Identification of key genes in hepatitis B associated hepatocellular carcinoma based on WGCNA

Chang Liu[1*†] , Qinghai Dai[2†], Qian Ding[1], Min Wei[1,2*] and Xiaohong Kong[1*]

Abstract

Chronic Infection of Hepatitis B virus (HBV) is one risk factor of hepatocellular carcinoma (HCC). Much effort has been made to research the process of HBV-associated HCC, but its molecular mechanisms of carcinogenesis remain vague. Here, weighted gene co-expression network analysis (WGCNA) was employed to explore the co-expressed modules and hub/key genes correlated to HBV-associated HCC. We found that genes of the most significant module related to HBV-associated HCC were enriched in DNA replication, p53 signaling pathway, cell cycle, and HTLV-1 infection associated pathway; these cellular pathways played critical roles in the initiation and development of HCC or viral infections. Furthermore, seven hub/key genes were identified based on the topological network analysis, and their roles in HCC were verified by expression and Kaplan-Meier survival analysis. Protein-protein interaction and KEGG pathway analysis suggested that these key genes may stimulate cellular proliferation to promote the HCC progression. This study provides new perspectives to the knowledge of the key pathways and genes in the carcinogenesis process of HBV-associated HCC, and our findings provided potential therapeutic targets and clues of the carcinogenesis of HBV-associated HCC.

Keywords: HBV, HCC, WGCNA, Mitosis, Cell cycle, K-M survival

Introduction

Hepatocellular carcinoma (HCC), which is also known as primary liver cancer, is one of the common malignant tumors, and the 2nd leading cause of cancer-related death worldwide [1, 2]. The risk factors of HCC are chronic infections of hepatitis viruses, diet polluted with aflatoxin, obesity, type two diabetes, use and abuse of alcohol and tobacco [3–5]. Among the hepatitis viruses associated with HCC, hepatitis B virus (HBV) is responsible for about 80% of virus-related HCC cases, particularly in East Asia and Africa [6]. HBV causes acute and chronic liver infections, which can then lead to liver cirrhosis and HCC. The

carcinogenesis of HBV-associated HCC is a complex process, which can be summarized into the following stages: the host inflammatory reaction against HBV, interaction with endogenous mutagens, integration of viral genome DNA into cellular DNA genome, and alternating gene expression in multiple ways [7–9]. Studies of the carcinogenesis molecular process and comparisons of the difference of genetic expression between HBV-associated HCC tissues and adjacent normal tissues, will help us find interesting and important clues of the carcinogenesis, and may provide potential therapeutic targets of HBV-associated HCC. Up to date genomic technologies such as gene-array or next-generation sequence facilitate the detection of the whole genome expression.

High-throughput hybridization array- and sequencing-based experiments generate vast amount of data on the molecular abundance of RNAs (mRNAs, miRNAs,

lncRNAs, cirRNAs), genomic DNAs, and proteins in absolute or relative terms. Public databases such as GEO and ArrayExpress store these data and provide information for the research community to reuse [10–12], and many data analysis tools and methods have been developed to reanalyze or reuse these data in order to get some new interesting results. Among these methods, weighted gene co-expression network analysis (WGCNA) is a powerful systems biological method to find co-expressed modules and hub/key genes in transcriptomics, proteomic and metabolomic studies [13–15]. Gene co-expression network analysis enables us to systematically analyze large, high-dimensional data sets. WGCNA groups genes into a module/network according to pairwise correlations between genes of their similar expression profile; furthermore, these models may correlate to some special clinical traits of interest, such as tumor stages, ages, gender and other biological characteristics or traits that we are interested in.

The objective of the present study is to explore novel genes or pathways relating HBV-related HCC through the data mining of public databases. In this study, we reanalyzed a GEO data set (GSE121248) of the HBV induced HCC and adjacent normal tissues, and then constructed a gene co-expression network based on WGCNA and then identified 21 modules based on the gene expression data sets. According to the results of WGCNA, there were six modules significantly correlated to the tumor trait in our study; especially the turquoise module, which is distinguished in WGCNA, is the most significant module correlated to the tumor trait. The co-expression network of these genes in the turquoise module was analyzed by Cytoscape network topological analysis tool "cytoHubba" plugin to get the hub or key genes in the network. Finally, seven hub/key genes were found correlating to HBV-associated HCC tumor trait; based on TCGA database tools, these key genes' expression levels showed significant differences between people without HCC and HCC patients, and the expression levels of these hub/key genes also influence the survival of HCC patients significantly. With the help of WGCNA and net topological algorithms, several new key genes correlating to HBV associated HCC were found, while these genes have not been paid close attention to in the original paper that the GEO data set came from; we also predicted the functions of these genes and hope to provide some useful information to interested researchers. These genes or other cellular factors associated to these key genes could be the biomarkers and potential therapeutic targets of HBV-associated HCC.

Material and methods
Data processing
The gene expression data set GSE121248 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121248), which is provided by Hui KM [16], was downloaded from the Gene Expression Omnibus (GEO) database [17]. Briefly, 107 tissue samples included 70 chronic hepatitis B induced HCC and 37 adjacent normal tissues. All the tissues were obtained from patients who underwent partial hepatectomy as a treatment for HCC. To assess recurrence, all treated HCC patients were monitored by routine clinical follow up once every 3 months. The detailed information of the study population was provided in supplementary files of Hui KM's original article [16] and our supplementary files. The RNAs were extracted from all tissue samples, and the reversed transcription DNAs were hybridized on the human U133 plus 2.0 arrays (Affymetrix, Santa Clara, CA, USA). Affyand-related R packages were used to process all the raw data. Robust Multi-array Average approach was used to normalize the background of raw data according to the package instruction. The expression set was processed through the *nsFilter* function to filter features exhibiting little variation or a consistently low signal across samples [18].

Weighted gene co-expression network analysis (WGCNA)
The freely accessible R package WGCNA (v 1.66) was taken to co-expression analysis [15]. According to the instruction, one-step network construction and module detection was taken; 15,139 annotated genes were used to construct the network. According to the software instruction, the module eigengene expression, adjacency matrix heatmap, Module-Trait relationships, and other related parameters/results were calculated and visualized. Briefly, the network type is chosen as an unsigned network, and the correlation method is Pearson correlation.

GO term and KEGG pathway enrichment analysis
Based on Bioconductor packages "clusterProfiler", GO term enrichment analysis including biological process, cellular component and molecular function, was used to explore the biologic significance of selected module genes. With the same Bioconductor packages "clusterProfiler", we also performed the KEGG pathway enrichment analysis of the selected module genes [19].

Key genes identification of the selected module
To identify hub/key genes of the selected module, the Cytoscape software (3.7.1) was utilized to construct of the network of the module genes. The important nodes (key genes) were predicted and explored by "cytoHubba" plugin [20]. The topological algorithms of "cytoHubba" consist of Degree, Edge Percolated Component (EPC), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Maximal Clique Centrality (MCC) and centralities based on shortest paths, such as Bottleneck (BN), EcCentricity, Closeness, Radiality, Betweenness, and Stress, were

applied to get respective top 20 ranked genes set. Then, the intersection of nine top 20 ranked genes sets were identified as the key genes.

Expression on box plots

The website GEPIA (http://gepia.cancer-pku.cn/index.html) was taken to box plot the expression of the key genes [21]. According to the web tutorial, the liver hepatocellular carcinoma (LIHC) dataset was chosen; $\log_2(TPM + 1)$ was used as log scale; jitter size was 0.4; the normal data for control was match GTEx data and TCGA normal.

Kaplan-Meier survival analysis

KM-plotter (http://kmplot.com/analysis/) was employed to perform the survival analysis of the combination and respective key genes [22]. Briefly, Liver Cancer RNA-seq data set was selected; patient groups were split by median expression of the gene (auto select best cutoff); overall survival was applied.

Prediction of the functions of key genes

The GeneMANIA Cytoscape plugin was used to predict the functions of key genes [23, 24]. The identified key genes were inputted as a query gene set. Based on *Homo sapiens* database of GeneMANIA, a network of query genes and result genes was constructed and visualized by Cytoscape. The distinctive relationships between the genes, including co-expression, co-localization, genetic interaction, pathway, physical interaction, shared protein domains, and predicted relations, were indicated by distinct color lines. In order to show the KEGG pathways of the key/hub genes, an R/Bioconductor package named Pathview was applied to visualize the concerned KEGG pathways [25].

Results

Data processing

GSE121248 raw files (.CEL format), which contains a total of 107 tissue sample were downloaded from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE121nnn/GSE121248/). The files were then transferred to expression matrix using the RMA algorithm based on R language, including background correction, normalization and summarization. (Supplementary Figure 1: (A) Box plot of relative log expression (RLE) and (B) Box plot of normalized unscaled standard errors (NUSE)). After annotation and *nsFilter* processing, there were 15,139 genes from all 54,675 probes for further WGCNA analysis. In order to give an outline of our study design, a workflow is shown in Fig. 1.

Weighted gene co-expression network identification of modules construction
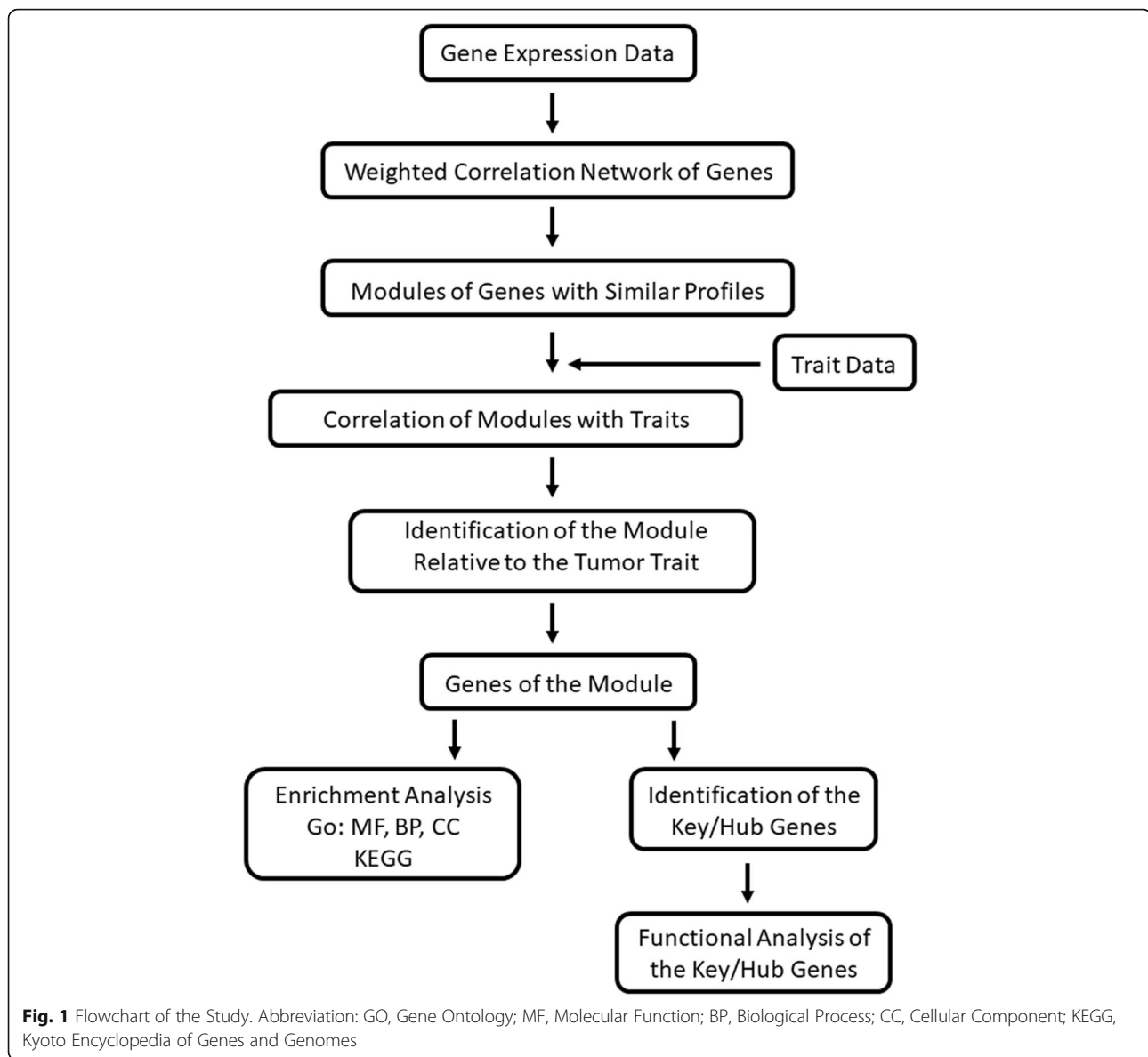
The co-expression network was constructed from the filtered annotated genes, in which 21 modules were identified. Before the net construction, the samples were clustered to see if there are any obvious outliers, and it appears that there was no outlier (see supplementary Figure 2). As shown in Fig. 2(a) and (b), the soft threshold power 5 was chosen to define the adjacency matrix based on the criterion of approximate scale-free topology, and the minimum module size was 21. The modules with different colors were shown in Fig. 2(c); the module grey (MEgrey) is reserved for genes outside of all modules. To show the co-expression relationship between the genes on genome level, 400 genes were randomly selected to plot the network heatmap as shown in supplementary Figure 3.

Correlation between each module and clinical traits

To figure out the interactions among these 20 co-expressed modules (except MEgrey, which is reserved for genes outside of all modules by WGCNA), we analyzed the connectivity of eigengenes. As shown in Fig. 3(a), a cluster analysis was performed; 20 modules were classified into two clusters, and each cluster contains two main branches. Figure 3(a) and (c) showed a significant difference among the 21 modules. There are multiple modules related to the clinical trait between tumor samples and adjacent normal samples. The module-trait heatmap represents the correlations of the module eigengenes with traits. When that correlation is positive, it means the eigengene increases with increasing trait. Generally, in a signed network, where all genes in a module are positively correlated with the eigengene, it means that all genes in the module should follow the same pattern of increasing expression with increasing trait values; on the contrary, in an unsigned network, which is actually the case in our study, there are also some genes that have the opposite behavior compared with the eigengene. It means that we do not know for sure if the expression of genes in the modules actually increases or decreases, but we know the expression definitely changes. As shown in Fig. 3, the midnight blue, magenta, turquoise, royal blue modules were positively related to the tumor trait, especially the turquoise module (MEturquoise) was the most significantly relative module to the tumor trait (correlations 0.8, *p* value 5×10^{-25}). On the other hand, the blue, tan, yellow modules were negatively related to the tumor trait. There was no module significantly related to the gender trait, and it is releasable. Interestingly, the salmon module (MEsalmon) was slightly related to the age trait.

Functional enrichment and pathway analysis of the key module genes

Because MEturquoise is the most relevant module related to the tumor trait, we did the further functional

Fig. 1 Flowchart of the Study. Abbreviation: GO, Gene Ontology; MF, Molecular Function; BP, Biological Process; CC, Cellular Component; KEGG, Kyoto Encyclopedia of Genes and Genomes

assay of the genes in turquoise module. As shown in Fig. 4(a), the higher module membership genes in MEturquoise are the more significant genes for the tumor trait; they have a strong positive correlated relationship. Then, we did GO term enrichment analysis of MEturquoise genes. As to molecular functions in Fig. 4(b), the MEturquoise genes considerably enriched in ATPase activity, catalytic activity (acting on RNA or DNA), helicase activity, macromolecule binding (such as tubulin, ribonucleoprotein complex, single-stranded DNA, damaged DNA). For biological process as shown in Fig. 4(c), the MEturquoise genes significantly enriched in organelle fission, nuclear division, chromosome functions (such as segregation, organization). With regard to cellular component in Fig. 4(d), the MEturquoise genes significantly enriched in a chromosomal and microtubule region, including normal or condensed chromosome, centromere, kinetochore, telomere. In order to get the pathway involved MEturquoise genes, KEGG pathway analysis was taken. As shown in Fig. 4(e), several significant enriched pathways were found, including HTLV1 infection, RNA transport, cell cycle, spliceosome, pyrimidine metabolism, p53 signaling pathway, DNA replication and so on.

Network analysis of MEturquoise genes
In order to identify the hub genes or key genes of the module turquoise, the MEturquoise network file was imported into the Cytoscape. The module net, which has 3,600,506 edges, is too huge to be analyzed on a personal computer. The net was firstly cut off by edge weight (more than 0.1) into 92,188 edges, and then the
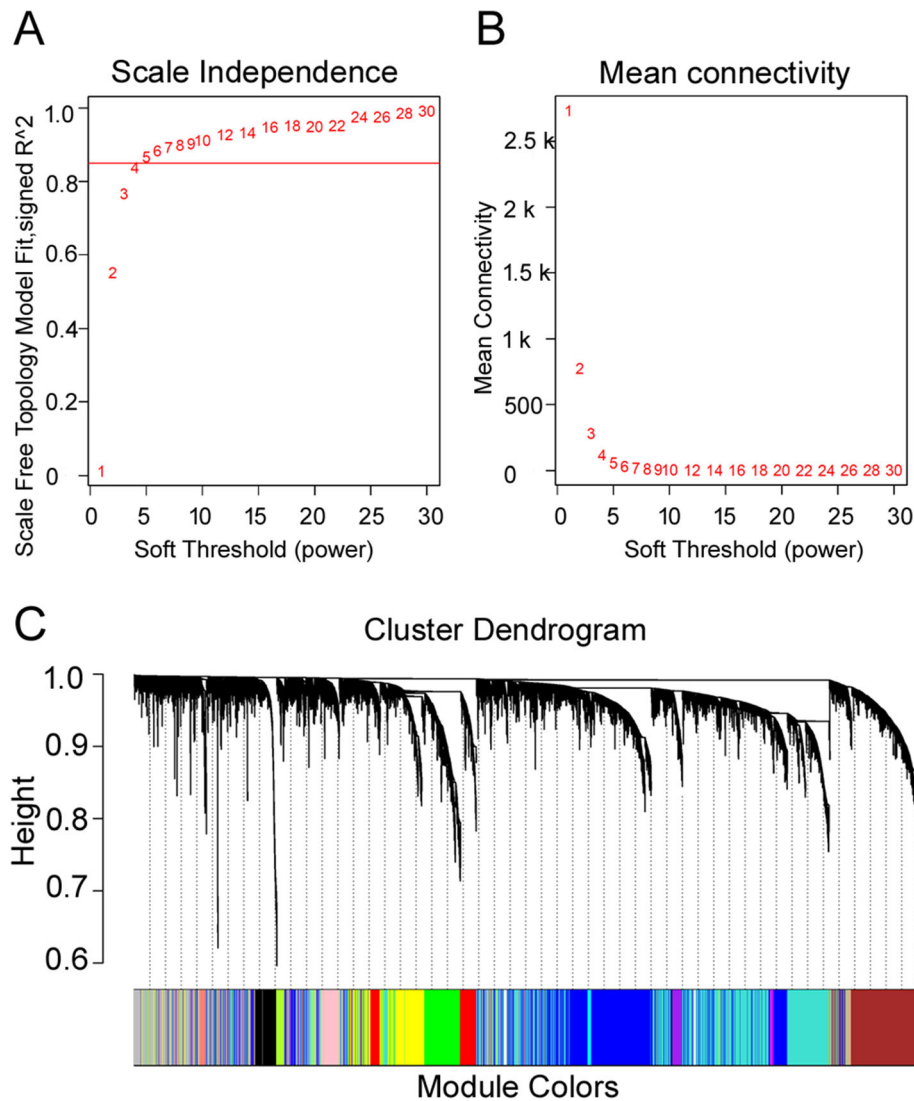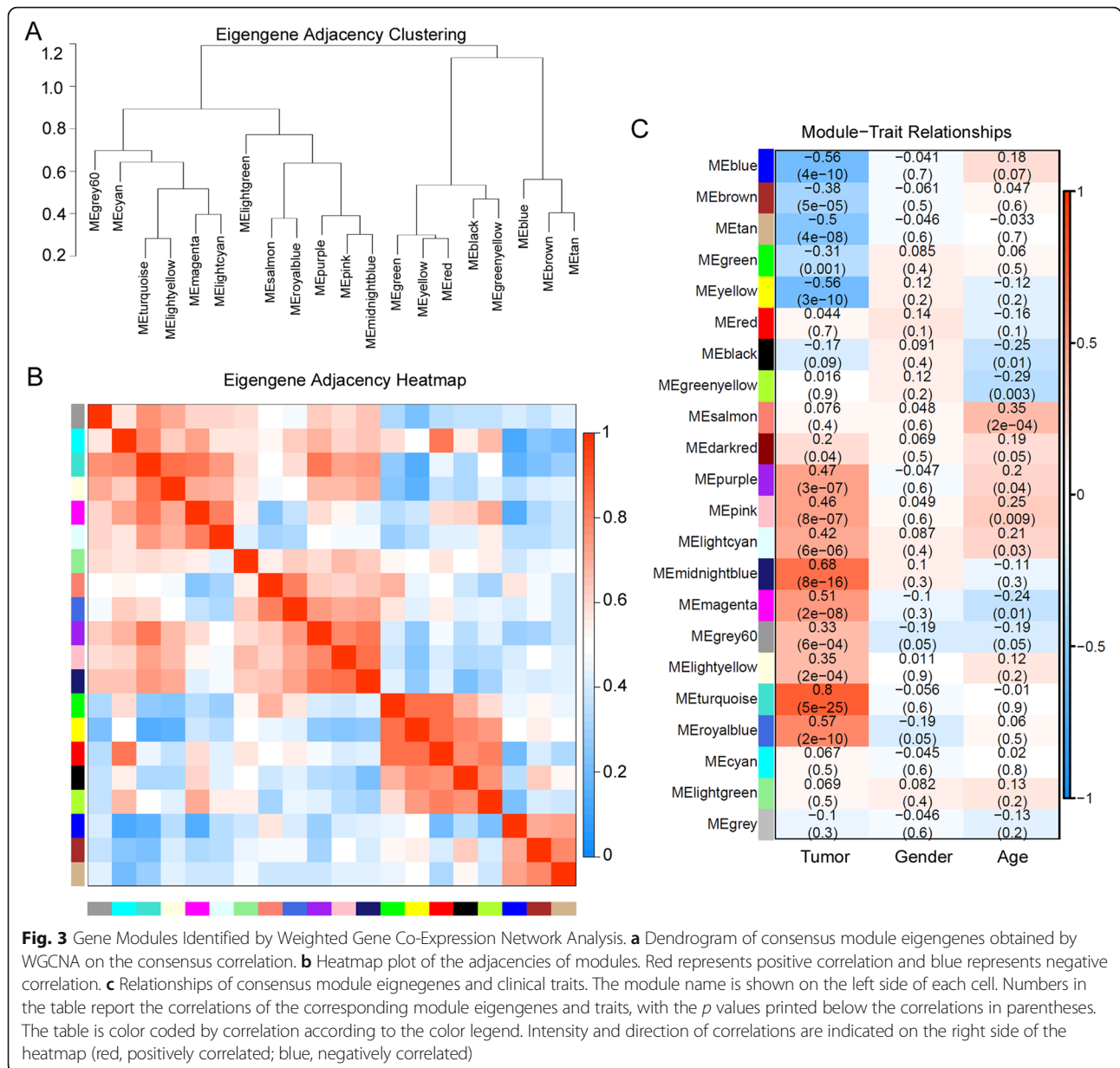
Fig. 2 Construction of Weighted Gene Co-Expression Network Identification of Modules. **a** Scale independence of network topology for different soft-thresholding powers. The approximate scale-free topology can be attained at the soft-thresholding power of 5. **b** Mean connectivity of network topology for different soft-thresholding powers. Numbers in the plots indicate the corresponding soft thresholding powers. **c** DEGs clustering and module screening based on gene expression pattern. The top was gene dendrogram and the bottom was genes' modules with different colors. A total of 21 modules were identified

cutoff net was analyzed by "cytoHubba" Cytoscape plugin. There are 12 topologic algorithms available in cytoHubba plugin. The top ranking 20 genes' sets of each topologic algorithm were all obtained, and then the intersections of all the sets were taken. There were three sets getting none or too few genes with other sets, so they were discarded. The sub networks of the top 20 genes of other nine topological algorithms were shown in Fig. 5. The intersections of all nine genes' sets were these seven genes: *CCNB1, GINS1, PRC1, KIF20A, NUSAP1, NEK2, BUB1B*. These seven genes were considered as hub genes or key genes involved in HBV associated HCC in our study, and Table 1 showed the detail

information of these genes. The annotation of these 7 key genes was given by GEPIA website [21].

Roles of the key genes in the process of HHC

To analyze the functions of key genes, publicly available data and tools from TCGA and GTEx databases were applied to analyze whether these expression levels of 7 key genes were substantially different between healthy people and HCC patients and whether the expression levels of 7 key genes may influence the survival of HCC patient. As showing in Fig. 6(a), the expression levels of all seven key genes in 369 HCC tumor tissues are significantly higher than the respective gene expression level

Fig. 3 Gene Modules Identified by Weighted Gene Co-Expression Network Analysis. **a** Dendrogram of consensus module eigengenes obtained by WGCNA on the consensus correlation. **b** Heatmap plot of the adjacencies of modules. Red represents positive correlation and blue represents negative correlation. **c** Relationships of consensus module eigngenes and clinical traits. The module name is shown on the left side of each cell. Numbers in the table report the correlations of the corresponding module eigengenes and traits, with the *p* values printed below the correlations in parentheses. The table is color coded by correlation according to the color legend. Intensity and direction of correlations are indicated on the right side of the heatmap (red, positively correlated; blue, negatively correlated)
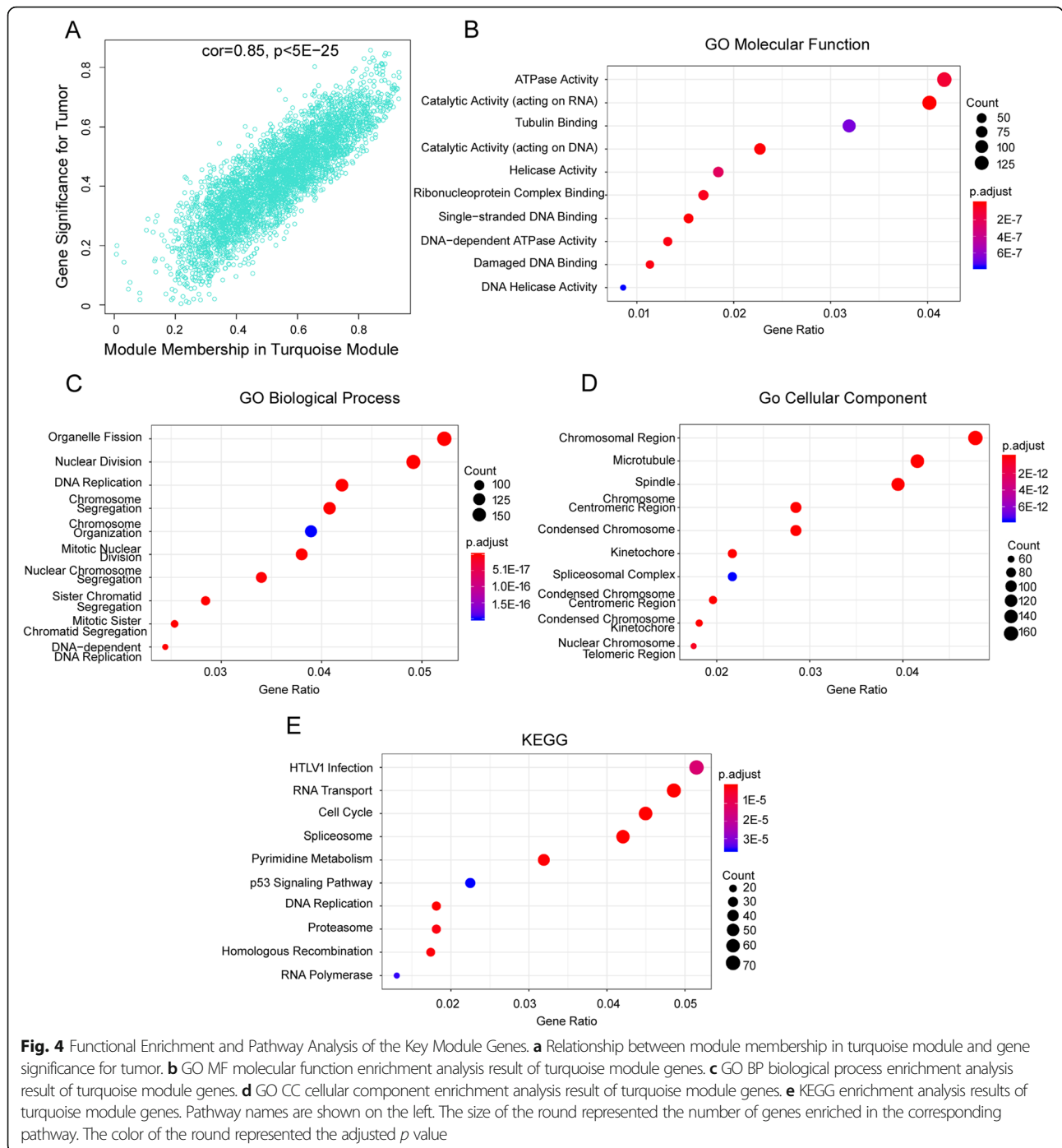
in 160 normal tissues. As shown in Fig. 6(b), the patients who have higher levels of expression of the key genes show shorter overall survival periods compared with lower expression patients.

Functional prediction of the key genes involved in HBV associated HCC

To predicting the functions of 7 key genes, GeneMANIA plugin (version 3.5.0) was applied to find the result genes and construct the PPI net. As showing in Fig. 7(a), 20 result genes were found to relate to the key genes/query genes; they are *CENPF, KIF23, CCNF, CCNA2, CENPE, NDC80,*
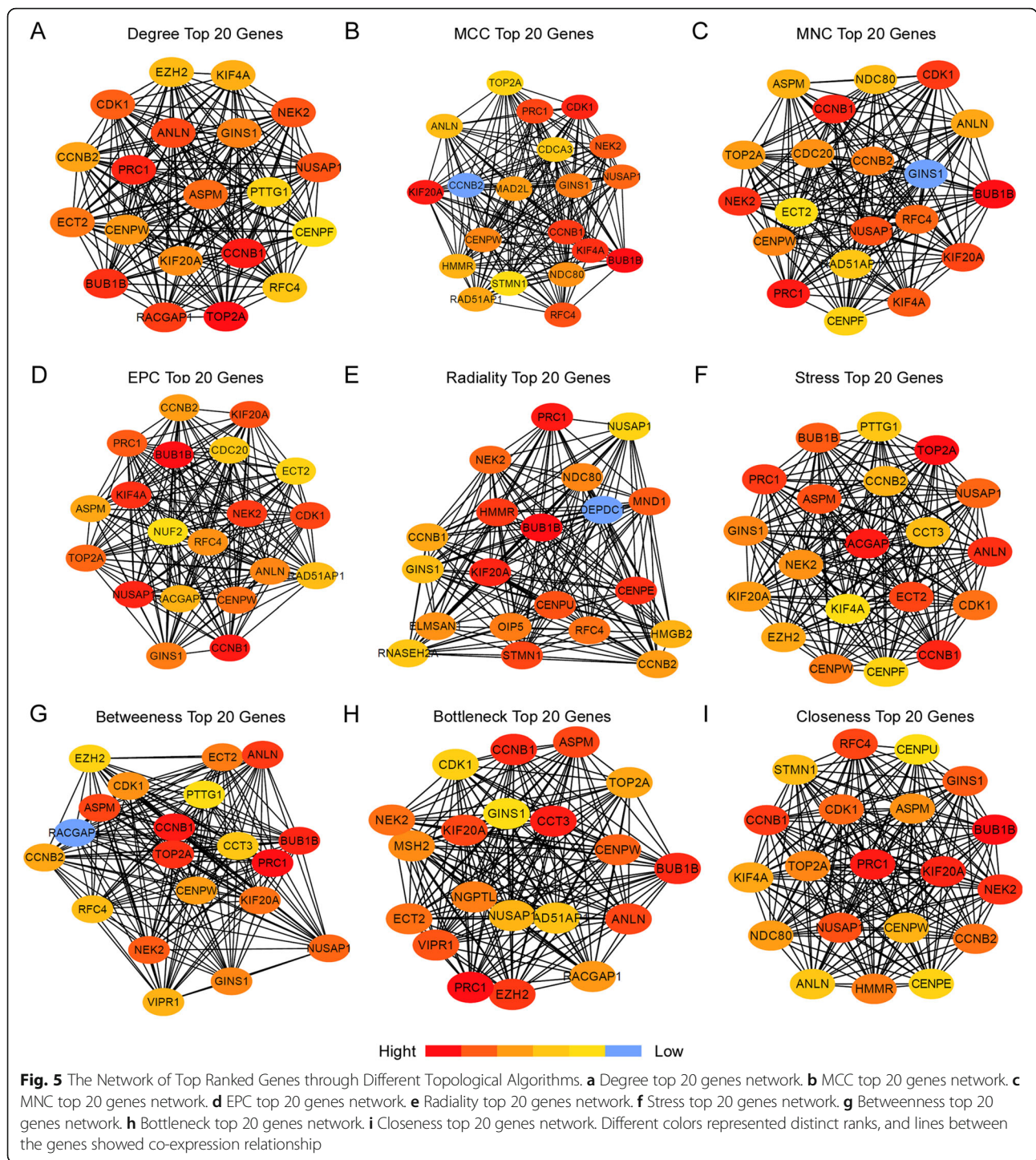
MKI67, AURKA, TOP2A, AURKB, KIF11, CDK1, CDCA3, HMMR, ZWINT, KIF4A, DEPDC1, CDC25C, SMC4, ASPM. Most of the result genes are centromere proteins or involved in cell cycle processing. GeneMANIA predicts seven different relationships between genes/proteins based on published papers, including co-expression, physical interactions, co-localization, predicted relations, shared protein domains, genetic interactions and pathway. These relationships were indicted as distinct colors as shown in Fig. 7(a). From the Fig. 7(a), the co-expression is the main relationships among key genes and result genes. The result is intelligible, because all the key genes are come from the same co-expression

Fig. 4 Functional Enrichment and Pathway Analysis of the Key Module Genes. **a** Relationship between module membership in turquoise module and gene significance for tumor. **b** GO MF molecular function enrichment analysis result of turquoise module genes. **c** GO BP biological process enrichment analysis result of turquoise module genes. **d** GO CC cellular component enrichment analysis result of turquoise module genes. **e** KEGG enrichment analysis results of turquoise module genes. Pathway names are shown on the left. The size of the round represented the number of genes enriched in the corresponding pathway. The color of the round represented the adjusted *p* value

module; besides the co-expression, there are other important relations predicted for us, such as co-localization, physical interactions, predicted, shared protein domains, genetic interactions; you can get more detailed relations in the supplementary file (supplementary Tables 1, 2, 3). Interestingly, the predicted relation was based on Stein's work about protein interaction network on cancer data analysis, so it indicated these genes played critical roles in a carcinogenesis process [26].

Key pathway analysis of the genes involved in HBV associated HCC

Based on the results of key genes' PPI net as shown in Fig. 7(a), seven key/hub genes and 20 results genes were used to search the key KEGG pathways; as the result of Fig. 7(b), the CELL CYCLE (hsa04110) pathway was regarded as the key pathway in our study. The genes *BUB1B, CCNB1, CDK1, CDC25C* and *CCNA2*, which were also in PPI net, take essential places in the CELL

Fig. 5 The Network of Top Ranked Genes through Different Topological Algorithms. **a** Degree top 20 genes network. **b** MCC top 20 genes network. **c** MNC top 20 genes network. **d** EPC top 20 genes network. **e** Radiality top 20 genes network. **f** Stress top 20 genes network. **g** Betweenness top 20 genes network. **h** Bottleneck top 20 genes network. **i** Closeness top 20 genes network. Different colors represented distinct ranks, and lines between the genes showed co-expression relationship

CYCLE (hsa04110) pathway. This result implicated that HBV might influence the cell cycle pathway through those above-mentioned genes, and eventually effect the process of HCC. Two genes in PPI net, namely *CDK1* and *CycA* (*CCNA2*), were also found in the VIRAL CARCINOGENESIS (hsa05203) KEGG pathways. As we can see in Fig. 7(c), non-structural proteins of HCV, EBV or HPV can stimulate the cellular proliferation through two above genes. The genes in our PPI net were not found in the VIRAL CARCINOGENESIS pathway of HBV, but our results implicated that HBV might promote proliferation of HCC tumor cells in a similar way.

Table 1 Key Genes Involved in HBV Associated HCC

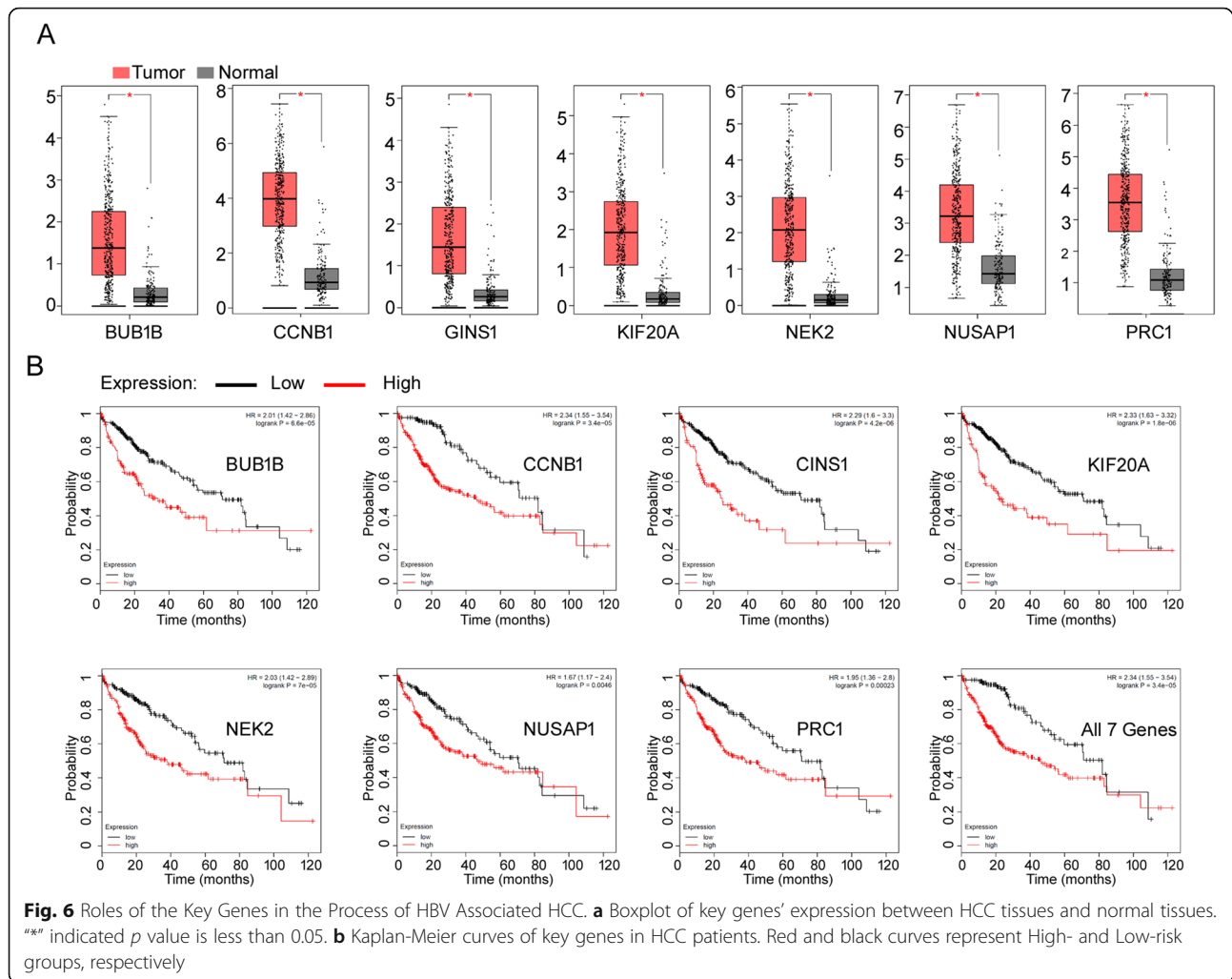| Gene	Ensembl ID	Description	Alias	Summary
CCNB1	ENSG00000134057.14	cyclin B1	CCNB	The protein encoded by this gene is a regulatory protein involved in mitosis. The gene product complexes with p34(cdc2) to form the maturation-promoting factor (MPF). Two alternative transcripts have been found, a constitutively expressed transcript and a cell cycle-regulated transcript, that is expressed predominantly during G2/M phase. The different transcripts result from the use of alternate transcription initiation sites.
GINS1	ENSG00000101003.9	GINS complex subunit 1	PSF1	The yeast heterotetrameric GINS complex is made up of Sld5 (GINS4; MIM 610611), Psf1, Psf2 (GINS2; MIM 610609), and Psf3 (GINS3; MIM 610610). The formation of the GINS complex is essential for the initiation of DNA replication in yeast and Xenopus egg extracts.
PRC1	ENSG00000198901.13	protein regulator of cytokinesis 1	ASE1	This gene encodes a protein that is involved in cytokinesis. The protein is present at high levels during the S and G2/M phases of mitosis but its levels drop dramatically when the cell exits mitosis and enters mitosis G1 phase. It is located in the nucleus during interphase, becomes associated with mitotic spindles in a highly dynamic manner during mitosis, and localizes to the cell midbody during cytokinesis. This protein has been shown to be a substrate of several cyclin-dependent kinases (CDKs). It is necessary for polarizing parallel microtubules and concentrating the factors responsible for contractile ring assembly. Alternative splicing results in multiple transcript variants.
KIF20A	ENSG00000112984.11	kinesin family member 20A	MKLP2, RAB6KIFL	–
NUSAP1	ENSG00000137804.12	nucleolar and spindle associated protein 1	ANKT, BM037, LNP, NUSAP, PRO0310p1, Q0310, SAPL	NUSAP1 is a nucleolar-spindle-associated protein that plays a role in spindle microtubule organization.
NEK2	ENSG00000117650.12	–	–	–
BUB1B	ENSG00000156970.12	BUB1 mitotic checkpoint serine/threonine kinase B	BUB1beta, BUBR1, Bub1A, MAD3L, MVA1, SSK1, hBUBR1	This gene encodes a kinase involved in spindle checkpoint function. The protein has been localized to the kinetochore and plays a role in the inhibition of the anaphase-promoting complex/cyclosome (APC/C), delaying the onset of anaphase and ensuring proper chromosome segregation. Impaired spindle checkpoint function has been found in many forms of cancer.

Discussion

In our studies, WGCNA was employed to explore gene expression alternation between HBV associated HCC tissues and adjacent normal tissues. There are many advantages of WGCNA over other traditional methods for differential expression analysis, because it focuses on co-expression patterns and the functional relevant modules that consist of related genes will be discovered. Key/hub genes in the modules related to some specific traits may serve as clinical detective biomarkers or therapeutic targets [27–29].

HCC is a common consequence of HBV chronic infection. HBV's prolonged infection can change the expression of genes of hepatic cells in different ways. HBV-DNA integrations randomly distributed among chromosomes into host chromosomes, and the integrations of HBV will alternate the expression of cellular genes near the integrating sites [30]. Noncoding RNAs, such as miRNAs, lncRNAs and circRNAs, also take parts in the pathogenesis of HBV-associated HCC [31–33]. Viral protein HBx plays important roles in hepatocarcinogenesis by interfering with telomerase activity [34], affecting hepatocellular apoptosis [35, 36], and up-regulating the transcriptional activation of human telomerase transcriptase [37]. Interesting, our GO cellular component enrichment results of genes of module turquoise, which is the co-expression module most positively related to the tumor trait, showed that some genes of this module were enriched in the telomeric region.

In our study, modules changed significantly between HBV associated HCC tissues and normal adjacent tissues consisted of the midnight blue, magenta, turquoise, royal blue modules which were up-regulated in HCC, whereas the blue, tan, yellow modules which were down-regulated in HCC. These up-regulated modules and down-regulated modules mentioned above were classified into two main groups through eigengene adjacency clustering as shown in Fig. 2(a). Among these modules, the turquoise module is the most significant module related to the tumor trait. According to the KEGG results, the turquoise module, enriched in DNA replication, p53 signaling pathway, cell cycle, especially

Fig. 6 Roles of the Key Genes in the Process of HBV Associated HCC. **a** Boxplot of key genes' expression between HCC tissues and normal tissues. "*" indicated *p* value is less than 0.05. **b** Kaplan-Meier curves of key genes in HCC patients. Red and black curves represent High- and Low-risk groups, respectively

HTLV-1 infection associated pathway, increased in HBV-associated HCC tissues. These important processes, including the acceleration of genome DNA replication, the misregulation of tumor suppressor p53, and the abnormal cell cycle-associated pathway, all played critical roles in the initiation and development of HCC. Furthermore, the enrichment of HTLV-1 infection associated pathway indicated that some signaling pathways associated with viral infections were also significantly activated in HBV-associated HCC tissues.

According to the topological network analysis from the turquoise model, seven key genes were identified playing critical roles in the network. They are *CCNB1, GINS1, PRC1, KIF20A, NUSAP1, NEK2,* and *BUB1B*. *CCNB1* is the gene, which encodes a regulatory protein involved in mitosis, and it is expressing predominantly during G2/M phase [38]. It was reported that *CDK1-CCNB1* enables MPS1 kinetochore localization to create a spindle checkpoint-permissive state [39].*GINS1* encodes subunit 1 of the GINS

complex, and the complex is essential for the initiation of DNA replication [40]. It has been reported that the high expression of *KIF20A* is associated with poor prognosis of glioma patients, and *KIF20A* can be a potential immunotherapeutic target for glioma [41, 42]. *PRC1* encodes a protein involved in the cytokinesis process, and the protein maintains a high level during the S and G2/M phases of mitosis [43]. *NUSAP1* encodes a nucleolar-spindle-associated protein playing a role in spindle microtubule organization [44]. *NEK2* has shown it is involved in some different cancers: *NEK2* promotes aerobic glycolysis in multiple myeloma [45]; targeting *NEK2* attenuates glioblastoma growth and radioresistance [46]; *NEK2* can be a prognostic biomarker of hepatocellular carcinoma [47]. *BUB1B* encodes a kinase playing a role in spindle checkpoint function. The kinase localizes to the kinetochore and plays a role in the inhibition of the anaphase-promoting complex/cyclosome (APC/C) [48]. It has been reported that individuals having

Fig. 7 (See legend on next page.)
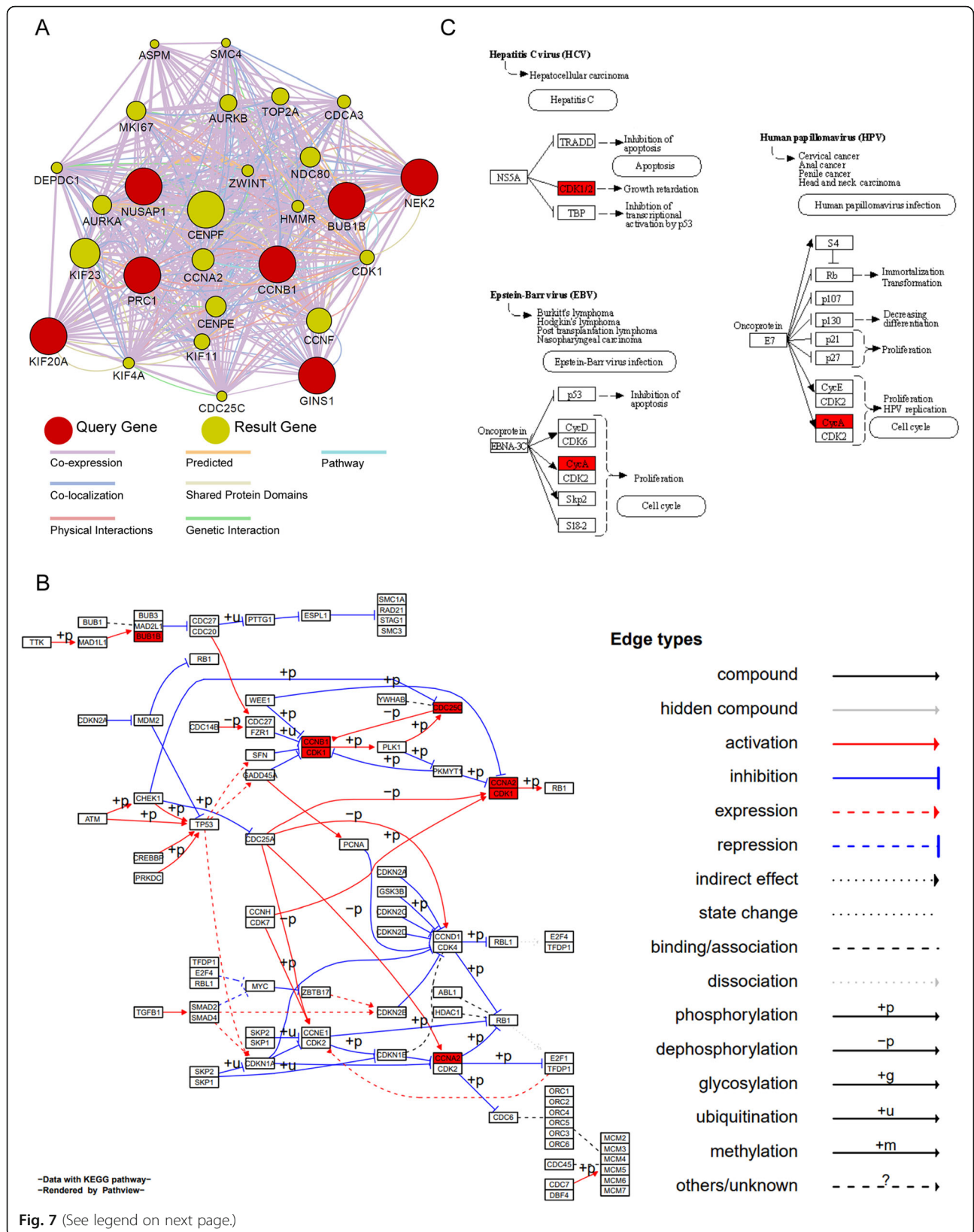
(See figure on previous page.)

Fig. 7 Functional Prediction of the Key Genes Involved in HBV Associated HCC. **a** Protein-Protein-Interaction (PPI) Network of the Key Genes Involved in HBV Associated HCC. GeneMANIA was applied to predict the function of key genes and construct the PPI network. The red rounds indicated key genes (query genes), and the khaki rounds indicated result genes. The different color lines between the rounds indicated distinct relationships between corresponding genes/proteins, such as co-expression, co-localization, physical interactions, predicted relations, shared protein domains, genetic interactions and pathway. **b** Visualization of the CELL CYCLE (hsa04110) KEGG pathway. The genes in the PPI network are highlighted in red. The different relations between proteins in the pathway are indicated in distinct types of lines. **c** The VIRAL CARC INOGENESIS pathways (hsa05203) related to the genes in PPI network. The genes in the PPI network are also highlighted in red

biallelic *TRIP13* or *BUB1B* mutations are prone to having embryonal tumors, and their cells display severe spindle assembly checkpoint (SAC) impairment [49]. In our expression and survival analyses, all 7 key/hub genes, which were identified in our selected module, can be the biomarkers and potential therapeutic targets of HBV-associated HCC.

According to genes'/proteins' multifunctional relations, we constructed a protein-protein interaction net consisting of 27 genes (our seven key/hub genes as query genes, and 20 result genes). For total 27 genes, we search KEGG for their pathway information. Finally, we focused on CELL CYCLE (hsa04110) pathway and VIRAL CARCINOGENESIS (hsa05203) KEGG pathways. Genes *BUB1B*, *CCNB1*, *CDK1*, *CDC25C* and *CCNA2* play important roles in cell cycle regulation; HCV, EBV and HPV can stimulate cell proliferation through *CDK1* or *CCNA2*. These genes were not found in HBV carcinogenesis pathways, but our results implied that there should be an unknown regulation of cell cycle in HBV-related HCC. Specially, the function of *BUB1B* and *CCNB1* in HBV-related HCC has attracted the attention of some researchers in recent years [50–52]; HBV may regulate these two genes to influence cell cycle progression promoting the development of HCC.

In summary, we provide a systematic biological interpretation of gene expression data derived from HBV associated HCC tissues and adjacent normal tissues. Based on WGCNA, there were 21 modules identified, and the turquoise module which was the most significant module relating to the tumor trait was selected to be analyzed in detail. Our results showed that the turquoise module, enriched in DNA replication, p53 signaling pathway, cell cycle, and HTLV-1 infection associated pathway, was activated in HBV-associated HCC tissues. Seven hub/key genes were identified; pathway analysis implicates that these key genes may stimulate cellular proliferation to promote the HBV-related HCC progression. All of our findings provide new perspectives to the understanding of pathways and genes underlying HBV-associated HCC, and experimental verification is needed to validate our predictions.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s13027-021-00357-4.

Additional file 1.

Additional file 2.

Authors' contributions
CL, MW and XHK conceived and designed the study. CL, QHD and QD performed the data analysis. CL and QHD wrote the paper. MW and XHK reviewed and edited the manuscript. The author(s) read and approved the final manuscript.

Availability of data and materials
All data, models, or codes generated or used during the study are available from the corresponding author (Chang Liu: changliu@nankai.edu.cn) by request.

Declarations

Ethics approval and consent to participate
The authors declare not applicable in this section.

Consent for publication
The authors declare that the work described has not been published before, and its publication has been approved by all co-authors.

Competing interests
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.

Author details
[1]School of Medicine, Nankai University, Tianjin, China. [2]Nankai University Second People's Hospital, Nankai University, Tianjin, China.

References
1. Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends--An update. Cancer Epidemiol Biomark Prev. 2016;25(1):16.

2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65(2):87.

3. Wallace MC, Preen D, Jeffrey GP, Adams LA. The evolving epidemiology of hepatocellular carcinoma: a global perspective. Expert Rev Gastroenterol Hepatol. 2015;9(6):765.

4. Yang W, Ma Y, Liu Y, Smith-Warner SA, Simon TG, Chong DQ, Qi Q, Meyerhardt JA, Giovannucci EL, Chan AT, Zhang X. Association of intake of whole grains and dietary fiber with risk of hepatocellular carcinoma in US adults. JAMA Oncol. 2019;5(6):879–86.

5. Mittal S, El-Serag HB, et al. J Clin Gastroenterol. 2013;47(Suppl):S2.

6. Petruzziello A. Epidemiology of hepatitis B virus (HBV) and hepatitis C virus (HCV) related hepatocellular carcinoma. Open Virol J. 2018;12:26.

7. Feitelson MA, Bonamassa B, Arzumanyan A. The roles of hepatitis B virus-encoded X protein in virus replication and the pathogenesis of chronic liver disease. Expert Opin Ther Targets. 2014;18(3):293.

8. Hsieh YH, Hsu JL, Su IJ, Huang W. Genomic instability caused by hepatitis B virus: into the hepatoma inferno. Front Biosci (Landmark Ed). 2011;16:2586.

9. Sitia G, Aiolfi R, Di Lucia P, Mainetti M, Fiocchi A, Mingozzi F, Esposito A, Ruggeri ZM, Chisari FV, Iannacone M, Guidotti LG. Antiplatelet therapy prevents hepatocellular carcinoma and improves survival in a mouse model of chronic hepatitis B. Proc Natl Acad Sci U S A. 2012; 109(32):E2165.

10. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991.

11. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007;35(Database issue):D747.

12. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. ArrayExpress update--simplifying data submissions. Nucleic Acids Res. 2015;43(Database issue):D1113.

13. Pei G, Chen L, Zhang W. WGCNA application to proteomic and metabolomic data analysis. Methods Enzymol. 2017;585:135.

14. Luo Y, Coskun V, Liang A, Yu J, Cheng L, Ge W, Shi Z, Zhang K, Li C, Cui Y, Lin H, Luo D, Wang J, Lin C, Dai Z, Zhu H, Zhang J, Liu J, Liu H, deVellis J, Horvath S, Sun YE, Li S. Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. Cell. 2015;161(5):1175.

15. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

16. Wang SM, Ooi LL, Hui KM. Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma. Clin Cancer Res. 2007;13(21):6275.

17. Clough E, Barrett T. The gene expression omnibus database. Methods Mol Biol. 2016;1418:93.

18. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A. 2010;107(21):9546.

19. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284.

20. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8(Suppl 4):S11.

21. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98.

22. Menyhart O, Nagy A, Gyorffy B. Determining consistent prognostic biomarkers of overall survival and vascular invasion in hepatocellular carcinoma. R Soc Open Sci. 2018;5(12):181006.

23. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q, Bader GD. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics. 2010;26(22):2927.

24. Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, Morris Q. GeneMANIA update 2018. Nucleic Acids Res. 2018;46(W1):W60.

25. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830.

26. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11(5):R53.

27. Levine AJ, Miller JA, Shapshak P, Gelman B, Singer EJ, Hinkin CH, Commins D, Morgello S, Grant I, Horvath S. Systems analysis of human brain gene expression: mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer's disease. BMC Med Genet. 2013;6:4.

28. Tang X, Huang X, Wang D, Yan R, Lu F, Cheng C, Li Y, Xu J. Identifying gene modules of thyroid cancer associated with pathological stage by weighted gene co-expression network analysis. Gene. 2019;704:142.

29. Sezin T, Vorobyev A, Sadik CD, Zillikens D, Gupta Y, Ludwig RJ. Gene expression analysis reveals novel shared gene signatures and candidate molecular mechanisms between pemphigus and systemic lupus erythematosus in CD4(+) T cells. Front Immunol. 2017;8:1992.

30. Mason WS, Gill US, Litwin S, Zhou Y, Peri S, Pop O, Hong ML, Naik S, Quaglia A, Bertoletti A, Kennedy PT. HBV DNA integration and clonal hepatocyte expansion in chronic hepatitis B patients considered immune tolerant. Gastroenterology. 2016;151(5):986.

31. Jiang X, Kanda T, Wu S, Nakamura M, Miyamura T, Nakamoto S, Banerjee A, Yokosuka O. Regulation of microRNA by hepatitis B virus infection and their possible association with control of innate immunity. World J Gastroenterol. 2014;20(23):7197.

32. Qiu L, Wang T, Xu X, Wu Y, Tang Q, Chen K. Long Non-coding RNAs in hepatitis B virus-related hepatocellular carcinoma: regulation, functions, and underlying mechanisms. Int J Mol Sci. 2017;18(12):2505.

33. Wang S, Cui S, Zhao W, Qian Z, Liu H, Chen Y, Lv F, Ding HG. Screening and bioinformatics analysis of circular RNA expression profiles in hepatitis B-related hepatocellular carcinoma. Cancer Biomark. 2018;22(4):631.

34. Wang M, Xi D, Ning Q. Virus-induced hepatocellular carcinoma with special emphasis on HBV. Hepatol Int. 2017;11(2):171.

35. Kanda T, Yokosuka O, Imazeki F, Yamada Y, Imamura T, Fukai K, Nagao K, Saisho H. Hepatitis B virus X protein (HBx)-induced apoptosis in HuH-7 cells: influence of HBV genotype and basal core promoter mutations. Scand J Gastroenterol. 2004;39(5):478.

36. Liu H, Shi W, Luan F, Xu S, Yang F, Sun W, Liu J, Ma C. Hepatitis B virus X protein upregulates transcriptional activation of human telomerase reverse transcriptase. Virus Genes. 2010;40(2):174.

37. Zou SQ, Qu ZL, Li ZF, Wang X. Hepatitis B virus X gene induces human telomerase reverse transcriptase mRNA expression in cultured normal human cholangiocytes. World J Gastroenterol. 2004;10(15):2259.

38. Saldivar JC, Hamperl S, Bocek MJ, Chung M, Bass TE, Cisneros-Soberanis F, Samejima K, Xie L, Paulson JR, Earnshaw WC, Cortez D, Meyer T, Cimprich KA. An intrinsic S/G2 checkpoint enforced by ATR. Science. 2018;361(6404):806.

39. Hayward D, Alfonso-Perez T, Cundell MJ, Hopkins M, Holder J, Bancroft J, Hutter LH, Novak B, Barr FA, Gruneberg U. CDK1-CCNB1 creates a spindle checkpoint-permissive state by enabling MPS1 kinetochore localization. J Cell Biol. 2019;218(4):1182.

40. Ueno M, Itoh M, Kong L, Sugihara K, Asano M, Takakura N. PSF1 is essential for early embryogenesis in mice. Mol Cell Biol. 2005;25(23):10528.

41. Saito K, Ohta S, Kawakami Y, Yoshida K, Toda M. Functional analysis of KIF20A, a potential immunotherapeutic target for glioma. J Neuro-Oncol. 2017;132(1):63.

42. Duan J, Huang W, Shi H. Positive expression of KIF20A indicates poor prognosis of glioma patients. Onco Targets Ther. 2016;9:6741.

43. Zhou Q, Lee KJ, Kurasawa Y, Hu H, An T, Li Z. Faithful chromosome segregation in Trypanosoma brucei requires a cohort of divergent spindle-associated proteins with distinct functions. Nucleic Acids Res. 2018;46(16):8216.

44. Raemaekers T, Ribbeck K, Beaudouin J, Annaert W, Van Camp M, Stockmans I, Smets N, Bouillon R, Ellenberg J, Carmeliet G. NuSAP, a novel microtubule-associated protein involved in mitotic spindle organization. J Cell Biol. 2003; 162(6):1017.

45. Gu Z, Xia J, Xu H, Frech I, Tricot G, Zhan F. NEK2 promotes aerobic glycolysis in multiple myeloma through regulating splicing of pyruvate kinase. J Hematol Oncol. 2017;10(1):17.

46. Wang J, Cheng P, Pavlyukov MS, Yu H, Zhang Z, Kim SH, Minata M, Mohyeldin A, Xie W, Chen D, Goidts V, Frett B, Hu W, Li H, Shin YJ, Lee Y, Nam DH, Kornblum HI, Wang M, Nakano I. Targeting NEK2 attenuates glioblastoma growth and radioresistance by destabilizing histone methyltransferase EZH2. J Clin Invest. 2017;127(8):3075.

47. Li G, Zhong Y, Shen Q, Zhou Y, Deng X, Li C, Chen J, Zhou Y, He M. NEK2 serves as a prognostic biomarker for hepatocellular carcinoma. Int J Oncol. 2017;50(2):405.

48. Gui L, Homer H. Spindle assembly checkpoint signalling is uncoupled from chromosomal position in mouse oocytes. Development. 2012;139(11):1941.
49. Yost S, de Wolf B, Hanks S, Zachariou A, Marcozzi C, Clarke M, de Voer R, Etemad B, Uijttewaal E, Ramsay E, Wylie H, Elliott A, Picton S, Smith A, Smithson S, Seal S, Ruark E, Houge G, Pines J, Kops G, Rahman N. Biallelic TRIP13 mutations predispose to Wilms tumor and chromosome missegregation. Nat Genet. 2017;49(7):1148.
50. Zhang L, Makamure J, Zhao D, Liu Y, Guo X, Zheng C, Liang B. Bioinformatics analysis reveals meaningful markers and outcome predictors in HBV-associated hepatocellular carcinoma. Exp Ther Med. 2020;20(1):427.
51. Zhang X, Wang L, Yan Y. Identification of potential key genes and pathways in hepatitis B virus-associated hepatocellular carcinoma by bioinformatics analyses. Oncol Lett. 2020;19(5):3477.
52. Chen Z, Chen J, Huang X, Wu Y, Huang K, Xu W, Xie L, Zhang X, Liu H. Identification of potential key genes for hepatitis B virus-associated hepatocellular carcinoma by bioinformatics analysis. J Comput Biol. 2019; 26(5):485.

Publisher's Note