DNA AND CELL BIOLOGY
Volume 38, Number 10, 2019
Mary Ann Liebert, Inc.
Pp. 1112–1124
DOI: 10.1089/dna.2019.4869

Identifying Hepatocellular Carcinoma Driver Genes by Integrative Pathway Crosstalk and Protein Interaction Network

Wenbiao Chen,[1] Jingjing Jiang,[1] Peizhong Peter Wang,[2] Lan Gong,[3] Jianing Chen,[1] Weibo Du,[1] Kefan Bi,[1] and Hongyan Diao[1]

In this study, we mined out hepatocellular carcinoma (HCC) driver genes from MEDLINE literatures by bioinformatics methods of pathway crosstalk and protein interaction network. Furthermore, the relationship between driver genes and their clinicopathological characteristics, as well as classification effectiveness was verified in the public databases. We identified 560 human genes reported to be associated with HCC in 1074 published articles. Functional analysis revealed that biological processes and biochemical pathways relating to tumor pathogenesis, cancer disease, tumor cell molecule, and hepatic disease were enriched in these genes. Pathway crosstalk analysis indicated that significant pathways could be divided into three modules: cancer disease, virus infection, and tumor signaling pathway. The HCC-related protein–protein interaction network comprised 10,212 nodes, and 56,400 edges were mined out to identify 18 modules corresponding to 14 driver genes. We verified that these 14 driver genes have high classification effectiveness to distinguish cancer samples from normal samples and the classification effectiveness was better than that of randomly selected genes. Present study provided pathway crosstalk and protein interaction network for understanding potential tumorigenesis genes underlying HCC. The 14 driver genes identified from this study are of great translational value in HCC diagnosis and treatment, as well as in clinical study on the pathogenesis of HCC.

Keywords: hepatocellular carcinoma, pathway crosstalk, protein interaction network, driver genes

Introduction

Hepatocellular carcinoma (HCC) is the fifth most common cancer worldwide and remains the third most frequent cause of cancer death, with nearly 321,200 deaths and 366,100 new cases reported in China (Torre *et al.*, 2015; Chen *et al.*, 2016). Despite modern management, including the introduction of improved surgical techniques, comprehensive treatment, and targeted therapies, the survival rate of patients is still quite low, largely attributable to late diagnosis, resistance to treatment, tumor recurrence, and metastasis (Forner *et al.*, 2012). HCC has become a comprehensive health problem, not only affecting on the HCC and their families but also bringing a heavy burden to community (Jinjuvadia *et al.*, 2017).

Although much effort has been dedicated to research, the gene loci are associated with pathogenesis of HCC and clinical therapeutic targets via various approaches, including gene expression (Wei *et al.*, 2013), autophagy (Liu *et al.*, 2017), exosome (Liu and Li, 2018), gut microbiota (Tao *et al.*, 2015), epigenetic dysregulation (Nakamura *et al.*, 2018), and immunologic mechanisms (Harding *et al.*, 2016). However, the HCC pathogenesis-related genes and biomarkers of genes were far from being explored. It is generally admitted that HCC as a result of the multifactorial and multistep complex process, is influenced by both environmental and genetic factors (Chuang *et al.*, 2009). Importantly, it is well known that disruption of the genetic machinery is closely associated with liver carcinogenesis (Zhang, 2015). Currently, some fully proven

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Disease, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China.
[2]Division of Community Health and Humanities, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada.
[3]St George and Sutherland Clinical School, University of New South Wales, Sydney, Australia.

HCC-related driver genes, such as *MET* (Boix *et al.*, 1994), *AXIN1* (Satoh *et al.*, 2000), *CTNNB1* (Devereux *et al.*, 2001), and tumor protein p53 (*TP53*) (Hsu *et al.*, 1991), have shown clinical implication as biomarker for diagnosis, in clinical trials of epigenetic drugs, as well as pathogenic research. However, those driver genes are only responsible for a minority of population, specific experimental cell lines, or animal models. Moreover, genetic analyses have suggested that, a complicated pathogenesis may be under the influence of other genes, and that individual differences can be caused by many genes and their variants. Genes with different biological functions may work together to promote the tumorigenesis of HCC, with a moderate or small effect exerted by each gene (Devereux *et al.*, 2001).

Thus, a comprehensive analysis of potential causal genes within a pathway and/or a network framework might provide many important insights beyond the conventional single-gene analyses (Wang, 2013). Lin *et al.* detected four lncRNAs gathered as a single prognostic signature, which could act as an indicator for HCC patient outcome and a potential independent biomarker for prognosis prediction of HCC (Sui *et al.*, 2018). Kim *et al.* reported that a novel gene expression signature involving four epithelial–mesenchymal transition genes was associated with the prognosis of HCC patients, and complement prognostic assessment based on important clinicopathologic parameters (Kim *et al.*, 2010). Chen *et al.* identified six hub genes in association with HCC metastasis risk and prognosis, which might improve the prognosis by influencing

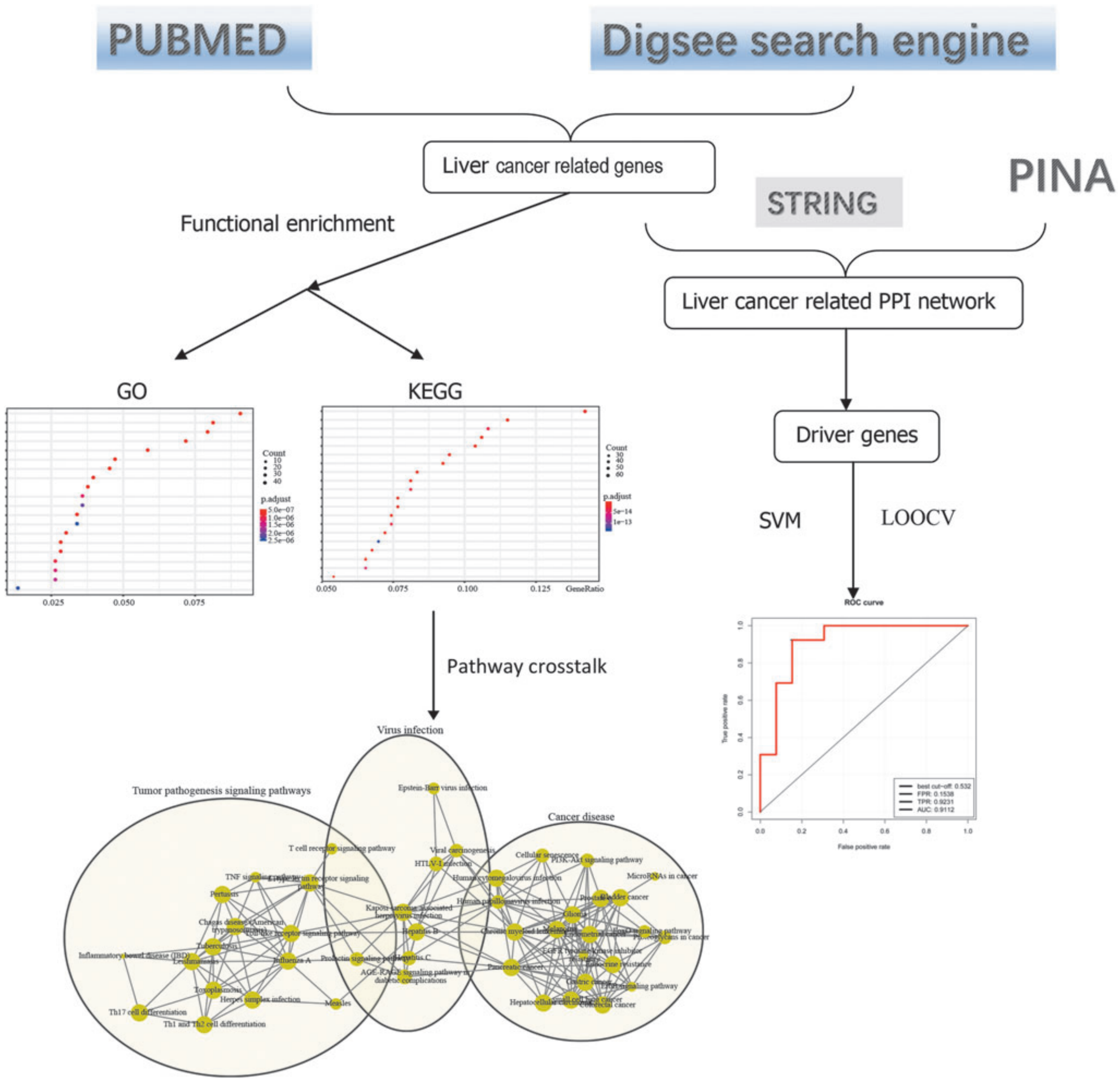

FIG. 1. Flow diagram of the analysis procedure: data collection, preprocessing, analysis, and validation.

amino acid metabolism and oxidation (Chen *et al.*, 2017). Existing studies are mainly target-based selected genes, and there is a paucity of framework guided and comprehensive examinations of driver genes.

In this study, we conducted a comprehensive collection of HCC-related genes from genetic association literatures. Also, the bioinformatics analysis of biochemical pathways was then performed to reveal the important functional themes within these genetic factors and to identify the interaction and correlation between the pathways by pathway crosstalk analysis. Furthermore, the protein–protein interaction (PPI) associated with HCC-related genes was constructed to mine out modules corresponding to driver genes. In addition, the relationship between driver genes and clinicopathological characteristics was verified by chi-squared test. Finally, we used leave-one-out cross validation (LOOCV) algorithm to verify the classification effectiveness of driver genes in other public databases (Fig. 1). The set of driver genes which we mined out has a high value for the diagnosis and treatment of HCC. Besides, driver genes offered useful insights for understanding the molecular pathogenesis of HCC from a perspective of systems biology. Also, the frame of the research methods could be applied to other disease models.

Materials and Methods

Identification of driver genes

The study was approved by the Clinical Research Ethics Committee of College of Medicine, Zhejiang University (2018983). We searched for genes genetically associated with HCC by DigSee search engine, which searches MEDLINE abstracts for evidence sentences describing that ''genes'' are involved in the development of ''cancer'' through ''biological events'' (Kim *et al.*, 2013). Since the epigenetic changes in the molecular mechanism of genes play an important role in the development of HCC. Therefore, the key words, including HCC, as well as mutation, gene expression, regulation, protein catabolism, phosphorylation, localization, binding, transcription, hydroxylation, ubiquitination, DNA methylation, glycosylation, acetylation, methylation, and catalysis, were used to research the genes associated with HCC by DigSee search engine

Functional analysis of HCC-related genes

Gene ontology (GO) analyses were performed to investigate HCC-related genes attributes in any organism, including molecular function, biological processes, and cellular components. Besides, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed to determine the pathways associated with HCC-related genes. Org.Hs.eg.db R package was used to convert HCC-related genes into Entrez Gene Identifiers (Carlson *et al.*, 2016). The clusterProfiler R package offers a gene classification method (groupGO) to classify genes based on the projection at a specific level of the GO corpus, and provides enrichGO and enrichKEGG based on hypergeometric distribution to calculate enrichment test for GO terms and KEGG pathways (Yu *et al.*, 2012). Therefore, we used the enrichment analysis method clusterProfiler to calculate enrichment test for GO and KEGG, and the *p* value was corrected by false discovery rate (FDR). Thereafter, either the GO biological process terms or KEGG pathways with FDR <0.05 were considered to be significantly enriched.

Pathway crosstalk analysis

Analysis of crosstalk of relationships among pathways was used to investigate interlinks and interactions of the significant pathways, especially those with overlapping coefficients of two significant pathway analysis results. The overlap between two pathways was determined based on overlap coefficient (OC) and the Jaccard coefficient (JC) formulas: $JC = |(A \cap B)/(A \cup B)|$ and $OC = |A \cap B|/\min(|A|,|B|)$, where A and B are the lists of genes of the two examined pathways. Pathway pairs that were used to build the pathways crosstalk and the overlap significance of each pathway pair were measured based on the average scores of JC and OC. Then, we performed the following procedures to construct the pathway crosstalk: (1) pathways with five or more candidate genes were included, resulted from the biological significance of the association between pathways that was low if the number of overlapping genes between pathway pairs was too small; (2) counting the number of common candidate genes of pathway pairs, in which only the pathways pairs with more than six overlapped genes were taken into account; (3) measuring the overlap of all pathway pairs by the value of JC and OC algorithm; and (4) the pathways and correlations between pathways were considered as node and edge, respectively. We visualized the pathway crosstalk via software Cytoscape (Shannon *et al.*, 2003).

Construction of PPI network

We constructed a PPI network to explore the correlation and interaction among the HCC-related genes. First, the PPI data of *Homo sapiens* were downloaded from protein interaction network analysis (PINA) database* (Cowley *et al.*, 2012). Meanwhile, we used UniProt Retrieve/ID mapping tool to transfer protein identifier to gene symbol (Pundir *et al.*, 2016). In addition, another human PPI datum was selected from STRING database of Homos sapiens.** The STRING database was aimed to collect and integrate this information, by consolidating known and predicted protein–protein association data for a large number of organisms. The associations in STRING include physical interactions as well as functional interactions (Szklarczyk *et al.*, 2017). For each protein–protein association stored in STRING, a score was provided, which indicated the estimated likelihood that a given interaction was biologically meaningful, specific, and reproducible, given the supporting evidence. Thus, we selected the PPI data with score >900. Finally, we merged the two interactome databases by excluding the self-interacting and redundant pairs. The interaction pairs containing HCC-related genes were retained in the PPI network.

Data mining of driver genes

Several molecular alterations are known to occur in the genes that encode signaling proteins critical for tumorigenesis, cellular proliferation, tumor growth, diffusion, and survival. These genes, which contain driver mutations, have been defined as ''driver genes'' (Bailey *et al.*, 2018). Thus, the

*(http://omics.bjcancer.org/pina/)

**(https://string-db.org/cgi/input.pl)

‘‘specific genes,’’ which were mined out from HCC-related genes in this research could be defined as driver genes because they have multiple biological associations with tumorigenesis. The HCC-related PPI network was imported into Cytoscape platforms. ClusterONE algorithm was used to discover densely connected and possibly overlapping models within the Cytoscape network. ClusterONE is a method for detecting potentially overlapping protein complexes from PPI data based on the score of matching partial protein complex, geometric precision prediction, and maximum matching rate. The algorithm was built on the concept of the cohesiveness score and used a greedy growth process to find groups in a PPI network that are likely to correspond to protein complexes (Nepusz *et al.*, 2012). The number of genes in the models was calculated, and minimum number >20 was used as the screening criteria. Finally, we selected the highest degree genes in the HCC-related PPI network of each model as HCC driver genes. The higher the degree was, the more connected the gene was, and the genes were also associated with each other. The more the degrees, the more likely the genes were to participate in the occurrence of HCC.

Correlation between clinicopathological characteristics and driver genes

The RNA-Seq was downloaded from The Cancer Genome Atlas Cancer Genome (TCGA) database, including 421 liver cancer patient samples (corresponding to 371 pathological tissue samples). In the meantime, 378 clinically clinicopathological characteristics were also downloaded. We standardized the expression values of these driver genes, and obtained the standardized genes expression matrix, which contained N rows and M columns. N and M stood for the number of samples and the number of driver genes, respectively. Then the mean value of driver genes expression values of each sample was obtained, and a vector of length of N was received. Following, we calculated the mean of this vector to obtain the mean value of all the samples. If the expression value of a sample was higher than the mean value, the sample

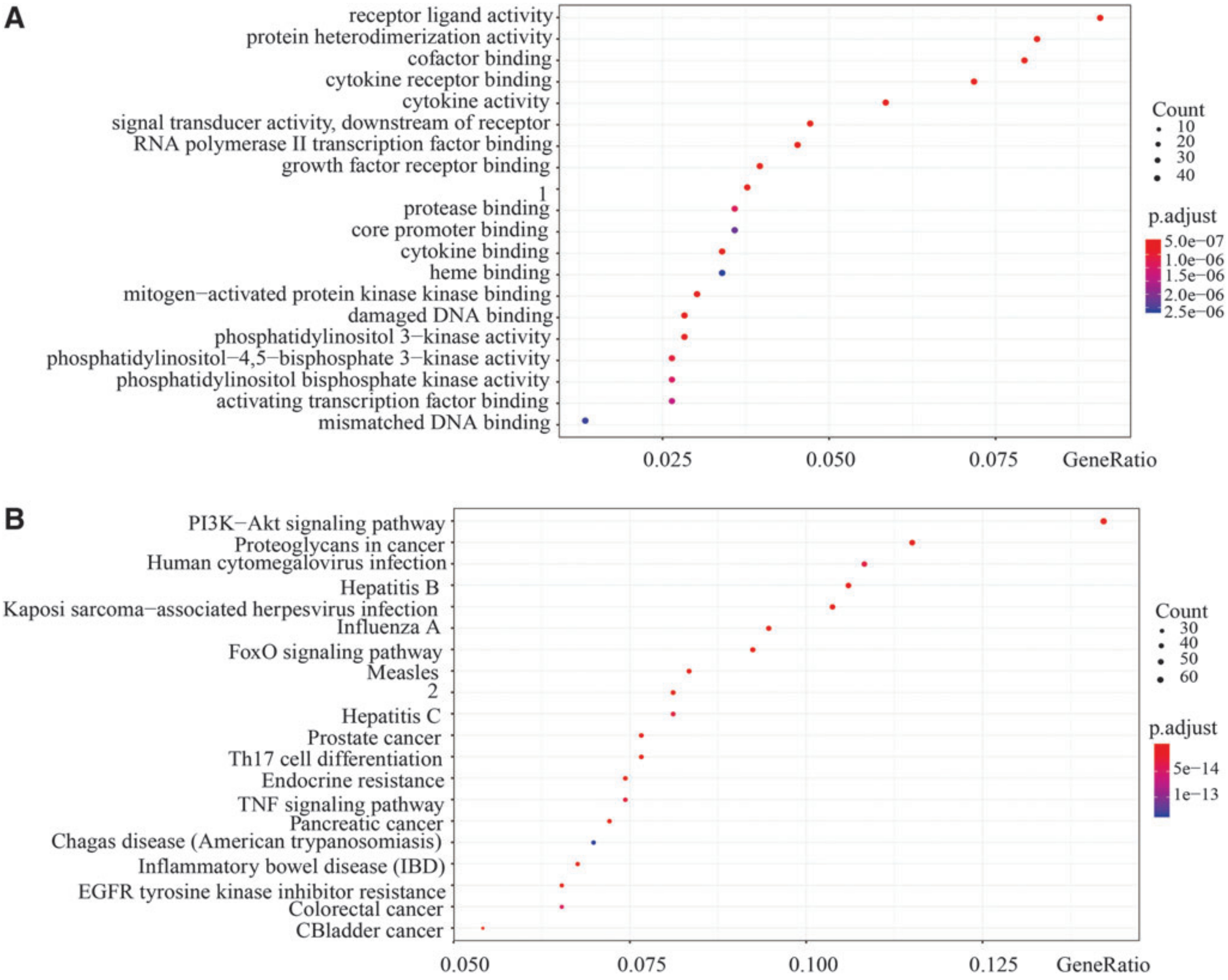

FIG. 2. Functional analysis of HCC-related genes. **(A)** GO analysis and **(B)** KEGG pathway analysis. (1) Signal transducer, downstream of receptor, with serine/threonine kinase activity. (2) AGE–RAGE signaling pathway in diabetic complications. GO, gene ontology; HCC, hepatocellular carcinoma; KEGG, Kyoto Encyclopedia of Genes and Genomes.

was classified as high expression group. The sample was classified as low expression group when the expression value of a sample was lower than the mean value. The correlation between clinicopathological characteristics (tumor grading, tumor staging, tumor, node, metastasis [TNM], age, gender, and cancer-normal) was analyzed by Pearson chi-squared. Data analyses were conducted using SPSS software (SPSS 22.0, Chicago, IL). Data measured using continuous variables are expressed as means ± standard deviations. The *p* values <0.05 were considered statistically significant.

Verification of the driver genes

We used support vector machines method to construct classifiers by R e1071 package (parameter used default value), which were further used to classify cancer samples and normal samples with the characteristics of the driver genes identified in this study. The LOOCV approach was used to evaluate the classification effectiveness and to verify the accuracy of the classification results. LOOCV is one of the most commonly used methods of evaluating predictive performances of a model, which is given *a priori* or developed by a modeling procedure. Under cross-validation, the available data are divided into *k* disjoint sets; *k* models are then trained, each on a different combination of *k*–1 partitions and tested on the remaining partition. The *k*-fold cross-validation estimate of a given performance statistic is then simply the mean of *k* models over the corresponding test partitions of the data (Cawley, 2006). Finally, the corresponding receiver operating characteristic curve was drawn and the area under curve (AUC) value under the curve was used to evaluate the classification effectiveness. We made comparison between the targeted driver genes and same number of random genes to verify the classification effectiveness of driver genes based on AUC. The random gene sets were randomly selected in TCGA expression profile with sample command in R language package. In addition, the same methods were used to verify the external two datasets from Gene Expression Omnibus (GEO) database (GSE73708, GSE14520).

Results

Identification of HCC-related genes reported to be associated with HCC

Understanding the HCC-related genes could be further enhanced by identifying biological events (e.g., gene expression, regulation, epigenetic modification, localization, and protein catabolism), in which the genetic effect is valid for the HCC development. We used DigSee, a search engine to find explicit association between genes and cancer through biological events in evidence sentences of MEDLINE abstracts. DigSee is a robust and accessible search engine, which using fine-grained information extraction techniques to mine out the specific information from the literature (Kim *et al.*, 2013). According to the user's request, it can serve the sentences of the identified triple relationship, which requires ''what genes'' to participate in the ''what kind of disease'' through ''what biological events.'' A gene was considered a disease-associated gene if it was directly or indirectly related to the cause of the disease or helps to increase or decrease the properties of the disease in the cell. Then, the DigSee algorithm collected and sorted the sentences, called evidence

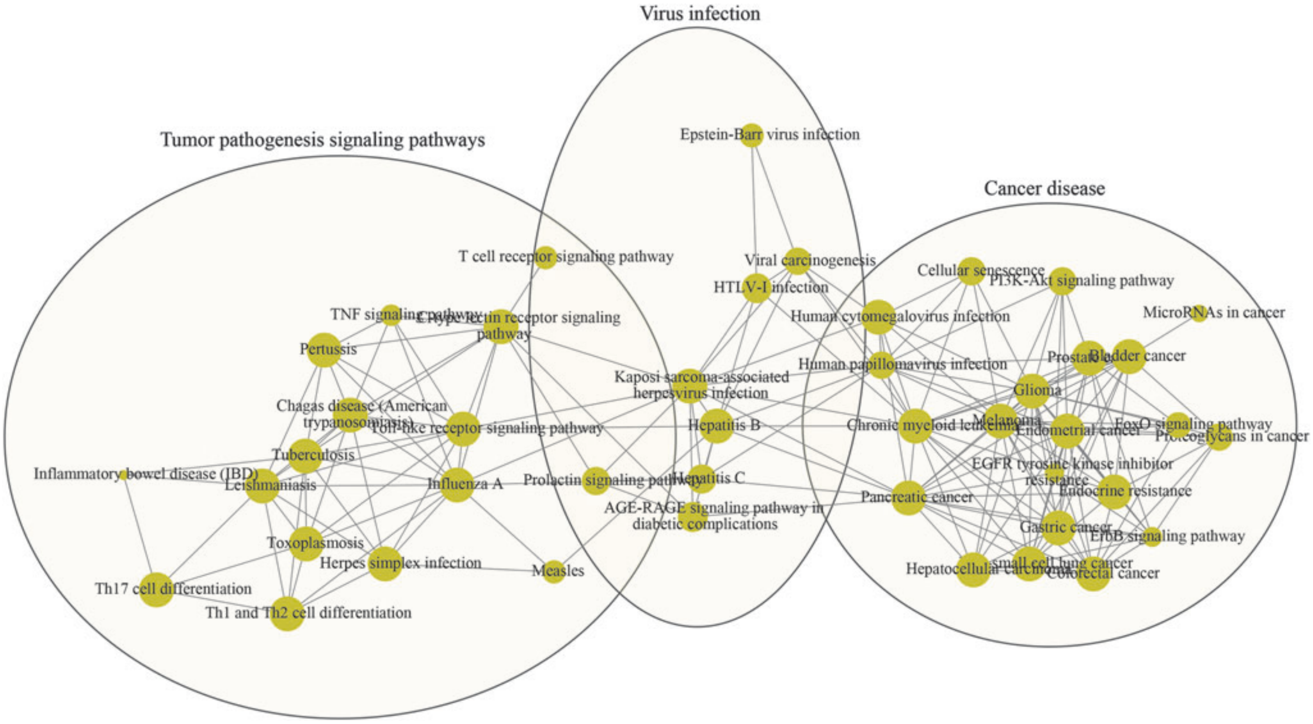
Virus infection
Tumor pathogenesis signaling pathways
Cancer disease
Epstein-Barr virus infection
T cell receptor signaling pathway
Viral carcinogenesis
HTLV-I infection
Human cytomegalovirus infection
Human papillomavirus infection
Cellular senescence
PI3K-Akt signaling pathway
MicroRNAs in cancer
TNF signaling pathway
C-type lectin receptor signaling pathway
Pertussis
Kaposi sarcoma-associated herpesvirus infection
Prostate cancer
Bladder cancer
Glioma
Chagas disease (American trypanosomiasis)
Toll-like receptor signaling pathway
Hepatitis B
Chronic myeloid leukemia
Melanoma
Endometrial cancer
FoxO signaling pathway
Proteoglycans in cancer
Tuberculosis
Inflammatory bowel disease (IBD)
Leishmaniasis
Influenza A
Prolactin signaling pathway
Hepatitis C
EGFR tyrosine kinase inhibitor resistance
Endocrine resistance
AGE-RAGE signaling pathway in diabetic complications
Pancreatic cancer
Toxoplasmosis
Herpes simplex infection
Measles
Gastric cancer
ErbB signaling pathway
Th17 cell differentiation
Th1 and Th2 cell differentiation
Hepatocellular carcinoma
small cell lung cancer
Colorectal cancer

FIG. 3. Pathway crosstalk analysis amid HCC pathways. *Yellow nodes* represent pathways, and edges represent crosstalk between pathways. The three *big circles* were used to delineate the main functional areas. Each big circular path represented a functional module and the function pathways in each *circle* would perform the function intersection.

sentences, to clearly indicate that the disease gene alters the characteristics of the diseased cells through biological events. In this study, we used HCC as disease to research the associational gene in the search engine website. Since the occurrence of epigenetic changes of genes contributed to the tumorigenesis, gene epigenetic changes, such as mutation, gene expression, regulation, protein catabolism, phosphorylation, localization, binding, transcription, hydroxylation, ubiquitination, DNA methylation, glycosylation, acetylation, methylation, and catalysis, were used as key words to find out the HCC-related genes from literatures. Thereafter, we identified 560 HCC-related genes corresponding to 1074 published research articles that were found out to be significantly associated with HCC. These HCC-related genes are highly variable in functions, such as mutation, gene expression, protein catabolism, DNA methylation, transcription, and localization. This underscored the complexity which indicates that the HCC-related genes are involved in the complex process of tumorigenesis.

Biological function enrichment and biochemical pathway of HCC-related genes

The biological function enrichment analysis enabled us to produce a more specific function spectrum of HCC-related genes. As a result, we identified 560 HCC-related genes that were significantly enriched in 211 GO terms. We selected the top 20 items of GO with the lowest *p* value to show in the Figure 2A. Among them, some GO terms were previously studied to be associated with the pathogenesis of HCC and

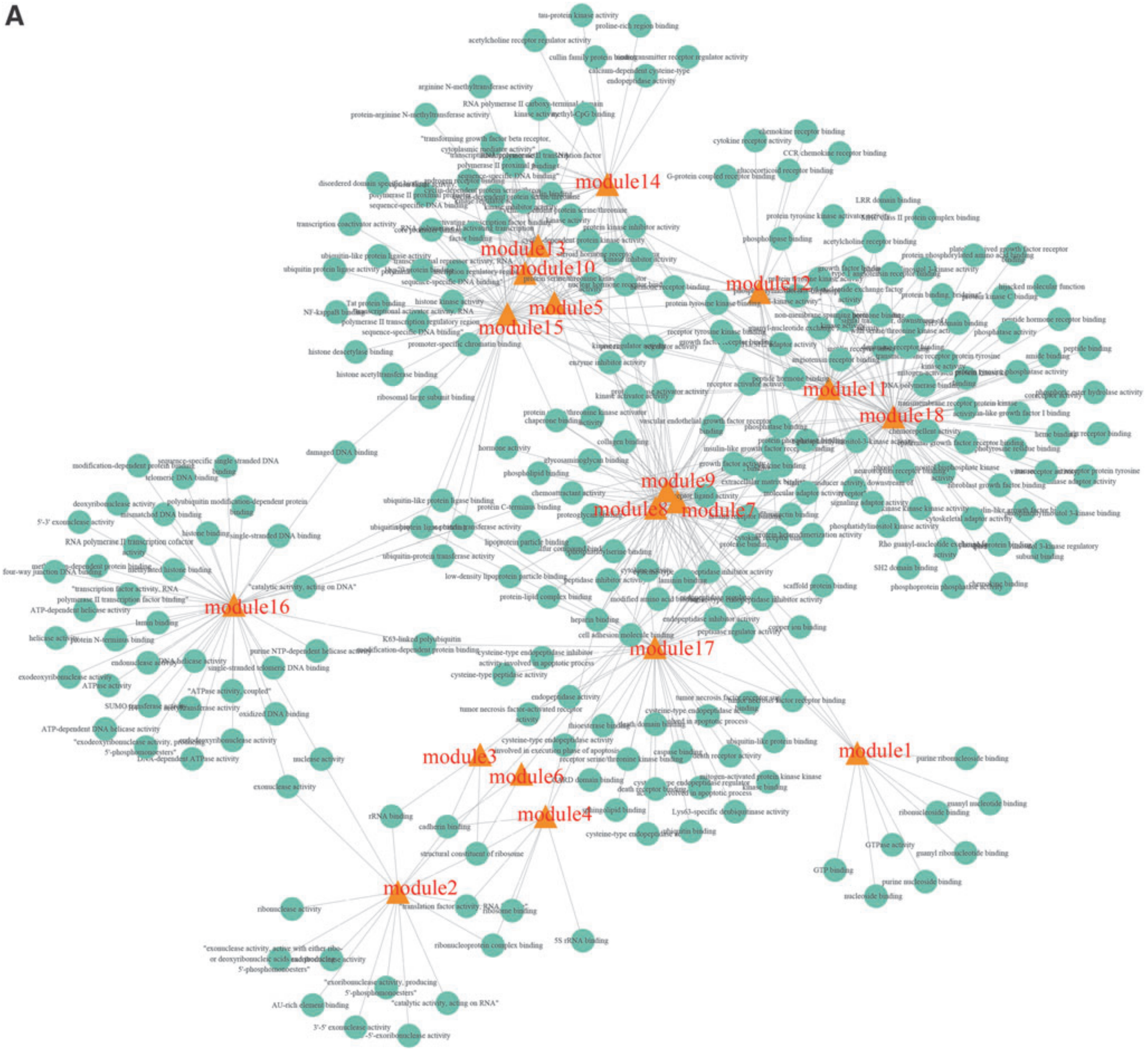
A
module14
module13
module10
module5
module15
module12
module11
module18
module9
module8
module7
module16
module17
module3
module6
module4
module1
module2
tau-protein kinase activity
proline-rich region binding
acetylcholine receptor regulator activity
cullin family protein binding
neurotransmitter receptor regulator activity
calcium-dependent cysteine-type endopeptidase activity
arginine N-methyltransferase activity
protein-arginine N-methyltransferase activity
chemokine receptor binding
cytokine receptor activity
CCR chemokine receptor binding
G-protein coupled receptor binding
glucocorticoid receptor binding
LRR domain binding
transcription coactivator activity
disordered domain specific binding
protein tyrosine kinase activator activity
phospholipase binding
acetylcholine receptor binding
histone deacetylase binding
histone acetyltransferase binding
ribosomal large subunit binding
damaged DNA binding
modification-dependent protein binding
telomeric DNA binding
deoxyribonuclease activity
5'-3' exonuclease activity
four-way junction DNA binding
ATP-dependent helicase activity
helicase activity
exodeoxyribonuclease activity
ATPase activity
SUMO transferase activity
ATP-dependent DNA helicase activity
DNA-dependent ATPase activity
oxidized DNA binding
nuclease activity
exonuclease activity
hormone activity
glycosaminoglycan binding
phospholipid binding
collagen binding
ubiquitin-like protein ligase binding
ubiquitin-protein transferase activity
low-density lipoprotein particle binding
protein-lipid complex binding
heparin binding
cell adhesion molecule binding
copper ion binding
scaffold protein binding
SH2 domain binding
phosphoprotein phosphatase activity
peptide binding
amide binding
heme binding
tumor necrosis factor-activated receptor activity
endopeptidase activity
death receptor activity
Lys63-specific deubiquitinase activity
ubiquitin binding
purine ribonucleoside binding
guanyl nucleotide binding
ribonucleoside binding
GTPase activity
GTP binding
guanyl ribonucleotide binding
purine nucleoside binding
nucleoside binding
rRNA binding
cadherin binding
structural constituent of ribosome
ribonuclease activity
ribosome binding
5S rRNA binding
ribonucleoprotein complex binding
AU-rich element binding
3'-5' exonuclease activity
"catalytic activity, acting on RNA"
"catalytic activity, acting on DNA"

FIG. 4. Functional analysis of modules and acquirement of driver genes. **(A)** GO analysis and **(B)** KEGG pathway analysis. **(C)** Distribution of driver genes based on degree. *Orange triangles* and *green circles* represented 18 modules and GO/KEGG items, respectively. The detail information of GO/KEGG items (*green circles*) in the figure could be found in Supplementary Table S1 and Supplementary Table S2.

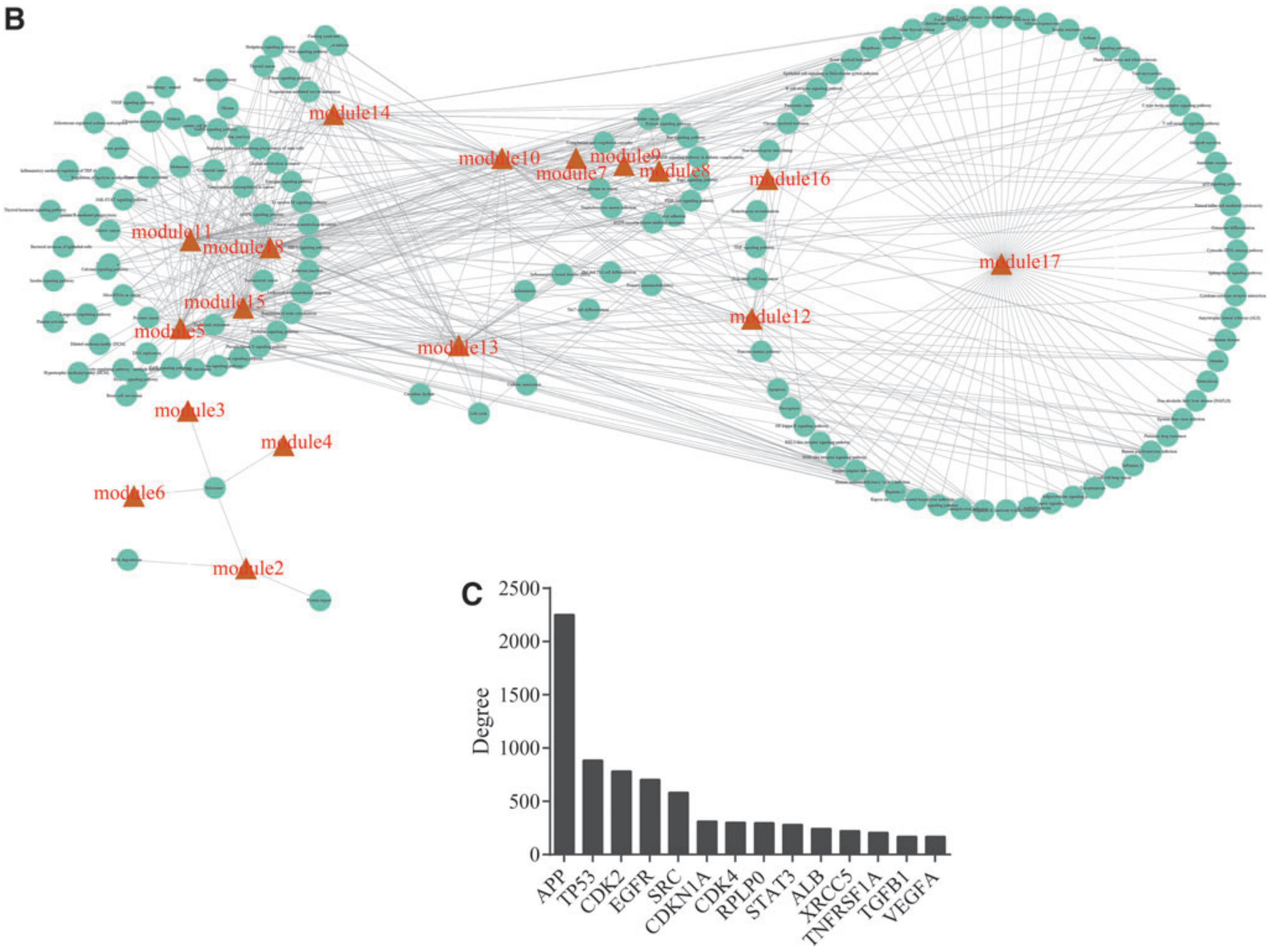

FIG. 4. (*Continued*)

liver disease, such as cytokine activity (Diao *et al.*, 2012), RNA polymerase II transcription factor binding (Stefanska *et al.*, 2013), signal transducer, downstream of receptor, with serine/threonine kinase activity (Hou *et al.*, 2019), and mitogen-activated protein kinase binding (Yang *et al.*, 2014). These results were consistent with the finding that complicated connections exist within the mechanism of HCC. Furthermore, we searched for enriched pathways of HCC-related genes and identified 144 significant enrichment pathways for HCC. The top 20 items of KEGG with the lowest *p* value are represented in the Figure 2B. Consistent with previous research, some pathways are related to HCC signaling pathway, such as transforming growth factor-beta signaling pathway (Chen *et al.*, 2018), NF-κB signaling pathway (Lu *et al.*, 2018), Hippo signaling pathway (Wang *et al.*, 2018), and JAK-STAT signaling pathway (Shen *et al.*, 2017). Several pathways involved in cellular physiological process that were generally admitted to link to tumor cell proliferation were included, such as apoptosis, necroptosis, adherens junction, and cell cycle. Also, we found HCC-related genes enriched in other cancer types, included PTEN for pancreatic cancer (Gu *et al.*, 2019), MYC of bladder cancer (Wu *et al.*, 2016), CYP1B1 of prostate cancer (Gu *et al.*, 2016), and CXL12 of endometrial cancer (Krikun, 2018). These results indicated that the HCC-related genes participate in the pathogenic molecular mechanism underlying HCC, and also proved that the identified HCC-related genes are relatively reliable for further bioinformatics analysis.

Crosstalk among significant enrichment pathways

Crosstalk among significant enrichment pathway analysis was not only to identify significantly enriched pathways but also to understand the interactions among them. Thus, we performed a pathway crosstalk analysis for the top 50 enriched pathways with 256 genes. Based on their crosstalk, those pathways could be roughly divided into three major modules, each module has more interactions between the pathways within the module than with those outside of this module, likely to be associated with the same or similar biological procedure (Fig. 3). The first module mainly is consisted of tumor pathogenesis signaling pathways, such as T cell receptor signaling pathway, TNF signaling pathway, Toll-like receptor signaling pathway, and prolactin signaling pathway. The second module is primarily dominated by theme of virus infection, including viral carcinogenesis, Human T lymphocyte leukemia virus I type (HTLV-I) infection, Kaposi sarcoma-associated herpesvirus infection, human cytomegalovirus infection, human papillomavirus infection, hepatitis B, and hepatitis C. The major contents of third modules are cancer disease, such as bladder cancer, pancreatic cancer, non-small cell lung cancer, HCC, colorectal cancer, endometrial cancer, and gastric cancer. In the meantime, the three modules are interlinked with each other via a couple of pathway interactions.

Acquirement of driver genes

The acquirement of driver genes was based on the construction of PPI and data mining of module. At first, 166,776 interactions, corresponding to 5211 proteins were obtained from PINA database (last update: May 21, 2014). At the same time, we selected the human-related protein interactions with score >900 from STRING database. We received PPI network comprised 17,170 proteins and 360,061 interactions after merging the two interactome databases by excluding the self-interacting and removing redundant pairs. ''Merging'' referred to the merging of network edges that were obtained from two databases, while ''excluding the self-interacting'' referred to remove the edges from which interacted with themselves. ''Removing redundant pairs'' meant the retention of only one edge that was common to both databases. Then, the PPI network comprised 10,212 proteins and 56,400 interactions, which is associated with HCC-related genes that were retained to further mine out modules. We imported the PPI network with 10,212 proteins and 56,400 interactions related to HCC into Cytoscape software. Next, the mining modules were carried out using the ClusterONE plug-in Cytoscape software. The minimum gene number threshold value of parameter selection module was 20, and all other parameters were default. Totally, 18 modules were selected with each containing at least 20 genes. These 18 modules went through GO and KEGG functional analysis (Fig. 4A, B). We can see that 18 modules (orange triangles) were associated with all kinds of biologically functional items and they were also interconnected with each other (Supplementary Tables S1 and S2). Functional analysis revealed a more specific function of modules related to tumorigenesis, such as TNF-activated receptor activity, mismatched DNA binding, protein kinase inhibitor activity, and activity involved in apoptotic process in GO terms, as well as p53 signaling pathway, NF-κB signaling pathway, Ras signaling pathway, and mTOR signaling pathway in KEGG pathways. In each module, we selected the highest ''degree'' genes, which were regarded as driver genes from HCC-related PPI network. After deleting the repeated genes, we acquired 14 driver genes, included *APP*, *TP53*, *CDK2*, *EGFR*, *SRC*, *CDKN1A*, *CDK4*, *RPLP0*, *STAT3*, *ALB*, *XRCC5*, *TNFRSF1A*, *TGFB1*, and *VEGFA* (Fig. 4C).

Verification of driver genes

We validated the average value of expression of driver genes and investigated the relationship between the 14 driver genes and clinicopathological characteristics. Using median expression level as the cutoff point, the 14 driver genes were categorized into high-expression group and low-expression group. As shown in Table 1, the significant association was represented between driver genes expression and cancer-normal ($p = 1.63E\text{-}07$), grading ($p = 0.0257$). There was no significant relationship between 14 driver genes expression and the other clinicopathological characteristics ($p > 0.05$). The result indicated that the level of the 14 driver genes expression could be used to distinguish cancer samples from normal samples.

TABLE 1. THE RESULT OF CHI-SQUARE TEST ON THE RELATIONSHIP BETWEEN 14 DRIVER GENES AND CLINICOPATHOLOGICAL CHARACTERISTICS

Clinicopathological characteristics	*Driver gene expression (No. of patients)*		p
	High	*Low*	
Cancer-normal			1.63E-07
Cancer	156	181	
Normal	30	1	
T			0.2217
T1	99	82	
T2	39	55	
T3	40	40	
T4	6	7	
Stage			0.2596
S1	93	78	
S2	36	50	
S3	40	45	
S4	2	3	
Grading			0.0257
G1	34	21	
G2	94	83	
G3	52	70	
G4	3	9	
Age			0.1451
Young (age <60)	81	96	
Old (age >60)	104	89	
Sex			1
Male	125	125	
Female	61	60	

Furthermore, we made comparison between the 14 driver genes and same number of random genes to verify the classification effectiveness of driver genes by LOOCV algorithm. The classification effectiveness of 14 driver genes and same number of random genes is represented in Figure 5A and B, respectively. We can see that the classification effectiveness of 14 driver genes was 0.929 (AUC) larger than any of the 6 groups of randomly selected same number of genes (Fig. 5C). This result indicated that the classification effectiveness of the 14 driver genes was better than that of randomly selected genes. Moreover, to further confirm the reliability of the 14 driver genes, the same methods were used to verify the external dataset from GEO database (GSE73708). As expected, the classification effectiveness of the 14 driver genes was 1 (AUC), which was larger than any of the six groups of randomly selected same number of genes (Fig. 6). Consistently, we could obtain the same result on the other GEO database (GSE14520) (Supplementary Fig. S1). This result further confirmed the reliability of classification effectiveness of the 14 driver genes, which could be used to distinguish cancer samples from normal samples.

Discussion

Although more and more genes potentially involved in HCC have been identified with the improvement of sequencing

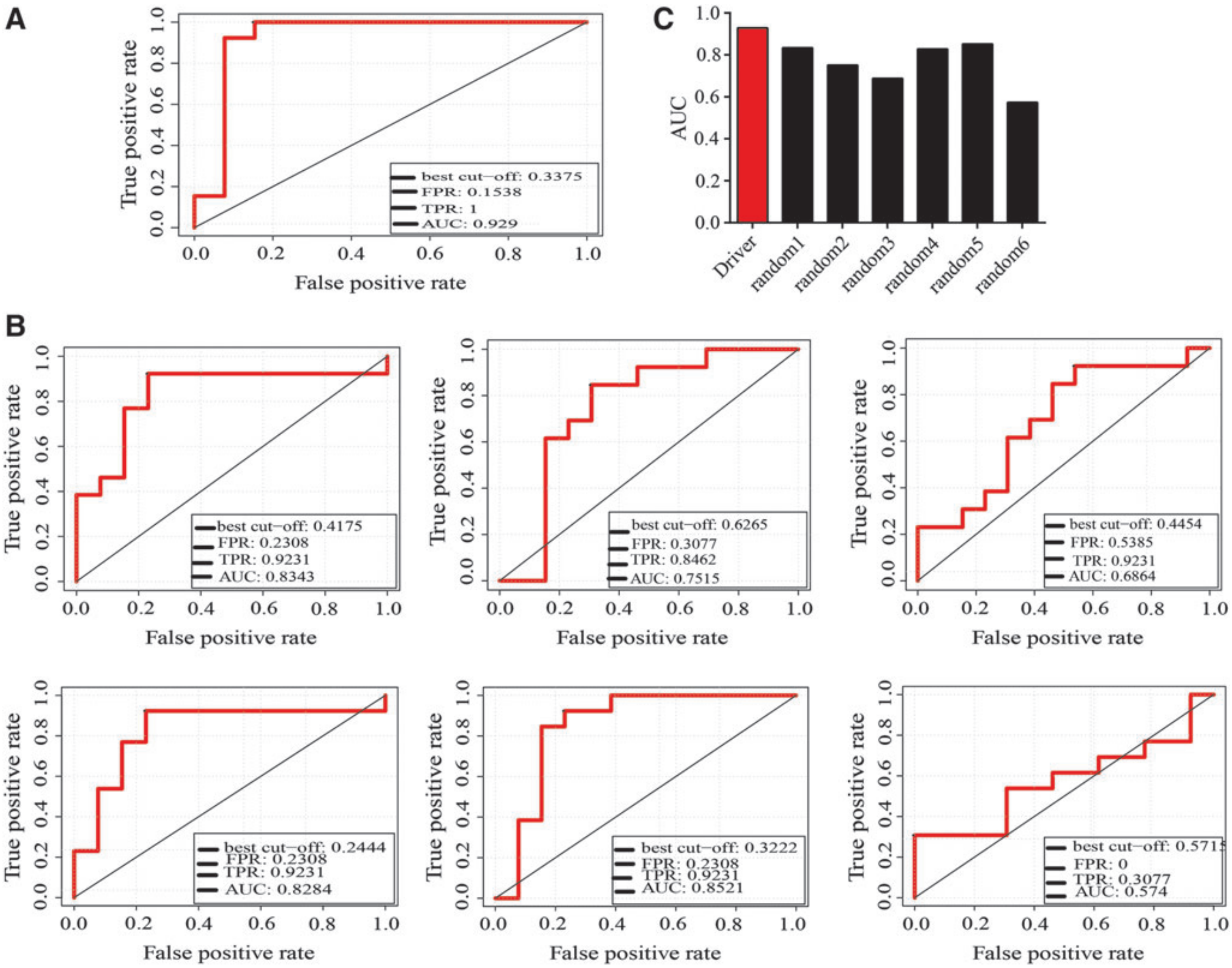

FIG. 5. Verification of classification effectiveness on driver genes. Classification effectiveness of **(A)** 14 driver genes and **(B)** random same number of genes. **(C)** Comparison of classification effectiveness between 14 driver genes and random genes.

platform and microarrays technology, a thorough analysis of the biochemical processes associated with the HCC driver genes is still uncompleted. Thus, it was urgent to dissect the tumorigenesis genes underlying HCC at systems biology level. In this study, we collected HCC-related genes in published literatures, and systematically delineated the interaction of those genes by means of pathway crosstalk and protein interaction network to discover 14 HCC driver genes, which have robust classification effectiveness to distinguish HCC samples from normal samples.

Our study conducted a comprehensive and systematic framework to analyze HCC-related genes and mine out driver genes, which has its own significant advantages. At first, the HCC-related genes came from genetic association literatures on HCC that have been proved or believed to have important relationship with pathogenesis of HCC. Therefore, we could obtain reliable HCC-related gene source for further bioinformatics analysis. Besides, to understand the biological interaction of HCC-related genes, functional enrichment and pathway crosstalk were taken into account, it not only verifies HCC-related genes involved in HCC tumorigenesis but also provides view of the molecular mechanisms underlying HCC. Furthermore, the driver genes were mined out based on PPI network. PPI network represents an essential aspect of cellular systems biology. Identification of key genes players and their interaction networks provides crucial insights into the regulation of cellular developmental processes and into physical connections between gene products (VanderSluis *et al.*, 2018). Fourteen driver genes came from HCC-related PPI network, which supplied a close relationship between the 14 driver genes and HCC molecular mechanisms. Moreover, as alterations in different types of genes were responsible for tumorigenesis, and only when several genes were mutated does an invasive cancer develop. We mined out several driver genes collectively showing connection with HCC, which indicated significantly genetic association with HCC (Vogelstein and Kinzler, 2004). Finally, the robust classification effectiveness of 14 driver genes was verified twice (in TCGA and GEO databases) by LOOCV algorithm to confirm its accuracy. The result suggested the potentially clinical implication of 14 driver genes in the diagnosis, treatment, and study of HCC.

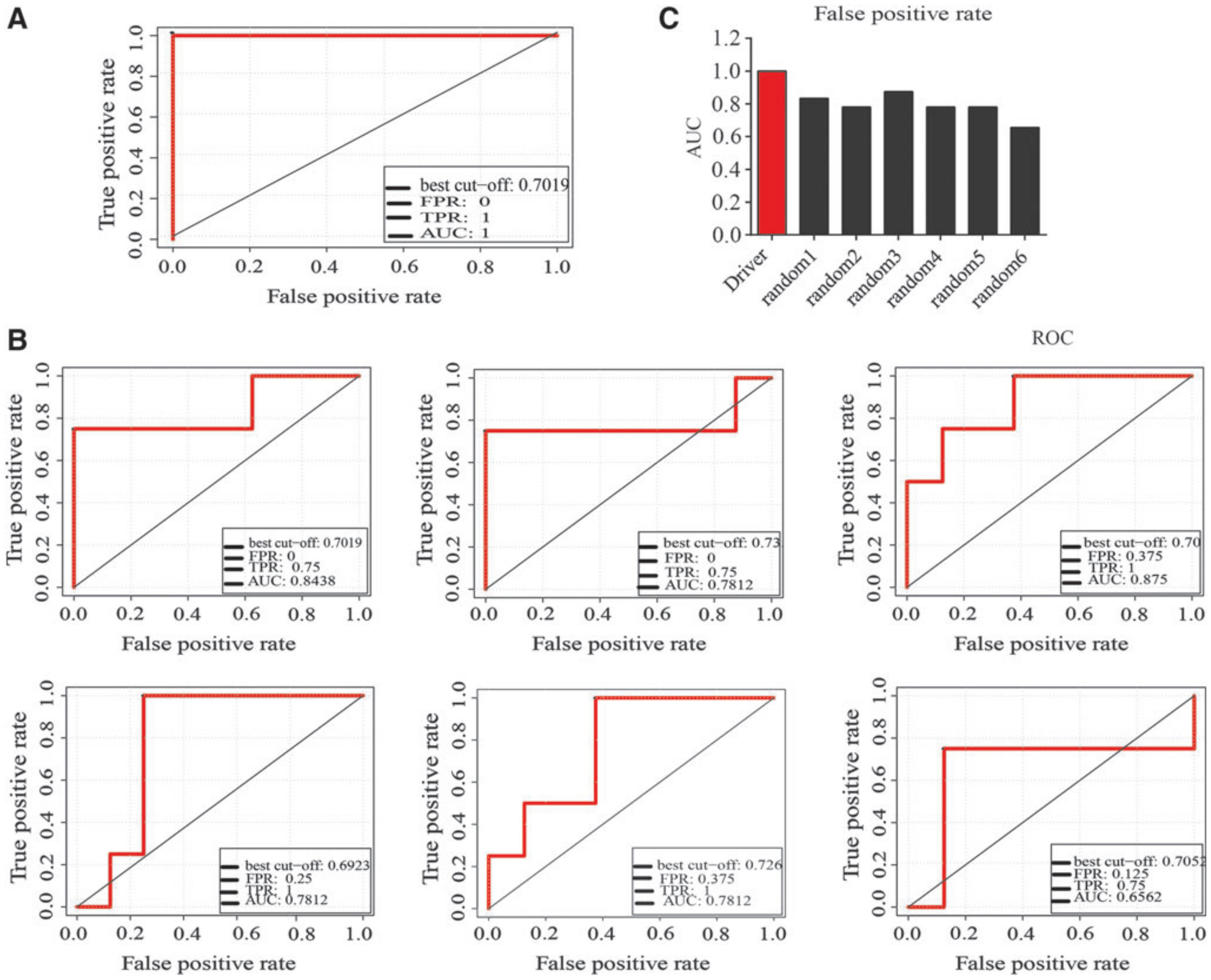

FIG. 6. Verification of driver genes in external database. Classification effectiveness of **(A)** 14 driver genes and **(B)** random same number of genes. **(C)** Comparison of classification effectiveness between 14 driver genes and random genes.

Of importance, we detected three major modules in pathways crosstalk. One module mainly involved in the pathways associated with tumor pathogenesis signaling pathways. Among these pathways, T cell receptor signaling pathway (Matsumoto *et al.*, 2014), TNF signaling pathway (Sayers, 2011), Toll-like receptor signaling pathway (Chen *et al.*, 2007), and prolactin signaling pathway (Ben-Jonathan *et al.*, 2002) have been well proved to be related to molecular mechanisms of tumorigenesis. These molecular mechanisms included signal transmission, immunologic suppression, cellular metabolism, and hormonal factors and indicated that the pathways significantly associated with HCC were diverse in function and consistent with the complexity of tumorigenesis. The second module was mainly dominated by the pathways of virus infection. It was generally admitted the associations of hepatitis B and C virus with liver cancer, human papillomaviruses with cervical cancer, and a subset of head and neck cancers (Cantalupo *et al.*, 2018). For third module, the major contents of pathways were different cancer diseases, hinted that different cancer diseases were caused by the altered expression of a lot of genes, which acted in concert to affect the same biological functions of pathways that eventually contributed to the different types of cancer. In addition, we found that the three modules were interacted via multiple edges formed by pathways, indicated each modules and pathways, and acted as a concerted manner to lead to the tumorigenesis of HCC.

We constructed HCC-related PPI network to mine out driver genes. As demonstrated by the results described above, this PPI network approach not only understands the protein functions in biological systems of HCC but also possesses the potentiality to detect promising relevant genes. Fourteen driver genes identified from HCC-related PPI network have close interaction with pathogenesis related to the biological processes involved in the HCC. Notably, these 14 driver genes possess robust classification effectiveness to distinguish cancer samples from normal samples. The result was consistent with previous findings that complicated connections existed between the pathophysiology of HCC and the 14 driver genes. The functions of 14 driver genes are not alone.

Instead, they interact with each other to promote HCC tumorigenesis. Of note, cyclin-dependent kinase 2 (*CDK2*) and cyclin-dependent kinase 4 (*CDK4*) encode members of a family of serine/threonine protein kinases that in association with cyclin E and cyclin D promote the G1/S phase transition (Padmakumar *et al.*, 2009). *CDK2* and *CDK4* play an important role in the development of HCC from cirrhosis by the molecular mechanism underlying perturbation of cell cycle regulation during hepatocarcinogenesis (Masaki *et al.*, 2003). However, the function of *CDK2* and *CDK4* could be inhibited by cyclin-dependent kinase inhibitor 1A (*CDKN1A*), which is functioned as a regulator of cell cycle progression at G1 to suppress the tumorigenesis (Cazzalini *et al.*, 2010; Jalili *et al.*, 2012). Importantly, the expression of *CDKN1A* is tightly controlled by the *TP53* (Soto *et al.*, 2005), which was multifunctional transcription factor that, along with number of other functions, regulates genes involved in cell cycle arrest, apoptosis, and senescence in response to various types of stress (Kandoth *et al.*, 2013). Vascular endothelial growth factor A (*VEGFA*) is upregulated in many known tumors and its expression is correlated with tumor stage and progression (Claesson-Welsh and Welsh, 2013). However, it was reported that *TP53* can inhibit *VEGFA* expression by regulating the transcriptional activity of Sp1 and also by downregulating the Src kinase activity to arrest angiogenesis and tumor growth (Pal *et al.*, 2001). It cannot be denied that epidermal growth factor receptor (*EGFR*) is commonly expressed in a variety of malignant epithelial cells. The methods of resistance to *EGFR*-targeted therapy were major research hotspots in HCC treatment (Wang *et al.*, 2016). Besides, one of the important signaling mediators downstream of normally and abnormally activated *EGFR* is signal transducer and activator of transcription 3 (*STAT3*). *STAT3* is regarded as oncogene, which is latent transcription factors that mediate cytokine- and growth factor-directed transcription. In many human cancers (including HCC) and transformed cell lines, *STAT3* is persistently activated, and in cell culture, active *STAT3* is either required for transformation, enhances transformation, or blocks apoptosis (Gao *et al.*, 2007). Many studies found that *STAT3* is activated by *EGFR* tyrosine kinases. *EGFR* is upstream of *STAT3* and is frequently overexpressed or overactivated in tumor cells (Yu *et al.*, 2009). As specified by the results detailed above, these 14 driver genes had close interaction with each other to be related to the process of tumorigenesis and development involved in HCC, they may also provide a list of potential candidates for further tumorigenesis pathogenesis exploration. Thus, the set of 14 driver genes could be used as targets of clinical utilities of diagnosis and treatment.

Conclusion

In this study, we investigated the integrative pathway crosstalk and protein interaction network related to HCC based on the genes associated with the disease by systems biology framework. By using integrating analysis of biological function, biochemical process, and pathway crosstalk analyses, we identified the biological processes and pathways associated with tumorigenesis underlying HCC. Moreover, HCC pathological PPI network was constructed to mine out 14 HCC driver genes, which were proved to have high classification effectiveness to distinguish cancer samples from normal samples. Such comprehensive analysis of genes involved in HCC will not only enhance our understanding of the genetic factors and their interaction with the pathogenesis of HCC but also improved our knowledge to capability to identify potential targets for HCC diagnosis and treatment. In the meantime, the framework represented in our study can be used to investigate the integrative pathway crosstalk, protein interaction network, and corresponding genes related to other disease models.

Disclosure Statement

No competing financial interests exist.

Funding Information

This work was supported by the National Natural Science Foundation of China (No. 81571953), the major national S&T projects for infectious diseases (2018ZX10301401-005), Key Research and Developement Plan of Zhejiang Province (2019C04005), the National Key Research and Development Program of China (2018YFC2000500).

Supplementary Material

Supplementary Figure S1
Supplementary Table S1
Supplementary Table S2

References

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., *et al.* (2018). Comprehensive characterization of cancer driver genes and mutations. Cell **174**, 1034–1035.

Ben-Jonathan, N., Liby, K., McFarland, M., and Zinger, M. (2002). Prolactin as an autocrine/paracrine growth factor in human cancer. Trends Endocrinol Metab **13,** 245–250.

Boix, L., Rosa, J.L., Ventura, F., Castells, A., Bruix, J., Rodes, J., *et al.* (1994). c-met mRNA overexpression in human hepatocellular carcinoma. Hepatology **19,** 88–91.

Cantalupo, P.G., Katz, J.P., and Pipas, J.M. (2018). Viral sequences in human cancer. Virology **513,** 208–216.

Carlson, M.R., Pages, H., Arora, S., Obenchain, V., and Morgan, M. (2016). Genomic annotation resources in R/bioconductor. Methods Mol Biol **1418,** 67–90.

Cazzalini, O., Scovassi, A.I., Savio, M., Stivala, L.A., and Prosperi, E. (2010). Multiple roles of the cell cycle inhibitor p21(CDKN1A) in the DNA damage response. Mutat Res **704,** 12–20.

Chen, J., Zaidi, S., Rao, S., Chen, J.S., Phan, L., Farci, P., *et al.* (2018). Analysis of genomes and transcriptomes of hepatocellular carcinomas identifies mutations and gene expression changes in the transforming growth factor-beta pathway. Gastroenterology **154,** 195–210.

Chen, P., Wang, F., Feng, J., Zhou, R., Chang, Y., Liu, J., *et al.* (2017). Co-expression network analysis identified six hub genes in association with metastasis risk and prognosis in hepatocellular carcinoma. Oncotarget **8,** 48948–48958.

Chen, R., Alvero, A.B., Silasi, D.A., and Mor, G. (2007). Inflammation, cancer and chemoresistance: taking advantage of the toll-like receptor signaling pathway. Am J Reprod Immun **57,** 93–107.

Chen, W., Zheng, R., Zeng, H., and Zhang, S. (2016). The incidence and mortality of major cancers in China, 2012. Chin J Cancer **35,** 73.

Chuang, S.C., La Vecchia, C., and Boffetta, P. (2009). Liver cancer: descriptive epidemiology and risk factors other than HBV and HCV infection. Cancer Lett **286,** 9–14.

Claesson-Welsh, L., and Welsh, M. (2013). VEGFA and tumour angiogenesis. J Intern Med **273,** 114–127.

Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., *et al.* (2012). PINA v2.0: mining interactome modules. Nucleic Acids Res **40,** D862–D865.

Devereux, T.R., Stern, M.C., Flake, G.P., Yu, M.C., Zhang, Z.Q., London, S.J., *et al.* (2001). CTNNB1 mutations and beta-catenin protein accumulation in human hepatocellular carcinomas associated with high exposure to aflatoxin B1. Mol Carcinog **31,** 68–73.

Diao, H., Liu, X., Wu, Z., Kang, L., Cui, G., Morimoto, J., *et al.* (2012). Osteopontin regulates interleukin-17 production in hepatitis. Cytokine **60,** 129–137.

Forner, A., Llovet, J.M., and Bruix, J. (2012). Hepatocellular carcinoma. Lancet **379,** 1245–1255.

Gao, S.P., Mark, K.G., Leslie, K., Pao, W., Motoi, N., Gerald, W.L., *et al.* (2007). Mutations in the EGFR kinase domain mediate STAT3 activation via IL-6 production in human lung adenocarcinomas. J Clin Invest **117,** 3846–3856.

Gu, C.Y., Qin, X.J., Qu, Y.Y., Zhu, Y., Wan, F.N., Zhang, G.M., *et al.* (2016). Genetic variants of the CYP1B1 gene as predictors of biochemical recurrence after radical prostatectomy in localized prostate cancer patients. Medicine **95,** e4066.

Gu, J., Wang, D., Zhang, J., Zhu, Y., Li, Y., Chen, H., *et al.* (2019). Corrigendum to "GFRα2 prompts cell growth and chemoresistance through down-regulating tumor suppressor gene PTEN via Mir-17-5p in pancreatic cancer." Cancer Lett **452,** 270.

Harding, J.J., El Dika, I., and Abou-Alfa, G.K. (2016). Immunotherapy in hepatocellular carcinoma: primed to make a difference? Cancer **122,** 367–377.

Hou, X., Yang, Y., Chen, J., Jia, H., Zeng, P., Lv, L., *et al.* (2019). TCRβ repertoire of memory T cell reveals potential role for Escherichia coli in the pathogenesis of primary biliary cholangitis. Liver Int **39,** 956–966.

Hsu, I.C., Metcalf, R.A., Sun, T., Welsh, J.A., Wang, N.J., and Harris, C.C. (1991). Mutational hotspot in the p53 gene in human hepatocellular carcinomas. Nature **350,** 427–428.

Jalili, A., Wagner, C., Pashenkov, M., Pathria, G., Mertz, K.D., Widlund, H.R., *et al.* (2012). Dual suppression of the cyclin-dependent kinase inhibitors CDKN2C and CDKN1A in human melanoma. J Natl Cancer Inst **104,** 1673–1679.

Jinjuvadia, R., Salami, A., Lenhart, A., Jinjuvadia, K., Liangpunsakul, S., and Salgia, R. (2017). Hepatocellular carcinoma: a decade of hospitalizations and financial burden in the United States. Am J Med Sci **354,** 362–369.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., *et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature **502,** 333–339.

Kim, J., Hong, S.J., Park, J.Y., Park, J.H., Yu, Y.S., Park, S.Y., *et al.* (2010). Epithelial-mesenchymal transition gene signature to predict clinical outcome of hepatocellular carcinoma. Cancer Sci **101,** 1521–1528.

Kim, J., So, S., Lee, H.J., Park, J.C., Kim, J.J., and Lee, H. (2013). DigSee: disease gene search engine with evidence sentences (version cancer). Nucleic Acids Res **41,** W510–W517.

Krikun, G. (2018). The CXL12/CXCR4/CXCR7 axis in female reproductive tract disease: review. Am J Reprod Immunol **80,** e13028

Liu, H., and Li, B. (2018). The functional role of exosome in hepatocellular carcinoma. J Cancer Res Clin Oncol **144,** 2085–2095.

Liu, L., Liao, J.Z., He, X.X., and Li, P.Y. (2017). The role of autophagy in hepatocellular carcinoma: friend or foe. Oncotarget **8,** 57707–57722.

Lu, X., Wo, G., Li, B., Xu, C., Wu, J., Jiang, C., *et al.* (2018). The anti-inflammatory NHE-06 restores antitumor immunity by targeting NF-kappaB/IL-6/STAT3 signaling in hepatocellular carcinoma. Biomed Pharmacother **102,** 420–427.

Nakamura, M., Chiba, T., Kanayama, K., Kanzaki, H., Saito, T., Kusakabe, Y., *et al.* (2019). Epigenetic dysregulation in hepatocellular carcinoma: an up-to-date review. Hepatol Res 49, 3–13.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods **9,** 471–472.

Masaki, T., Shiratori, Y., Rengifo, W., Igarashi, K., Yamagata, M., Kurokohchi, K., *et al.* (2003). Cyclins and cyclin-dependent kinases: comparative study of hepatocellular carcinoma versus cirrhosis. Hepatology **37,** 534–543.

Matsumoto, A., Takeishi, S., and Nakayama, K.I. (2014). p57 regulates T-cell development and prevents lymphomagenesis by balancing p53 activity and pre-TCR signaling. Blood **123,** 3429–3439.

Meijer, R.J., and Goeman, J.J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. Biom J **55,** 141–155.

Padmakumar, V.C., Aleem, E., Berthet, C., Hilton, M.B., and Kaldis, P. (2009). Cdk2 and Cdk4 activities are dispensable for tumorigenesis caused by the loss of p53. Mol Cell Biol **29,** 2582–2593.

Pal, S., Datta, K., and Mukhopadhyay, D. (2001). Central role of p53 on regulation of vascular permeability factor/vascular endothelial growth factor (VPF/VEGF) expression in mammary carcinoma. Cancer Res **61,** 6952–6957.

Pundir, S., Martin, M.J., and O'Donovan, C. (2016). UniProt tools. Curr Protoc Bioinformatics **53,** 1.29.1–1.29.15.

Satoh, S., Daigo, Y., Furukawa, Y., Kato, T., Miwa, N., Nishiwaki, T., *et al.* (2000). AXIN1 mutations in hepatocellular carcinomas, and growth suppression in cancer cells by virus-mediated transfer of AXIN1. Nat Genet **24,** 245–250.

Sayers, T.J. (2011). Targeting the extrinsic apoptosis signaling pathway for cancer therapy. Cancer Immunol Immunother **60,** 1173–1180.

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.

Shen, W., Chen, C., Guan, Y., Song, X., Jin, Y., Wang, J., *et al.* (2017). A pumpkin polysaccharide induces apoptosis by inhibiting the JAK2/STAT3 pathway in human hepatoma HepG2 cells. Int J Biol Macromol **104,** 681–686.

Soto, J.L., Cabrera, C.M., Serrano, S., and Lopez-Nevot, M.A. (2005). Mutation analysis of genes that control the G1/S cell cycle in melanoma: TP53, CDKN1A, CDKN2A, and CDKN2B. BMC Cancer **5,** 36.

Stefanska, B., Suderman, M., Machnes, Z., Bhattacharyya, B., Hallett, M., and Szyf, M. (2013) Transcription onset of genes critical in liver carcinogenesis is epigenetically regulated by methylated DNA-binding protein MBD2. Carcinogenesis **34,** 2738–2749.

Sui, J., Miao, Y., Han, J., Nan, H., Shen, B., Zhang, X., *et al.* (2018). Systematic analyses of a novel lncRNA-associated signature as the prognostic biomarker for Hepatocellular Carcinoma. Cancer Med [Epub ahead of print]; DOI: 10.1002/cam4.1541.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., *et al.* (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res **45,** D362–D368.

Tao, X., Wang, N., and Qin, W. (2015). Gut microbiota and hepatocellular carcinoma. Gastrointest Tumors **2,** 33–40.

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.

VanderSluis, B., Costanzo, M., Billmann, M., Ward, H.N., Myers, C.L., Andrews, B.J., *et al.* (2018). Integrating genetic and protein-protein interaction networks maps a functional wiring diagram of a cell. Curr Opin Microbiol **45,** 170–179.

Vogelstein, B., and Kinzler, K.W. (2004). Cancer genes and the pathways they control. Nat Med **10,** 789–799.

Wang, E. (2013). Understanding genomic alterations in cancer genomes using an integrative network approach. Cancer Lett **340,** 261–269.

Wang, S., Song, Y., Yan, F., and Liu, D. (2016). Mechanisms of resistance to third-generation EGFR tyrosine kinase inhibitors. Front Med **10,** 383–388.

Wang, T., Qin, Z.Y., Wen, L.Z., Guo, Y., Liu, Q., Lei, Z.J., *et al.* (2018). Epigenetic restriction of Hippo signaling by MORC2 underlies stemness of hepatocellular carcinoma cells. Cell Death Differ **25,** 2086–2100.

Wei, Y.F., Cui, G.Y., Ye, P., Chen, J.N., and Diao, H.Y. (2013). MicroRNAs may solve the mystery of chronic hepatitis B virus infection. World J Gastroenterol **19,** 4867–4876.

Wu, X., Liu, D., Tao, D., Xiang, W., Xiao, X., Wang, M., *et al.* (2016). BRD4 regulates EZH2 transcription through upregulation of C-MYC and represents a novel therapeutic target in bladder cancer. Mol Cancer Ther **15,** 1029–1042.

Yang, F., Deng, R., Qian, X.J., Chang, S.H., Wu, X.Q., Qin, J., *et al.* (2014). Feedback loops blockade potentiates apoptosis induction and antitumor activity of a novel AKT inhibitor DC120 in human liver cancer. Cell Death Dis **5,** e1114.

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.

Yu, H., Pardoll, D., and Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer **9,** 798–809.

Zhang, Y. (2015). Detection of epigenetic aberrations in the development of hepatocellular carcinoma. Methods Mol Biol **1238,** 709–731.

Address correspondence to:
Hongyan Diao, PhD
State Key Laboratory for Diagnosis and Treatment of Infectious Diseases
National Clinical Research Center for Infectious Disease
Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases
The First Affiliated Hospital
School of Medicine
Zhejiang University
Hangzhou 310003
China

E-mail: diaohy@zju.edu.cn

Received for publication June 1, 2019; received in revised form July 26, 2019; accepted August 5, 2019.