



OPEN Integrating single-cell RNA sequencing, WGCNA, and machine learning to identify key biomarkers in hepatocellular carcinoma

Gang Wang^{1,3,4}, Jiaying Zhang^{2,3,4}, Yirong Li^{1,3}, Yuyu Zhang^{2,3}, Weiwei Dong^{2,3}, Hengquan Wu^{2,3}, Jinglan Wang^{1,3}, Peiqing Liao^{2,3}, Ziqiang Yuan^{1,3}, Tao Liu^{1,2,3}✉ & Wenting He^{1,2,3}✉

The microarray and single-cell RNA-sequencing (scRNA-seq) datasets of hepatocellular carcinoma (HCC) were downloaded from the Gene Expression Omnibus (GEO) database. Differential expression analysis and weighted gene co-expression network analysis (WGCNA) were used to identify HCC-related biomarkers. Based on an analysis of scRNA-seq data, several marker genes expressed on tumor cells have been identified. Three machine-learning algorithms were used to identify shared diagnostic genes. Furthermore, logistic regression analysis was conducted to re-evaluate and identify essential biomarkers, which were then employed to develop a diagnostic prediction model. Additionally, AutoDockTools were used for molecular docking to investigate the association between the most sensitive drug and the core proteins. 44 genes were obtained by intersecting the WGCNA results, marker genes from scRNA-seq data, and up-regulated DEGs. Three machine-learning algorithms refined CDKN3, PPIA, PRC1, GMNN, and CENPW as hub biomarkers. GMNN and PRC1 were further selected by logistic regression analysis to build a nomogram. The molecular docking results showed that the drug NPK76-II-72-1 had a good binding ability with the GMNN and PRC1 proteins. The results highlighted CDKN3, PPIA, PRC1, GMNN, and CENPW as potential detection biomarkers for HCC patients. Our research offers novel insights into the diagnosis and treatment of HCC.

Keywords Machine learning, Molecular docking, WGCNA, Biomarker, Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer death worldwide, accounting for 7.8% of all cancer deaths¹. The main risk factors of HCC include cirrhosis, alcohol, nonalcoholic fatty liver disease, diabetes, obesity, aflatoxin, and genetic diseases². There are many therapeutic options for HCC patients, including liver transplantation, surgical resection, radiation therapies, percutaneous ablation, and transarterial and systemic treatment approaches³. However, the overall prognosis for patients with HCC remains poor, and the therapeutic outcomes are still unsatisfactory. Therefore, it is urgent to explore new treatment methods to overcome the difficulties of HCC treatment. With advancements in treatment modalities, targeted therapy and immunotherapy have emerged as new treatment methods and have gradually become research hotspots for researchers. Targeted therapies for patients with intrahepatic cholangiocarcinoma, including those targeting (fibroblast growth factor receptor-2, FGFR-2) fusions, (isocitrate dehydrogenase-1, IDH-1) mutations, (neurotrophic tyrosine receptor kinase, NTRK) fusions, and (human epidermal growth factor-2, HER-2) amplifications, are rapidly changing the treatment landscape⁴. Also, Immune checkpoint inhibitors are important for cancer treatment. The biological behavior of various immune checkpoints, such as (cytotoxic T lymphocyte-associated antigen-4, CTLA-4), (programmed cell death protein 1, PD-1), and (programmed cell death ligand 1, PDL-1), has been demonstrated in HCC⁵. Consequently, searching for effective biomarkers may present a potential therapeutic strategy to prevent the progression of HCC.

Over the years, researchers have been diligently committed to elucidating the mechanisms of HCC and discovering novel diagnostic biomarkers, aiming to offer insights into potential diagnostic and therapeutic

¹School of Basic Medical Sciences, Lanzhou University, Lanzhou 730000, Gansu Province, China. ²The Second Hospital & Clinical Medical School, Lanzhou University, Lanzhou 730030, Gansu Province, China. ³Gansu Provincial Key Laboratory of Environmental Oncology, Lanzhou 730000, Gansu Province, China. ⁴Gang Wang and Jiaying Zhang contributed equally. ✉email: liut@lzu.edu.cn; hewt@lzu.edu.cn

strategies. For instance, Li Yang et al.⁶ used bioinformatics methods to screen for two genes that played an essential role in HCC prognosis prediction. Tao Wang et al.⁷ identified several biomarkers for immunotherapy in hepatocellular carcinoma based on WGCNA and the LASSO Algorithm. Chenkai Huang et al.⁸ utilized WGCNA to screen co-expressed genes and identified three genes associated with the prognosis of HCC. Nevertheless, the underlying molecular mechanisms that govern the initiation and progression of HCC remain largely uncharacterized, highlighting the need for further research in this area. In the present study, an integrated analysis of scRNA-seq data analysis, WGCNA, and machine learning was conducted to identify potential biomarkers involved in the occurrence and progression of HCC. This research aims to leverage these genes to develop a novel diagnostic model or contribute to the advancement of targeted therapy and immunotherapy for HCC patients.

Materials and methods

Data acquisition and preprocessing

The workflow chart is presented in Supplementary Fig. 1. We obtained two microarrays (GSE214846 and GSE77314) from the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/>). Specifically, the GSE214846 dataset, comprising 65 pairs of HCC tissues and adjacent non-HCC tissues, was analyzed using the GPL24676 platform. Meanwhile, the GSE77314 dataset, encompassing 50 pairs of normal and tumor samples, was processed on the GPL9052 platform. The “limma” R package was employed to assess the differentially expressed genes (DEGs) between HCC tissue and normal tissue in the datasets GSE214846 and GSE77314. The criteria for selecting DEGs were set as $|\log_2 \text{fold change (FC)}|$ greater than 1 and p -value less than 0.05. The R packages “ggplot2” and “pheatmap” were used to visualize DEGs as volcano plots and heat maps. Specifically, the R packages “clusterProfiler” and “org.Hs.eg.db” were utilized to explore the up-regulated genes’ function analysis, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) enrichment analysis^{9,10}. GO enrichment analysis encompassed three main categories: biological processes (BP), cellular components (CC), and molecular functions (MF). Finally, the GSE214846 and GSE77314 datasets were integrated to form a comprehensive dataset for the following analysis. The Combat function from the “sva” R package was used to eliminate batch effects¹¹.

Weighted Gene Co-expression Network Analysis (WGCNA)

The “WGCNA” R package was utilized to identify modules with highly similar co-expression patterns¹². The parameters were carefully set as follows: the minModuleSize was set to 30 and the mergeCutHeight was set to 0.25. Next, we chose an optimal soft threshold to convert the correlation matrix into an adjacency matrix, which was applied to generate a topological overlap matrix (TOM). $\beta = 5$ ($R^2 = 0.9$) was chosen as the better soft-threshold power based on the scale independence and mean connectivity. Using the TOM-based dissimilarity indicators, positive or negative correlations of genes with similar expression patterns were clustered into the same modules through average linkage hierarchical clustering. The association of module genes with tumor or normal was quantified by calculating gene significance (GS). Moreover, the correlation between module eigengenes and gene expression patterns was assessed by module membership (MM). Pearson correlation analysis was employed to assess the relationship between modules and groups, with a threshold of $P < 0.05$. Finally, the exportNetworkToCytoscape function was used to export a network as edge and node list files, with the threshold set to 0.2. Unless otherwise specified, other calculations were performed using the default parameters implemented in the WGCNA package. This method helped us determine which genes had a high association level with HCC growths.

Single-cell RNA-sequencing dataset quality control and processing

The single-cell transcriptome dataset GSE149614 was also downloaded from the GEO database. From this dataset, we obtained 10 HCC samples, including HCC01T, HCC02T, HCC03T, HCC04T, HCC05T, HCC06T, HCC07T, HCC08T, HCC09T, and HCC10T. The scRNA sequence data were processed with the “Seurat” R package. Firstly, the Seurat standardization analysis workflow was applied, whereby cells identified as low-quality (nFeature_RNA < 200, percent.mt \geq 20) were excluded. After that, the data were normalized using the LogNormalize method. The top 2000 genes with highly variable features were found using the FindVariableFeatures algorithm. Furthermore, the “RunPCA” function was employed to conduct principal component analysis (PCA) of the highly variable genes. The “Harmony” R package was used for batch effects correction¹³. Subsequently, the Seurat “FindAllMarkers” function (logfc.threshold = 0.25, min.pct = 0.25) was adopted to compare various groups and identify cluster-specific expressed genes. Importantly, the “inferCNV” R package was applied to identify malignant cells with clonal large-scale chromosomal copy number variations (CNV)¹⁴. Finally, the “venn” R package was used to obtain common genes by overlapping the DEGs, module genes, and marker genes.

Screening of candidate diagnostic genes by machine learning

This study employed three machine learning techniques, specifically, Support Vector Machine Recursive Feature Elimination (SVM-RFE), Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest (RF), to narrow down the candidate biomarkers. RF classifier constructs multiple decision trees, and the optimal classification result is selected based on the output values of all decision trees¹⁵. The “randomForest” R package was employed to perform the RF algorithm, and core genes were defined as MeanDecreaseGini (MDG) greater than 1.0¹⁶. The SVM-RFE algorithm was developed for the selection of a small subset of genes from broad patterns of gene expression data¹⁷. The “e1071” R package was employed to perform the SVM-RFE algorithm¹⁸, which enabled us to rank candidate feature genes by their importance and utilize ten-fold cross-validation for error assessment. LASSO regression can handle a larger set of potential biomarkers and effectively selects the variables most strongly associated with liver hepatocellular carcinoma (LIHC). The LASSO algorithm was implemented

using the “glmnet” R package¹⁹, with the response variable of “family” set as “binomial” and the alpha parameter set to 1. This approach incorporated ten-fold cross-validation to evaluate the model’s performance and ultimately obtained the candidate genes based on the best λ . Ultimately, the diagnostic accuracy of the biomarkers was assessed utilizing the receiver operating characteristic (ROC) curves, with a focus on the corresponding area under the curve (AUC) values. The LIHC data from the TCGA database was used for external validation of the diagnostic efficacy of these genes.

The construction of the nomogram and the assessment of the diagnostic marker prediction model

The integrated dataset from GSE214846 and GSE77314 was divided into training and testing sets at a 70:30 ratio. This ensured that one subset was used for model training while the other was reserved for performance evaluation. In addition, we retrieved LIHC data from The Cancer Genome Atlas Program (TCGA) database (<https://www.cancer.gov/ccg/research/genome-sequencing/tcga>) in the UCSC Xena platform (<https://xena.broadinstitute.org/xena/app/browser/datapages/>) for external validation, thereby assessing the model’s predictive accuracy. Initially, the “compareGroups” R package was used to assess the differences between the training and testing groups. The univariable and multivariate logistic regression analyses were employed to screen further genes utilizing “autoReg” R package. The expression data of the candidate genes from this dataset were extracted to build the nomogram using the “rms” R package. The ROC analysis was established to evaluate the diagnosis model’s ability using the “pROC” package. The Decision curve analysis (DCA) and calibration curves were used further to evaluate the nomogram’s clinical value and predictive accuracy. The “rmda” R package was used to plot the DCA curves.

Drug sensitivity and molecular docking analysis

The Gene Set Cancer Analysis (GSCA: <http://bioinfo.life.hust.edu.cn/GSCA/#/>) platform was used to examine the sensitivity of the five hub genes to various drugs²⁰. The 3D structure of the proteins was obtained from the AlphaFold Protein Structure Database (<https://www.alphafold.ebi.ac.uk/entry/AF-O75354>)^{21,22}. The three-dimensional structure of the drug was downloaded from the PubChem database (<https://pubchem.ncbi.nlm.nih.gov/>)²³. Molecular docking experiments were performed using AutoDockTools (version 1.5.7), followed by visualization of the docking outcomes with PyMOL (version 3.0.3) software.

The expression profile and prognostic analysis

The differential expression profile of the marker genes between tumor and normal tissues was compared. The Wilcoxon rank-sum test was used to assess differences between normal and tumor groups. In addition, patients were divided into high and low groups based on genes’ expression levels, with the median value as the threshold. The R “survival” and “survminer” packages were used for the survival analysis of hub genes in LIHC patients. The Kaplan-Meier curves were performed to plot the overall survival (OS) the disease-specific survival (DSS), progression-free interval (PFI), and disease-free interval (DFI) curves.

Functional enrichment analysis

The Genes positively correlated with the five biomarkers were downloaded from The University of Alabama at Birmingham CANcer data analysis Portal (<https://ualcan.path.uab.edu/analysis.html>)²⁴. Functional enrichment analysis, including GO and KEGG pathways, was performed using the “clusterProfiler” and “and org.Hs.eg.db” packages.

Statistical analysis

All data processing and statistical analysis were performed using R software (version 4.4.2). The statistical significance was defined as a p -value < 0.05.

Results

DEGs, GO, and KEGG analysis for the GSE214846 dataset

In our detailed examination of the GSE214846 dataset, we identified a comprehensive list of 1,407 DEGs, further categorized into 983 down-regulated genes and 424 up-regulated genes (Fig. 1A). The corresponding volcano plot and the cluster heatmap for the microarray datasets were displayed in Fig. 1B, respectively. GO and KEGG pathway enrichment analyses were used to explore the biological characteristics and pathways of the up-regulated genes. The KEGG pathway analysis results showed that up-regulated genes concentrated on the cell cycle, DNA replication, Oocyte meiosis, Progesterone-mediated oocyte maturation, Motor proteins, and p53 signaling pathway (Fig. 1C). The enriched BP included the responses to chromosome segregation, nuclear chromosome segregation, sister chromatid segregation, and mitotic sister chromatid segregation. The enriched CC were chromosomal region, chromosome, centromeric region, condensed chromosome, and spindle. The enriched MF were microtubule binding, single-stranded DNA helicase activity, tubulin binding, ATP-dependent activity, acting on DNA, and ATP hydrolysis activity (Fig. 1D).

DEGs, GO, and KEGG analysis for the GSE77314 dataset

In the GSE77314 dataset, we identified a total of 1929 DEGs, including 1,160 down-regulated and 769 up-regulated genes. (Fig. 2A). The cluster heatmap was shown based on all DEGs of the dataset (Fig. 2B). We also discussed the GO and KEGG analysis of up-regulated genes. It was demonstrated that these genes were associated with multiple signaling pathways. Many of them overlapped with the GSE214846 dataset, but some showed inconsistency, such as Coronavirus disease-COVID-19 and Fanconi anemia pathway (Fig. 2C). The

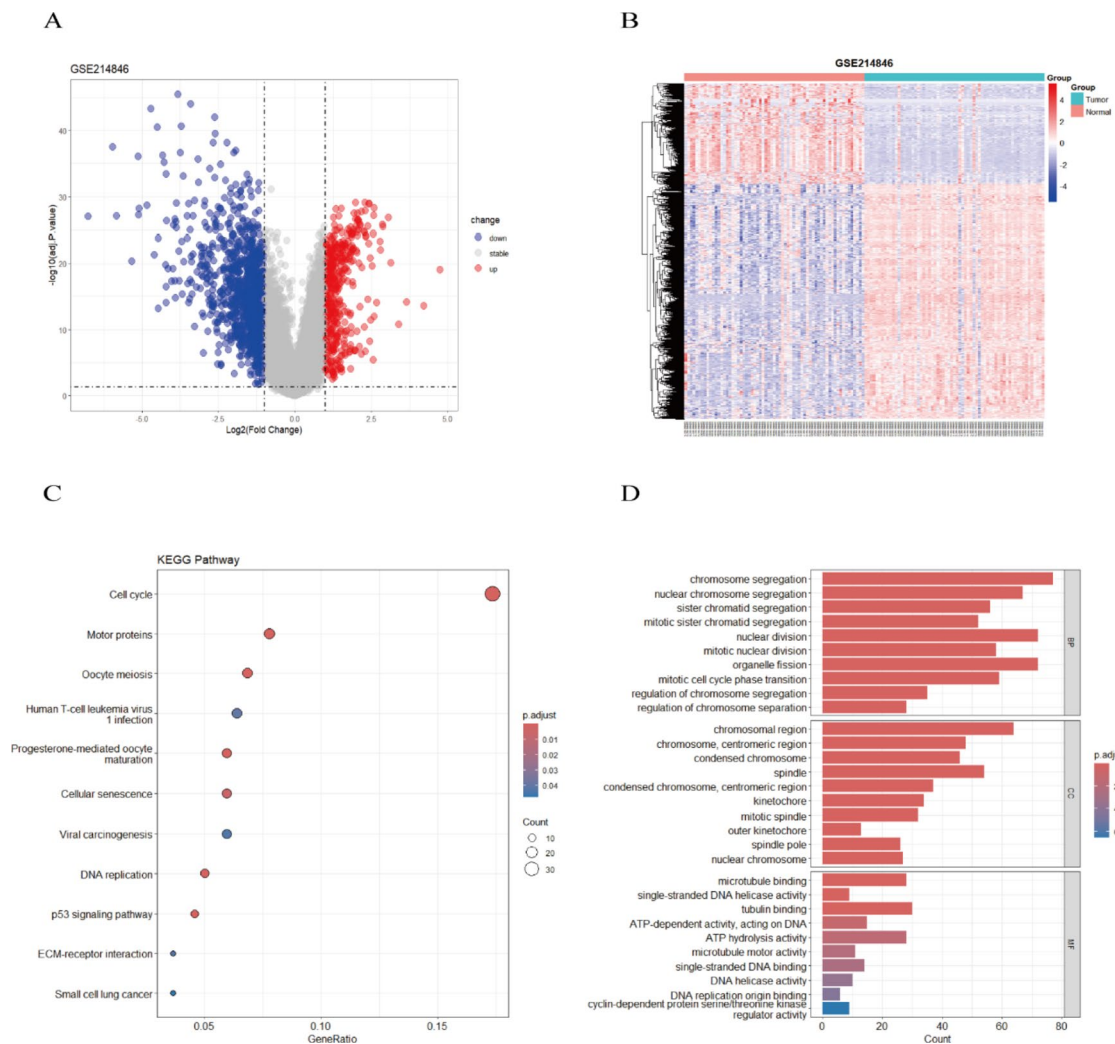


Fig. 1. DEGs analysis for the GSE214846 dataset. Volcano plots of DEGs in the GSE214846 dataset (A). Red dots represent up-regulated genes, blue dots represent down-regulated genes. The differential analysis results for both the normal and HCC groups are represented by a heatmap (B). KEGG pathway analysis (C). GO analysis for up-regulated DEGs in BP, MF, and CC terms (D).

results obtained from BP, CC, and MF analyses were highly consistent with those from the GSE214846 dataset analysis, highlighting the reliability of our analysis results (Fig. 2D).

WGCNA analysis for the merged dataset

WGCNA was performed to determine the most correlated gene modules in LIHC tissues. The 115 samples were separated into normal and tumor groups (Fig. 3A). The optimal soft-power β was set at 5 (scale-free $R^2 = 0.9$) to establish a gene co-expression network (Fig. 3B). Subsequently, 39 modules were constructed through WGCNA (Fig. 3C). Furthermore, correlation analysis of the module-trait relationships revealed that the blue module had a strong positive correlation with HCC ($r = 0.85$, $p = 2e-65$). In contrast, the turquoise module showed a highly negative correlation with HCC ($r = -0.71$, $p = 5e-36$) (Fig. 3D). Then, we primarily focused on the blue module in the subsequent analysis, which could provide a more precise indication of HCC. In the blue module, key components MM and GS showed a significant positive correlation ($r = 0.96$, $P < 1e-200$) (Fig. 3E), indicating these genes had a strong association with the HCC characteristic. Finally, a threshold of 0.2 was established in the blue module, yielding 740 significant feature genes.

Single-cell RNA-sequencing dataset analysis for finding marker genes

To investigate cellular diversity and molecular signatures in HCC tissues, we performed marker gene identification and cell classification using the “Seurat” package. We identified and visualized 24 clusters using the T-distributed stochastic neighbor embedding (t-SNE) method (Fig. 4A). Cells can be classified into nine main types based on typical marker genes: natural killer (NK) cells (KLRD1 and KLRF1), epithelial cells (EPCAM and KRT19), B cells (MS4A1, CD19, and CD79A), T cells (CD2, CD3D, and CD3E), endothelial cells (VWF and PECAM1), hepatic stellate cells (ACTA2 and PDGFRB), myeloid cells (CD14, LYZ, and C1QB), mast cells (CPA3, TPSAB1,

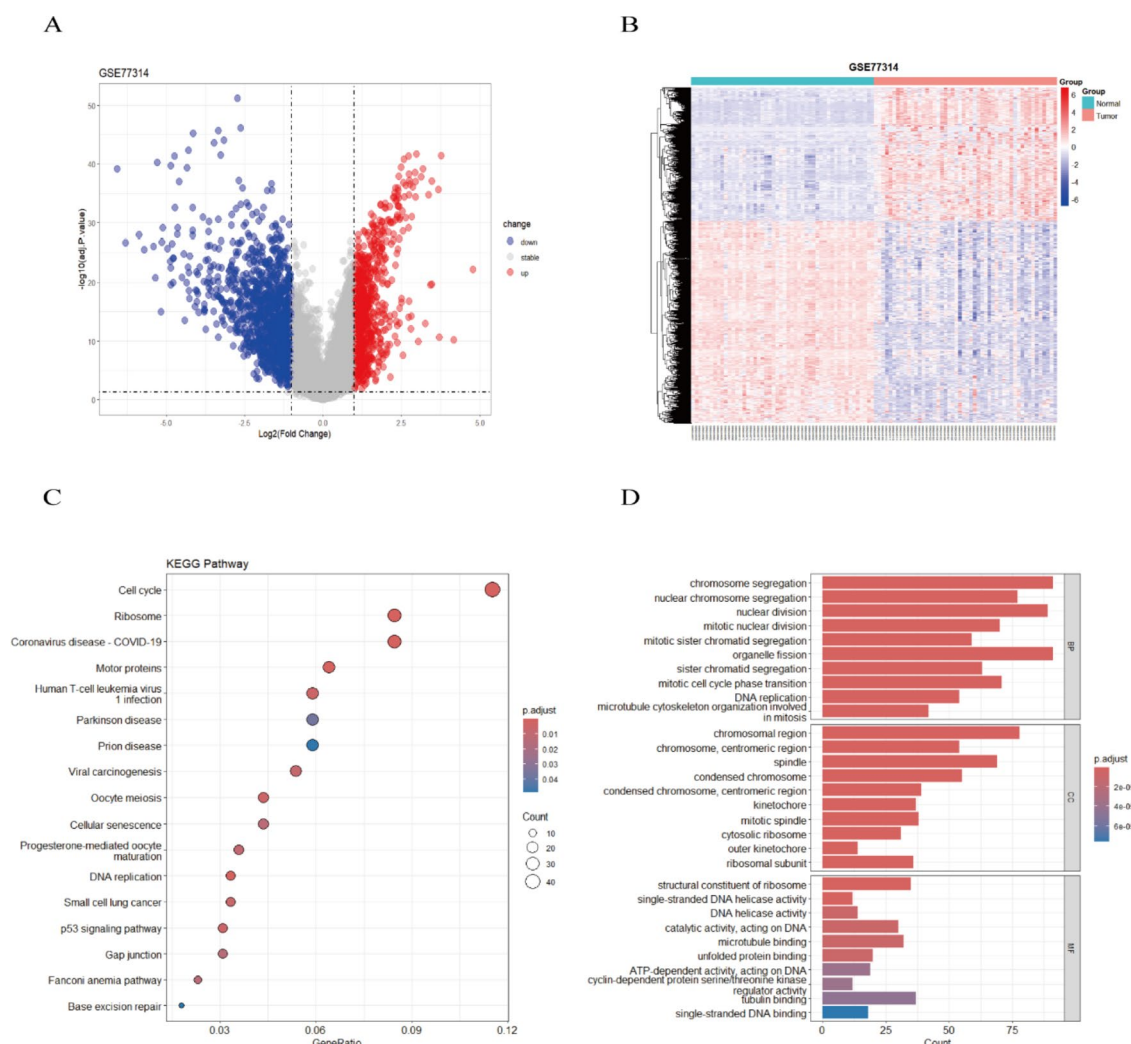


Fig. 2. DEGs analysis for the GSE77314 dataset. Volcano plots of DEGs in the GSE77314 dataset (A). Red dots represent up-regulated genes, blue dots represent down-regulated genes. The differential analysis results for both the normal and HCC groups are represented by a heatmap (B). KEGG pathway analysis (C). GO analysis for up-regulated DEGs in BP, MF, and CC terms (D).

and TP52), and malignant cells (Fig. 4B). The heatmap showed the average expression of typical marker genes for each of the clusters (Fig. 4C). It is worth noting that we found the expression of myeloid cell marker genes (CD14 and LYZ) in the malignant cells (Fig. 4D). Therefore, it may contain non-malignant cells. Next, the cluster of malignant cells was classified into 13 clusters again (Fig. 4E). Then, we conducted a comprehensive analysis of CNV variations in each chromosome across these clusters using the inferCNV package. The inferCNV result displayed the different CNV scores among 13 identified cell clusters (Fig. 4F). It highlighted that clusters 3, 12, and 10 manifested relatively lower CNV scores (Fig. 4G), meaning that these clusters might not be associated with the occurrence and development of malignant tumors. In contrast, the remaining clusters demonstrated a strong correlation with the presence and development of tumors. The “FindAllMarkers” function identified 661 markers across these clusters (excluding clusters 3, 10, and 12) when the logfc.threshold was set to ≥ 1 .

Identification of candidate hub biomarkers via machine learning

First, a total of 44 overlapped genes were obtained from WGCNA, marker genes for single-cell data analysis, and up-regulated DEGs (Fig. 5A). To further identify core genes with significant efficacy in the diagnosis, three machine-learning algorithms were adopted to explore potential candidate biomarkers. Utilizing the RF algorithm, we identified 24 feature genes with MeanDecreaseGini scores exceeding 1.0, indicating their significant impact on the model’s predictive power (Fig. 5BC). According to the SVM-RFE algorithm coupled with ten-fold cross-validation, we identified 14 genes with a minimal error rate of 0.051 (Fig. 5D). In addition, the results from the LASSO regression analysis revealed that the presence of HCC was associated with the expression levels of 6 specific genes, including CDKN3, PPIA, SNRPB, PRC1, GMNN, and CENPW (Fig. 5EF). Finally, we identified the intersecting genes from the three algorithms to serve as feature genes, which included CDKN3, PPIA, PRC1, GMNN, and CENPW, as depicted in Fig. 5G. These genes were particularly significant in terms of their diagnostic

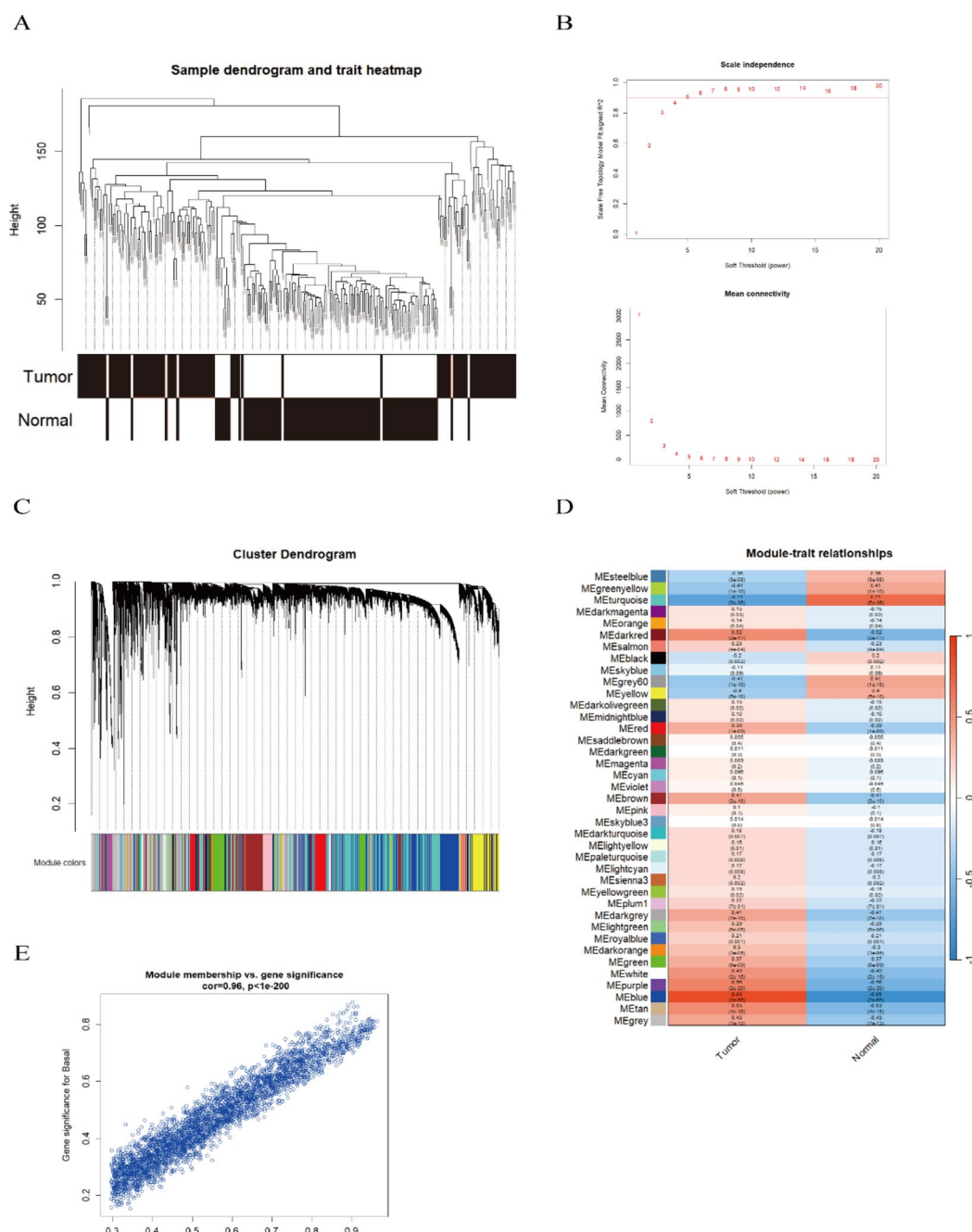


Fig. 3. Identification of key module genes by WGCNA. The sample dendrogram and trait heatmap (A). $\beta = 5$ ($R^2 = 0.9$) was chosen as the soft-threshold power based on the scale independence and mean connectivity (B). Hierarchical cluster analysis of co-expression clusters with corresponding colour assignments (C). The heatmap of the relationship between module eigengenes and clinical traits (D). Each row corresponds to a specific module, while each column corresponds to a particular trait. Each box contains the corresponding correlation and P-value. Scatter plot between gene significance for basal and module members in blue modules (E).

value. To ascertain the diagnostic utility of these genes, we conducted an evaluation utilizing ROC curves. In the GSE214846 and GSE77314 datasets, PRC1 (AUC=0.957), PPIA (AUC=0.936), GMNN (AUC=0.926), CENPW (AUC=0.954), and CDKN3 (AUC=0.961) exhibited high efficacy in diagnostic value, as indicated by their respective AUC scores (Fig. 5H). In the LIHC validation dataset, the AUC values of PRC1, PPIA, GMNN, CENPW, and CDKN3 were 0.986, 0.979, 0.974, 0.969, and 0.982 respectively. These results suggest that these genes exhibited great diagnostic utility (Fig. 5H).

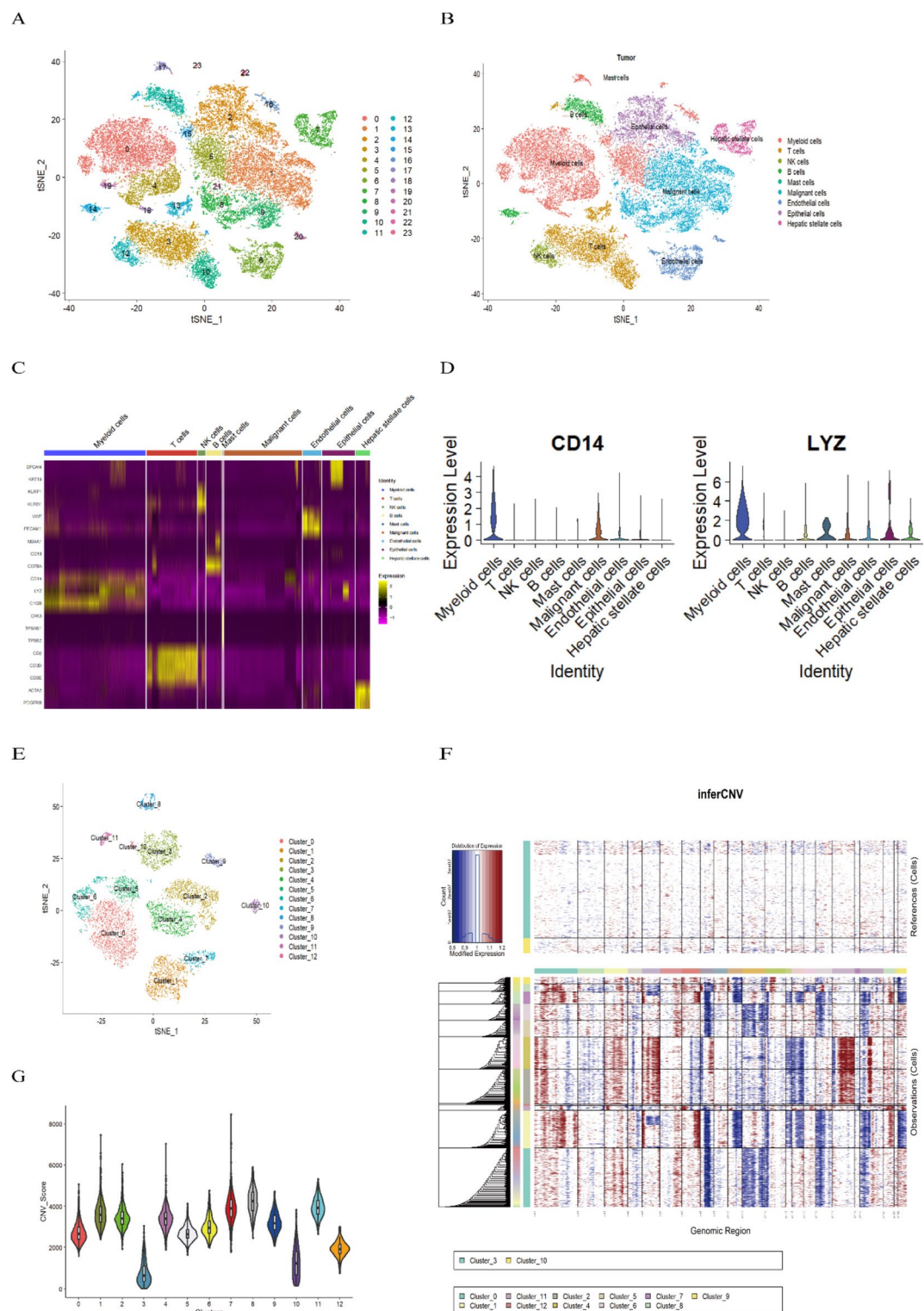


Fig. 4. Single-cell RNA-sequencing dataset analysis. t-SNE representation of the 24 cell clusters (A). t-SNE plot showing the annotation cell types in the HCC ecosystem (B). Different color clusters represent different cell types. A heatmap of typical marker gene expression (C). Brightness is positively correlated with cell expression. The expression levels of CD14 and LYZ in various cells (D). Malignant cells were reclassified into 13 clusters (E). The results of inferCNV analysis (F). Red and blue colors represent high and low CNV levels, respectively. The total score of CNV (G).

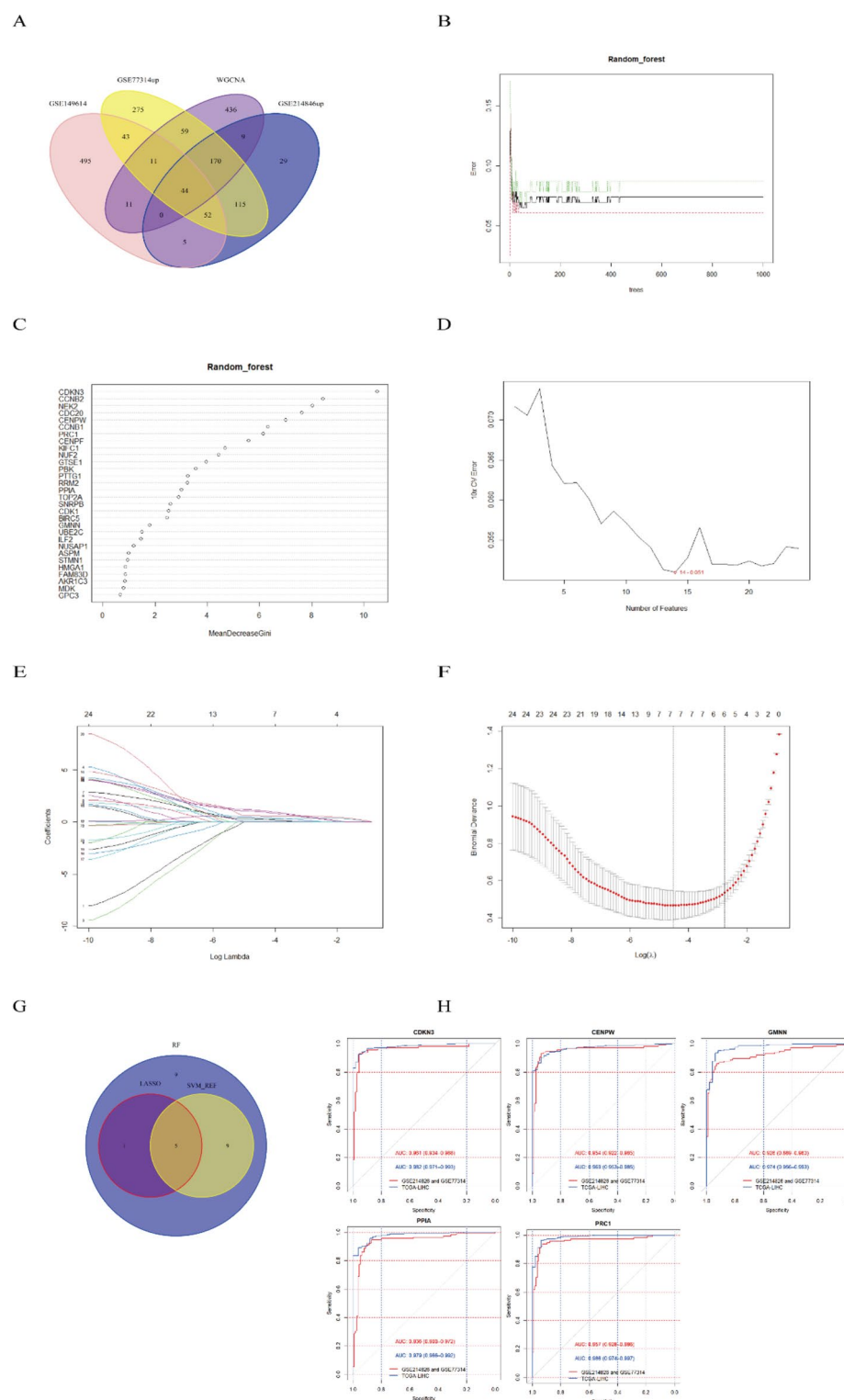


Fig. 5. Identification of hub genes via the LASSO model, the SVM-RFF model, and the RF model. A Venn diagram illustrating the overlapping genes (A). The correlation between the number of trees in the random forest and the error rates (B). The top 25 significant genes were identified from the random forest. MeanDecreaseGini shows the rank of genes based on their relative importance (C). The error of 10-fold cross-validation (CV) in SVM-RFE algorithms (D). The plot of LASSO coefficient profiles (E). Each curve represents a single gene. The plot of partial likelihood deviance (F). Vertical dashed lines were plotted at the optimal lambda values. The intersection of key genes from three machine learning (G). The ROC curves and AUC values of the hub genes (H).

Gene	Testing	Training	P Value
	N=69	N=161	
CDKN3	2.14 (1.54)	2.02 (1.56)	0.603
CENPW	1.29 (0.84)	1.21 (0.88)	0.510
GMNN	3.24 (0.92)	3.17 (0.93)	0.619
PPIA	7.89 (0.71)	7.92 (0.67)	0.800
PRC1	2.31 (1.46)	2.32 (1.49)	0.962

Table 1. Summary descriptives table of the testing and training groups.

Gene	Normal (N=83) Mean \pm SD	Tumor (N=78) Mean \pm SD	OR (univariable)	OR (multivariable)
GMNN	2.5 \pm 0.4	3.9 \pm 0.8	35.27 (12.02–103.50, $p < 0.001$)	5.00 (1.42–17.56, $p = 0.012$)
PRC1	1.1 \pm 0.7	3.6 \pm 1.0	10.17 (5.29–19.57, $p < 0.001$)	5.49 (2.70–11.13, $p < 0.001$)
PPIA	7.4 \pm 0.4	8.4 \pm 0.5	55.19 (17.54–173.58, $p < 0.001$)	1.52 (0.27–8.62, $p = 0.638$)
CDKN3	0.8 \pm 0.7	3.3 \pm 1.1	9.02 (4.85–16.78, $p < 0.001$)	0.93 (0.21–4.18, $p = 0.921$)
CENPW	0.6 \pm 0.4	1.9 \pm 0.7	54.82 (16.48–182.34, $p < 0.001$)	1.64 (0.30–8.86, $p = 0.567$)

Table 2. The results of logistic regression analysis in the training group.

Construction and validation of the nomogram

Initially, the baseline characteristics of the testing and training groups were presented in Table 1. The testing group contained 69 samples and the training group contained 161 samples. Then, univariate and multivariate logistic regression analyses were performed on the five genes. The results of univariate logistic regression analysis showed that all five genes are meaningful ($p < 0.05$). However, multivariate logistic regression analysis showed that only GMNN (OR = 5.00 (1.42–17.56, $p = 0.012$)) and PRC1 (OR = 5.49 (2.70–11.13, $p < 0.001$)) demonstrated statistical significance, with a p -value less than 0.05 (Table 2). We then developed a nomogram to elucidate the diagnostic significance of the two genes in HCC (Fig. 6A). The calibration curves showed that the predictive model, the testing group, and the validation group exhibited a good fit, indicating the good predictive accuracy of the nomogram (Fig. 6B). Moreover, we also conducted ROC analysis to ascertain the model's sensitivity and specificity and subsequently employed DCA to evaluate the nomogram's performance. The model demonstrated a strong performance, with an AUC value of 0.968 in the training set, 0.970 in the testing set, and 0.989 in the validation set, indicating its great effectiveness (Fig. 6C). The DCA curves were higher than the None and All lines under the High-Risk Threshold of 0.4 to 0.9, suggesting that the nomogram could be a valuable tool for clinical utility (Fig. 6D).

Drug sensitivity and molecular docking analysis for hub genes

To uncover potential therapeutic targets, we performed a drug sensitivity analysis. The GSCA results revealed significant negative correlations between GMNN/PRC1 expression and NPK76-II-72-1/Navitoclax (Fig. 7A). This suggests that higher expression of these genes may confer enhanced drug sensitivity. In contrast, a positive correlation was observed between GMNN/PRC1 expression and Trametinib, selumetinib, and RDEA119, suggesting that higher expression levels of GMNN and PRC1 could potentially lead to drug resistance (Fig. 7A). To explore the interactions between NPK76-II-72-1 and these two proteins, we employed molecular docking technology to predict the preferred orientation of proteins and drugs when they bind to form stable complexes. According to molecular docking results, the binding energies between GMNN and PRC1 and NPK76-II-72-1 were -7.4 kcal/mol and -6.82 kcal/mol, respectively (Fig. 7B,C).

The expression profile and prognostic analysis

To verify the expression of biomarkers in LIHC, we analyzed TCGA-LIHC data, revealing that CDKN3, PRC1, GMNN, CENPW, and PPIA were significantly overexpressed in LIHC samples relative to normal tissues ($P < 0.05$) (Fig. 8A). To investigate the prognostic benefit of hub genes, we compared survival differences between high and low gene expression groups in LIHC. Patients with high expression of CDKN3 exhibited significantly lower OS and PFI compared to those with low expression ($P < 0.05$) in LIHC (Fig. 8C, E). Additionally, higher expression levels of PRC1, CENPW, and PPIA were associated with worse OS, DFI, PFI, and DSS (Fig. 8B–E). These findings indicated that these genes were linked with poor outcomes in LIHC. Therefore, they had the potential to be a predictive biomarker.

Functional enrichment analysis

KEGG pathway⁹ and GO enrichment analyses of the hub genes were performed to explore the functional roles in HCC progression and potential mechanisms. GO enrichment analysis showed that the most significantly enriched BP of genes positively correlated with CDKN3, GMNN, PRC1, and CENPW was the RNA splicing (Fig. 9A–D). In contrast, the most significantly enriched BP of genes positively correlated with PPIA was the

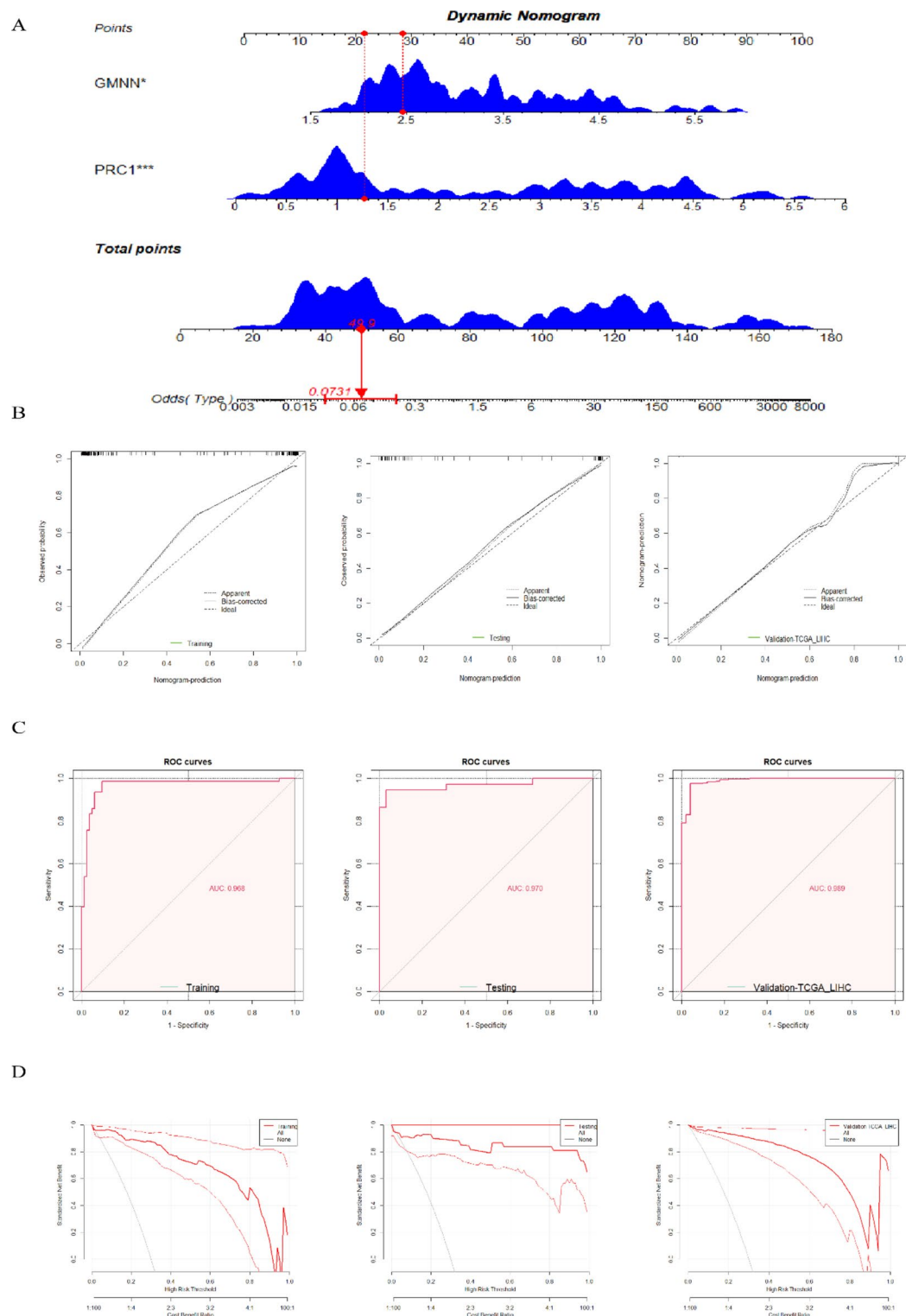
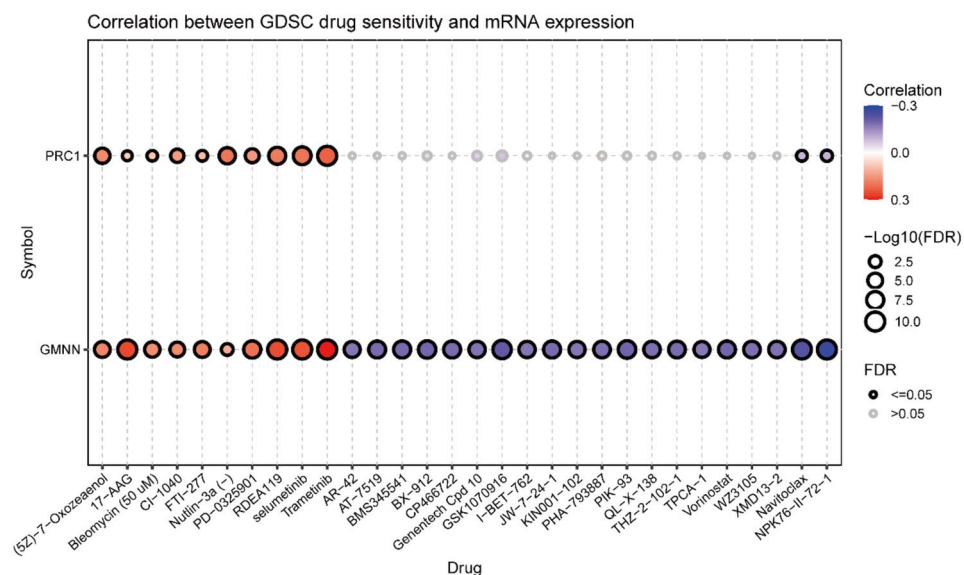


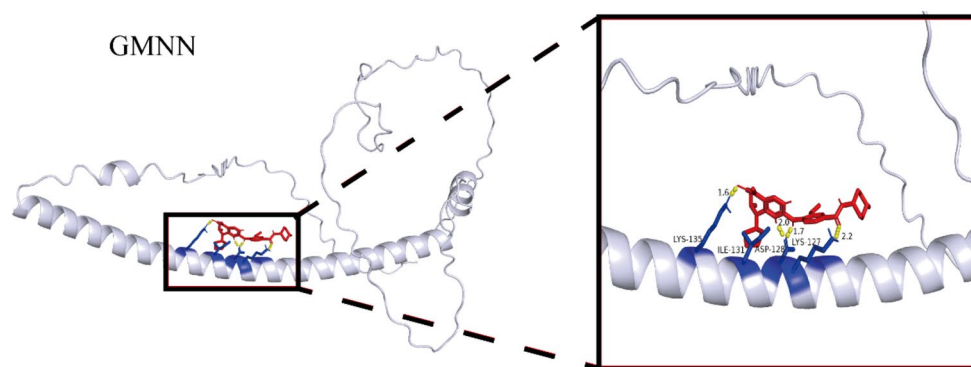
Fig. 6. The nomogram construction and validation. The nomogram model was built based on GMNN and PRC1 expression (A). The calibration curves of training, testing, and validation dataset (B). The ROC curves of training, testing, and validation dataset (C). The DCA results of training, testing, and validation dataset (D). (* $p < 0.05$, *** $p < 0.001$).

ribonucleoprotein complex biogenesis (Fig. 9E). The CC of the genes positively correlated with CDKN3, GMNN, PRC1, CENPW, and PPIA exhibited the most significant enrichment in the chromosomal region, nuclear speck, spindle, chromosomal region, and ribosome, respectively (Fig. 9A-E). The results of MF analysis revealed that the genes positively correlated with CDKN3, GMNN, PRC1, CENPW, and PPIA were involved in ATP hydrolysis

A



B



C

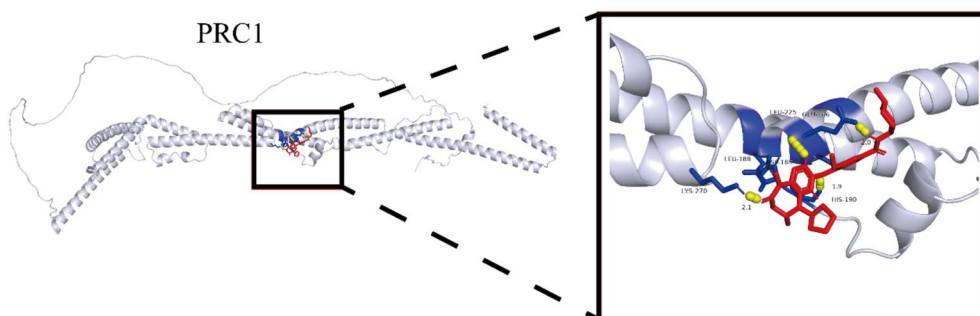


Fig. 7. Molecular docking analysis for GMNN and PRC1. Analysis of drug sensitivity of GMNN and PRC1 performed online by GSCA (A). Molecular docking analysis between NPK76-II-72-1 and GMNN (B). Molecular docking analysis between NPK76-II-72-1 and PRC1 (C). Red molecules represent drugs, silver-white molecules represent GMNN, and molecules are connected by hydrogen bonds.

activity, DNA-binding transcription factor binding, ubiquitin-like protein transferase activity, catalytic activity, acting on DNA, and structural constituent of ribosome (Fig. 9A-E). In addition, the KEGG pathway analysis revealed that the genes positively correlated with the five hub genes exhibited significant enrichment in Cell cycle and Nucleocytoplasmic transport pathways (Fig. 9A-E). The genes positively correlated with CDKN3,

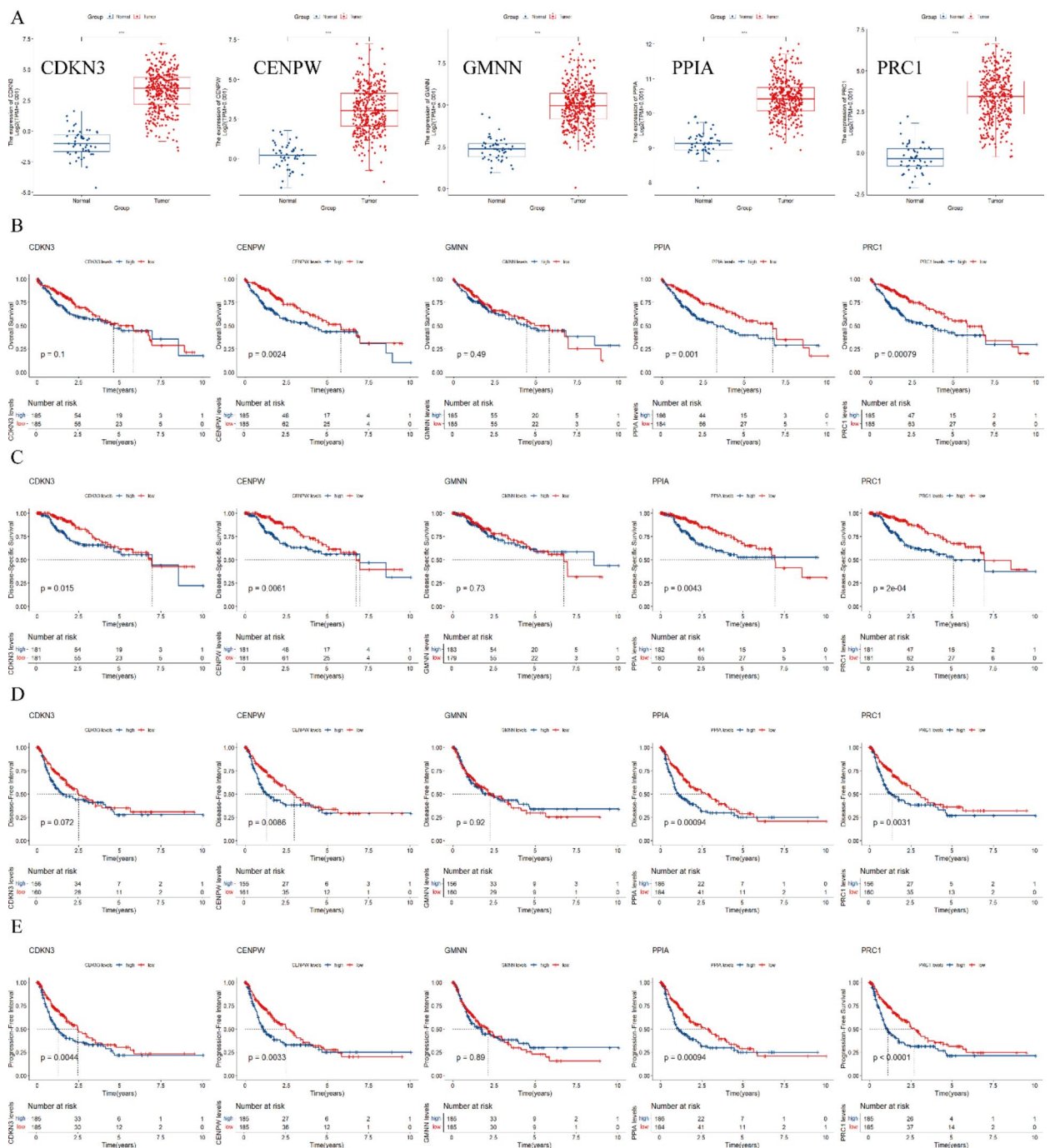


Fig. 8. The expression profile and survival analysis of the five genes. CDKN3, GMNN, PRC1, CENPW, and PPIA expression in LIHC samples compared to normal tissues(A). The survival curves of OS (B), DSS (C), DFI (D), PFI (E) (** $p < 0.001$).

PRC1, and CENPW were significantly enrichment in mRNA surveillance pathway, Base excision repair, and DNA replication (Fig. 9A, C, D).

Discussion

Our study provides a comprehensive analysis of biomarkers involved in HCC progression and evaluates their potential as diagnostic indicators. Initially, through analysis using the ‘limma’ package, up-regulated genes were identified and found to be associated with key biological processes, including the cell cycle, DNA replication, and the p53 signaling pathway. More evidence showed that the tumor suppressor protein p53 participated in a wide range of cellular processes, including cellular metabolism, tumor microenvironment modulation, and autophagy²⁵. Mutant p53 is highly associated with advanced malignancies and poor prognosis²⁶. Our results

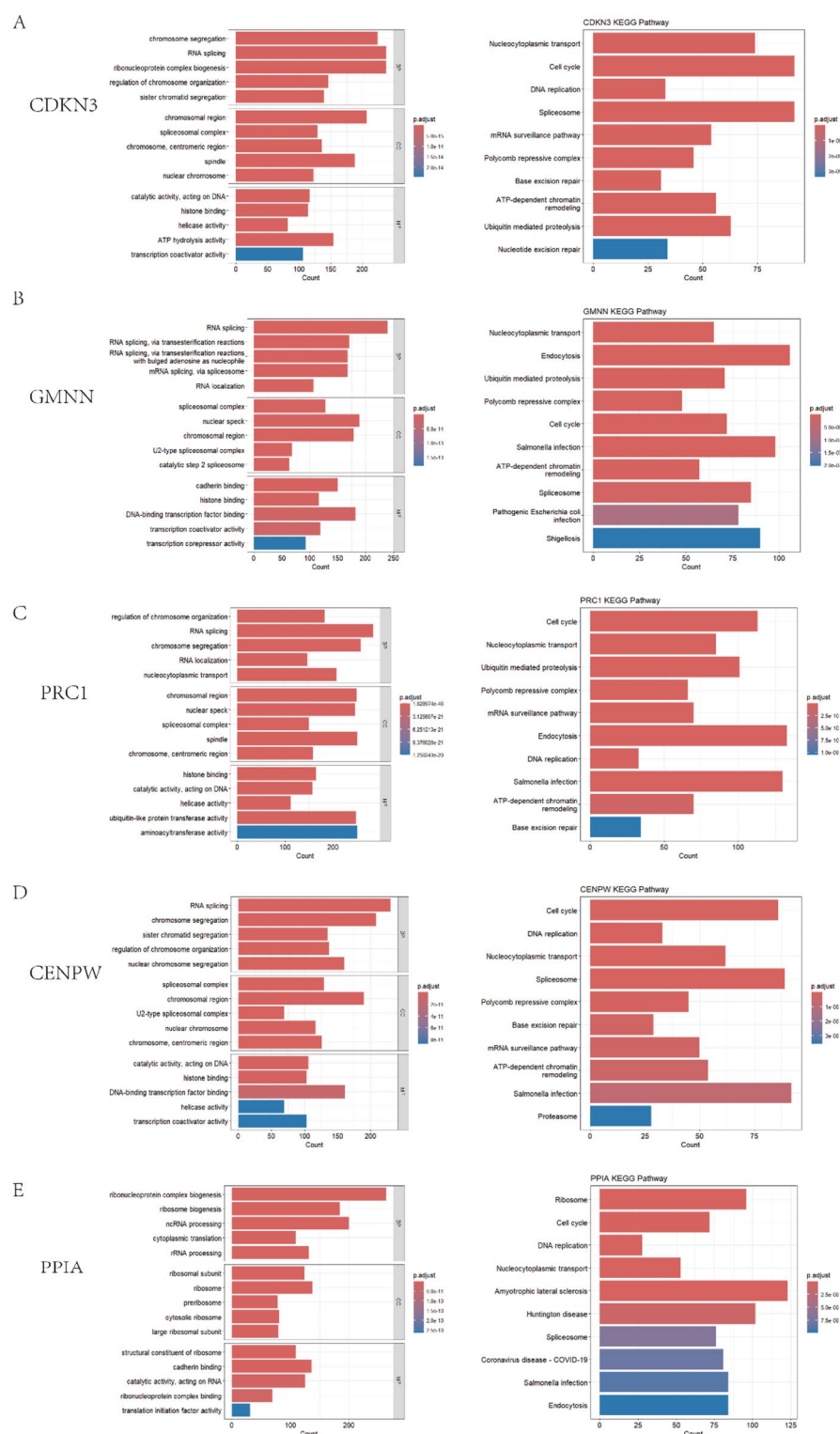


Fig. 9. GO and KEGG analyses of the five genes. The charts show the top 5 most significant GO enrichment results in each category of GO enrichment analysis and the top 10 most significant results of KEGG enrichment analysis ($P < 0.05$). CDKN3, GMNN, PRC1, CENPW, and PPIA correspond to A, B, C, D, and E, respectively.

revealed that up-regulated genes were closely associated with the mutation of p53, thus playing a significant role in regulating the occurrence and development of HCC.

WGCNA is a powerful bioinformatics analysis tool that can identify functionally relevant gene modules associated with clinical outcomes¹². By applying WGCNA, we revealed co-regulated gene modules and potential

candidate genes that were significantly associated with HCC. Through investigation, it was found that most of the extracted genes were associated with the poor prognosis of HCC, suggesting the credibility of our experimental results.

Microarrays have been widely used as a key technology for mapping diverse transcript types, such as mRNAs, non-coding RNAs, and small RNAs²⁷. One of the applications of microarray technology is to identify diagnostic markers through thousands of tested genes and then compare gene expression profiles from normal and malignant tissues²⁸. However, this technology has always been limited by the quality and amount of RNA²⁸. The emergence of RNA sequencing (RNA-Seq) provides a more accurate measurement of gene expression levels and is gradually becoming the mainstream for screening marker genes²⁹. A previous study has combined a novel approach for analyzing scRNA-seq data and identified a marker associated with poor prognosis in HCC³⁰. Pioneering research has successfully integrated scRNA-seq data with bulk transcriptomic profiles, identifying and validating disulfidptosis-related prognostic signatures specifically for HCC³¹. In our research, we combined single-cell data analysis to isolate 661 mutated genes that were expressed in malignant cells. These findings suggest that these candidate genes might play a crucial role in driving cancer occurrence and progression, providing a comprehensive approach for HCC biomarker screening.

Machine learning is an artificial intelligence wherein computers utilize data to learn and subsequently make predictions or decisions without explicit programming³². This method can analyze extensive gene expression matrices to identify core genes closely associated with diseases, thereby delivering crucial guidance for clinical research. We have identified five core genes (CDKN3, PPIA, PRC1, GMNN, and CENPW) that were closely related to HCC utilizing three machine learning algorithms. There is a possibility that these genes play a significant part in the development of HCC.

The cyclin-dependent kinase inhibitor 3 (CDKN3), a member of the dual-specificity protein phosphatase family, plays an essential role in cell cycle regulation, especially in the mitotic phase³³. Previous studies have revealed that this gene was the core gene of HCC and promoted tumor cell proliferation^{34,35}. CDKN3 was frequently overexpressed in both HCC clinical samples and cell lines, and this overexpression was associated with poor tumor differentiation and advanced tumor stage³⁶. However, there is limited research on the signaling pathway of this gene in HCC. In our study, KEGG pathway analysis revealed that PRC1 may contribute to the disease process by participating in the process of Cell cycle and Nucleocytoplasmic transport signaling pathways and affecting cell differentiation, proliferation, and apoptosis.

The peptidylprolyl isomerase A (PPIA), which belongs to the peptidyl-prolyl cis-trans isomerase family, serves as a biomarker and therapeutic molecular target of HCC³⁷. The high expression level of PPIA is correlated with a poor prognosis of HCC³⁸. Yuwei Gu et al.³⁹ revealed that PPIA could play a crucial role in the development of HCC by regulating mRNA metabolism and modulating tumor immunity. The MF analysis revealed that the genes positively correlated with PPIA exhibited significant enrichment in the catalytic activity, acting on RNA (Fig. 9E), suggesting that this gene may be involved in RNA regulation. These findings further confirm the efficacy and reliability of employing a multi-omics analysis integrated with machine learning in our study.

The centromere protein-w(CENPW) was highly expressed in many cancers compared to normal tissues, including lung cancer, stomach cancer, and HCC⁴⁰. Abnormal upregulation of the CENPW gene was associated with the clinical features and prognosis of HCC patients⁴¹. Our research also confirmed this. CENPW knockdown has been shown to suppress HCC cell progression by inhibiting cell proliferation, migration, and invasion, as well as inducing G0/G1 phase arrest and promoting cell apoptosis, potentially via E2F signaling pathway regulation⁴². Our research has found that this gene appears associated with RNA splicing. This offers novel insights for future research on this gene in HCC.

The protein Regulator of Cytokinesis 1 (PRC1) was detected in HCC stem cells⁴³. Deregulation of PRC1 led to cytokinesis defects that promoted tumor heterogeneity and cancer evolution⁴⁴. Previous research revealed that PRC1 overexpression played a role in the genome instability and the suppressive immune microenvironment of LIHC⁴⁵. A recent study showed that PRC1 was involved in the occurrence and development of HBV-associated liver cancer⁴⁶. In our study, PRC1 has been identified as a biomarker in the progression and development of HCC. The MF of genes positively correlated with PRC1 was closely associated with the ubiquitin-like protein transferase activity and aminoacyltransferase activity. The specific molecular regulatory mechanism involving this gene is worth further exploration in the future.

Geminin (GMNN), a 25 kDa protein, was an inhibitor of DNA replication during the S, G2, and M phases⁴⁷. Prem Prakash Kushwaha et al.⁴⁸ have provided a detailed summary of the structure, function, and significance of this protein in cancer. Interestingly, Kenichi Yoshida⁴⁹ proposed that this protein could serve as a target for anti-tumor drugs. Based on these findings and our experimental results, it is reasonable to speculate that GMNN may be a prospective drug target. New research has found that it may be a biomarker for multiple cancers, including oral cancer⁵⁰ and adrenocortical carcinoma⁵¹. Our research combines computer algorithms to analyze this protein as a novel biomarker in HCC from multiple perspectives and aspects. Except for the PRC1, there is limited research on the regulatory mechanisms of these genes in HCC, and it is worth further exploration in the future.

Notably, based on the expression of GMNN and PRC1, our study constructed a nomogram, which is a practical and valuable tool for diagnosing and treating patients. This integration of molecular information enhances our study's clinical translational application. Additionally, molecular docking was performed to verify the interaction between NPK76-II-72-1 and these two key genes. We found that NPK76-II-72-1 had strong binding interaction with GMNN and PRC1 proteins. NPK76-II-72-1 has been found as a potential drug for treating refractory glioblastoma multiforme⁵². However, there have been no reports on the application of this drug in HCC research. Further investigation into its therapeutic potential and the underlying mechanisms could yield valuable insights for future treatment strategies.

Our research also has some limitations. Firstly, the data analyzed were sourced from the GEO and TCGA databases, which may be subject to potential biases and confounding factors. Additionally, our study relied on computational analyses. Future research efforts will be needed to explore the molecular mechanism of these genes in HCC in depth and their potential implications in targeted therapy and immunotherapy.

Conclusion

We identified five hub genes (CDKN3, PPIA, PRC1, GMNN, and CENPW), which were closely related to the pathogenesis of HCC. GMNN and PRC1 were further screened through logistic regression analysis to establish a nomogram, which had good diagnostic performance and potential utility in the clinical decision-making of HCC. Nonetheless, the clinical potential of NPK76-II-72-1 as a targeted therapeutic agent requires further investigation. This study may provide a novel insight into the targeted therapy of HCC.

Data availability

The publicly available database analyzed in this study can be found in the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>) and the UCSC Xena platform (<https://xenabrowser.net/datapages/>).

Received: 2 January 2025; Accepted: 21 March 2025

Published online: 01 April 2025

References

- Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263. <https://doi.org/10.3322/caac.21834> (2024).
- Cao, L. Q., Xie, Y., Fleishman, J. S., Liu, X. & Chen, Z. S. Hepatocellular carcinoma and lipid metabolism: novel targets and therapeutic strategies. *Cancer Lett.* **597**, 217061. <https://doi.org/10.1016/j.canlet.2024.217061> (2024).
- Vogel, A., Meyer, T., Sapisochin, G., Salem, R. & Saborowski, A. Hepatocellular carcinoma. *Lancet* **400**, 1345–1362. [https://doi.org/10.1016/s0140-6736\(22\)01200-4](https://doi.org/10.1016/s0140-6736(22)01200-4) (2022).
- Lamarca, A., Edeline, J. & Goyal, L. How I treat biliary tract cancer. *Esmo Open*. **7** <https://doi.org/10.1016/j.esmoop.2021.100378> (2022).
- Jin, H. et al. New insights into checkpoint inhibitor immunotherapy and its combined therapies in hepatocellular carcinoma: from mechanisms to clinical trials. *Int. J. Biol. Sci.* **18**, 2775–2794. <https://doi.org/10.7150/ijbs.70691> (2022).
- Yang, L., Zhang, W. & Yan, Y. Identification and characterization of a novel molecular classification based on disulfidptosis-related genes to predict prognosis and immunotherapy efficacy in hepatocellular carcinoma. *Aging (Albany NY)*. **15**, 6135–6151. <https://doi.org/10.18632/aging.204809> (2023).
- Wang, T. et al. Comprehensive molecular analyses of a Macrophage-Related gene signature with regard to prognosis, immune features, and biomarkers for immunotherapy in hepatocellular carcinoma based on WGCNA and the LASSO algorithm. *Front. Immunol.* **13**, 843408. <https://doi.org/10.3389/fimmu.2022.843408> (2022).
- Huang, C. et al. A new finding in the key prognosis-related proto-oncogene FYN in hepatocellular carcinoma based on the WGCNA hub-gene screening strategy. *BMC Cancer*. **22**, 380. <https://doi.org/10.1186/s12885-022-09388-5> (2022).
- Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–d677. <https://doi.org/10.1093/nar/gkac909> (2025).
- The Gene Ontology Resource. 20 Years and still going strong. *Nucleic Acids Res.* **47**, D330–d338. <https://doi.org/10.1093/nar/gky1055> (2019).
- Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127. <https://doi.org/10.1093/biostatistics/kxj037> (2007).
- Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559. <https://doi.org/10.1186/1471-2105-9-559> (2008).
- Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*. **16**, 1289–1296. <https://doi.org/10.1038/s41592-019-0619-0> (2019).
- Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401. <https://doi.org/10.1126/science.1254257> (2014).
- Belgiu, M. & Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogrammetry Remote Sens.* **114**, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011> (2016).
- Degenhardt, F., Seifert, S. & Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* **20**, 492–503. <https://doi.org/10.1093/bib/bbx124> (2019).
- Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422. <https://doi.org/10.1023/A:1012487302797> (2002).
- Sanz, H., Valim, C., Vegas, E., Oller, J. M. & Reverter, F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinform.* **19**, 432. <https://doi.org/10.1186/s12859-018-2451-4> (2018).
- Engelbrechtsen, S. & Böhlin, J. Statistical predictions with Glimnet. *Clin. Epigenetics*. **11**, 123. <https://doi.org/10.1186/s13148-019-0730-1> (2019).
- Liu, C. J. et al. GSCA: an integrated platform for gene set cancer analysis at genomic, Pharmacogenomic and Immunogenomic levels. *Brief. Bioinform.* **24** <https://doi.org/10.1093/bib/bbac558> (2023).
- Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589. <https://doi.org/10.1038/s41586-021-03819-2> (2021).
- Varadi, M. et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–d444. <https://doi.org/10.1093/nar/gkab1061> (2022).
- Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–d1380. <https://doi.org/10.1093/nar/gkac956> (2023).
- Chandrashekar, D. S. et al. An update to the integrated cancer data analysis platform. *Neoplasia* **25**, UALCAN, 18–27. <https://doi.org/10.1016/j.neo.2022.01.001> (2022).
- Wang, H., Guo, M., Wei, H. & Chen, Y. Targeting p53 pathways: mechanisms, structures, and advances in therapy. *Signal. Transduct. Target. Ther.* **8**, 92. <https://doi.org/10.1038/s41392-023-01347-1> (2023).
- Chen, X. et al. Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death Dis.* **13**, 974. <https://doi.org/10.1038/s41419-022-05408-1> (2022).
- Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63. <https://doi.org/10.1038/nrg2484> (2009).
- Russo, G., Zegar, C. & Giordano, A. Advantages and limitations of microarray technology in human cancer. *Oncogene* **22**, 6497–6507. <https://doi.org/10.1038/sj.onc.1206865> (2003).

29. Wilhelm, B. T. & Landry, J. R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257. <https://doi.org/10.1016/j.ymeth.2009.03.016> (2009).
30. Ma, D. et al. A novel approach for the analysis of single-cell RNA sequencing identifies TMEM14B as a novel poor prognostic marker in hepatocellular carcinoma. *Sci. Rep.* **13**, 10508. <https://doi.org/10.1038/s41598-023-36650-y> (2023).
31. Tian, Z. et al. Constructing a disulfidptosis-related prognostic signature of hepatocellular carcinoma based on single-cell sequencing and weighted co-expression network analysis. *Apoptosis* **29**, 1632–1647. <https://doi.org/10.1007/s10495-024-01968-z> (2024).
32. Du, H. et al. Explainable machine learning models for early gastric cancer diagnosis. *Sci. Rep.* **14**, 17457. <https://doi.org/10.1038/s41598-024-67892-z> (2024).
33. Cress, W. D., Yu, P. & Wu, J. Expression and alternative splicing of the cyclin-dependent kinase inhibitor-3 gene in human cancer. *Int. J. Biochem. Cell. Biol.* **91**, 98–101. <https://doi.org/10.1016/j.biocel.2017.05.013> (2017).
34. Jiang, C. H. et al. Bioinformatics-based screening of key genes for transformation of liver cirrhosis to hepatocellular carcinoma. *Journal of Translational Medicine*. **18**; (2020). <https://doi.org/10.1186/s12967-020-02229-8>
35. Xing, C. Y. et al. Cyclin-dependent kinase inhibitor 3 is overexpressed in hepatocellular carcinoma and promotes tumor cell proliferation. *Biochem. Biophys. Res. Commun.* **420**, 29–35. <https://doi.org/10.1016/j.bbrc.2012.02.107> (2012).
36. Xing, C. et al. Cyclin-dependent kinase inhibitor 3 is overexpressed in hepatocellular carcinoma and promotes tumor cell proliferation. *Biochem. Biophys. Res. Commun.* **420**, 29–35. <https://doi.org/10.1016/j.bbrc.2012.02.107> (2012).
37. Cheng, S. et al. Downregulation of Peptidylprolyl isomerase A promotes cell death and enhances doxorubicin-induced apoptosis in hepatocellular carcinoma. *Gene* **591**, 236–244. <https://doi.org/10.1016/j.gene.2016.07.020> (2016).
38. Wang, S. L., Li, M. H., Xing, L. G. & Yu, J. M. High expression level of Peptidylprolyl isomerase A is correlated with poor prognosis of liver hepatocellular carcinoma. *Oncol. Lett.* **18**, 4691–4702. <https://doi.org/10.3892/ol.2019.10846> (2019).
39. Gu, Y. W. et al. A critical role of Peptidylprolyl isomerase A pseudogene 22/microRNA-197-3p/Peptidylprolyl isomerase A axis in hepatocellular carcinoma. *Front. Genet.* **12** <https://doi.org/10.3389/fgene.2021.604461> (2021).
40. Zhang, P. et al. CENPW knockdown inhibits progression of bladder cancer through inducing cell cycle arrest and apoptosis. *J. Cancer*. **15**, 858–870. <https://doi.org/10.7150/jca.90449> (2024).
41. Ho, C. M. et al. Prognostic comparative genes predict targets for Sorafenib combination therapies in hepatocellular carcinoma. *Comput. Struct. Biotechnol. J.* **20**, 1752–1763. <https://doi.org/10.1016/j.csbj.2022.04.008> (2022).
42. Zhou, Y. et al. Knockdown of CENPW inhibits hepatocellular carcinoma progression by inactivating E2F signaling. *Technol. Cancer Res. Treat.* **20**, 15330338211007253. <https://doi.org/10.1177/15330338211007253> (2021).
43. Yao, F. W. et al. Single-Cell RNA sequencing reveals the role of Phosphorylation-Related genes in hepatocellular carcinoma stem cells. *Front. Cell. Dev. Biology.* **9** <https://doi.org/10.3389/fcell.2021.734287> (2022).
44. Li, J., Dallmayer, M., Kirchner, T., Musa, J. & Grunewald, T. G. P. PRC1: linking cytokinesis, chromosomal instability, and cancer evolution. *Trends Cancer*. **4**, 59–73. <https://doi.org/10.1016/j.trecan.2017.11.002> (2018).
45. Zhang, C. et al. Protein Regulator of Cytokinesis 1 (PRC1) Upregulation Promotes Immune Suppression in Liver Hepatocellular Carcinoma. *J Immunol Res.* 7073472; (2022). <https://doi.org/10.1155/2022/7073472> (2022).
46. Huang, J., Cheng, X., Wang, C. & Gong, F. Protein regulator of cytokinesis 1 regulates autophagy in hepatitis B virus-associated liver cancer development. *Oncol. Rep.* **53** <https://doi.org/10.3892/or.2025.8869> (2025).
47. McGarry, T. J. & Kirschner, M. W. Geminin, an inhibitor of DNA replication, is degraded during mitosis. *Cell* **93**, 1043–1053. [https://doi.org/10.1016/s0092-8674\(00\)81209-x](https://doi.org/10.1016/s0092-8674(00)81209-x) (1998).
48. Kushwaha, P. P., Rapalli, K. C. & Kumar, S. Geminin a multi task protein involved in cancer pathophysiology and developmental process: A review. *Biochimie* **131**, 115–127. <https://doi.org/10.1016/j.biochi.2016.09.022> (2016).
49. Yoshida, K. Geminin as a molecular target for the development of new anticancer drugs. *Mini Rev. Med. Chem.* **6**, 461–462. <https://doi.org/10.2174/138955706776361411> (2006).
50. Abraham, N. M. M. et al. Evaluation of BUBR1, MCM2, and GMNN as oral cancer biomarkers. *Eur. J. cancer Prevention: Official J. Eur. Cancer Prev. Organisation (ECP)*. <https://doi.org/10.1097/cej.0000000000000932> (2024).
51. Zhao, X. Z. et al. High expression of GMNN predicts malignant progression and poor prognosis in ACC. *European Journal of Medical Research*. **27**; (2022). <https://doi.org/10.1186/s40001-022-00950-2>
52. Chen, W. et al. The signature of immune-subtype specific driving transcription factors suggest potential drugs for refractory glioblastoma. *Am. J. Cancer Res.* **13**, 1278–1294 (2023).

Author contributions

G.W. and JX. Z. analyzed and interpreted the data and were major contributors to writing the manuscript. YR. L., WW. D., YY. Z., and HQ. W. collected the data and validated the results. JL. W., PQ. L., and ZQ. Y. reviewed the manuscript. T. L. and WT. H. supervised the study and revised the manuscript. All authors read and approved the final manuscript.

Funding

The research was funded by the Science and Technology Program of Gansu Province, No. 23JRRA1015.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-95493-x>.

Correspondence and requests for materials should be addressed to T.L. or W.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025