known HFD-related genes were not observed. Subsequently, we identified all gene mutations in the known Caroli syndrome-causing gene *PKHD*1 in 3 CHF patients with NCPH (**Supplementary Table S1**). All missense mutations in *PKHD1* had GERP scores higher than 5.4, indicating that these mutations were in evolutionarily highly conserved regions. Moreover, all three patients harbored an additional truncation mutation, which were not reported in any public databases. Therefore, we assumed that the patients carried recessive mutations in *PKHD1*, and validated the mutations by Sanger sequencing (**Figure 2A**). In addition, analysis of *PKHD1* mutation distributions in different protein domains showed that all mutations were located in the extracellular domain (**Figure 2B**). Further, by using scRNA-seq data to explore *PKHD1* expression, we found that *PKHD1* was highly expressed in ureteric bud cells and stromal cells and moderately expressed in metanephric cells in the kidney. In contrast, *PKHD1* was highly expressed on hepatoblasts in the liver (**Figure 2C**). Moreover, we also detected *PKD1* compound mutations, which may explain that the patient P1 had early disease onset at the age of five and large renal cysts (38 mm).

Potential Candidate Genes Associated With HFDs Phenotype

To narrow the potential HFD phenotype-associated genes, we retrieved all known cilia genes from the literature and the European project SYSCILIA (Boldt et al., 2016). After

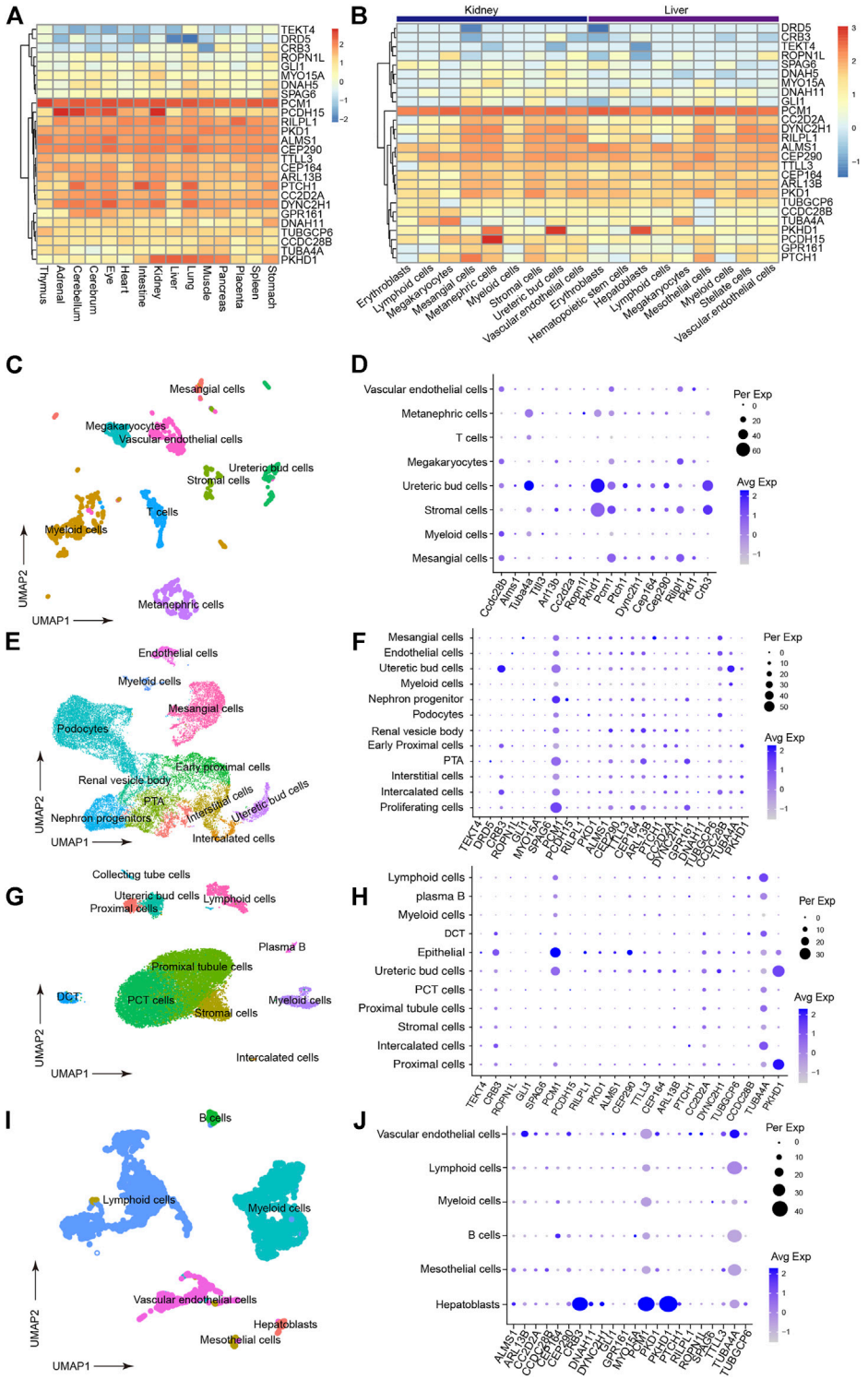


FIGURE 3 | Prioritization of detected cilia genes with scRNA-seq. **(A)** Heatmap of detected cilia gene expression in multiple human fetal organs using data under accession GSE156793. Gene expression have been scaled with normalized read counts for each gene. **(B)** Heatmap of detected cilia gene expression in different cell types from human fetal kidney and liver tissues (from GSE156793). Gene expression have been scaled with normalized read counts for each gene. **(C)** UMAP plot showing the clustering of different cell types in mouse kidneys (from GSE140023). **(D)** Dot plot of cilia genes in mouse kidney (from GSE140023). The expression inferred Pkhd1, Crb3 and Tuba4a had similar expression pattern. Dot size indicated percentage of cell which expressed gene of interest and color indicated expression level. **(E)** UMAP plot showing the clustering of different cell types in human fetal kidney (from GSE114530). **(F)** Dot plot reflecting the cell type expression patterns of the detected cilia genes (from GSE114530). This result supported that PKHD1 and CRB3 had similar expression pattern. **(G)** UMAP plot showing the clustering of different cell types in human fetal kidney (from GSE131685). **(H)** Dot plot reflecting the cell type expression patterns of the detected cilia genes (from GSE131685). **(I)** UMAP plot showing the clustering of different cell types in the adult human liver (from GSE159929). **(J)** Dot plot reflecting the cell type expression patterns of the detected cilia genes (from GSE159929).

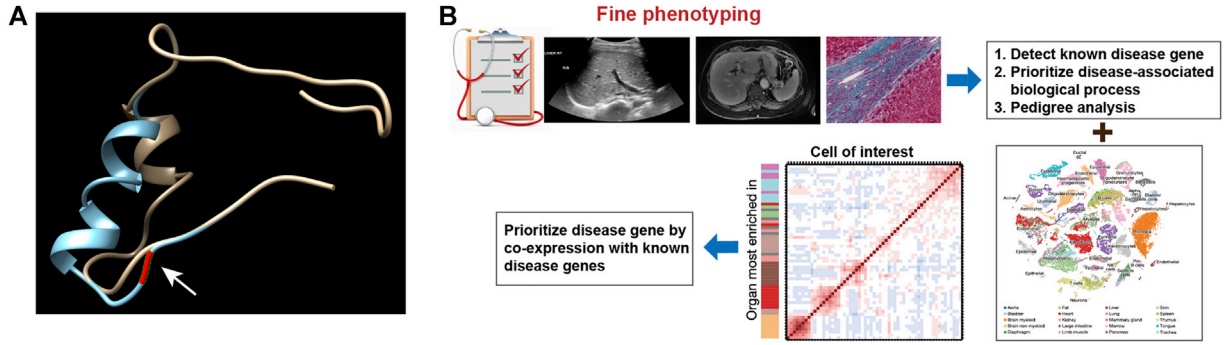


FIGURE 4 | (A) Prediction of the conformation of CRB3 protein. The yellow chain shows the wild-type protein, and the blue chain shows the mutated protein. Residue 114 is pointed by arrow and highlighted in red on the protein backbone. **(B)** Schematic illustration of the novel analytic strategy to identify disease-related genes.

prioritizing candidate disease-associated cilia genes by scRNA-seq analysis, the cilia-related mutations derived from the enrolled patients were listed in **Supplementary Table S1**, such as *CRB3* p.P114L. We first plotted the expression for candidate cilia genes in multiple human fetal organs, and annotated different cell types, particularly in the kidneys and liver (**Figures 3A,B**). We further assessed the cilia genes of cell clustering and annotated different cell type in adult mice kidney and adult human liver using uniform manifold approximation and projection (UMAP). We found that *CRB3, TUBA4A, PTCH1* and *CEP290* co-expressed with *PKHD1* (**Figures 3C,D**). Similar co-expression pattern were also seen in human fetal kidney (**Figures 3E,F**) and adult kidney (**Figures 3G,H**). Followed by cell clustering and annotation in adult human liver, the *CRB3* and *PKHD1* co-expression were spotted in hepatoblasts (**Figures 3I,J**). To further investigate the functional effect of the *CRB3* p.P114L variants, we performed protein structure remodeling and predicted that the mutation would lead to protein conformational alterations in the PDZ-domain (**Figure 4A**). The bona fide tight junction assembly acquired the capacity for *PRKC1/PARD6A* complex translocation to the apical surface by interacting with the CRB3 C-terminus. Hence, we speculated that this mutation might ultimately blocking CRB3 function. With additional scRNA-seq data, we believe that this newly proposed analytical strategy may help clinicians and geneticists to map disease-related genes (**Figure 4B**).

DISCUSSIONS

Currently, the etiologies and genetic pathogenesis of NCPH with CHF, especially idiopathic non-cirrhotic portal hypertension (INCPH), have not been fully elucidated (Lanktree et al., 2021). The MRC is more common in CHF and NCPH (Vilarinho et al., 2016; Boëlle et al., 2019; McConnachie et al., 2021). We have reasons to believe that MRC, as a cilia, may have genetic variation disorders, especially non-classical genetic mutations of HFD phenotype in INCPH with MRC patients. In this study, the prevalence of MRC in NCPH patients was accounted almost for 19.5%, which was significantly higher than the cirrhotic portal hypertension patients with known etiologies. The clinical characteristics of NCPH with MRC from these data were older-onset of the complications of portal hypertension, obvious manifestations of portal hypertension, such as EGVB, and having preserved liver functions.

Genome-wide single-cell analysis represents the ultimate frontier of genomics research (Boldt et al., 2016). In particular, scRNA-seq studies have been boosted in the last few years of new technologies enabling the study of the transcriptomic landscape of thousands of single cells in complex pathogenesis of diseases (Ying et al., 2021). Owing to the dramatic improvement in scRNA-seq technology, especially integrating WGS and scRNA-seq, tissue-specific expression at the single-cell level has improved our understanding of biological processes (Zeggini et al., 2019). In this study, WGS and scRNA-seq were performed to survey potential genetic modifiers or candidate disease genes in NCPH patients with MRC. The results showed that genes also expressed in ureteric bud cells, stromal cells, and hepatoblasts may have additive effects on NCPH with MRC. We also found that *CRB3, TUBA4A, PTCH1* and *CEP290* co-expressed with *PKHD1* at hepatoblasts in liver using UMAP. Interestingly, we discovered that patient P5 carried a homozygous candidate mutation in *CRB3* without family history of MRC. The *CRB3* encodes an apical transmembrane protein that regulates the morphogenesis of tight junctions in mammalian epithelial cells (Lemmers et al., 2004). CRB3 protein plays an important role in apicobasal polarity formation, such as cyst formation (Hurd et al., 2003). Furthermore, CRB3 participates in interactions with *TAZ/YAP*, thereby affecting transforming growth factor (TGF)-β signaling. Disruption of CRB3 function enhances TGF-β signaling and predisposes cells to TGF-β-mediated epithelial-to-mesenchymal transition (Varelas et al., 2010). Therefore, loss of function of CRB3 could potentially be linked to cyst formation and/or fibrosis. Importantly, further narrowing of the candidate gene selection showed that *CRB3* could be a novel disease risk gene for HFDs. Although patients carrying a homozygous mutation in *CRB3* showed late disease onset, this mutation affects PDZ domain conformation and might alleviate protein

function rather than cause complete loss of function. However, further studies are needed for functional validation of the pathogenicity of this gene.

One of the limitations of this study is the lack of parental genomic materials and family pedigree of MRC, making it challenging to further prioritize selected candidates. In addition, the present study lacks WGS analysis in MRC patients with mutations in known pathogenic genes, such as ARPKD phenotypes associated genes and *PKHD1* or *PKD* genes because of small sample size of NCPH with MRC and hepatitis B cirrhosis patients without MRC. Furthermore, the novel identified *CRB3* p.P114L variant has not been undergone biological function study, and we will conduct the research in the future.

CONCLUSION

CRB3 gene is commonly co-expressed with *PKHD1* in NCPH with MRC. The homozygous variant in *CRB3* may be associated with genetic pathogenesis of NCPH with MRC. Therefore, we speculate that there may be non-classical genetic mutations in NCPH patients with MRC. *CRB3* may be a novel homozygous candidate gene mutation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee Review Board of You'an Hospital, Capital Medical University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

YaW, YoW, and HD designed the study; JH and HD supervised the study; YoW, HL, and SW performed experiments and clinical examinations; KL and JH performed data analysis; YaW performed data interpretation; YoW and KL wrote the manuscript. JH and HD made revisions. HD providing funding. All authors read the manuscript and approved it before submission.

FUNDING

This work was supported by the National Natural Science Foundation (grant no. 81970525), and Digestive Medical Coordinated Development Center of Beijing Hospitals Authority (grant no. XXZ0801).

ACKNOWLEDGMENTS

We sincerely appreciated the enrolled patients for their support and understanding. Additionally, we also would like to thank Editage for language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.775470/full#supplementary-material

REFERENCES

Abdul Majeed, N., Font-Montgomery, E., Lukose, L., Bryant, J., Veppumthara, P., Choyke, P. L., et al. (2020). Prospective Evaluation of Kidney and Liver Disease in Autosomal Recessive Polycystic Kidney Disease-Congenital Hepatic Fibrosis. *Mol. Genet. Metab.* 131, 267–276. doi:10.1016/j.ymgme.2020.08.006

Boëlle, P. Y., Debray, D., Guillot, L., Clement, A., and Corvol, H. (2019). Cystic Fibrosis Liver Disease: Outcomes and Risk Factors in a Large Cohort of French Patients. *Hepatology* 69, 1648–1656. doi:10.1002/hep.30148

Boldt, K., van Reeuwijk, J., van Reeuwijk, J., Lu, Q., Koutroumpas, K., Nguyen, T.-M. T., et al. (2016). An Organelle-Specific Protein Landscape Identifies Novel Diseases and Molecular Mechanisms. *Nat. Commun.* 7, 11491. doi:10.1038/ncomms11491

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., et al. (2020). A Human Cell Atlas of Fetal Gene Expression. *Science* 370, eaba7721. doi:10.1126/science.aba7721

Conway, B. R., O'Sullivan, E. D., Cairns, C., O'Sullivan, J., Simpson, D. J., Salzano, A., et al. (2020). Kidney Single-Cell Atlas Reveals Myeloid Heterogeneity in Progression and Regression of Kidney Disease. *J. Am. Soc. Nephrol.* 31, 2833–2854. doi:10.1681/ASN.2020060806

Gao, Z.-Q., Han, Y., Li, L., and Ding, H.-G. (2020). Pharmacological Management of Portal Hypertension: Current Status and Future. *Chin. Med. J. (Engl)* 133, 2362–2364. doi:10.1097/CM9.0000000000001004

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., et al. (2020). Construction of a Human Cell Landscape at Single-Cell Level. *Nature* 581, 303–309. doi:10.1038/s41586-020-2157-4

He, S., Wang, L.-H., Liu, Y., Li, Y.-Q., Chen, H.-T., Xu, J.-H., et al. (2020). Single-Cell Transcriptome Profiling of an Adult Human Cell Atlas of 15 Major Organs. *Genome Biol.* 21, 294. doi:10.1186/s13059-020-02210-0

Hochane, M., van den Berg, P. R., Fan, X., Bérenger-Currias, N., Adegeest, E., Bialecka, M., et al. (2019). Single-Cell Transcriptomics Reveals Gene Expression Dynamics of Human Fetal Kidney Development. *Plos Biol.* 17, e3000152. doi:10.1371/journal.pbio.3000152

Hurd, T. W., Gao, L., Roh, M. H., Macara, I. G., and Margolis, B. (2003). Direct Interaction of Two Polarity Complexes Implicated in Epithelial Tight junction Assembly. *Nat. Cel Biol.* 5, 137–142. doi:10.1038/ncb923

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* 10, 845–858. doi:10.1038/nprot.2015.053

Khanna, R., and Sarin, S. K. (2014). Non-Cirrhotic Portal Hypertension - Diagnosis and Management. *J. Hepatol.* 60, 421–441. doi:10.1016/j.jhep.2013.08.013

Khanna, R., and Sarin, S. K. (2019). Noncirrhotic Portal Hypertension. *Clin. Liver Dis.* 23, 781–807. doi:10.1016/j.cld.2019.07.006

Lanktree, M. B., Haghighi, A., di Bari, I., Song, X., and Pei, Y. (2021). Insights into Autosomal Dominant Polycystic Kidney Disease from Genetic Studies. *Clin. J. Am. Soc. Nephrol.* 16, 790–799. doi:10.2215/CJN.02320220

Lasagni, A., Cadamuro, M., Morana, G., Fabris, L., and Strazzabosco, M. (2021). Fibrocystic Liver Disease: Novel Concepts and Translational Perspectives. *Transl Gastroenterol. Hepatol.* 6, 26. doi:10.21037/tgh-2020-04

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: A Probabilistic Framework for Structural Variant Discovery. *Genome Biol.* 15, R84. doi:10.1186/gb-2014-15-6-r84

Lemmers, C., Michel, D., Lane-Guermonprez, L., Delgrossi, M.-H., Médina, E., Arsanto, J.-P., et al. (2004). CRB3 Binds Directly to Par6 and Regulates the Morphogenesis of the Tight Junctions in Mammalian Epithelial Cells. *Mol. Biol. Cel. Mar.* 15, 1324–1333. doi:10.1091/mbc.e03-04-0235

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: Recent Updates, New Developments and Status in 2020. *Nucleic Acids Res.* 49, D458–D460. doi:10.1093/nar/gkaa937

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Liao, J., Yu, Z., Chen, Y., Bao, M., Zou, C., Zhang, H., et al. (2020). Single-Cell RNA Sequencing of Human Kidney. *Sci. Data* 7, 4. doi:10.1038/s41597-019-0351-8

McConnachie, D. J., Stow, J. L., and Mallett, A. J. (2021). Ciliopathies and the Kidney: A Review. *Am. J. Kidney Dis.* 77, 410–419. doi:10.1053/j.ajkd.2020.08.012

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, 1297–1303. doi:10.1101/gr.107524.110

Myram, S., Venzac, B., Lapin, B., Battistella, A., Cayrac, F., Cinquin, B., et al. (2021). A Multitubular Kidney-On-Chip to Decipher Pathophysiological Mechanisms in Renal Cystic Diseases. *Front. Bioeng. Biotechnol.* 9, 624553. doi:10.3389/fbioe.2021.624553

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera? A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25, 1605–1612. doi:10.1002/jcc.20084

Varelas, X., Samavarchi-Tehrani, P., Narimatsu, M., Weiss, A., Cockburn, K., Larsen, B. G., et al. (2010). The Crumbs Complex Couples Cell Density Sensing to Hippo-Dependent Control of the TGF-β-SMAD Pathway. *Develop. Cel.* 19, 831–844. doi:10.1016/j.devcel.2010.11.012

Vilarinho, S., Sari, S., Yilmaz, G., Stiegler, A. L., Boggon, T. J., Jain, D., et al. (2016). Recurrent Recessive Mutation in Deoxyguanosine Kinase Causes Idiopathic Noncirrhotic Portal Hypertension. *Hepatology* 63, 1977–1986. doi:10.1002/hep.28499

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Wu, H., Vu, M., Dhingra, S., Ackah, R., Goss, J. A., Rana, A., et al. (2019). Obliterative Portal Venopathy without Cirrhosis Is Prevalent in Pediatric Cystic Fibrosis Liver Disease with Portal Hypertension. *Clin. Gastroenterol. Hepatol.* 17, 2134–2136. doi:10.1016/j.cgh.2018.10.046

Xu, X.-Y., Ding, H.-G., Li, W.-G., Xu, J.-H., Han, Y., Jia, J.-D., et al. (2020). Chinese Guidelines on the Management of Liver Cirrhosis (Abbreviated Version). *World J. Gastroenterol.* 26, 7088–7103. doi:10.3748/wjg.v26.i45.7088

Ying, P., Huang, C., Wang, Y., Guo, X., Cao, Y., Zhang, Y., et al. (2021). Single-cell RNA Sequencing of Retina:new Looks for Gene Marker and Old Diseases. *Front. Mol. Biosci.* 8, 699906. doi:10.3389/fmolb.2021.699906

Zeggini, E., Gloyn, A. L., Barton, A. C., and Wain, L. V. (2019). Translational Genomics and Precision Medicine: Moving from the Lab to the Clinic. *Science* 365, 1409–1413. doi:10.1126/science.aax4588

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 February 2022
doi: 10.3389/fgene.2021.787208

Identifying Potential Biomarkers of Prognostic Value in Colorectal Cancer *via* Tumor Microenvironment Data Mining

Lei Li[1,2], *Xiao Du*[2,3]* *and Guangyi Fan*[2,3]*

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [2]BGI-Qingdao, BGI-Shenzhen, Qingdao, China, [3]BGI-Shenzhen, Shenzhen, China

Colorectal cancer (CRC) is a common cancer that has increased rapidly worldwide in the past decades with a relatively high mortality rate. An increasing body of evidence has highlighted the importance of infiltrating immune and stromal cells in CRC. In this study, based on gene expression data of CRC patients in TCGA database we evaluated immune and stromal scores in tumor microenvironment using ESTIMATE method. Results showed there was potential correlation between these scores and the prognosis, and that patients with higher immune score and lower stromal score had longer survival time. We found that immune score was correlated with clinical characteristics including tumor location, tumor stage, and survival time. Specifically, the right-sided colon cancer had markedly elevated immune score, compared to left-sided colon cancer and rectal cancer. These results might be useful for understanding tumor microenvironment in colorectal cancer. Through the differential analysis we got a list of genes significantly associated with immune and stromal scores. Gene Set Enrichment and protein-protein interaction network analysis were used to further illustrate these differentially expressed genes. Finally, 15 hub genes were identified, and three (CXCL9, CXCL10 and SELL) of them were validated with favorable outcomes in CRC patients. Our result suggested that these tumor microenvironment related genes might be potential biomarkers for the prognosis of CRC.

Keywords: immune, stromal, hub genes, colorectal cancer, survival analysis, tumor location

OPEN ACCESS

Edited by:
Peng Wang,
Harbin Medical University, China

Reviewed by:
Feng Gao,
The Sixth Affiliated Hospital of Sun Yat-sen University, China
Xin Wang,
The Chinese University of Hong Kong, China

***Correspondence:**
Guangyi Fan
fanguangyi@genomics.cn
Xiao Du
duxiao@genomics.cn

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

Received: *30 September 2021*
Accepted: *16 December 2021*
Published: *03 February 2022*

Citation:
Li L, Du X and Fan G (2022) Identifying Potential Biomarkers of Prognostic Value in Colorectal Cancer via Tumor Microenvironment Data Mining. Front. Genet. 12:787208. doi: 10.3389/fgene.2021.787208

INTRODUCTION

Colorectal cancer (CRC) is one of the most commonly occurring cancers, whose incidence occupies 10% of all cancer diagnoses (Sung et al., 2021; Wong et al., 2021). As the second most common cause of cancer death, Colorectal cancer has been increasing rapidly in the past decades with over 1.9 million new cases reported in 2020 (Arnold et al., 2017; Sawicki et al., 2021). Colorectal cancer may develop on either the proximal colon (right side), the distal colon (left side) or the rectum. Right-sided colon cancer (RCC) differs from the left-sided colon cancer (LCC) and rectal cancer (RC) in pathogenesis and prognosis, exhibiting distinct molecular characteristics and histology (Baran et al., 2018; Imperial et al., 2018; Siegel et al., 2020). Presently, CRC screening is not common and the diagnosis is usually made after the onset of symptoms. Because the tumor status and TNM stage at diagnosis have a fundamental role in CRC prognosis, early symptom investigation and diagnosis are of high importance (Bosch et al., 2011; Kawakami et al., 2015). However, although CRC prevalence is

high, the awareness of colorectal cancer and its symptoms is relatively low. Due to wide variation in colorectal cancer and complexity in treatment outcome prediction, investigation for new strategies and new biomarkers is necessary in CRC for improving prognosis.

It has been documented that tumor microenvironment (TME) has a great impact on tumor cells and clinical outcomes (Turley et al., 2015; Lim et al., 2018). Apart from tumor cells, TME also comprises a variety of nontumor components including endothelial cells, immune cells, inflammatory mediators, and extracellular matrix (ECM) molecules (Lorusso and Rüegg, 2008; Bolouri, 2015). The cells and molecules in the TME are in a dynamic process, jointly promoting tumor immune escape, tumor growth and metastasis (Quail and Joyce, 2013). Accumulating evidence suggests that the stromal and immune cells, which constitute two main nontumor components in the TME, are valuable in investigating tumor diagnosis and clinical outcome (Kalluri and Zeisberg, 2006; Hanahan and Weinberg, 2011; Fridman et al., 2012). Recent evidence has indicated that tumor microenvironment plays a significant role in colorectal carcinogenesis, metastasis and the choosing of therapy strategies (Peddareddigari et al., 2010; Pedrosa et al., 2019). T cells, a major part of the immune system, were described to be of major importance for tumor growth, invasion, early metastasis and prognosis in colorectal cancer (Pagès et al., 2005; Mlecnik et al., 2011). Calon et al. suggested that high expression of mesenchymal genes associated with poor outcomes in CRC patients is primarily caused by stromal cells instead of epithelial cancer cells (Calon et al., 2015). To promote the understanding of cancer prognosis, efforts have been made in studying tumor microenvironment components and developing novel immunotherapeutic strategies in recent years. Algorithms such as ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) (Yoshihara et al., 2013) have been developed to predict tumor purity and levels of infiltrating stromal and immune cells in tumor, such as gastric cancer, hepatocellular carcinoma and colorectal cancer (Mao et al., 2018; Deng et al., 2019; Wang et al., 2019).

To promote the understanding of CRC microenvironment and prognosis, in this study we took use of ESTIMATE algorithm and public database to evaluate the tumor-infiltrating immune and stromal cells of TME. By performing survival analysis and correlation analysis, we explored the relationship between immune/stromal score and clinical factors in CRC. Moreover, we aim to extract a list of tumor microenvironment associated genes of prognostic value, through the differential analysis, network construction and survival analysis. We hope to provide insights to investigate stromal and immune cells of CRC and offer evidence to potential prognostic markers.

MATERIALS AND METHODS

Data Collection and Preprocessing

In this study, gene expression profiles of colorectal cancer were downloaded and collected from The Cancer Genome Atlas (TCGA) data portal (https://portal.gdc.cancer.gov/) using TCGAbiolinks (Colaprico et al., 2016) R package. Relevant clinical information including age, gender, survival time, pathologic stage, and tissue or organ of origin were also obtained from TCGA database. Patients with primary tumor expression and survival information were retained in this study. Before further analysis, TCGA gene expression profiles were normalized using R package DESeq2 (Love et al., 2014).

GSE41258 expression and clinical data were also downloaded from the Gene Expression Omnibus (GEO) database as the validation set. The GSE41258 dataset was processed *via* the Affymetrix MAS5 background correction algorithm using affy package (Gautier et al., 2004) in R and log2 transformation. Probe sets were transformed into gene sets by retaining only the probes with the highest expression levels if one gene corresponds to multiple probes. When multiple genes per probe, this probe would be discarded.

Estimation of Immune and Stromal Scores

The normalized expression data was analyzed by the ESTIMATE algorithm for calculating the Immune and Stromal Score. We used ESTIMATE to calculate the fraction of immune and stromal cells in tumor using the gene expression data. In our study v.1.0.13 estimate R package (Yoshihara et al., 2013) was used to predict the level of infiltrating immune cells (immune score) and the level of infiltrating stromal cells (stromal score).

Survival Analysis Based on Immune and Stromal Scores

Survival analysis was performed by R package survival (Therneau, 2019) and survminer (Kassambara et al., 2019) to assess the association of immune and stromal score with prognosis. The best cut-off value of immune/stromal score was inferred using R program surv_cutpoint. Subsequently, patients were divided into two groups (high *vs*. low) based on the cut-off value. The Kaplan-Meier (KM) method was used to estimate the likelihood of survival based on the observed overall survival time. Overall Survival differences between high and low score groups were compared by log-rank test.

Differential Gene Expression Analysis

We analyzed differentially expressed genes (DEGs) between high score and low score groups using R package DESeq2 (Love et al., 2014), which based on the negative binomial distribution algorithm. And $|\log2 \text{ fold change (FC)}| > 2$ and p value < 0.01 were selected as the criteria to select the significantly different genes. R package pheatmap (Kolde, 2019) was used to visualize the DEGs.

Function Analysis

To explore the potential function of DEGs, function analysis was carried out by using the Gene Set Enrichment Analysis (GSEA) web server (Mootha et al., 2003; Subramanian et al., 2005). Enrichment analyses of hallmark gene sets, ontology gene

TABLE 1 | Summary and Cox Regression Analysis of overall survival for TCGA CRC study dataset.

Characteristics	Count	Univariate Cox		Multivariate Cox	
		Hazard ratio (95% CI)	P-value	Hazard ratio (95% CI)	P-value
Age	613	1.03 (1.015-1.049)	<0.001	1.04 (1.021-1.067)	<0.001
Gender					
Female	286	1	-	1	-
Male	327	1.02 (0.710-1.454)	0.93	0.88 (0.554-1.400)	0.59
Location					
Right	189	1	-	1	-
Left	132	0.70 (0.435-1.134)	0.15	0.58 (0.349-0.974)	0.04
Rectum	85	0.72 (0.386-1.326)	0.29	0.55 (0.279-1.094)	0.09
Stage					
Stage I	103	1	-	1	-
Stage II	227	1.72 (0.712-4.150)	0.23	1.03 (0.384-2.775)	0.95
Stage III	177	3.19 (1.345-7.580)	0.01	2.31 (0.870-6.151)	0.09
Stage IV	86	8.62 (3.647-20.370)	<0.001	7.36 (2.803-19.327)	<0.001
Stromal score					
High	230	1	-	1	-
Low	383	0.69 (0.483-0.998)	0.05	0.66 (0.332-1.312)	0.24
Immune score					
High	425	1	-	1	-
Low	188	1.44 (1.001-2.071)	0.05	2.07 (1.060-4.043)	0.03

terms (cellular component, molecular function, and biological process), and KEGG gene sets were selected to extract biological insight in different risk groups. The top 20 biological functional terms with False discovery rate (FDR) q-value below 0.01 were selected.

PPI Network Construction and Hub Gene Selection

To further investigate the relationship between different genes, the protein-protein interaction (PPI) network analysis was performed *via* the version 11.5 STRING (Search Tool for the Retrieval of Interacting Genes, https://string-db.org/) (Szklarczyk et al., 2015), an online tool and database of protein-protein interaction. A minimum required interaction score > 0.7 were selected and reconstructed in the Cytoscape (Shannon et al., 2003) software. In a gene candidate module, one gene with high correlation with other genes is called a hub gene. In this study, We used CytoHubba plugin (Chin et al., 2014) in Cytoscape v3.7.1 to find hub genes in PPI network. The top 15 genes with the highest prediction scores calculated by the Maximal Clique Centrality (MCC) algorithm were defined as the hub genes.

Statistical Analysis

All statistical analyses were performed in R statistical environment version ≥ 3.5.0. Cox proportional hazard regression survival analysis was applied to overall survival time with different clinical features including age, gender, tumor location, tumor stage, immune score and stromal score. Correlations between the clinical factors and immune/stromal score were also calculated in this study. Kruskal-Wallis Test for three or more groups and Wilcoxon Test for two groups were used to estimate the P value.

RESULTS

Tumor Immune and Stromal Scores Significantly Associated With Prognosis in CRC

HTSeq-Counts based gene expression profiles and clinical information of 613 CRC patients were downloaded from TCGA database. In this cohort, patients were diagnosed with colorectal cancer between 1998 and 2013 and their sequencing and clinical information were collected into the TCGA database. Among them, 286 (46.7%) patients were female and 327 (53.3%) patients were male. The ages ranged from 31 to 90. Clinical diagnosis included 189 (30.8%) cases with right-sided colon cancer, 132 (21.5%) cases with left-side colon cancer, and 85 (13.9%) cases with rectal cancer. The Pathologic stage I, stage II, stage III and stage IV accounted for 16.8% (n = 103), 37.0% (n = 227), 28.9% (n = 177) and 14.0% (n = 86) of the total number (**Table 1**). In addition, based on ESTIMATE algorithm immune and stromal scores were obtained. Stromal scores for the analyzed CRC cohort ranged from -2,531.36 to 1,481.74, and immune scores were distributed between −1,724.23 and 1,856.93, respectively. The average immune score was −600.92 and the median was −658.63. The average stromal score was −966.83 and the median was −1,026.69.

To explore the potential correlation of overall survival time with stromal and immune scores, 613 CRC cases were divided into high- and low-score groups according to the cut-off of stromal/immune scores. As shown in **Figure 1**, survival analysis indicated that both the immune and stromal scores were significantly correlated with overall survival time, and that patients with high immune score or low stromal score significantly correlated with better overall survival time (**Figures 1A,B**, p-value = 0.048 for immune

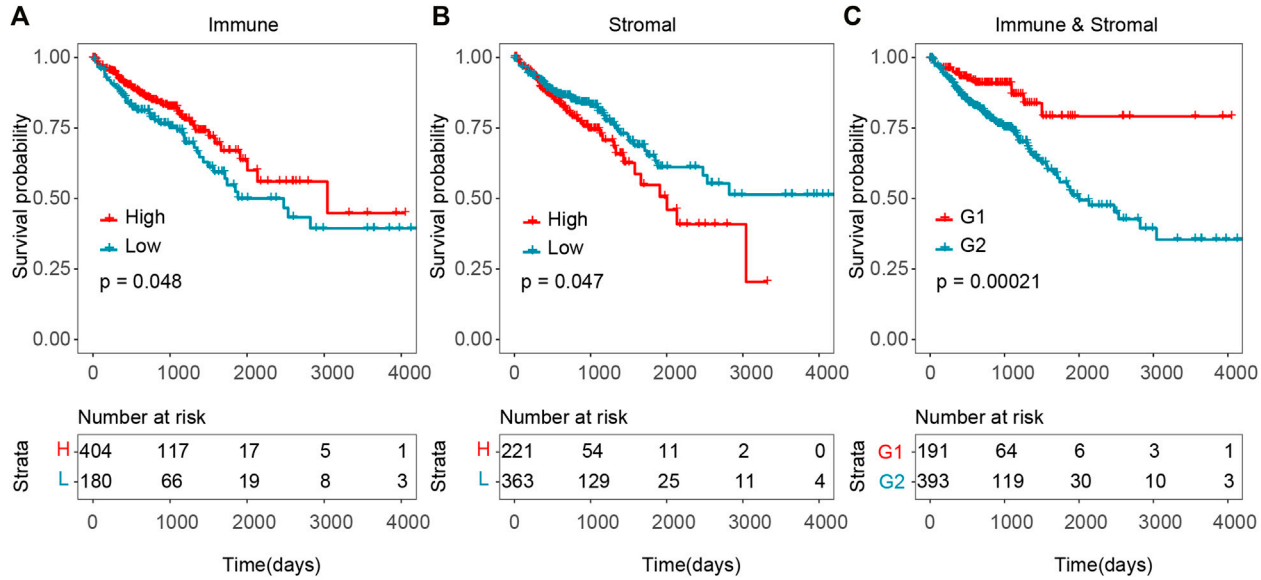


FIGURE 1 | Association between tumor microenvironment and overall survival time in TCGA CRC cohort. **(A)** Kaplan–Meier curves of high and low immune score groups. **(B)** Kaplan–Meier curves of high and low stromal score groups. **(C)** Kaplan–Meier curves of G1 (high immune score and low stromal score group) *versus* G2 (low immune and stromal score group, high immune and stromal score group, low immune score and high stromal score group).

score and p-value = 0.047 for stromal score, log-rank test). Patients with high immune score had a median overall survival time of 101.4 months, while patients with low immune score had a median survival of 62.7 months. Patients with lower stromal score also had a longer median overall survival compared to those with high stromal score. Especially, patients with combined high immune score and low stromal score have a significantly better overall survival time than others (**Figure 1C**, p-value = 0.00021, log-rank test).

In order to validate these results which were obtained from the TCGA database, we downloaded and analyzed another independent cohort in GEO database. We retrieved 182 CRC patients' gene expression data and clinical information from GSE41258 cohort as validation set. Although the difference was not statistically significant, Patients with high immune score displayed a longer median survival (**Supplementary Figure S1A**, high- vs. low-score = 91 *vs.* 86 months). And patients with lower stromal score showed a longer median survival (**Supplementary Figure S1B**, high- *vs.* low-score = 72 *vs.* 113 months). Consistently, patients with high immune score and low stromal score in the validation cohort had a longer survival time (**Supplementary Figure S1C**, p-value = 0.021, log-rank test). These results indicated that higher level of immune score and lower level of stromal score in CRC might mean the favorable survival outcome, which might provide potential prognosis stratification factors for clinical predictions.

Immune Scores Correlated With Tumor Location and Tumor Stage in CRC

To determine the clinical significance of immune and stromal scores, we investigated the association between immune/stromal score and clinical features, and the results suggested that the right-sided colon cancer have a significantly higher immune score. Immune score significantly correlated with tumor stage and tumor location (**Figures 2A,B**, p-value < 0.01). The median immune score of the RCC patients ranked the highest of all three tumor location subtypes, and the LCC subtype cases had the lowest immune scores (RCC *vs* RC = −571.65 *vs* −838.1, p-value = 0.019, RCC *vs* LCC = −571.65 *vs* −860.61, p-value = 0.00012, LCC *vs* RC = −860.61 *vs* −838.1, p-value = 0.23, Wilcoxon Test) (**Figure 2B**). Similarly, the rank order of immune scores across tumor stage from highest to lowest was Stage I > Stage II > Stage III > Stage IV (**Figure 2A**). What's more, we found immune score was also significantly associated with tumor location and the RCC also had the highest immune score in GSE41258 dataset (**Supplementary Figure S2B**, p-value = 0.032), which indicated that immune score might be predictive in the classification of CRC tumor location. However, we found no significant differences between stromal scores with CRC tumor stage or location (**Figures 2C,D**, p-value > 0.05). Consequently, immunotherapy is likely to be more effective for right-sided colon cancer with more immune infiltration and activation in CRC.

Differential Expressed Genes Revealed by Immune and Stromal Scores in CRC

To reveal the correlation of gene expression profiles with immune and stromal scores, we performed differential expression genes analysis using DESeq2, and 318 DEGs were screened out in total. By comparing immune scores between high- and low-score groups, 188 genes were identified to be differentially expressed genes. A total of 150

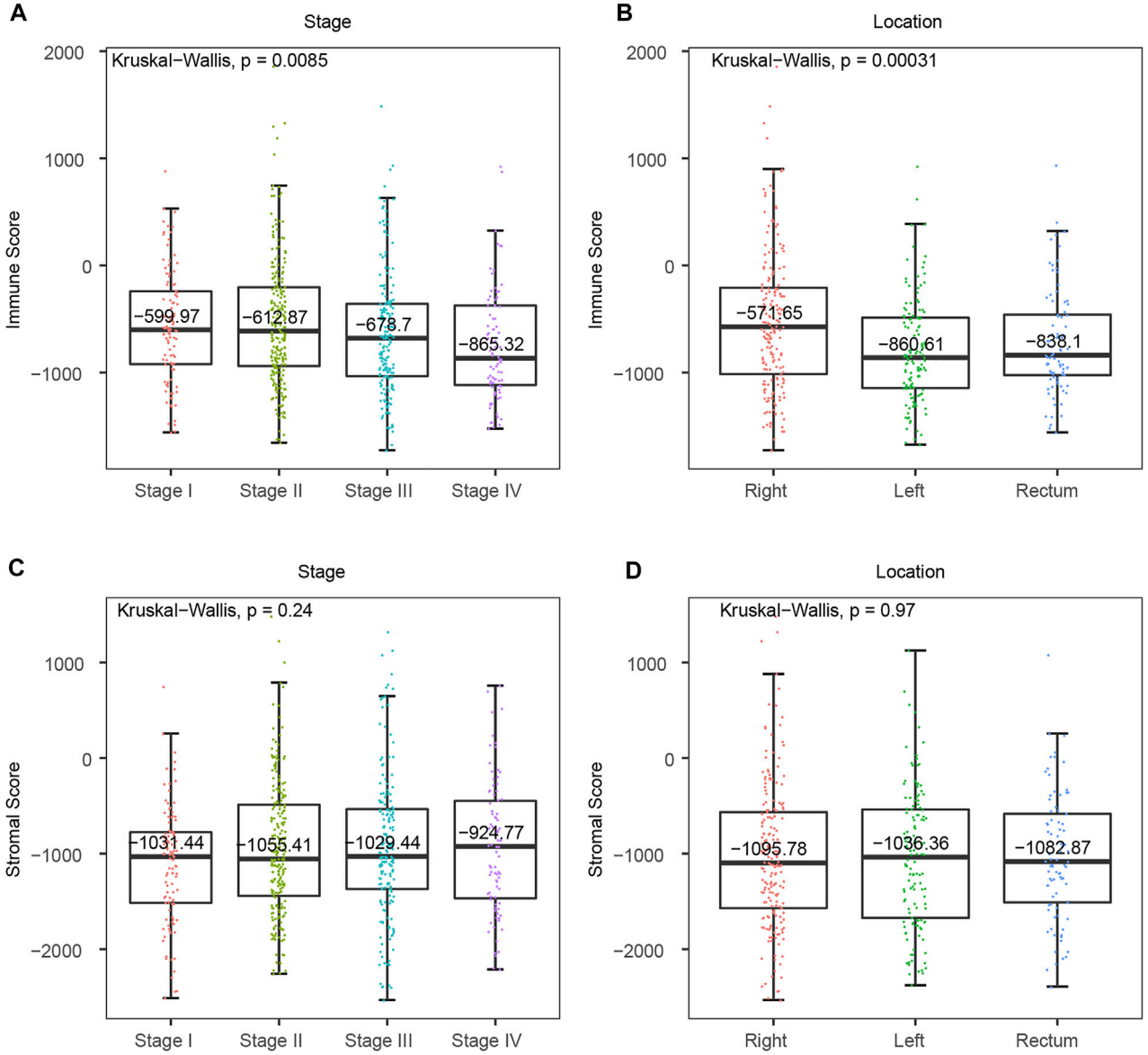


FIGURE 2 | Association between tumor microenvironment scores and clinical features in TCGA CRC cohort. **(A)** Distribution of immune scores in consecutive CRC tumor stages. **(B)** Distribution of immune scores from distinct CRC primary tumor locations. **(C)** Distribution of stromal scores in consecutive CRC tumor stages. **(D)** Distribution of stromal scores from different CRC primary tumor locations.

DEGs were found for high stromal score as compared to low stromal score. What's more, we got 43 DEGs when high immune and low stromal score patients were compared to the rest. The expression level of the DEGs in each group was displayed in heatmap (**Figure 3**). The subsequent analysis in our study were based on these DEGs.

To better understand the potential biological functions and mechanisms of DEGs in different immune and stromal score groups, Gene Set Enrichment Analysis was used to annotate the biological roles of these DEGs. GO: BP, GO: CC, GO: MF, KEGG pathways and hallmark gene sets were included in the functional enrichment analysis. The top 20 functional terms of DEGs in each group were shown in **Figure 4**. For the immune score group, the DEGs were mostly enriched in the regulation of immune system process and defense response. For the stromal score group and combined group, the top biological terms were external encapsulating structure and muscle system process. Moreover, circulatory system development, collagen containing extracellular matrix, external encapsulating structure, intrinsic component of plasma membrane, and skeletal system development were enriched in at least 2 groups. According to the result of GSEA, it could be concluded that these 318 DEGs were mostly involved in the immune regulation biological process that modulates the

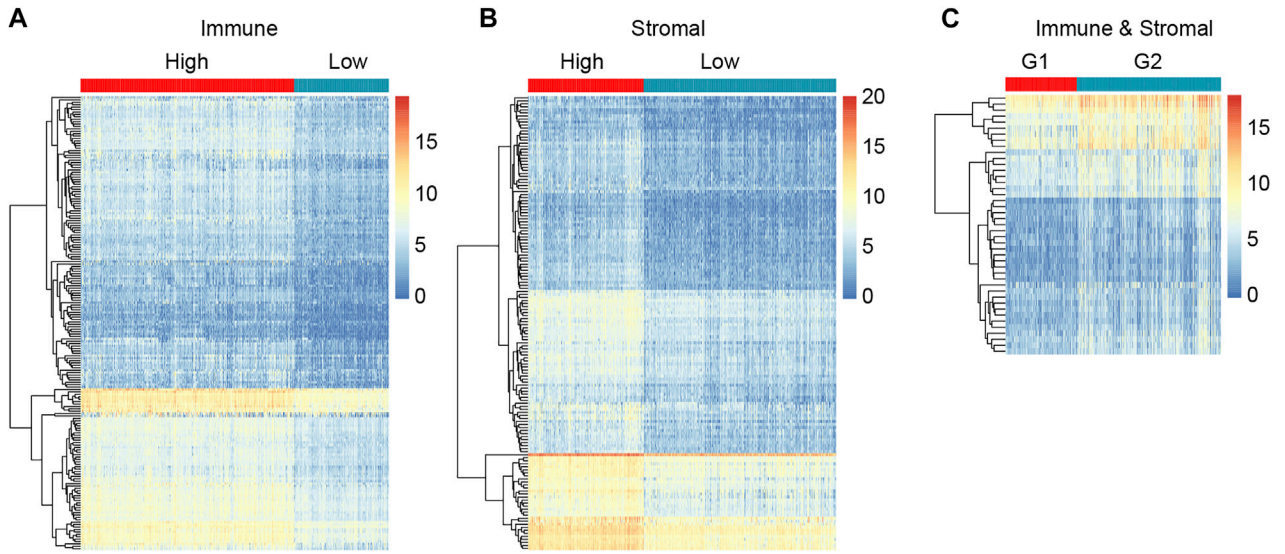


FIGURE 3 | Compare of gene expression profiles in different immune/stromal score groups to identify DEGs. **(A)** Heatmap showing DEGs in high *versus* low Immune score groups. **(B)** Heatmap showing DEGs in high *versus* low Stromal score groups. **(C)** Heatmap showing DEGs in G1 (high immune score and low stromal score group) *versus* G2 (low immune and stromal score group, high immune and stromal score group, low immune score and high stromal score group).

frequency, rate, extent of an immune system process, and cytokine-cytokine receptor interaction pathway.

Hub Gene Selection Based on PPI Network

In order to evaluate the protein interactive relationships among DEGs, PPI network was constructed based on STRING database and nodes that reported high scores in the network were screened as hub genes. A total of 318 differential expressed genes comprised 318 nodes and 372 edges based on STRING database, and result was visualized in **Figure 5** after hided disconnected nodes in the network. Following STRING analysis, the network was reconstructed in the Cytoscape. According to the calculation of CytoHubba plugin module, we identified a list of important genes, from which the top fifteen genes identified by the MCC algorithm were used for further analysis. Finally, 15 genes were selected as hub genes (CD86, ITGAM, PTPRC, FCGR3A, FCGR3B, MRC1, CD163, CCR2, SELL, CD69, CXCL10, CXCL8, CXCL9, CCL19 and CCL4), which were marked with red color in the PPI network (**Figure 5**). And we found that these genes were significantly enriched in the external side of plasma membrane, cell surface and chemokine receptor binding according to Gene Set Enrichment Analysis (**Supplementary Table S1**).

Identification and Validation of Survival Related Hub Genes

We performed survival analysis between the 15 hub genes and the overall survival time to identify potential prognostic or predictive markers for CRC. Colorectal cancer samples were splited into high- and low- expression groups according to the optimal survival cut-off. We found that 11 hub genes were correlated with survival in TCGA dataset (**Figures 6A–C** and **Supplementary Figure S3**, p-value < 0.05, log-rank test). As shown in **Figure 6**, CXCL9 and CXCL10 were significantly correlated with the overall survival time in TCGA dataset (**Figures 6A,B**, p-value < 0.05, log-rank test), and a higher expression of them might correspond to better survival. Importantly, similar result was observed in the validation set GSE41258 (**Figures 6D,E**, p-value < 0.05, log-rank test). Moreover, high expression of SELL also showed longer overall survival in TCGA dataset (**Figure 6C**, p-value < 0.05, log-rank test), even though this pattern was not statistically significant in GSE41258 cohort (**Figure 6F**, p-value = 0.053, log-rank test). Higher expression of PTPRC and CCL4 had a better survival time in TCGA dataset (**Supplementary Figure S3**, p-value < 0.05, log-rank test) and showed a longer median survival time in GSE41258, but this correlation was not statistically significant (**Supplementary Figure S4**, 0.05 < p-value < 0.1, log-rank test).

DISCUSSION

Colorectal cancer is one of the most common pathological types of cancer. Previous research have demonstrated that tumor microenvironment play an important role in the occurrence and development of CRC (Peddareddigari et al., 2010; Kamal et al., 2020). Data from previous studies indicated that the infiltration of immune cells into the tumor bed may be a valuable prognostic factor in the treatment of colorectal tumor (Pagès et al., 2005; Galon et al., 2006; Galon et al., 2007; Ganesh et al., 2019). Research showed that the high density of infiltrating memory CD45RO+ T cells, one type of immune cell, was associated with the absence of signs of early tumor lymphovascular and perineural invasion, a less advanced tumor stage, and a good clinical outcome (Pagès et al., 2005).

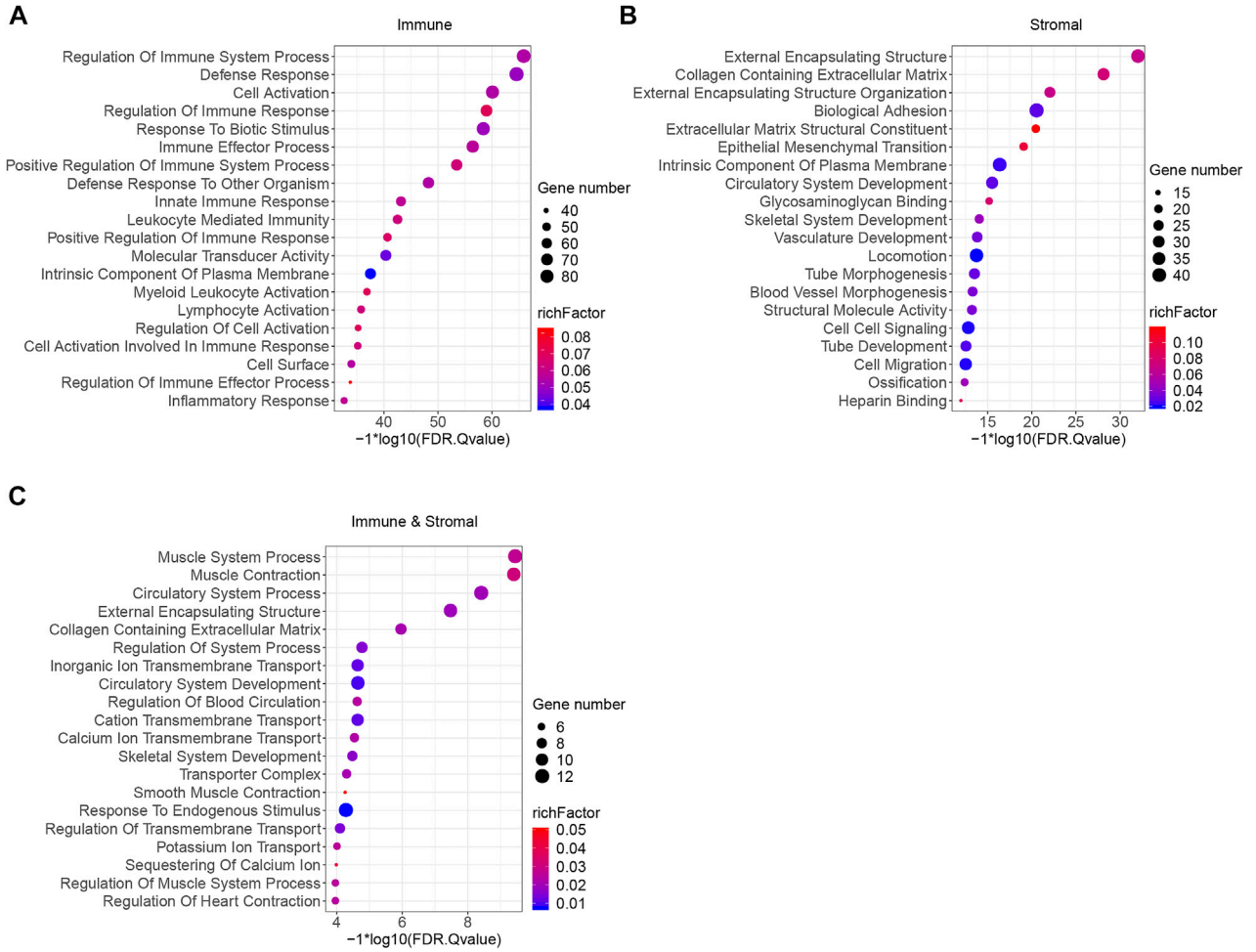


FIGURE 4 | Functional enrichment analysis of immune and stromal scores related DEGs. **(A)** GSEA analysis results of immune score related DEGs. **(B)** GSEA analysis results of stromal score related DEGs. **(C)** GSEA analysis results of immune and stromal score related DEGs. Top 20 terms were exhibited according to the significance of FDR q-value.

Cancer-associated fibroblasts (CAFs) are one of the most abundant and key components of the tumor mesenchyme among all the stromal cells (Liu et al., 2019). According to the study of Isella et al., the presence of high levels of CAFs was associated with poor prognosis in untreated CRC (Isella et al., 2015). Understanding the relationship between tumor microenvironment and patients' clinical features is vital in figuring out cancer recurrence and metastasis mechanisms. However, this mechanism is not well-understood yet.

In this study, we used the ESTIMATE algorithm to evaluate the infiltration degree of immune and stromal cells in colorectal cancer. A total of 613 CRC patients were divided into two groups based on the immune and stromal scores calculated by the R function ESTIMATE. As a result, we found high immune score was related with prolonged survival time. This observation was in general agreement with the study of Mlecnik et al. (Mlecnik et al., 2016). Besides, we found lower stromal score indicated a longer overall survival time, which further confirmed previous work by Calon et al. (Calon et al., 2015). More importantly, when patients had high immune and low stromal scores, they displayed a significantly better clinical outcome. The similar trends were also observed in another independent dataset GSE41258. These results from our study may help elucidate the underlying mechanisms in colorectal cancer microenvironment and prognosis.

Apart from that, we found clinical factors including primary tumor location and tumor stage were significantly correlated with immune score in CRC. It is worth noting that right-sided colon cancer had significantly higher immune score, as compared to left-sided colon cancer or rectum cancer. These findings might explain why right-sided colon cancer, presenting a high level of neoantigens, responded well to immunotherapies rather than adjuvant chemotherapies (Ribic et al., 2003; Wang et al., 2015; Passardi et al., 2017; Baran et al., 2018). To the best of our knowledge, previous researches mainly focused on the difference between right-sided and left-sided colon cancer (Petrelli et al., 2017; Mao et al., 2018; Zhang et al., 2018). Our study provides a more comprehensive analysis about right-sided, left-sided, and

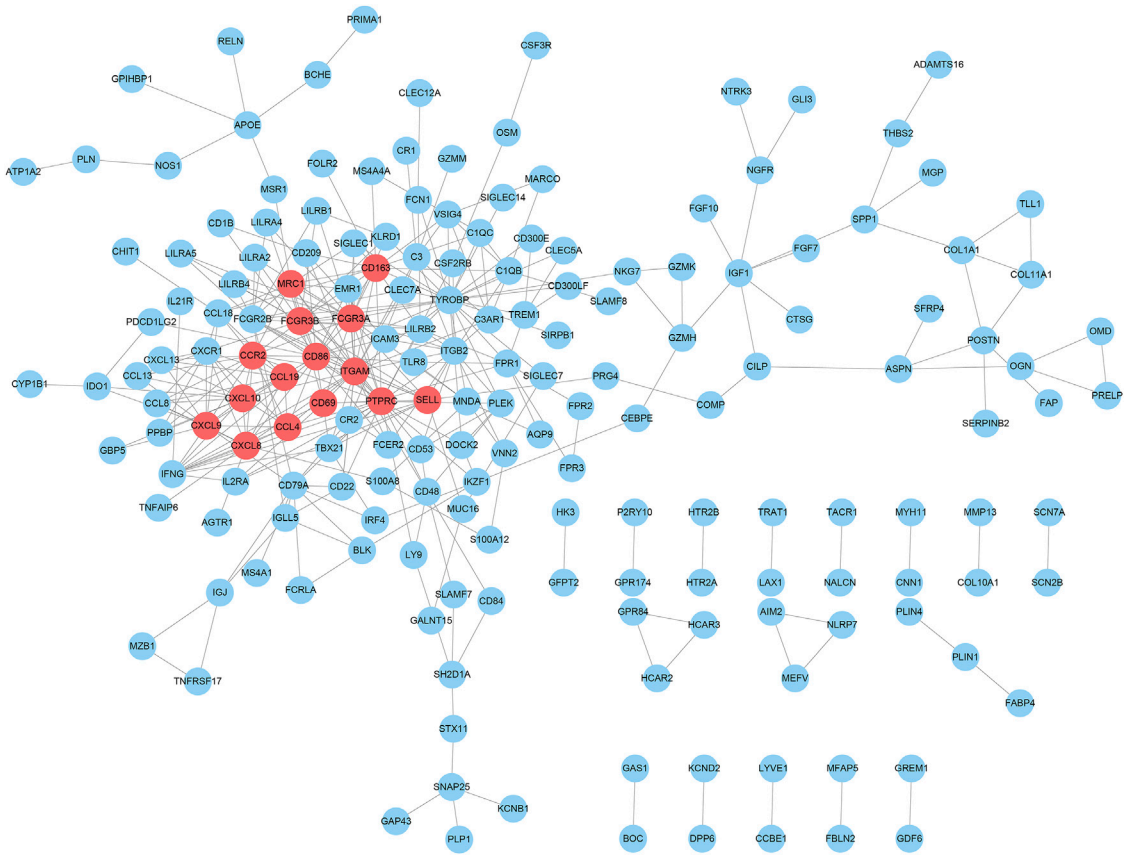


FIGURE 5 | PPI network analysis of DEGs and their hub genes screen. The hub gene nodes were highlighted in red.

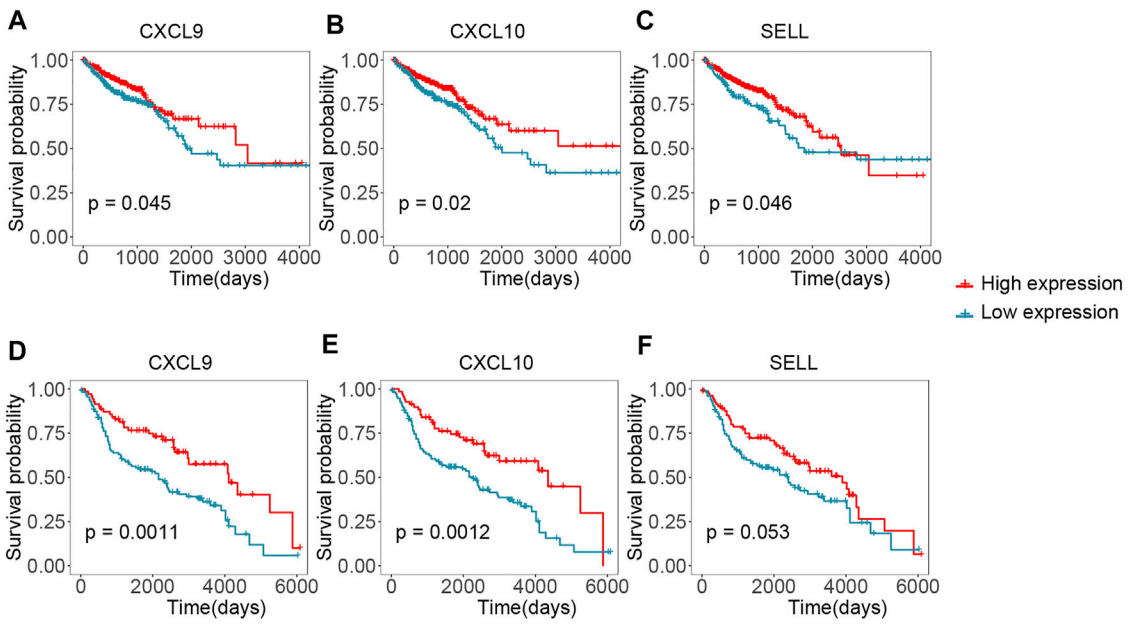


FIGURE 6 | Validating the hub genes by survival time in TCGA and GEO cohorts. **(A–C)** Kaplan-Meier plots reflecting the overall survival status of CRC patients in TCGA cohort. **(D-F)** Kaplan-Meier plots showing the overall survival status of CRC patients in GSE41258 cohort.

rectum in CRC patients. Our results further indicated that immune infiltration was different among right, left, and rectal CRCs. These immune infiltrating differences might contribute to the different survival time of CRC patients and providing a potential explanation for prognostic survival associated with primary tumor location (Petrelli et al., 2017).

Through the immune and stromal scores related DEGs analysis, a total of 318 DEGs were screened out and many of them were involved in tumor microenvironment related biological processes and pathways. Specifically, based on the DEGs analysis and GSEA annotation results, 188 DEGs were significantly correlated with immune score and most of them were involved in function that modulates the frequency or extent of an immune system process. Based on the analysis of DEGs and annotation of GSEA, 150 genes were significantly correlated with stromal score and mainly enriched in a structure that lies outside the plasma membrane and surrounds the entire cells.

Via PPI network construction, 15 genes (CD86, ITGAM, PTPRC, FCGR3A, FCGR3B, MRC1, CD163, CCR2, SELL, CD69, CXCL10, CXCL8, CXCL9, CCL19 and CCL4) were selected as hub genes. Especially, three genes (CXCL9, CXCL10, and SELL) were detected to be correlated with overall survival time both in the TCGA dataset and the validation GEO dataset. As shown in **Figure 6**, their higher expression was associated with an increased survival rate, indicating that they might be potential prognostic targets of CRC.

C-X-C motif chemokine ligand 9 (CXCL9, also known as CMK and MIG) and C-X-C Motif Chemokine Ligand 10 (CXCL10, also known INP10 and SCYB10) are mainly involved in selective and non-covalent interaction with the CXCR3 chemokine receptor and cytokine activity according to the Gene Ontology annotation. The protein encoded by CXCL9 is a member of CXC chemokine family that participates in T cell trafficking. Previous study suggested that CXCL9 plays an important role in different types of tumors (Ding et al., 2016). CXCL9 can be a tumor suppressor in breast cancer, non-small cell lung carcinoma, and colorectal cancer (Addison et al., 2000; Denkert et al., 2010; Wu et al., 2016). Conversely, it acts as tumor promoter in various types of cancer such as hepatocellular carcinoma, oral cavity squamous cell carcinoma, squamous cell cervical cancer, and chronic lymphocytic leukemia (Yan et al., 2011; Chang et al., 2013; Zhi et al., 2014; Liu et al., 2015). CXCL10 which is an important paralog of CXCL9, binds CXCR3 receptor to induce a variety of processes including chemotaxis, regulation of cell growth and apoptosis, regulation of angiostasis, and activation of immune cells (Liu et al., 2011; Sidahmed et al., 2012). The study of Chen et al. revealed that lower expression of CXCL10 was significantly associated with unsatisfied survival time (Chen et al., 2020). Our result showed that high expression of CXCL9 and CXCL10 were correlated with a better prognosis, which is consistent with studies of colorectal cancer in recent years (Wu et al., 2016; Chen et al., 2020).

SELL, also known as CD62L and L-selectin, belongs to the selectin family of glycoprotein adhesion molecules (Lefer, 2000), which is expressed on multiple tumor-infltrating immune cells and abundant in the surface of neutrophils (Lefer, 2000; Kumari et al., 2021). Recent study suggest that L-selectin might be a favorable prognosis factor in breast cancer (Kumari et al., 2021). To the best of our knowledge, there are limited studies about SELL expression and overall survival time in colorectal cancer. In this study, the high level of SELL was found correlated with better survival of CRC patients, indicating that SELL might be a new potential prognostic biomarker in CRC.

In Summary, based on the tumor immune and stromal analysis, we found that tumor microenvironment was related to CRC survival outcome and clinical characteristics such as tumor stage and location. And we identified a series of candidate genes which might serve as prognostic biomarkers in CRC. However, there were some limitations in our study. All analysis was based on public data mining instead of experiments. More experiments need to be carried out in order to further verify our conclusion and have a comprehensive insight on the potential link between the tumor microenvironment and colorectal cancer. Our current findings might provide insights into understanding the potential role of tumor microenvironment in CRC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found in TCGA (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

LL, GF, and XD designed the study. LL collected and analyzed data. LL and XD wrote the manuscript. LL, XD, and GF contributed to writing this study. All authors had full access to the final version of the manuscript and agreed to its submission.

ACKNOWLEDGMENTS

We are grateful to the group of TCGA and GEO databases for the availability of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.787208/full#supplementary-material

REFERENCES

Addison, C. L., Arenberg, D. A., Morris, S. B., Xue, Y.-Y., Burdick, M. D., Mulligan, M. S., et al. (2000). The CXC Chemokine, Monokine Induced by Interferon-Gamma, Inhibits Non-small Cell Lung Carcinoma Tumor Growth and Metastasis. *Hum. Gene Ther.* 11 (2), 247–261. doi:10.1089/10430340050015996

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

Baran, B., Mert Ozupek, N., Yerli Tetik, N., Acar, E., Bekcioglu, O., and Baskin, Y. (2018). Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. *Gastroenterol. Res.* 11 (4), 264–273. doi:10.14740/gr1062w

Bolouri, H. (2015). Network Dynamics in the Tumor Microenvironment. *Semin. Cancer Biol.* 30, 52–59. doi:10.1016/j.semcancer.2014.02.007

Bosch, L. J. W., Carvalho, B., Fijneman, R. J. A., Jimenez, C. R., Pinedo, H. M., van Engeland, M., et al. (2011). Molecular Tests for Colorectal Cancer Screening. *Clin. Colorectal Cancer* 10 (1), 8–23. doi:10.3816/ccc.2011.n.002

Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., et al. (2015). Stromal Gene Expression Defines Poor-Prognosis Subtypes in Colorectal Cancer. *Nat. Genet.* 47 (4), 320–329. doi:10.1038/ng.3225

Chang, K.-P., Wu, C.-C., Fang, K.-H., Tsai, C.-Y., Chang, Y.-L., Liu, S.-C., et al. (2013). Serum Levels of Chemokine (C-X-C Motif) Ligand 9 (CXCL9) Are Associated with Tumor Progression and Treatment Outcome in Patients with Oral Cavity Squamous Cell Carcinoma. *Oral Oncol.* 49 (8), 802–807. doi:10.1016/j.oraloncology.2013.05.006

Chen, J., Chen, Q.-L., Wang, W.-H., Chen, X.-L., Hu, X.-Q., Liang, Z.-Q., et al. (2020). Prognostic and Predictive Values of CXCL10 in Colorectal Cancer. *Clin. Transl Oncol.* 22 (9), 1548–1564. doi:10.1007/s12094-020-02299-6

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: Identifying Hub Objects and Sub-Networks from Complex Interactome. *BMC Syst. Biol.* 8 (4), S11–S17. doi:10.1186/1752-0509-8-S4-S11

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data. *Nucleic Acids Res.* 44 (8), e71. doi:10.1093/nar/gkv1507

Deng, Z., Wang, J., Xu, B., Jin, Z., Wu, G., Zeng, J., et al. (2019). Mining TCGA Database for Tumor Microenvironment-Related Genes of Prognostic Value in Hepatocellular Carcinoma. *Biomed. Res. Int.* 2019 (3), 1–12. doi:10.1155/2019/2408348

Denkert, C., Loibl, S., Noske, A., Roller, M., Müller, B. M., Komor, M., et al. (2010). Tumor-Associated Lymphocytes as an Independent Predictor of Response to Neoadjuvant Chemotherapy in Breast Cancer. *J. Clin. Oncol.* 28 (1), 105–113. doi:10.1200/jco.2009.23.7370

Ding, Q., Lu, P., Xia, Y., Ding, S., Fan, Y., Li, X., et al. (2016). CXCL9: Evidence and Contradictions for its Role in Tumor Progression. *Cancer Med.* 5 (11), 3246–3259. doi:10.1002/cam4.934

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

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., et al. (2006). Type, Density, and Location of Immune Cells within Human Colorectal Tumors Predict Clinical Outcome. *Science* 313 (5795), 1960–1964. doi:10.1126/science.1129139

Galon, J., Fridman, W.-H., and Pagès, F. (2007). The Adaptive Immunologic Microenvironment in Colorectal Cancer: A Novel Perspective: Figure 1. *Cancer Res.* 67 (5), 1883–1886. doi:10.1158/0008-5472.can-06-4806

Ganesh, K., Stadler, Z. K., Cercek, A., Mendelsohn, R. B., Shia, J., Segal, N. H., et al. (2019). Immunotherapy in Colorectal Cancer: Rationale, Challenges and Potential. *Nat. Rev. Gastroenterol. Hepatol.* 16 (6), 361–375. doi:10.1038/s41575-019-0126-x

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy--Analysis of Affymetrix GeneChip Data at the Probe Level. *Bioinformatics* 20 (3), 307–315. doi:10.1093/bioinformatics/btg405

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

Imperial, R., Ahmed, Z., Toor, O. M., Erdoğan, C., Khaliq, A., Case, P., et al. (2018). Comparative Proteogenomic Analysis of Right-Sided colon Cancer, Left-Sided colon Cancer and Rectal Cancer Reveals Distinct Mutational Profiles. *Mol. Cancer* 17 (1), 177. doi:10.1186/s12943-018-0923-9

Isella, C., Terrasi, A., Bellomo, S. E., Petti, C., Galatola, G., Muratore, A., et al. (2015). Stromal Contribution to the Colorectal Cancer Transcriptome. *Nat. Genet.* 47 (4), 312–319. doi:10.1038/ng.3224

Kalluri, R., and Zeisberg, M. (2006). Fibroblasts in Cancer. *Nat. Rev. Cancer* 6 (5), 392–401. doi:10.1038/nrc1877

Kamal, Y., Schmit, S. L., Frost, H. R., and Amos, C. I. (2020). The Tumor Microenvironment of Colorectal Cancer Metastases: Opportunities in Cancer Immunotherapy. *Immunotherapy* 12 (14), 1083–1100. doi:10.2217/imt-2020-0026

Kassambara, A., Kosinsk, M., Biecek, P., and Fabian, S. (2019). Survminer: Drawing Survival Curves Using 'ggplot2'. R package version 0.4.6. Available at: https://CRAN.R-project.org/package=survminer (Accessed September 4, 2019).

Kawakami, H., Zaanan, A., and Sinicrope, F. A. (2015). Microsatellite Instability Testing and its Role in the Management of Colorectal Cancer. *Curr. Treat. Options. Oncol.* 16 (7), 30. doi:10.1007/s11864-015-0348-2

Kolde, R. (2019). Pheatmap: Pretty Heatmaps. R package version 1.0.12. Available at: https://CRAN.R-project.org/package=pheatmap (Accessed January 4, 2019).

Kumari, S., Arora, M., Singh, J., Chauhan, S. S., Kumar, S., and Chopra, A. (2021). L-Selectin Expression Is Associated with Inflammatory Microenvironment and Favourable Prognosis in Breast Cancer. *3 Biotech.* 11 (2), 1–13. doi:10.1007/s13205-020-02549-y

Lefer, D. J. (2000). Pharmacology of Selectin Inhibitors in Ischemia/reperfusion States. *Annu. Rev. Pharmacol. Toxicol.* 40, 283–294. doi:10.1146/annurev.pharmtox.40.1.283

Lim, C. J., Lee, Y. H., Pan, L., Lai, L., Chua, C., Wasser, M., et al. (2018). Multidimensional Analyses Reveal Distinct Immune Microenvironment in Hepatitis B Virus-Related Hepatocellular Carcinoma. *Gut* 68 (5), 916–927. doi:10.1136/gutjnl-2018-316510

Liu, M., Guo, S., Hibbert, J. M., Jain, V., Singh, N., Wilson, N. O., et al. (2011). CXCL10/IP-10 in Infectious Diseases Pathogenesis and Potential Therapeutic Implications. *Cytokine Growth Factor. Rev.* 22 (3), 121–130. doi:10.1016/j.cytogfr.2011.06.001

Liu, R.-X., Wei, Y., Zeng, Q.-H., Chan, K.-W., Xiao, X., Zhao, X.-Y., et al. (2015). Chemokine (C-X-C Motif) Receptor 3-Positive B Cells Link Interleukin-17 Inflammation to Protumorigenic Macrophage Polarization in Human Hepatocellular Carcinoma. *Hepatology* 62 (6), 1779–1790. doi:10.1002/hep.28020

Liu, T., Han, C., Wang, S., Fang, P., Ma, Z., Xu, L., et al. (2019). Cancer-Associated Fibroblasts: an Emerging Target of Anti-Cancer Immunotherapy. *J. Hematol. Oncol.* 12 (1), 86–15. doi:10.1186/s13045-019-0770-1

Lorusso, G., and Rüegg, C. (2008). The Tumor Microenvironment and its Contribution to Tumor Evolution toward Metastasis. *Histochem. Cel. Biol.* 130 (6), 1091–1103. doi:10.1007/s00418-008-0530-8

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Mao, Y., Feng, Q., Zheng, P., Yang, L., Liu, T., Xu, Y., et al. (2018). Low Tumor Purity Is Associated with Poor Prognosis, Heavy Mutation burden, and Intense Immune Phenotype in colon Cancer. *Cancer Manag. Res.* 10, 3569–3577. doi:10.2147/cmar.s171855

Mlecnik, B., Bindea, G., Angell, H. K., Maby, P., Angelova, M., Tougeron, D., et al. (2016). Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity* 44 (3), 698–711. doi:10.1016/j.immuni.2016.02.025

Mlecnik, B., Tosolini, M., Kirilovsky, A., Berger, A., Bindea, G., Meatchi, T., et al. (2011). Histopathologic-Based Prognostic Factors of Colorectal Cancers Are Associated with the State of the Local Immune Reaction. *J. Clin. Oncol.* 29 (6), 610–618. doi:10.1200/jco.2010.30.5425

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1α-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes. *Nat. Genet.* 34 (3), 267–273. doi:10.1038/ng1180

Pagès, F., Berger, A., Camus, M., Sanchez-Cabo, F., Costes, A., Molidor, R., et al. (2005). Effector Memory T Cells, Early Metastasis, and Survival in Colorectal Cancer. *N. Engl. J. Med.* 353 (25), 2654–2666. doi:10.1056/NEJMoa051424

Passardi, A., Canale, M., Valgiusti, M., and Ulivi, P. (2017). Immune Checkpoints as a Target for Colorectal Cancer Treatment. *Int. J. Mol. Sci.* 18 (6), 1324. doi:10.3390/ijms18061324

Peddareddigari, V. G., Wang, D., and DuBois, R. N. (2010). The Tumor Microenvironment in Colorectal Carcinogenesis. *Cancer Microenvironment* 3 (1), 149–166. doi:10.1007/s12307-010-0038-3

Pedrosa, L., Esposito, F., Thomson, T. M., and Maurel, J. (2019). The Tumor Microenvironment in Colorectal Cancer Therapy. *Cancers (Basel)* 11 (8), 1172. doi:10.3390/cancers11081172

Petrelli, F., Tomasello, G., Borgonovo, K., Ghidini, M., Turati, L., Dallera, P., et al. (2017). Prognostic Survival Associated with Left-Sided vs Right-Sided Colon Cancer. *JAMA Oncol.* 3 (2), 211–219. doi:10.1001/jamaoncol.2016.4227

Quail, D. F., and Joyce, J. A. (2013). Microenvironmental Regulation of Tumor Progression and Metastasis. *Nat. Med.* 19 (11), 1423–1437. doi:10.1038/nm.3394

Ribic, C. M., Sargent, D. J., Moore, M. J., Thibodeau, S. N., French, A. J., Goldberg, R. M., et al. (2003). Tumor Microsatellite-Instability Status as a Predictor of Benefit from Fluorouracil-Based Adjuvant Chemotherapy for colon Cancer. *N. Engl. J. Med.* 349 (3), 247–257. doi:10.1056/nejmoa022289

Sawicki, T., Ruszkowska, M., Danielewicz, A., Niedźwiedzka, E., Arłukowicz, T., and Przybyłowicz, K. E. (2021). A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis. *Cancers* 13 (9), 2025. doi:10.3390/cancers13092025

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Sidahmed, A. M. E., León, A. J., Bosinger, S. E., Banner, D., Danesh, A., Cameron, M. J., et al. (2012). CXCL10 Contributes to P38-Mediated Apoptosis in Primary T Lymphocytes *In Vitro*. *Cytokine* 59 (2), 433–441. doi:10.1016/j.cyto.2012.05.002

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

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102 (43), 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 (3), 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43 (D1), D447–D452. doi:10.1093/nar/gku1003

Therneau, T. M. (2019). Survival: A Package for Survival Analysis in R. R Package Version 3.1-8. Available at: https://CRAN.R-project.org/package=survival (Accessed December 3, 2019).

Turley, S. J., Cremasco, V., and Astarita, J. L. (2015). Immunological Hallmarks of Stromal Cells in the Tumour Microenvironment. *Nat. Rev. Immunol.* 15 (11), 669–682. doi:10.1038/nri3902

Wang, F., Bai, L., Liu, T. S., Yu, Y. Y., He, M. M., Liu, K. Y., et al. (2015). Right-Sided colon Cancer and Left-Sided Colorectal Cancers Respond Differently to Cetuximab. *Chin. J. Cancer* 34 (3), 384–393. doi:10.1186/s40880-015-0022-x

Wang, H., Wu, X., and Chen, Y. (2019). Stromal-Immune Score-Based Gene Signature: A Prognosis Stratification Tool in Gastric Cancer. *Front. Oncol.* 9, 1212. doi:10.3389/fonc.2019.01212

Wong, M. C. S., Huang, J., Lok, V., Wang, J., Fung, F., Ding, H., et al. (2021). Differences in Incidence and Mortality Trends of Colorectal Cancer Worldwide Based on Sex, Age, and Anatomic Location. *Clin. Gastroenterol. Hepatol.* 19 (5), 955–966. e61. doi:10.1016/j.cgh.2020.02.026

Wu, Z., Huang, X., Han, X., Li, Z., Zhu, Q., Yan, J., et al. (2016). The Chemokine CXCL9 Expression Is Associated with Better Prognosis for Colorectal Carcinoma Patients. *Biomed. Pharmacother.* 78, 8–13. doi:10.1016/j.biopha.2015.12.021

Yan, X.-J., Dozmorov, I., Li, W., Yancopoulos, S., Sison, C., Centola, M., et al. (2011), Identification of Outcome-Correlated Cytokine Clusters in Chronic Lymphocytic Leukemia. *Blood* 118 (19), 5201–5210. doi:10.1182/blood-2011-03-342436

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 (1), 2612. doi:10.1038/ncomms3612

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

Zhi, W., Ferris, D., Sharma, A., Purohit, S., Santos, C., He, M., et al. (2014). Twelve Serum Proteins Progressively Increase with Disease Stage in Squamous Cell Cervical Cancer Patients. *Int. J. Gynecol. Cancer* 24 (6), 1085–1092. doi:10.1097/IGC.0000000000000153

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, orclaim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 February 2022
doi: 10.3389/fgene.2021.803257

Integrated Analysis of ceRNA Network to Reveal Potential Prognostic Biomarkers for Glioblastoma

Ruifei Liu[1†], Zhengzheng Gao[2†], Qiwei Li[3†], Qiang Fu[3], Dongwei Han[3], Jixi Wang[4], Ji Li[4], Ying Guo[5]* and Yuchen Shi[3]**

[1]Second Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin, China, [2]College of Basic Medicine, Inner Mongolia Medical University, Hohhot, China, [3]Heilongjiang University of Chinese Medicine, Harbin, China, [4]Jiaxing University, Jiaxing, China, [5]First Affiliated Hospital of Heilongjiang University of Chinese Medicine, Harbin, China

Glioblastoma (GBM), originating in the brain, is a universally aggressive malignant tumor with a particularly poor prognosis. Therefore, insight into the critical role of underlying genetic mechanisms is essential to developing new therapeutic approaches. This study aims to identify potential markers with clinical and prognostic significance in GBM. To this end, increasing numbers of differentially expressed RNA have been identified used to construct competitive endogenous RNA networks for prognostic analysis via comparison and analysis of RNA expression levels of tumor and normal tissues in glioblastoma. This analysis demonstrated that the RNA expression patterns of normal and tumor samples were significantly different. Thus, the resulting differentially expressed RNAs were used to construct competitive endogenous RNA (competing endogenous RNA, ceRNA) networks. The functional enrichment indicated mRNAs in the network are critically involved in a variety of biological functions. Additionally, the prognostic analysis suggested 27 lncRNAs, including LOXL1-AS1, AL356414.1, etc., were significantly associated with patient survival. Given the prognostic significance of these 27 lncRNAs in GBM, we sought to classify the samples. Importantly, Kaplan-Meier analysis revealed that survival times varied significantly among the different categories. Overall, these results identify that the candidate lncRNAs are potential prognostic markers of GBM and its corresponding mRNAs may be a potential target for therapy.

Keywords: glioblastoma, lncRNA, ceRNA, network, prognostic biomarker

OPEN ACCESS

Edited by:
Xinyi Liu,
University of Illinois at Chicago, United States

Reviewed by:
Chunquan Li,
Harbin Medical University, China
Dongguo Li,
Capital Medical University, China

****Correspondence:***
Yuchen Shi
hmudrs@163.com
Ying Guo
hzygy2021@163.com
Ji Li
hljucmlj@163.com

†These authors share first authorship

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

Received: *27 October 2021*
Accepted: *17 December 2021*
Published: *14 February 2022*

Citation:
Liu R, Gao Z, Li Q, Fu Q, Han D, Wang J, Li J, Guo Y and Shi Y (2022) Integrated Analysis of ceRNA Network to Reveal Potential Prognostic Biomarkers for Glioblastoma. Front. Genet. 12:803257. doi: 10.3389/fgene.2021.803257

INTRODUCTION

Long non-coding RNAs (lncRNAs), a series of transcript RNAs longer than 200 nucleotides, plays a very crucial role in biological processes, such as cell proliferation, cell apoptosis, and cell cycle regulation (Zhang et al., 2020). Accumulating studies reported that lncRNA can be involved in the regulation of competitive endogenous RNA (ceRNAs) to communicate with other RNA transcripts (Calin et al., 2007; Arvey et al., 2010; Ebert and Sharp, 2010). LncRNA can function as an endogenous molecular sponge, indirectly regulating downstream mRNA expression levels by having shared microRNA response elements with reverse complementary binding seed regions competitively binding to miRNA, and subsequently involved in cancer development (Bai et al., 2019; Sun et al., 2020). In other words, lncRNA competes with miRNA target genes for miRNA molecules by sharing a common miRNA binding site with mRNA. It has been documented that ceRNAs play a regulatory role in gene expression and is involved in the pathogenesis of

diseases such as cancer (Tay et al., 2011). A growing body of evidence clarifies that molecular networks play an important role in a variety of human diseases (Silverman et al., 2020). Accordingly, it is valuable to dissect the ceRNA network for understanding the underlying molecular mechanisms of cancer development.

Glioblastoma (GBM), one of the most fatal and aggressive forms of brain tumors, is a prevalent malignant tumor that originates in the brain, currently accounting for more than half of all gliomas (Liang et al., 2005). GBM is characterized by its high invasiveness, poor clinical prognosis, and high mortality rates. Current therapeutic approaches include focal radiotherapy, chemotherapeutics, and surgical resection. The 5-years survival rate is less than 3% (De Leo et al., 2020). Over the past few years, little progress has been made in determining methods to predict which patients will better receive the current standards of care (Johnson et al., 2020). Although survival has improved with the optimization of treatment strategies, GBM prognosis remains poor (Wen and Kesari, 2008; Yuan et al., 2015). Consequently, investigating potential genetic mechanisms of GBM is of great significance. The development of alternative and suitable biomarkers to effectively diagnose and treat GBM remains one of the most pressing challenges in cancer therapy (Aldape et al., 2015; Zhou et al., 2019). Identification of prognostic markers of GBM also contributes to comprehending the mechanisms of metastasis, which may lead to the discovery of novel therapeutic targets. The exploration of ceRNA networks in GBM may provide new insight into understanding the biological mechanisms of the disease.

In this study, glioblastoma-specific ceRNA networks were constructed based on differentially expressed genes. In addition, we further derived and characterized the lncRNAs that were significantly associated with survival in the network, classified the samples based on the screened lncRNAs. We observed the significant differences in survival time among the types of samples, which could shed light on that the lncRNAs we screened are potential prognostic markers of GBM and its corresponding mRNA may be a potential target for therapy.

MATERIALS AND METHODS

Acquisition of Glioblastoma Transcriptome and Clinical Data

We obtained the transcriptome expression profile in glioblastoma with 154 tumor samples and 5 normal samples *via* The Cancer Genome Atlas (TCGA) database (https://tcga-data.nci.nih.gov/tcga). Moreover, we also retrieved the demographic information (age, gender, race and so on) and survival endpoint (vital status, days to death and days to last follow-up) of each patient.

Interactions of ceRNA

StarBasev2.0 (http://starbase.sysu.edu.cn/index.php) database is an open-source platform for decoding miRNA–ceRNA, miRNA–ncRNA, and protein–RNA interaction networks, stored the lncRNA related ceRNA interactions identified using hypergeometric tests (Li et al., 2014). The hypergeometric test (Sumazin et al., 2011) is executed for each ceRNA pair separately, which is defined by four parameters: 1) N is the total number of miRNAs used to predict targets; 2) K is the number of miRNAs that interact with the chosen gene of interest; 3) n is the number of miRNAs that interact with the candidate ceRNA of the chosen gene; and 4) c is the common miRNA number between these two genes. The test calculates the *p*-value by using the following formula:

$$P = \sum_{i=c}^{\min(K,n)} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \quad (1)$$

Multiple miRNAs belonging to the same family were every miRNA family only once, even if it had multiple binding sites at the same 3′-UTR of protein coding genes or transcript of non-coding genes. All *p*-values were subject to false discovery rate (FDR) correction.

In this study, starBase was utilized to downloaded and extracted the ceRNA-ceRNA interactions of lncRNA-mRNA.

Associations Between LncRNA and Cancer

We downloaded the relationships between lncRNA and cancer from the Lnc2Cancer 3.0 (Sumazin et al., 2011) (http://www.bio-bigdata.net/lnc2cancer/) database, which contains the associations verified by the literature of 2,659 human lncRNAs and 216 cancer subtypes.

Identification of Differentially Expressed LncRNAs and mRNAs

First, we screened out genes that expressed less than 2 in 20 % of the samples. Next, compared to the normal group with the tumor group, the R software (version 3.6.3) and limma package in Bioconductor were used to detect the differentially expressed lncRNAs (DElncRNAs) and mRNAs (DEmRNAs). DElncRNAs and DEmRNAs were identified using the selection criteria of adjusted *p*-value (FDR) < 0.01 and FC > 2.0 or FC < 0.5 calculated by the T-test and fold change algorithm. Then, the differentially expressed lncRNAs and mRNAs meeting the criteria were displayed in volcano plots.

Construction of Glioma-specific LncRNA-mRNA ceRNA Network

Ahead of analyzing the basic statistics, we downloaded information from starBase about the lncRNA-ceRNA interaction. All interactions are verified by the literature. The starBase database contains 83,916 lncRNA-ceRNA interactions, including 2,539 lncRNAs and 2079 mRNAs. After that, we mapped the DEmRNAs and DElncRNAs selected in the previous step to the lncRNA related ceRNA interactions. Subsequently, the interactions between DEmRNAs and DElncRNAs were singled out to construct a glioma-specific ceRNA regulatory network. Cytoscape (version 3.7.2) was used to visualize the ceRNA network.

Functional Enrichment Analysis

Gene Ontology (GO) is a universal tool for defining the biological process (BP), cellular component (CC), and molecular function (MF) of numerous genes. Kyoto Encyclopedia of Genes and Genomes

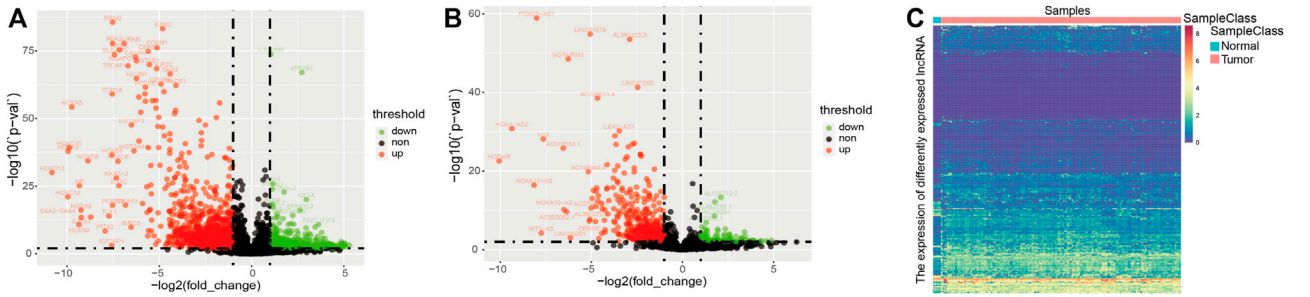


FIGURE 1 | (A) Volcano map of the differential mRNA. **(B)** Volcano map of the differential lncRNA. **(C)** Heatmap plots of the differentially expressed lncRNAs between normal and cancer samples.

(KEGG) pathway is a database that contains multiple biological pathways for several organisms (Kanehisa et al., 2017). The enrichment analyses of mRNAs on the glioma-specific ceRNA network were performed using the clusterProfiler package in Bioconductor, and a *p*-value less than 0.05 was considered as statistically significant (Yu et al., 2012). Furthermore, we performed a KEGG pathway enrichment for mRNAs connected to each lncRNA. GO and pathway analysis provided a deep insight into the relations of functions or pathways and the primary roles of these genes.

Survival Analysis

The Cox proportional hazards regression model has the function to process the truncated survival time while analyzing various variables with no requirement for the type of distribution of the survival function (Zhao et al., 2010). To assess the prognostic characteristics of all lncRNAs, the univariate Cox proportional hazards model was applied. We integrated all lncRNAs on the glioma-specific ceRNA network into the univariate Cox model to identify the lncRNAs significantly associated with survival. p values < 0.05 were regarded as significant.

Prognostic Analysis

\K-means clustering algorithm was used to classify the samples into four groups based on lncRNA that was significantly related to survival and R package "factoextra" was adopted to visualize it. To further determine the prognostic characteristics of lncRNAs, after combining the overall survival of 154 patients with GBM, the survival curves of these samples with classification information were plotted by using the "survival" package in R based on Kaplan-Meier curve analysis. Log-rank $p < 0.05$ was considered significant.

RESULT

Identification of Differentially Expressed Genes in Glioblastoma

In order to better explore the differences between glioma patients and normal samples at the gene transcriptome level, based on the dataset of 5 normal samples of glioblastoma and 154 cancer samples derived from TCGA, we performed a differential expression analysis to identify significantly differentially expressed lncRNAs and mRNAs. Then, as shown in **Figures 1A,B** , we compared the tumor group with the normal group to visualize significantly differentially expressed lncRNAs and mRNAs using volcano maps. Finally 2,326 DElncRNAs **(Figure 1A, B)** and 8,304 DEmRNAs were identified (**Figure 1C**).

Dissecting ceRNA Network Reveals lncRNA Functions

Recent studies have reported that lncRNAs can participate in competing endogenous RNAs (ceRNAs) regulations in order to communicate with other RNA transcripts. In order to better understand the regulatory relationship between differential mRNA and lncRNA, subsequently, we mapped the resulting DElncRNAs and DEmRNAs to the lncRNA-ceRNA relationship pairs downloaded from starbase and constructed a glioblastoma-specific ceRNA network (**Figure 2A**) which was composed of 343 lncRNAs, 1,427 mRNAs, and a total of 3,741 edges. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses revealed that the DElncRNAs involved in the ceRNA network were remarkably associated with a series of functions, namely, T cell activation, epidermal cell development, MAPK signaling, and cell apoptosis (**Figure 2B**). In addition, we counted the types of cancer associated with each lncRNA on the ceRNA network by using the lnc2Cancer database and performed functional enrichment of the interacting mRNAs *via* clusterProfiler, the results of functional enrichment analysis are listed in.

Screening for lncRNAs Significantly Associated With Survival Involved in ceRNA Network

To further analyze the relationship between lncRNA and glioblastoma prognosis in glioblastoma-specific ceRNA networks, all lncRNAs of the ceRNA network were incorporated into the univariate Cox model to spot lncRNAs significantly associated with survival based on the lncRNA expression and clinical information. As a result, using the threshold value of $p < 0.01$, 27 lncRNAs containing LOXL1-AS1 and HOTAIRM1 were revealed to be prominently associated with GBM prognosis among 343 lncRNAs (**Figure 3A**). The knockdown expression of LOXL1-AS1 has a functional inhibitory effect on the proliferation of GBM cells (Wang et al., 2018), which has been confirmed in the literature.

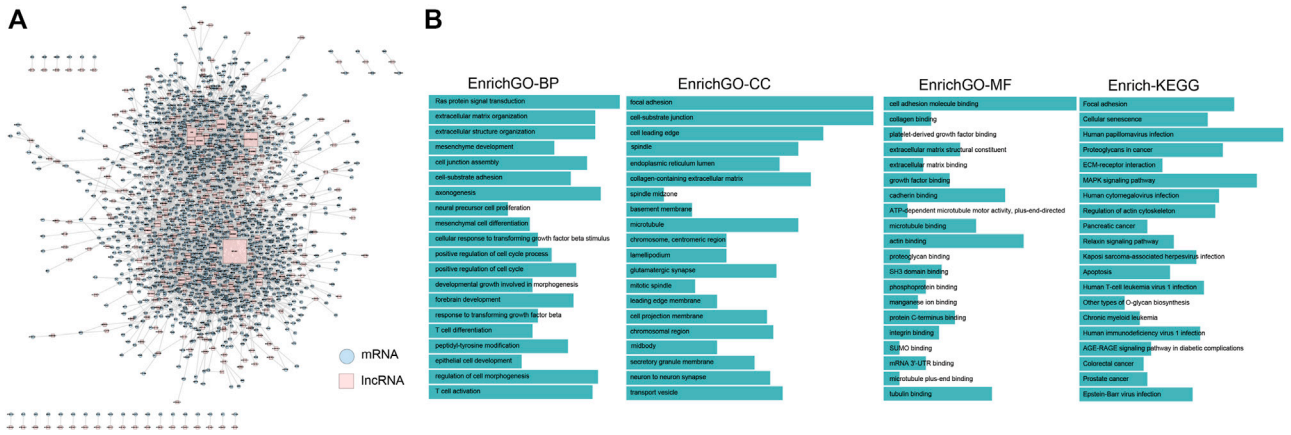


FIGURE 2 | (A) lncRNA-mRNA ceRNA network specific to glioblastoma. **(B)** The enrichment analyses of mRNAs on the glioma-specific ceRNA network, in which GO includes BP, MF and CC.

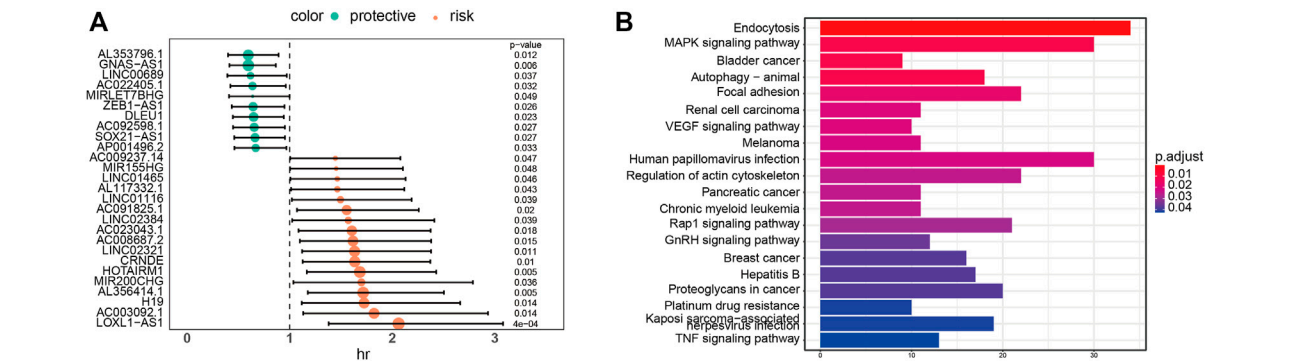


FIGURE 3 | (A) Univariate Cox proportional hazards model was applied to identify the 27 lncRNAs significantly associated with survival. **(B)** KEGG pathway enrichment analysis of 27 lncRNA-regulated mrnas significantly associated with survival.

Moreover, HOTAIRM1 knockdown has also been confirmed to significantly weaken the migration and invasion of GBM cells (Xie et al., 2020). Next, functional annotations were performed on the mRNAs regulated by identified lncRNAs significantly associated with survival in the ceRNA network, and we found that the mRNAs regulated by these lncRNAs significantly enriched MAPK signaling pathways and Focal adhesion (**Figure 3B**). It has been confirmed that MAPK signaling pathway is significantly correlated with survival and prognosis of glioma (Nicolas et al., 2019; Chen et al., 2020). In addition, GSK2256098 is a novel oral focal adhesion kinase (FAK) inhibitor. Preclinical studies demonstrate growth inhibition in glioblastoma cell lines (Swartz, 2020). These results indicate that the 27 lncRNAs identified by us can affect the survival prognosis of glioma patients.

Classification of Glioma Patients Based on LncRNA Related to Prognosis

After the above identification, it was known that the 27 lncRNAs related to survival can be used to judge the survival of patients. Continuously, based on the expression of 27 lncRNAs significantly related to survival, using the K-means clustering method, we gathered samples into four categories on the basis of the optimal number of clusters as k = 4 (**Figures 4A,B**). As results, 33 samples were divided into cluster 1, 11 samples were divided into cluster 2, and clusters 3 and 4 contained 16 and 94 samples, respectively.

Survival Analysis for the Glioblastoma Patients of the Four Subtypes

To determine lncRNAs whether have an impact on the overall survival of patients with GBM, survival analyses were conducted to investigate the Kaplan–Meier curves for GBM patients with the 27 lncRNAs (**Figure 5A**). The result exhibited an obvious difference in survival time between the four categories of samples ($p < 0.05$). It indicates that these 27 lncRNAs may be potential prognostic factors for glioblastoma. Then, we evaluated the differential expression of 27 lncRNAs significantly associated with survival in different subtypes. We found that lncRNA

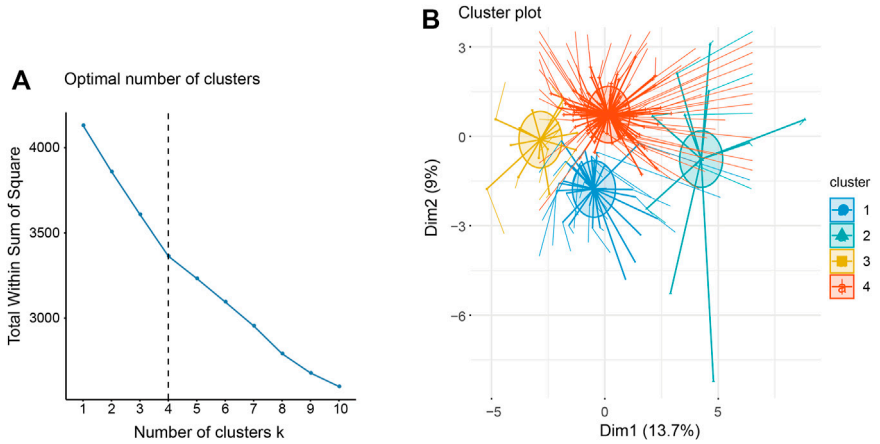


FIGURE 4 | (A) The optimal number of clusters for K-means clustering, that is, K = 4. **(B)** Cluster graph of K-means.

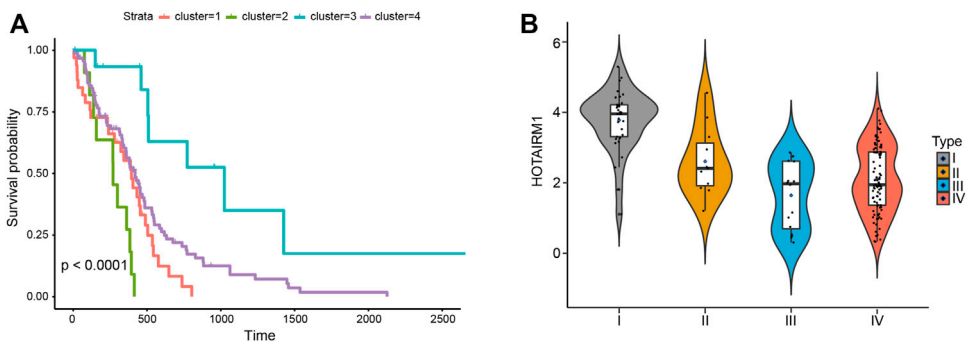


FIGURE 5 | (A) Kaplan–Meier survival curves of four types of samples. **(B)** The violin diagram shows the expression of lncRNA HOTAIRM1 in samples of four different subtypes.

HOTAIRM1 was highly expressed in subtype I compared with the other three subtypes. And it's been documented that Serum long noncoding RNA HOTAIR as a novel diagnostic and prognostic biomarker in glioblastoma multiforme. The higher the expression of HOTAIR, the worse the survival of patients (Tan et al., 2018). In our study, HOTAIRM1 was highly expressed in the samples of subtype 1 with the worst prognosis, while HOTAIRM1 expression was lowest in the samples of subtype 3 with a good prognosis (**Figure 5B**). This indicates that the lncRNAs identified by us can accurately classify patients and explain the clinical results of the corresponding subtypes.

Robustness Analysis of 27 lncrnas Significantly Associated With Prognosis

To verify the accuracy of patient classification based on the identification of 27 lncRNAs significantly associated with survival, we downloaded a set of transcriptome data

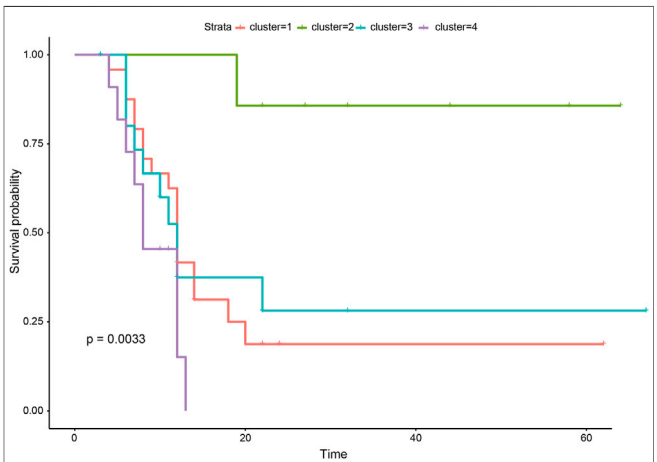


FIGURE 6 | Kaplan–Meier survival curves of four types of samples in independent validation set.

(GSE121720) from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) containing 60 glioma samples, and the survival time of patients was available. Similarly, we classified the samples based on the identified lncRNAs significantly associated with survival using k-means. In the independent validation set, we also divided the samples into four categories, and there were significant differences in the survival time of the four categories (**Figure 6**).

DISCUSSION

GBM is a common aggressive brain cancer which occurs in the central nervous system with a known poor prognosis and limited treatment options (Xiao et al., 2020). Searching for possible molecular mechanisms and potential biomarkers for GBM is a current urgent task (Rahaman et al., 2002). Increasing experimental evidence suggests that aberrant expression of ncRNA, including lncRNA and miRNA, are intimately associated with malignant progression and metastasis (Li et al., 2019; Liu et al., 2019). Since the ceRNA hypothesis was proposed, researchers have gained increasing interest in ceRNA networks, where lncRNA may influence mRNA transcription and expression by interacting with miRNA (Li et al., 2020). Competitive endogenous RNA (ceRNA) regulatory network has been confirmed to regulate expression based on competitive mechanisms and play a crucial part in multiple tumor pathological and physiological processes. ceRNAs are significant mechanisms by which lncRNAs regulating gene expression may exert huge influences on cancer. It has been extensively reported that the disorder of the ceRNA network is closely related to cancer progression (Salmena et al., 2011). For instance, a study showed that lncRNA ZEB1-AS1 functions as a ceRNA in BC, regulating the expression of the protein-coding gene fascin-1 via miR-200b (Gao et al., 2019). Thus, the ceRNA network might promote new tools for understanding the potential mechanisms of GBM and discovering potential new therapeutic targets. Here, relied on the RNA expression dataset, we proposed a ceRNA network by identifying significantly differentially expressed genes in normal samples and cancer samples.

The rapid development of bioinformatics methods provides methodological support for exploring high-throughput sequencing data (Zhong et al., 2021). The differential RNA expression observed in between the GBM and normal samples suggests that DERNAs may exert a critical role in cancer progression. In this study, we identified differentially expressed lncRNAs and mRNAs in GBM and normal brain tissue samples from the TCGA dataBase, and we further constructed a ceRNA network specific for glioblastoma combined with the lncRNA-ceRNA relationships attained in the starBase database. Functional enrichment analysis of the mRNA in the ceRNA network was performed to identify the notably enriched KEGG and GO terms. Based on the principle of the ceRNA network, lncRNA participates in biological processes by acting as endogenous molecular sponges that competitively bind to miRNAs and indirectly regulates the expression level of messenger RNA (mRNA). Hence, the potential functions and pathways of lncRNA may be similar to that of mRNA. The GO functional annotation mostly showed enrichment of mRNA related to several major regions, such as growth factor binding, Ras protein signal transduction, and positive regulation of cell cycle process. Moreover, several enriched pathways observed in the KEGG results have been reported in previous studies. MAPK is a key signaling pathway involved in GBM proliferation, apoptosis, migration, and infiltration (Vitucci et al., 2013). Finally, we assessed the survival time among the samples by clustering samples into four different subgroups based on K-means cluster analysis and the Kaplan–Meier survival curve showed remarkable differences in the survival time of the four categories of samples. This also indicates that the 27 selected lncRNAs that are significantly related to survival may be potential clinical prognostic factors for glioblastoma, and the mRNAs that interact with them may be potential therapeutic targets for glioblastoma.

Overall, we depicted a reliable prognostic ceRNA network using the differential lncRNAs and mRNAs involving GBM in the TCGA database and investigated the relevant clinical information. Our results provide a novel approach to discovering potential ceRNA networks in GBM, which will help to better understand the pathogenesis of GBM at the gene level and identify potential therapeutic agents for treating GBM.

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

YS and YG: study design. RL, ZG and QL: data analysis, manuscript writing, and resources. QF and DH: data analysis. JW and JL: funding and resources and resources. All authors have read, edited and approved of the final version of the manuscript.

REFERENCES

Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., and von Deimling, A. (2015). Glioblastoma: Pathology, Molecular Mechanisms and Markers. *Acta Neuropathol.* 129 (6), 829–848. doi:10.1007/s00401-015-1432-1

Arvey, A., Larsson, E., Sander, C., Leslie, C. S., and Marks, D. S. (2010). Target mRNA Abundance Dilutes microRNA and siRNA Activity. *Mol. Syst. Biol.* 6, 363. doi:10.1038/msb.2010.24

Bai, Y., Long, J., Liu, Z., Lin, J., Huang, H., Wang, D., et al. (2019). Comprehensive Analysis of a ceRNA Network Reveals Potential Prognostic Cytoplasmic lncRNAs Involved in HCC Progression. *J. Cel Physiol* 234 (10), 18837–18848. doi:10.1002/jcp.28522

Calin, G. A., Liu, C.-g., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., et al. (2007). Ultraconserved Regions Encoding ncRNAs Are Altered in Human Leukemias and Carcinomas. *Cancer Cell* 12 (3), 215–229. doi:10.1016/j.ccr.2007.07.027

Chen, X., Hao, A., Li, X., Ye, K., Zhao, C., Yang, H., et al. (2020). Activation of JNK and P38 MAPK Mediated by ZDHHC17 Drives Glioblastoma Multiforme Development and Malignant Progression. *Theranostics* 10 (3), 998–1015. doi:10.7150/thno.40076

De Leo, A., Ugolini, A., and Veglia, F. (2020). Myeloid Cells in Glioblastoma Microenvironment. *Cells* 10 (1), 18. doi:10.3390/cells10010018

Ebert, M. S., and Sharp, P. A. (2010). Emerging Roles for Natural microRNA Sponges. *Curr. Biol.* 20 (19), R858–R861. doi:10.1016/j.cub.2010.08.052

Gao, R., Zhang, N., Yang, J., Zhu, Y., Zhang, Z., Wang, J., et al. (2019). Long Non-coding RNA ZEB1-AS1 Regulates miR-200b/FSCN1 Signaling and Enhances Migration and Invasion Induced by TGF-B1 in Bladder Cancer Cells. *J. Exp. Clin. Cancer Res.* 38 (1), 111. doi:10.1186/s13046-019-1102-6

Johnson, R. M., Phillips, H. S., Bais, C., Brennan, C. W., Cloughesy, T. F., Daemen, A., et al. (2020). Development of a Gene Expression-Based Prognostic Signature for IDH Wild-type Glioblastoma. *Neuro Oncol.* 22 (12), 1742–1756. doi:10.1093/neuonc/noaa157

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/nar/gkw1092

Li, F., Guo, H., Liu, B., Liu, N., Xu, Z., Wang, Y., et al. (2020). Explore Prognostic Biomarker of Bladder Cancer Based on Competing Endogenous Network. *Biosci. Rep.* 40 (12). BSR20202463. doi:10.1042/BSR20202463

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, X., Lv, X., Li, Z., Li, C., Li, X., Xiao, J., et al. (2019). Long Noncoding RNA ASLNC07322 Functions in VEGF-C Expression Regulated by Smad4 during Colon Cancer Metastasis. *Mol. Ther. - Nucleic Acids* 18, 851–862. doi:10.1016/j.omtn.2019.10.012

Liang, Y., Diehn, M., Watson, N., Bollen, A. W., Aldape, K. D., Nicholas, M. K., et al. (2005). Gene Expression Profiling Reveals Molecularly and Clinically Distinct Subtypes of Glioblastoma Multiforme. *Proc. Natl. Acad. Sci.* 102 (16), 5814–5819. doi:10.1073/pnas.0402870102

Liu, Y., Sun, J., Yu, J., Ge, W., Xiao, X., Dai, S., et al. (2019). LncRNA CACS15 Accelerates the Malignant Progression of Ovarian Cancer through Stimulating EZH2-Induced Inhibition of APC. *Am. J. Transl Res.* 11 (10), 6561–6568. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6834498/pdf/ajtr0011-6561.pdf.

Nicolas, S., Abdellatef, S., Haddad, M. A., Fakhoury, I., and El-Sibai, M. (2019). Hypoxia and EGF Stimulation Regulate VEGF Expression in Human Glioblastoma Multiforme (GBM) Cells by Differential Regulation of the PI3K/Rho-GTPase and MAPK Pathways. *Cells* 8 (11), 1397. doi:10.3390/cells8111397

Rahaman, S. O., Harbor, P. C., Chernova, O., Barnett, G. H., Vogelbaum, M. A., and Haque, S. J. (2002). Inhibition of Constitutively Active Stat3 Suppresses Proliferation and Induces Apoptosis in Glioblastoma Multiforme Cells. *Oncogene* 21 (55), 8404–8413. doi:10.1038/sj.onc.1206047

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA Hypothesis: the Rosetta Stone of a Hidden RNA Language? *Cell* 146 (3), 353–358. doi:10.1016/j.cell.2011.07.014

Silverman, E. K., Schmidt, H. H. H. W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., et al. (2020). Molecular Networks in Network Medicine: Development and Applications. *Wires Syst. Biol. Med.* 12 (6), e1489. doi:10.1002/wsbm.1489

Sumazin, P., Yang, X., Chiu, H.-S., Chung, W.-J., Iyer, A., Llobet-Navas, D., et al. (2011). An Extensive microRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma. *Cell* 147 (2), 370–381. doi:10.1016/j.cell.2011.09.041

Sun, J.-R., Kong, C.-F., Xiao, K.-M., Yang, J.-L., Qu, X.-K., and Sun, J.-H. (2020). Integrated Analysis of lncRNA-Mediated ceRNA Network Reveals a Prognostic Signature for Hepatocellular Carcinoma. *Front. Genet.* 11, 602542. doi:10.3389/fgene.2020.602542

Swartz, L. (2020). Corrigendum. *Neuro Oncol.* 22 (6), 894. doi:10.1093/neuonc/noz100

Tan, S. K., Pastori, C., Penas, C., Komotar, R. J., Ivan, M. E., Wahlestedt, C., et al. (2018). Serum Long Noncoding RNA HOTAIR as a Novel Diagnostic and Prognostic Biomarker in Glioblastoma Multiforme. *Mol. Cancer* 17 (1), 74. doi:10.1186/s12943-018-0822-0

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S. M., Ala, U., et al. (2011). Coding-independent Regulation of the Tumor Suppressor PTEN by Competing Endogenous mRNAs. *Cell* 147 (2), 344–357. doi:10.1016/j.cell.2011.09.029

Vitucci, M., Karpinich, N. O., Bash, R. E., Werneke, A. M., Schmid, R. S., White, K. K., et al. (2013). Cooperativity between MAPK and PI3K Signaling Activation Is Required for Glioblastoma Pathogenesis. *Neuro-Oncology* 15 (10), 1317–1329. doi:10.1093/neuonc/not084

Wang, H., Li, L., and Yin, L. (2018). Silencing LncRNA LOXL1-AS1 Attenuates Mesenchymal Characteristics of Glioblastoma via NF-Kb Pathway. *Biochem. Biophysical Res. Commun.* 500 (2), 518–524. doi:10.1016/j.bbrc.2018.04.133

Wen, P. Y., and Kesari, S. (2008). Malignant Gliomas in Adults. *N. Engl. J. Med.* 359 (5), 492–507. doi:10.1056/NEJMra0708126

Xiao, K., Tan, J., Yuan, J., Peng, G., Long, W., Su, J., et al. (2020). Prognostic Value and Immune Cell Infiltration of Hypoxic Phenotype-related Gene Signatures in Glioblastoma Microenvironment. *J. Cel. Mol. Med.* 24 (22), 13235–13247. doi:10.1111/jcmm.15939

Xie, P., Li, X., Chen, R., Liu, Y., Liu, D., Liu, W., et al. (2020). Upregulation of HOTAIRM1 Increases Migration and Invasion by Glioblastoma Cells. *Aging* 13 (2), 2348–2364. doi:10.18632/aging.202263

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

Yuan, J., Xiao, G., Peng, G., Liu, D., Wang, Z., Liao, Y., et al. (2015). MiRNA-125a-5p Inhibits Glioblastoma Cell Proliferation and Promotes Cell Differentiation by Targeting TAZ. *Biochem. Biophysical Res. Commun.* 457 (2), 171–176. doi:10.1016/j.bbrc.2014.12.078

Zhang, Q., Sun, L., Zhang, Q., Zhang, W., Tian, W., Liu, M., et al. (2020). Construction of a Disease-specific lncRNA-miRNA-mRNA Regulatory Network Reveals Potential Regulatory Axes and Prognostic Biomarkers for Hepatocellular Carcinoma. *Cancer Med.* 9 (24), 9219–9235. doi:10.1002/cam4.3526

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., and Hovarth, S. (2010). Weighted Gene Coexpression Network Analysis: State of the Art. *J. Biopharm. Stat.* 20 (2), 281–300. doi:10.1080/10543400903572753

Zhong, Y., Xu, F., Wu, J., Schubert, J., and Li, M. M. (2021). Application of Next Generation Sequencing in Laboratory Medicine. *Ann. Lab. Med.* 41 (1), 25–43. doi:10.3343/alm.2021.41.1.25

Zhou, Y., Yang, L., Zhang, X., Chen, R., Chen, X., Tang, W., et al. (2019). Identification of Potential Biomarkers in Glioblastoma through Bioinformatic Analysis and Evaluating Their Prognostic Value. *Biomed. Res. Int.* 2019, 1–13. doi:10.1155/2019/6581576

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