Heliyon 10 (2024) e24854

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Construction of immune-related molecular diagnostic and predictive models of hepatocellular carcinoma based on machine learning

Hui Zheng [a], Xu Han [c], Qian Liu [a], Li Zhou [c], Yawen Zhu [c], Jiaqi Wang [a], Wenjing Hu [a], Fengcai Zhu [a,b,c,**], Ran Liu [a,*]

[a] *Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing, 210009, China*

[b] *National Health Commission Key Laboratory of Enteric Pathogenic Microbiology, Jiangsu Provincial Center for Disease Control and Prevention, Nanjing, Jiangsu Province, China*

[c] *School of Public Health, Nanjing Medical University, Nanjing, Jiangsu Province, China*

ARTICLE INFO

Keywords:
Hepatocellular carcinoma
WGCNA
Machine-learning
LASSO
Random forest
Biomarker analysis
Immune infiltration

ABSTRACT

Background: To exploit hepatocellular carcinoma (HCC) diagnostic substances, we identify potential predictive markers based on machine learning and to explore the significance of immune cell infiltration in this pathology.

Method: Three HCC gene expression datasets were used for weighted gene co-expression network analysis (WGCNA) and differential expression analysis. Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest were applied to identify candidate biomarkers. The diagnostic value of HCC diagnostic gene biomarkers was further assessed by the area under the ROC curve observed in the validation dataset. CIBERSORT was used to analyze 22 immune cell fractions from HCC patients and to analyze their correlation with diagnostic markers. In addition, the prognostic value of the markers and the sensitivity of the drugs were analyzed.

Result: WGCNA and differential expression analysis were used to screen 396 distinct gene signatures in HCC tissues. They were mostly engaged in cytoplasmic fusion and the cell division cycle, according to gene enrichment analyses. Five genes were shown to have a high diagnostic value for use as diagnostic biomarkers for HCC, including EFHD1 (AUC = 0.77), KIF4A (AUC = 0.97), UBE2C (AUC = 0.96), SMYD3 (AUC = 0.91), and MCM7 (AUC = 0.93). T cells, NK cells, macrophages, and dendritic cells were found to be related to diagnostic markers in HCC tissues by immune cell infiltration analysis, indicating that these cells are intimately linked to the onset and spread of HCC. Concurrently, these five genes and their constructed models have considerable prognostic value.

Conclusion: These five genes (EFHD1, KIF4A, UBE2C, SMYD3, and MCM7) may serve as new candidate molecular markers for HCC, providing new insights for future diagnosis, prognosis, and molecular therapy of HCC.

* Corresponding author.

** Corresponding author. National Health Commission Key Laboratory of Enteric Pathogenic Microbiology, Jiangsu Provincial Center for Disease Control and Prevention, Nanjing, 210009, China.

E-mail addresses: jszfc@vip.sina.com (F. Zhu), ranliu@seu.edu.cn (R. Liu).

https://doi.org/10.1016/j.heliyon.2024.e24854
Received 13 August 2023; Received in revised form 10 January 2024; Accepted 16 January 2024
Available online 17 January 2024

1. Introduction

Hepatocellular carcinoma accounts for 90 % of primary liver cancers and is the fifth most common cancer worldwide and the fourth most common cause of cancer-related deaths [1,2]. The current treatment of liver cancer mainly includes surgery and intervention. However, the high incidence and aggressiveness of hepatocellular carcinoma and the unclear underlying mechanisms of hepatocellular carcinoma are not conducive to an improved prognosis, with approximately 70 % of hepatocellular carcinomas recurring within five years of resection or ablation [3,4].

The current dilemma is that hepatocellular carcinoma does not show obvious clinical symptoms at an early stage, which is not conducive to early detection and treatment of hepatocellular carcinoma, nor to improving prognosis and reducing mortality [5]. In addition, the commonly used clinical methods of liver cancer diagnosis, such as serum tumor markers and imaging techniques, are not ideal in terms of their effectiveness [6,7]. Therefore, targeting features associated with the early diagnosis, invasion, and metastasis of hepatocellular carcinoma is greatly important in improving the treatment outcome and prognosis.

Increasingly, bioinformatics techniques are being applied to the medical field, such as aiding the diagnosis of diseases and uncovering pathogenic mechanisms. Weighted gene co-expression network analysis (WGCNA) is a powerful systems biology approach for analyzing network relationships and molecular mechanisms and is widely used to analyze large amounts of gene expression profiling data [8], widely used to identify gene function and associations between genes and clinical traits, as well as important modules associated with disease occurrence [9]. Machine learning is then used to extract and mine key molecular features for constructing diagnostic models with great accuracy and efficiency [10]. In addition, immune escape has a vital role in developing hepatocellular carcinoma. Analyzing the relationship between critical genes and tumor infiltration will help elucidate the immune escape mechanism of tumours [11].

In this study, a bioinformatics approach was used for differential gene enrichment analysis of RNA gene matrices of HCC tissues. Two machine learning algorithms were used to screen biomarkers associated with HCC and to further validate biomarkers closely associated with immune infiltration. *CIBERSORT* was employed to quantify the proportion of immune cells in HCC and normal tissue samples based on gene expression profiles and to explore the correlation between the identified biomarkers and infiltrating immune cells. The constructed Cox regression models of the markers have high prognostic value and provide new ideas for further prevention and treatment of HCC.

2. Materials and methods

2.1. Hepatocellular carcinoma data preparation

Two independent hepatocellular carcinoma gene expression profiles, GSE14520 and GSE112791, containing 427 hepatocellular carcinoma samples and 248 normal samples from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) were collected [12,13]. The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/) provides the TCGA-LIHC dataset with 50 normal and 371 hepatocellular carcinoma samples [14]. For the datasets, normalization was performed using Variance Stabilised Normalization (VSN) and Uniform Manifold Approximation and Projection (UMAP) for downscaling and quality control.

2.2. Differentially expressed gene analysis

To obtain differentially expressed genes (DEGs) between normal and tumor tissue, GSE14520 was analyzed using the R package *limma* [15]. Genes with a cut-off criterion of $|\log2FC| \geq 1.0$ and an adjusted p-value <0.05 were identified as DEGs after converting gene expression fold change (FC) to logarithmic values. The Benjamini-Hochberg method adjusted p-values to control for false discovery rate (FDR). DEGs for GSE14520 were visualized as volcano plots using the R package *ggplot2* [16].

2.3. Weighted gene Co-expression network analysis

The present study used the *WGCNA* package in R to convert gene expression data into a gene coexpression network to investigate highly correlated gene modules [17]. The function pickSoftThreshold selected the soft threshold for GSE14520. The formula then created the adjacency matrix: $a_{ij} = |S_{ij}|^{\beta}$ (a_{ij}: adjacency matrix between gene I and gene J, S_{ij}: similarity matrix calculated using Pearson correlation of all gene pairs, β: soft threshold). The adjacency matrix was then converted into a topological overlap matrix (TOM) and a corresponding difference matrix (1-TOM). Afterward, a hierarchical clustering dendrogram of 1-TOM matrices was used to classify related genes into various coexpression modules. To determine the association of modules with traits, the clinical traits of the samples were defined as normal and tumor tissue. Therefore, modules with high correlation coefficients were selected for the subsequent studies as they are likely to be closely associated with hepatocellular carcinoma.

2.4. GO enrichment analysis of differentially expressed genes and coexpression modules

The *clusterProfiler* is an available package in the R software for the analysis of GO ontologies, including biological processes (BP), molecular functions (MF), and cellular components (CC) [18]. Based on multiple Bioconductor annotation resources and the R package, the *clusterProfiler* package is widely used for bioinformatics analysis. P-value cut-offs were set at 0.05, and terms with p-values less than or equal to 0.05 were included. False discovery rates were applied to the adjustment of P-values.

2.5. Identification and enrichment analysis of gene signature

Intersections between DEG lists and co-expressed gene modules were used to identify gene signatures, displayed as Venn diagrams via the R package *VennDiagram*. The Metascape platform (http://metascape.prg/gp/index.html) has comprehensive annotation capabilities. Gene signatures were entered into the Metascape platform for enrichment analysis of their major biological processes (BPs).

2.6. Random forest analysis

Random Forest is a versatile and accessible machine-learning algorithm often used for classification and regression tasks [19]. As it has no restrictions on variable conditions, it has higher accuracy, sensitivity, and specificity than decision trees in predicting continuous variables and obtains predictions without significant bias. It is, therefore, a suitable prediction method for the data in this study. Here, dataset GSE14520 was used as the training set, and dataset GSE112791 as the validation set. The *randomForest* package in R was used to implement the random forest algorithm to select the feature variables [20].

2.7. Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) is a compression estimator known as the minimum absolute shrinkage and selection operator. It obtains a more refined model by constructing a penalty function that allows it to compress some of the regression coefficients, forcing the sum of the absolute values of the coefficients to be less than a fixed value while setting some of the regression coefficients to zero. It, therefore, retains the advantage of subset shrinkage, is a biased estimator for dealing with data with complex covariances, and is another predictive model suitable for our study data. In this work, the training dataset was GSE14520, and the validation dataset was GSE112791. The LASSO regression model was constructed using the *glmnet* package in R to pick the feature variables [21].

2.8. Prediction performance evaluation

To identify significant predictor variables, variables from two machine learning algorithms (random forest and LASSO) were selected for gene signatures. Their predictive performance was assessed using the subject operating curve ROC. Subsequently, to improve the diagnostic power of the model, we included the marker mentioned above in the logistic regression model and further assessed the model's accuracy in the TCGA-LIHC dataset as an additional validation set. In particular, the *pROC* package [22] in R was used to calculate the AUC, while the *glmnet* package in R was used to construct the logistic regression model.

2.9. Determination, evaluation, and correlation analysis of infiltrated immune cells

The infiltration of 22 immune cells in hepatocellular carcinoma tissue in the dataset GSE14520 was analyzed in R using the *CIBERSORT* package [23]. The relative abundance of infiltrating immune cells was obtained based on $P < 0.05$. Differential infiltration of immune cells in hepatocellular carcinoma and normal tissue was then explored using the Wilcoxon rank sum test. Finally, Spearman's relationship between marker genes and infiltrating immune cells were analyzed. In addition, tumor purity was compared between tumor and normal tissues using the R package *estimate* package [24]. The results were visualized by the *ggplot2* and *pheatmap* packages of R software.

2.10. Survival and mutation analysis of feature variables

Based on the Kaplan-Meier method, the overall survival of patients who donated hepatocellular carcinoma samples in the TCGA-LIHC dataset was assessed by the R package *survival* and *survminer*. The significance of survival differences between various groups was assessed using the log-rank test, with the median expression of the marker genes serving as the grouping criterion. We then used Cox regression models to analyze whether the characteristic variables independently predicted survival in patients with hepatocellular carcinoma. Using the R package *maftools*, marker genes in the TCGA-LIHC mutation data are analyzed and visualized for mutation status [25].

2.11. Consensus clustering of feature variables

To explore the prognostic value of five feature variables in hepatocellular carcinoma, we performed consensus clustering using the *ConsensusClusterPlus* package of R [26]. The study used the following parameters: 1000 repeats, $k = 10$, and agglomerative hierarchical clustering with ward criterion (Ward.D2) inner and complete outer linkage. Genes used for consensus clustering analyses are the five feature variables. Then, Kaplan-Meier curves were plotted to confirm the prognostic value of the cluster classification.

2.12. *Construction of survival models for feature variables*

Multivariate Cox proportional hazards regression analysis was performed to obtain the coefficients for these feature variables. The feature variables-related prognostic signature was constructed based on the coefficients of multivariate Cox regression analysis weighted with the expression of these selected genes. The detailed formula was shown as follows: Risk score = β1 * Exp1 + β2 * Exp2 + βi * Exp. β and Exp represent the coefficients from the multivariate Cox proportional hazards regression analysis and the expression levels of selected genes, respectively. According to the median risk score value, patients were then classified into high- and low-risk subgroups. Kaplan-Meier analysis and log-rank test for prognostic evaluation were performed using the *survminer* package.

2.13. *Chemotherapy sensitivity predictions*

The *oncoPredict* package (v2) in R was used to calculate the 198 drugs (including fluorouracil, epirubicin, camptothecin, Cisplatin,

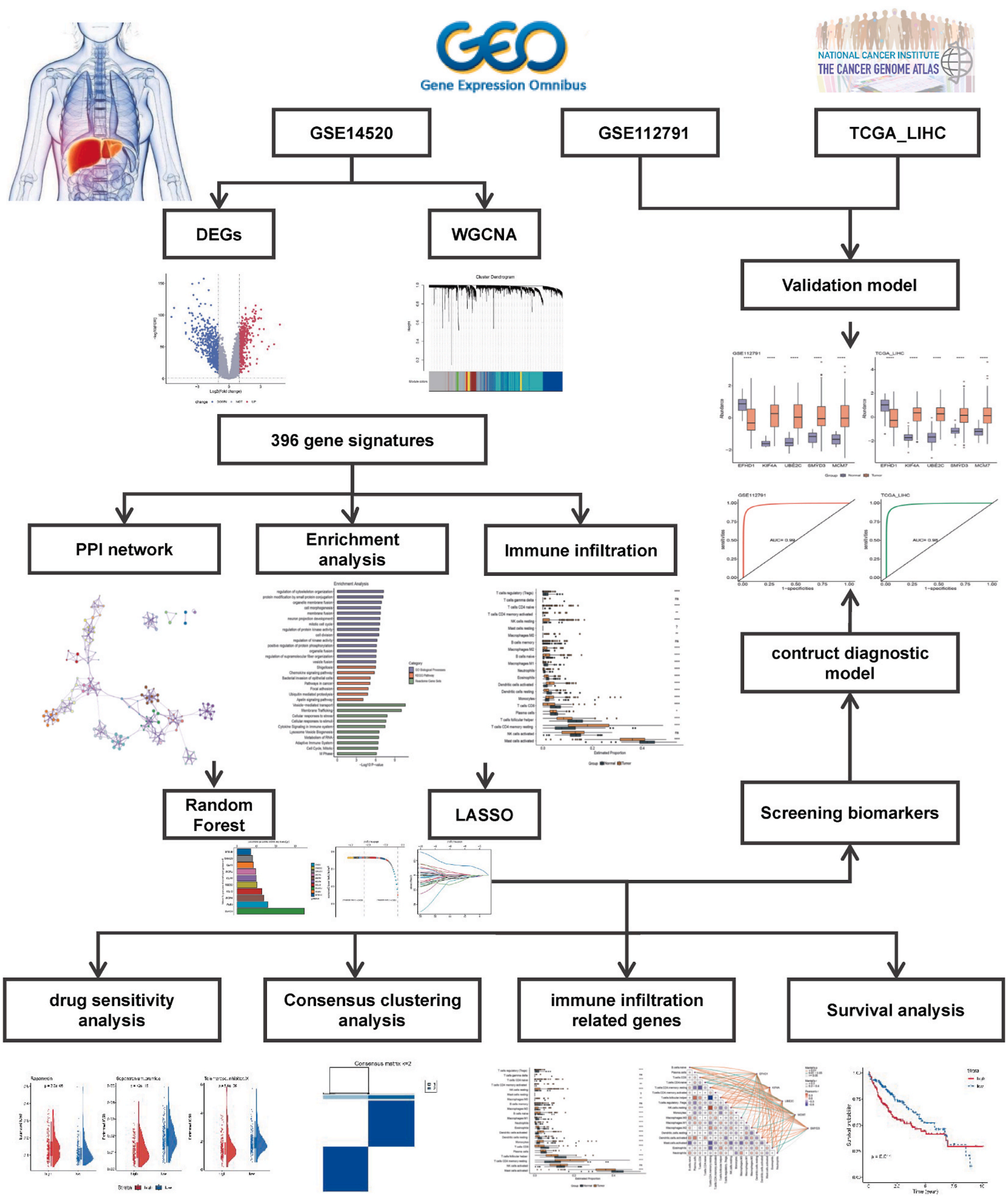

Fig. 1. Workflow of this study.

Cytarabine, Bortezomib, Dactinomycin, Staurosporine, Vinblastine, etc.) in each HCC sample by calculating the Half maximum inhibitory concentration (IC50) values to predict chemosensitivity [27]. The method was based on Genomics of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org/) using ridge regression. The prediction accuracy was also assessed using tenfold cross-validation.

3. Results

3.1. DEGs in hepatocellular carcinoma

The workflow of this research, including data extraction, processing, analysis, and validation, is shown in Fig. 1. To distinguish significant differences between normal and tumor samples in the dataset GSE14520, a UMAP was performed to reduce the dimensionality and assess the independence of the groups. The results showed significant differences between normal and tumor samples in the data, except for a few normal samples that were close to the tumor samples (Fig. 2A). The dataset GSE14520 contained 222 tumor samples and 212 normal samples. Between tumor and normal tissue, 972 DEGs were obtained (456 up-regulated and 516 down-regulated) (Fig. 2B and C), with 46.9 % of the genes up-regulated in tumours, while 53.1 % of the genes were down-regulated.

3.2. Coexpression modules identified by WGCNA

Coexpression analysis was performed on the dataset GSE14520 to construct a coexpression network. In this study, $\beta = 9$ was chosen as the soft threshold power for dataset GSE14520 to ensure a scale-free network (Fig. 3A). Subsequently, a total of 7 modules were identified in GSE14520, indicated by different colors, with the larger the module area, the greater the number of genes contained (Fig. 3B and C). Next, a heat map of module-trait relationships was created to assess the correlation between each module and clinical

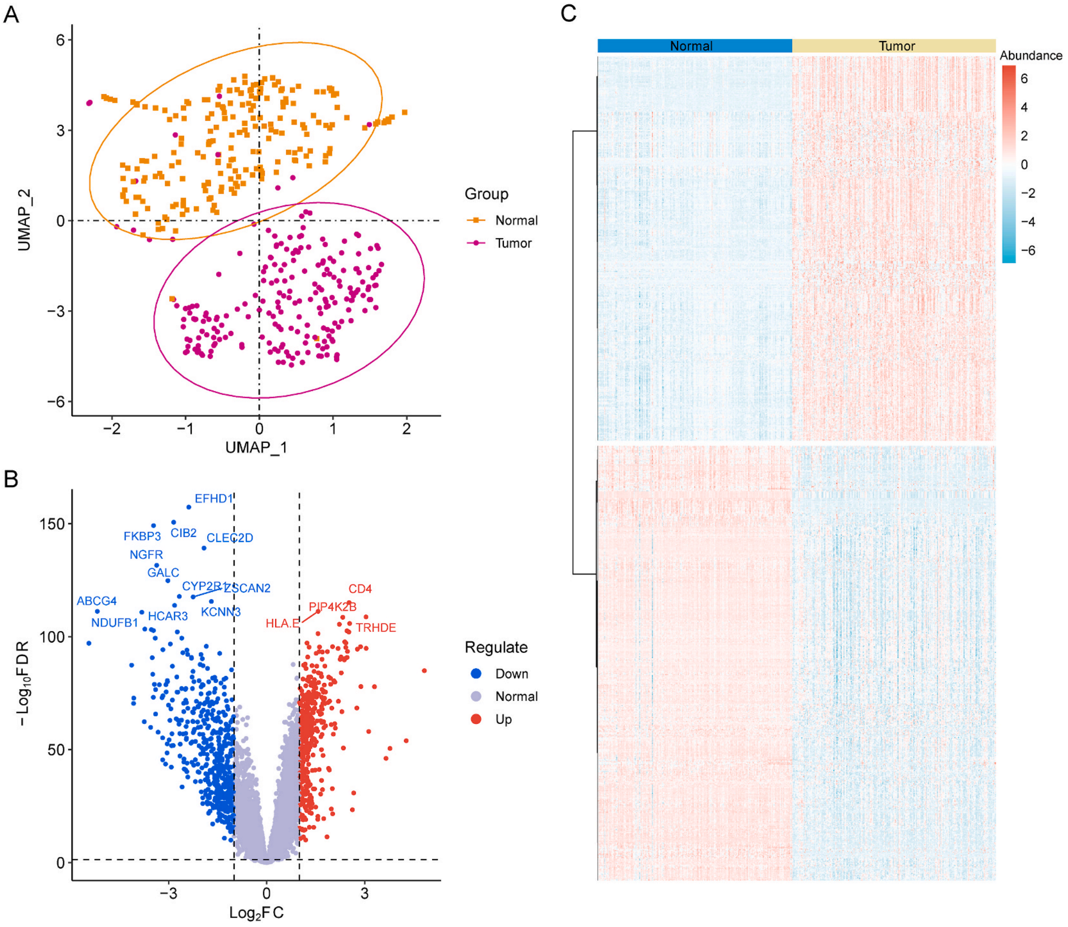

Fig. 2. Results of expression difference analysis of microarray studies (A) UMAP multivariate statistical model of the normal and tumor tissue. (B) Volcano plots showing DEGs in normal and tumor tissue in the GSE14520. Red dots indicate genes highly induced in tumor tissue, blue dots indicate genes significantly reduced in tumor tissue, and grey dots indicate non-DEGs. (C) Heatmap of 972 DEGs between normal and tumor tissue. Red represents increased expression and cyan represents decreased expression.

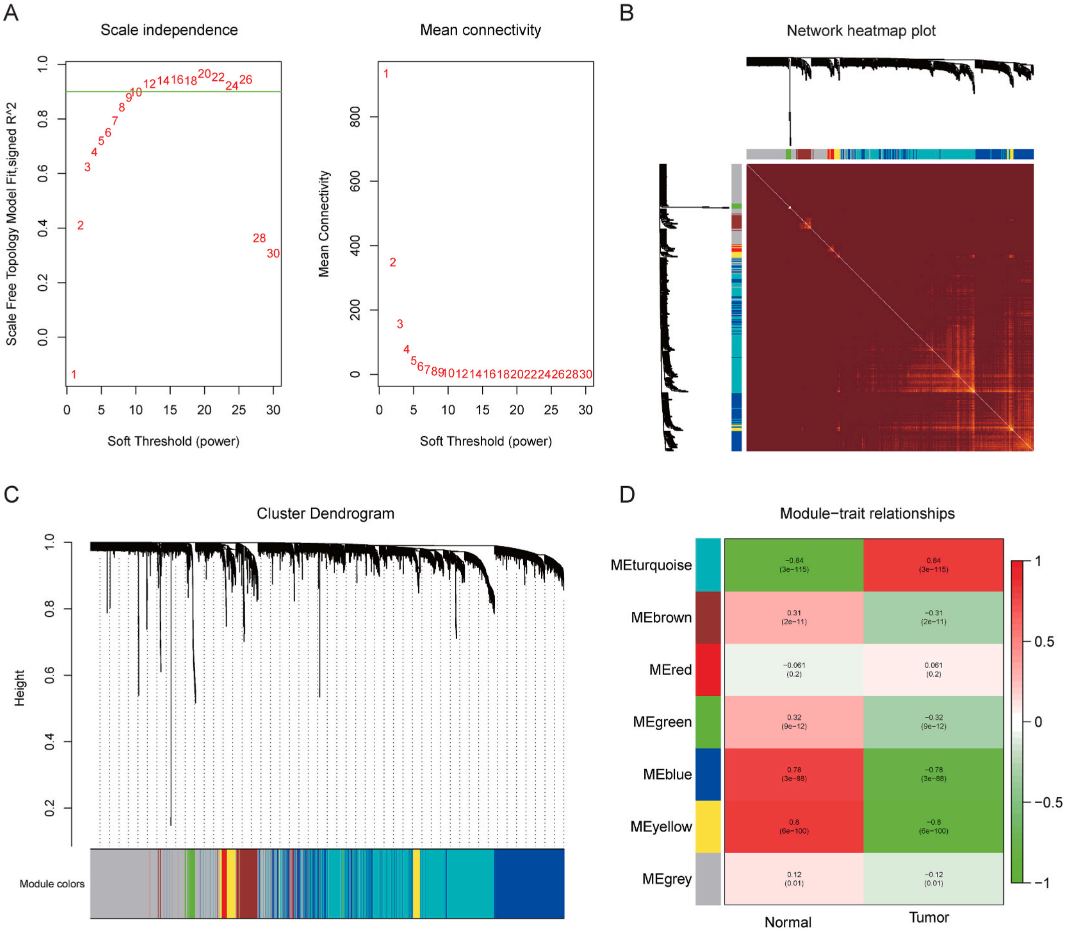

Fig. 3. WGCNA analysis was performed on the GSE14520 dataset to identify modules associated with clinical features (A) Soft-thresholding calculation of GSE14520; Left: scale-free fit indices using various soft-thresholding powers; Right: mean connectivity using various soft-thresholding powers. (B) A heatmap plot of topological overlap in the gene network was shown. (C) Cluster dendrograms of the coexpression network modules of dataset GSE14520 are ordered by hierarchical clustering of genes based on a 1-TOM matrix. Each module is colored differently. (D) Heatmap of module-trait relationships for datasets GSE14520. Each row corresponds to a module, and a column corresponds to a clinical trait (normal or tumor). Each cell includes the corresponding correlation and p-value.

traits (Fig. 3D). The turquoise module (P-value = 3E-115) in GSE14520 showed the most significant correlation with traits. Ultimately, this module was selected as clinically significant for further analysis.

3.3. *Enrichment analysis of DEGs, coexpression modules, and gene signature*

GO analysis consists of three sub-ontologies, including biological processes (BP), cellular components (CC), and molecular functions (MF). The GO analysis of 972 DEGs from dataset GSE14520 showed that in terms of BP, the prominent enrichment was in apoptotic signaling pathways and regulation of adiponectin activity; in terms of CC, the central enrichment was in the composition of the cell membrane; and in terms of MF, the primary enrichment was in the activity of nucleases and transcription factors (Figs. S1A, B, C). Furthermore, GO analysis of 1624 co-expressed genes in dataset GSE14520 showed that in BP, the prominent enrichment was in protein organization, catabolism, and modification processes; in CC, the apparent enrichment was in the composition of the cell membrane and associated vesicles. As for MF, it was mainly enriched in transcription factor binding (Figs. S1D, E, F).

Subsequently, a venogram analysis of the DEGs and co-expressed genes in the dataset yielded 396 genes (Fig. 4A). Of these genes, 22 were highly expressed in normal tissues and low in tumor tissues, while the other 374 genes were conversely low in normal tissues and high in tumor tissues (Fig. 4B). It also shows that genes closely related to tumours have increased expression in tumours and only some of them have decreased expression. Enrichment analysis of these 396 genes showed that they were mainly enriched in the fusion of the cytosol and the cell division cycle (Fig. 4C). Furthermore, network analysis revealed that these pathways were closely related (Fig. 4D).

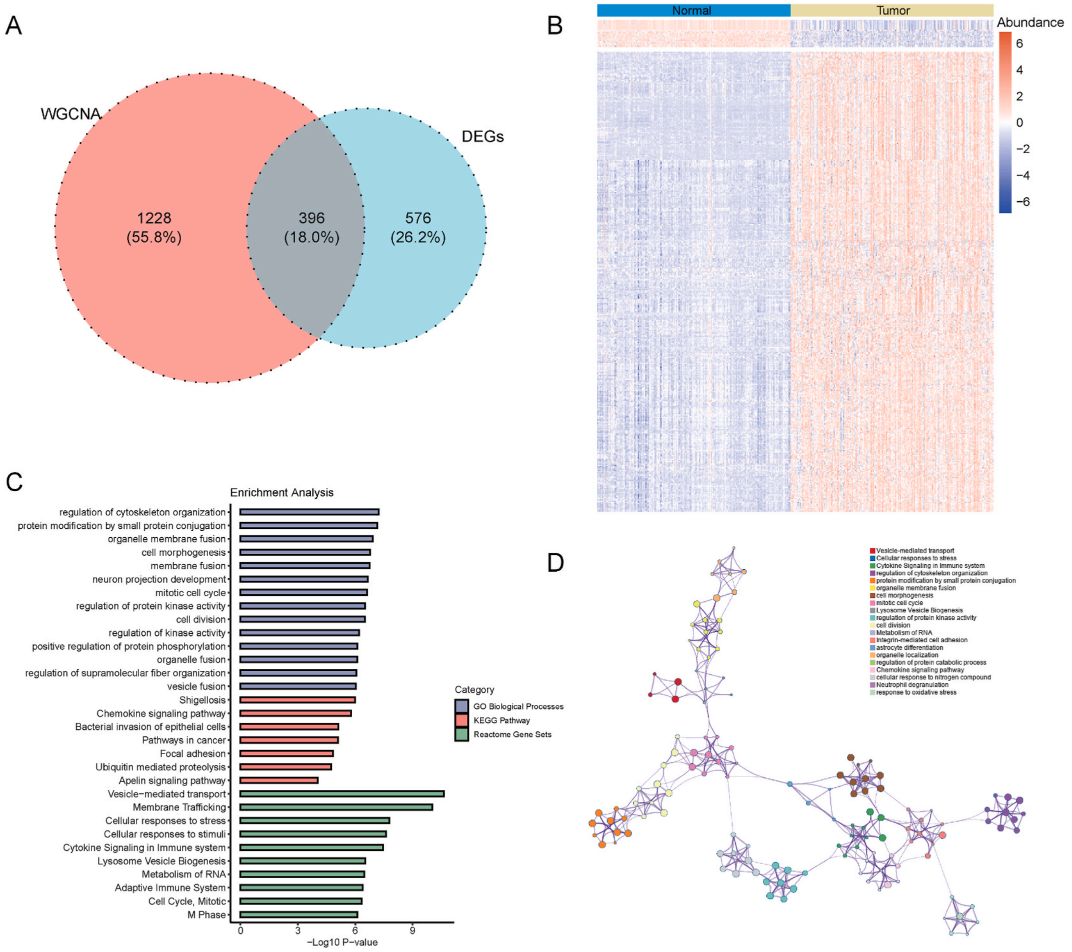

Fig. 4. Enrichment analysis of overlapping genes between differentially expressed genes (DEGs) and coexpression modules in the GSE14520 (A) The Venn diagram of genes from the DEGs lists and coexpression modules. (B) Heat map of 396 overlapping genes. The horizontal axis shows the samples, while the vertical axis shows the genes. Red represents increased expression and blue represents decreased expression. (C) Enrichment terms were identified by Metascape analysis of 396 overlapping genes. (D) Relationships among these enrichment terms are displayed as a network (Metascape). Each term is represented by a circular node whose size is proportional to the number of input genes under the term and whose color represents its clustering identity (i.e., nodes of the same color belong to the same cluster). An edge connects terms with a similarity greater than 0.3 (the thickness of the edge represents the similarity score). The network was visualized with Cytoscape using a 'force-directed' layout, with the edges bundled for clarity. One term from each cluster was selected, and its term description was displayed as a label.

3.4. Construction of a classification model for hepatocellular carcinoma

Classification prediction models were constructed by random forest to screen for combinations of variables with highly discriminatory patterns to distinguish known classifications. Random forest models were made using the R package *randomForest*, and the importance of 396 pivotal genes was assessed. After ten-fold cross-validation of the dataset GSE14520, we plotted the relationship between model error and the number of pivot genes used for fitting. We found that the error gradually increased as the number of pivot genes increased but remained relatively stable up to a pivot gene count of 10. This finding suggests that maintaining ten significant pivot genes gives the desired regression results, as the error is relatively small. The top 10 pivotal genes were selected as substantial variables from highest to lowest, based on the '% increase in mean squared error,' which was used to assess the importance of each pivotal gene (Fig. S2A).

Next, we attempt to select the critical variables using another algorithm, LASSO regression. Here, we use the R package *glmnet* to construct the model. When filtering the variables, the larger the lambda, the smaller the corresponding estimated parameters will be compressed until a fraction of the insignificant variables will be compressed to zero, representing that the variable has been removed from the model. Subsequently, a 10-fold cross-validation method was used to select the value of λ with the smallest mean error, resulting in a model with 11 pivotal genes (Figs. S2B and C).

3.5. Establishment and performance evaluation of a classification model for hepatocellular carcinoma

By filtering significant variables by random forest and LASSO as described above, we selected five pivotal genes (EFHD1, KIF4A, UBE2C, SMYD3, MCM7) (Fig. 5A–C). Then, ROC analysis of the subject operating curves was performed for these five pivotal genes. In the three datasets, their AUC values were higher than or equal to 0.77, indicating the high predictive value of these five essential genes (Fig. 5B). Subsequently, we included these five genes in the logistic regression model. The final model results showed an AUC value of 0.98, showing excellent predictive performance (Fig. 5D).

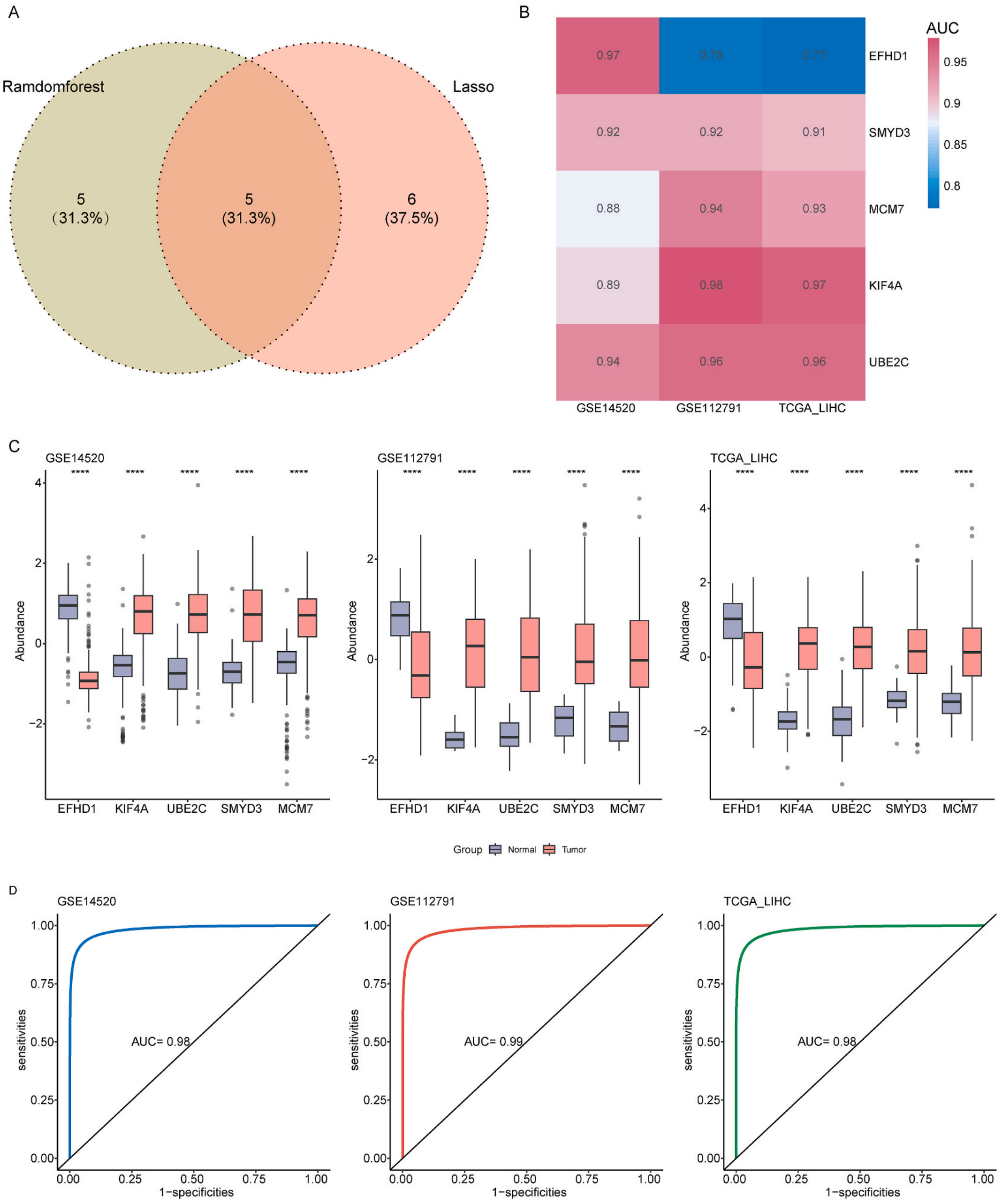

Fig. 5. Construction and performance evaluation of the predictive model (A) Venn diagram showing biomarkers based on Random Forest and LASSO regression screening. (B) AUC heatmap of 5 overlapping genes. (C) Expression levels of five gene biomarkers in normal and tumor tissue of the GSE14520, GSE112791, and TCGA-LIHC datasets. ****P-value<0.0001. (D) The area under the curve (ROC) was obtained by logistic regression fitting for five overlapping genes in the GSE14520, GSE112791, and TCGA-LIHC datasets.

3.6. Analysis of immune cell infiltration and tumor purity

Dysregulation of the immune system plays an integral role in cancer development, prompting us to explore the relationship between crucial genes and immune infiltration in hepatocellular carcinoma. The *CIBERSORT* algorithm was used to analyze 22 immune cell phenotypes in GSE14520. Significance analysis of the abundance of these 22 immune cells revealed that 14 immune cells, including T cells, NK cells, macrophages, and dendritic cells, were significantly associated with five genes each (P-value<0.001) (Fig. 6A and B, Fig. S3A). Four genes, KIF4A, UBE2C, SMYD3, and MCM7, were positively associated with the resting phase of T cells and NK cells and negatively related to the activation of Tregs and macrophages M1 (Fig. 6C). The opposite was true for the EFHD1 gene. Analysis of the tumor purity of the samples then showed that the proportion of tumor cells was significantly higher in the tumor samples than in the normal samples (P-value<0.001) (Fig. S3).

3.7. Survival and mutation analysis of five marker genes

The median expression of the five marker genes was used as a criterion to classify patients who donated hepatocellular carcinoma samples into high and low-expression groups. Based on the Kaplan-Meier method, four genes (EFHD1, UBE2C, MCM7, and KIF4A) were found to be significantly associated with survival in patients with hepatocellular carcinoma (P-value<0.05). Subsequent univariate Cox regression analysis further confirmed the negative effect of high expression of three of these genes (UBE2C, MCM7, and KIF4A) on patient survival (HR > 1) (Fig. 4). In addition, the mutation status of the five marker genes indicated that 2 % of patients had mutations in the MCM7 gene (Fig. S5).

3.8. Cluster classification of marker genes and their prognostic value

To explore the overall prognostic value of these genes, we performed a consensus clustering analysis to stratify HCC patients. It was found that k = 2 appeared to be a relatively stable value from k = 2 to 6 (Figs. S6A and B). Therefore, we classified HCC patients into two clusters. Kaplan-Meier curves showed that patients in cluster 1 had a worse prognosis than those in cluster 2 (Fig. S6C). In addition, there was a significant difference in the expression of these five genes between cluster 1 and cluster 2 (Fig. S6D).

3.9. Construction of a prognostic model based on five marker genes

Risk scores for the prognostic characteristics of HCC patients were calculated using the expression profiles of five marker genes multiplied by a multivariate Cox proportional hazards coefficient. The detailed formula is given below: Risk score =(−0.03150*EFHD1) + (0.15149*KIF4A) + (−0.04726*MCM7) + (−0.00927*SMYD3) + (0.09911*UBE2C). Patients were divided into high-risk and low-risk groups using the median risk score. The results showed that high-risk patients had a poorer prognosis compared to low-risk patients (Fig. 7A and B). The prognostic model assessed by ROC curves at different time points showed an AUC greater than 0.6 up to 5 years, which demonstrated a relatively promising diagnostic performance (Fig. 7C). These results demonstrate the considerable prognostic value of the five characterization genes and the models they constitute. In addition, drug sensitivity analysis revealed that six drugs (AZD7762, Dactolisib, Daporinad, Rapamycin, Sepantronium bromide, and Telomerase Inhibitor IX) had significantly different 50 % inhibitory concentrations (IC50s) between high- and low-risk groups, which could be valuable in guiding the use of medication or drug development for HCC patients (Fig. 7D).

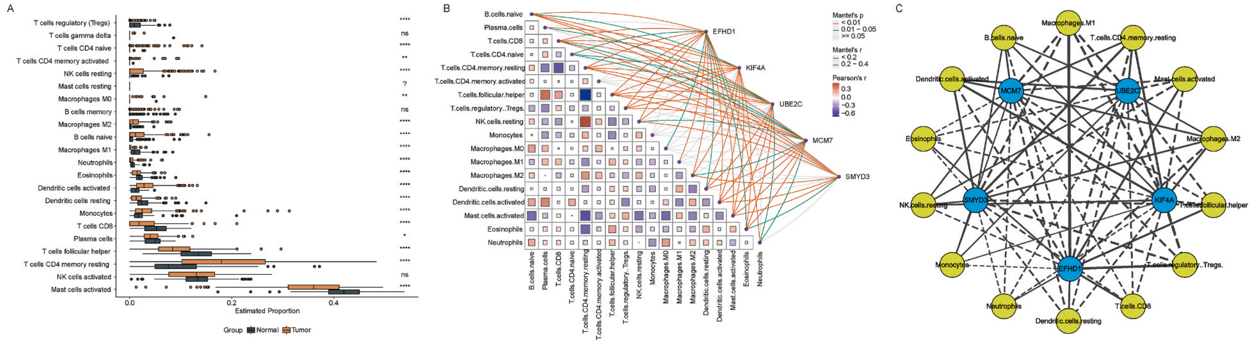

Fig. 6. Immune cell infiltration analysis and relationships between marker genes and immune cells in hepatocellular carcinoma (A) Box-plot of the proportion of 22 types of immune cells. *P-value<0.05, **P-value<0.01, *** P-value<0.001, ****P-value<0.0001, ns P-value≥0.05. (B) Heatmap of correlation in different immune cells. The color of the squares represents the strength of the correlation; red represents a positive correlation, and blue represents a negative correlation. Darker color implies a stronger association. (C) Network diagram of interactions between marker genes and immune cells. The orange circles represent immune cells, and the blue circles represent marker genes. The solid lines represent positive correlations, while the dashed lines represent negative correlations. The thicker the line, the stronger the correlation between them; conversely, the weaker the correlation.

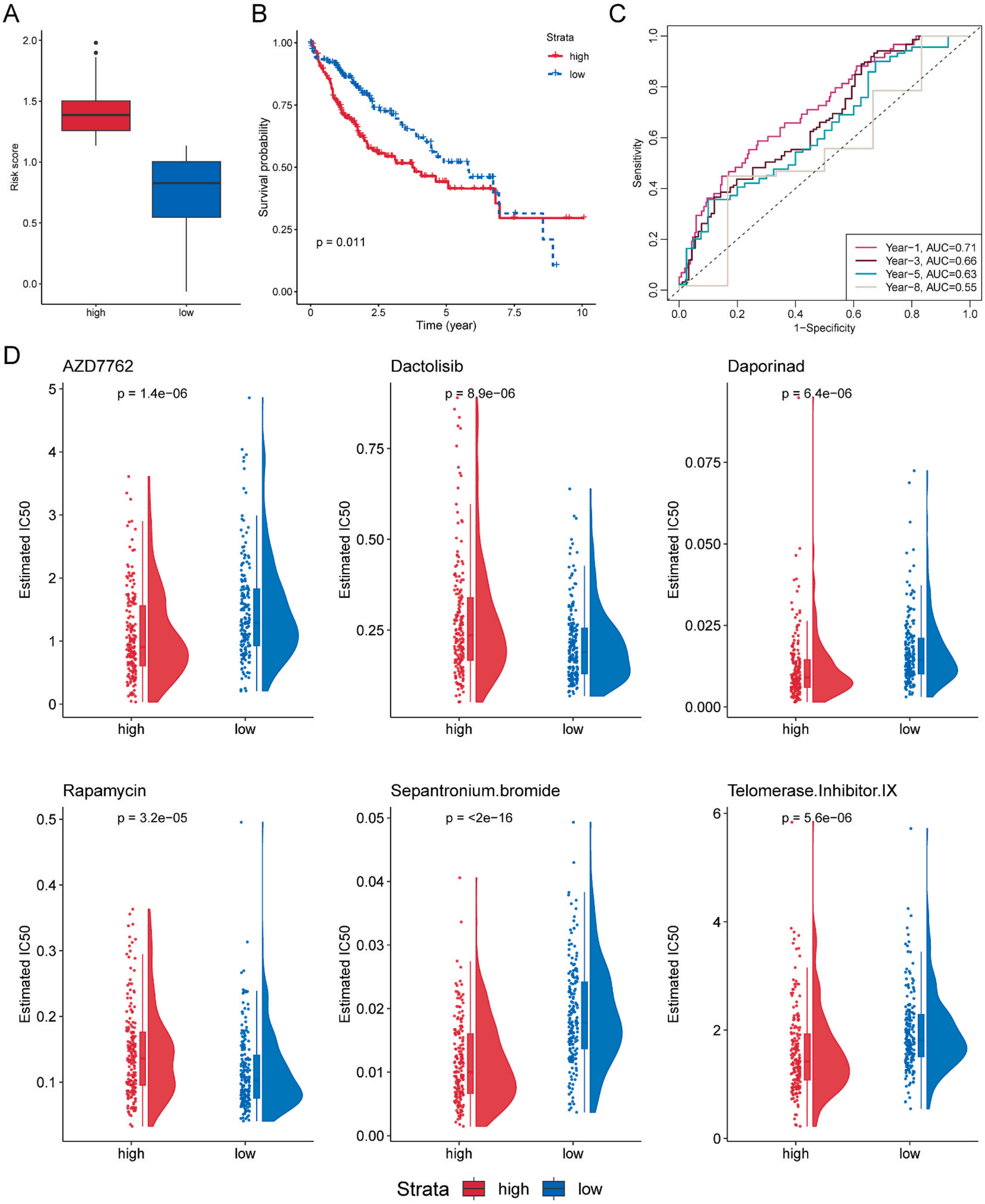

Fig. 7. Survival risk level based on five marker genes and its corresponding drug sensitivity for hepatocellular carcinoma (A) Boxplot of hepatocellular carcinoma risk score; (B) Kaplan-Meier curve of hepatocellular carcinoma risk score; (C) ROC curves to assess the diagnostic performance of prognostic models at different time points; (D) Drug sensitivity of six drugs in different risk score groups. IC50, half maximal inhibitory concentration.

4. Discussion

The development of hepatocellular carcinoma is a dynamic biological process involving multiple molecules, steps, and factors, and its mechanism is still unclear [28,29]. Tumourigenesis often involves changes in multiple genes, yet most traditional studies have focused on the possible effects of the single genes on tumours, thus resulting in a lack of in-depth understanding of the reciprocal network of tumor development involving multiple genes. Expression profiling by gene microarrays allows us to understand the expression levels of tens of thousands of genes simultaneously, which offers great potential for understanding the mechanisms of tumourigenesis and extracting markers.

This study identified 396 differentially co-expressed genes in the GSE14520 dataset using comprehensive bioinformatics analysis. Based on enrichment analysis, these genes were mainly enriched in cytoplasmic fusion and the cell division cycle. We screened and validated five marker genes by machine learning. Four genes (KIF4A, UBE2C, SMYD3, and MCM7) were up-regulated in liver cancer tissues, while only one (EFHD1) was down-regulated. T cells, NK cells, macrophages, and dendritic cells were associated with diagnostic markers in HCC tissues by immune cell infiltration analysis, suggesting that these cells are closely associated with the development and spread of HCC. The results of the ROC evaluation show that the models constructed from them have good diagnostic performance for hepatocellular carcinoma.

EFHD1 encodes a mitochondrial inner membrane protein that acts as a calcium sensor for mitochondrial flash activation. It is lowly expressed in various cancers (esophageal, clear cell renal cell carcinoma, and colorectal cancer). Its genetic variants are a potential risk factor for childhood glioblastoma multiforme [30–32]. The chromosome-associated kinesin, KIF4A, plays a role in mitotic chromosome positioning and bipolar spindle stabilization. KIF4A is overexpressed in human primary hepatocellular carcinoma, osteosarcoma, and esophageal cancer tissues. It promotes tumor cell proliferation and migration by regulating the phosphorylation levels of Hippo signaling pathway-related proteins [33–35]. DNA replication licensing factor MCM7, a component of the MCM2-7 complex (MCM complex), is a replication uncoupling enzyme. MCM7 promotes tumor proliferation and is a valuable biomarker in the early diagnosis of gastric cancer [36]. In addition, the effect of MCM7 on the survival of patients with hepatocellular carcinoma may be related to the occurrence of its mutation, which requires further observation.

Ubiquitination is one of the major post-translational modifications of proteins, playing a crucial role in many cellular functions, including protein degradation, interactions, and subcellular location. Ubiquitin-coupled enzyme E2C (UBE2C), a cell cycle-regulated ubiquitin ligase, is overexpressed in 27 common cancers, and its dysregulation contributes to the development of various cancers [37]. SMYD3, a histidine-lysine N-methyltransferase, is a regulator of epigenetic and signaling pathways in cancer and has been implicated in the development and progression of different cancer types (colorectal, hepatocellular, breast, gastric, etc.) [38,39]. The roles of SMYD3 in cancer include epithelial-mesenchymal transition, cell cycle alterations, promotion of cell proliferation, increased telomerase activity, and cell immortalization.

Understanding the pathogenesis of hepatocellular carcinoma is crucial for taking targeted therapeutic measures and improving prognosis, but the current situation is that we lack in-depth research on it. In hepatocarcinogenesis, it is often prompted by genetic lesions within the core cell cycle machinery, whose excessive activation accelerates the process of cell division, disrupting the normal order and successive production of tumor cells [40,41]. The aberrant expression of KIF4A, UBE2C, SMYD3, and MCM7 in hepatocellular carcinoma tissues suggests that they are closely related to the regulation of the cell division cycle of hepatocellular carcinoma and may serve as cell cycle target proteins to arrest the growth of hepatocellular carcinoma.

The interaction of multiple genetic changes in the cellular carcinogenesis process is essential for cancer's pathogenesis, not only for liver cancer. And an accurate diagnosis of the disease type helps select treatment strategies and improve prognosis to assist. The performance evaluation of the five genes associated with immune infiltration screened in this analysis using machine learning methods (LASSO and random forest) showed that these genes have solid diagnostic potential and may be able to be used in combination with existing liver cancer markers as a way to improve the diagnostic performance of liver cancer. And the survival model constructed by them also demonstrated their value in HCC prognostic analysis.

However, there are several limitations to this study. The evidence is based on publicly available data, and although we performed expression validation using two other datasets, further experiments are required to validate whether these markers still maintain a good diagnostic value, especially in the different stages of hepatocellular carcinoma (early, intermediate and advanced), before clinical application.

5. Conclusion

EFHD1, KIF4A, UBE2C, SMYD3, and MCM7 may be used as new molecular marker candidates for HCC. T cells, NK cells, macrophages, and dendritic cells may be involved in the pathogenesis and progression of HCC, and these immune cells may be new targets for immunotherapy in HCC patients.

Ethic statement

Informed consent was not required for this study because our study was based on the public databases GEO and TCGA, and the patients involved in the database had received ethical approval, so there were no ethical issues or other conflicts of interest.

Funding

This study was funded by the National Natural Science Foundation of China grants (grant numbers 81872579, 82173479).

Data availability statement

Gene expression data are available in public databases the Gene Expression Omnibus (GEO) database https://www.ncbi.nlm.nih.gov/geo/) and the Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/).

CRediT authorship contribution statement

Hui Zheng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Xu Han:** Methodology, Formal analysis, Conceptualization. **Qian Liu:** Validation, Software, Resources. **Li Zhou:** Visualization, Methodology. **Yawen Zhu:** Methodology, Investigation, Data curation. **Jiaqi Wang:** Software, Formal analysis, Data curation. **Wenjing Hu:** Visualization, Data curation, Conceptualization. **Fengcai Zhu:** Visualization, Supervision, Conceptualization. **Ran Liu:** Writing – review & editing, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e24854.

References

[1] J.D. Yang, P. Hainaut, G.J. Gores, A. Amadou, A. Plymoth, L.R. Roberts, A global view of hepatocellular carcinoma: trends, risk, prevention and management, Nat. Rev. Gastroenterol. Hepatol. 16 (10) (2019) 589–604.

[2] D.Q. Huang, H.B. El-Serag, R. Loomba, Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention, Nat. Rev. Gastroenterol. Hepatol. 18 (4) (2021) 223–238.

[3] S. Chidambaranathan-Reghupaty, P.B. Fisher, D. Sarkar, Hepatocellular carcinoma (HCC): epidemiology, etiology and molecular classification, Adv. Cancer Res. 149 (2021) 1–61.

[4] Y.C. Wang, Z.B. Tian, X.Q. Tang, Bioinformatics screening of biomarkers related to liver cancer, BMC Bioinf. 22 (Suppl 3) (2021) 521.

[5] J.D. Yang, J.K. Heimbach, New advances in the diagnosis and management of hepatocellular carcinoma, Bmj 371 (2020) m3544.

[6] Q.M. Anstee, H.L. Reeves, E. Kotsiliti, O. Govaere, M. Heikenwalder, From NASH to HCC: current concepts and future challenges, Nat. Rev. Gastroenterol. Hepatol. 16 (7) (2019) 411–428.

[7] S. Rebouissou, J.C. Nault, Advances in molecular classification and precision oncology in hepatocellular carcinoma, J. Hepatol. 72 (2) (2020) 215–229.

[8] J.A. Miller, R.L. Woltjer, J.M. Goodenbour, S. Horvath, D.H. Geschwind, Genes and pathways underlying regional and cell type changes in Alzheimer's disease, Genome Med. 5 (5) (2013) 48.

[9] J. Long, S. Huang, Y. Bai, J. Mao, A. Wang, Y. Lin, X. Yang, D. Wang, J. Lin, J. Bian, et al., Transcriptional landscape of cholangiocarcinoma revealed by weighted gene coexpression network analysis, Briefings Bioinf. 22 (4) (2021).

[10] P.S. Reel, S. Reel, E. Pearson, E. Trucco, E. Jefferson, Using machine learning approaches for multi-omics data analysis: a review, Biotechnol. Adv. 49 (2021) 107739.

[11] S. Wang, Q. Zhang, C. Yu, Y. Cao, Y. Zuo, L. Yang, Immune cell infiltration-based signature for prognosis and immunogenomic analysis in breast cancer, Briefings Bioinf. 22 (2) (2021) 2020–2031.

[12] S. Roessler, H.L. Jia, A. Budhu, M. Forgues, Q.H. Ye, J.S. Lee, S.S. Thorgeirsson, Z. Sun, Z.Y. Tang, L.X. Qin, et al., A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients, Cancer Res. 70 (24) (2010) 10202–10212.

[13] S. Shimada, K. Mogushi, Y. Akiyama, T. Furuyama, S. Watanabe, T. Ogura, K. Ogawa, H. Ono, Y. Mitsunori, D. Ban, et al., Comprehensive molecular and immunological characterization of hepatocellular carcinoma, EBioMedicine 40 (2019) 457–470.

[14] B.J. Erickson, S. Kirk, Y. Lee, O. Bathe, M. Kearns, C. Gerdes, K. Rieger-Christ, J. Lemmerman, The Cancer Genome Atlas Liver Hepatocellular Carcinoma Collection (TCGA-LIHC) (Version 5) the Cancer Imaging Archive, 2016.

[15] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.

[16] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016.

[17] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (2008) 559.

[18] G. Yu, L.G. Wang, Y. Han, Q.Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS 16 (5) (2012) 284–287.

[19] A. Ziegler, I.R. König, Mining data with random forests: current options for real-world applications, Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 4 (1) (2014) 55–63.

[20] J.H. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for generalized linear models via coordinate descent, J. Stat. Software 33 (1) (2010) 1–22.

[21] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[22] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinf. 12 (2011) 77.

[23] A.M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, C.D. Hoang, M. Diehn, A.A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457.

[24] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P.W. Laird, D.A. Levine, et al., Inferring tumour purity and stromal and immune cell admixture from expression data, Nat. Commun. 4 (2013) 2612.

[25] A. Mayakonda, D.C. Lin, Y. Assenov, C. Plass, H.P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer, Genome Res. 28 (11) (2018) 1747–1756.

[26] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics 26 (12) (2010) 1572–1573.

[27] D. Maeser, R.F. Gruener, R.S. Huang, oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data, Briefings Bioinf. 22 (6) (2021).

[28] O.O. Ogunwobi, T. Harricharran, J. Huaman, A. Galuza, O. Odumuwagun, Y. Tan, G.X. Ma, M.T. Nguyen, Mechanisms of hepatocellular carcinoma progression, World J. Gastroenterol. 25 (19) (2019) 2279–2293.

[29] A. Alqahtani, Z. Khan, A. Alloghbi, T.S. Said Ahmed, M. Ashraf, D.M. Hammouda, Hepatocellular carcinoma: molecular mechanisms and targeted therapies, Medicina (Kaunas) 55 (9) (2019).

[30] H. Huang, L. Zhu, C. Huang, Y. Dong, L. Fan, L. Tao, Z. Peng, R. Xiang, Identification of hub genes associated with clear cell renal cell carcinoma by integrated bioinformatics analysis, Front. Oncol. 11 (2021) 726655.

[31] K. Takane, Y. Midorikawa, K. Yagi, A. Sakai, H. Aburatani, T. Takayama, A. Kaneda, Aberrant promoter methylation of PPP1R3C and EFHD1 in plasma of colorectal cancer patients, Cancer Med. 3 (5) (2014) 1235–1245.

[32] S. Li, X. Gai, S.S. Myint, K. Arroyo, L. Morimoto, C. Metayer, A.J. de Smith, K.M. Walsh, J.L. Wiemels, Mitochondrial 1555 G>A variant as a potential risk factor for childhood glioblastoma, Neurooncol Adv 4 (1) (2022) vdac045.

[33] G. Hu, Z. Yan, C. Zhang, M. Cheng, Y. Yan, Y. Wang, L. Deng, Q. Lu, S. Luo, FOXM1 promotes hepatocellular carcinoma progression by regulating KIF4A expression, J. Exp. Clin. Cancer Res. 38 (1) (2019) 188.

[34] X. Sun, P. Chen, X. Chen, W. Yang, X. Chen, W. Zhou, D. Huang, Y. Cheng, KIF4A enhanced cell proliferation and migration via Hippo signaling and predicted a poor prognosis in esophageal squamous cell carcinoma, Thorac Cancer 12 (4) (2021) 512–524.

[35] D. Zhu, X. Xu, M. Zhang, T. Wang, Enhanced expression of KIF4A in osteosarcoma predicts a poor prognosis and facilitates tumor growth by activation of the MAPK pathway, Exp. Ther. Med. 22 (5) (2021) 1339.

[36] J.Y. Yang, D. Li, Y. Zhang, B.X. Guan, P. Gao, X.C. Zhou, C.J. Zhou, The expression of MCM7 is a useful biomarker in the early diagnostic of gastric cancer, Pathol. Oncol. Res. 24 (2) (2018) 367–372.

[37] H. Dastsooz, M. Cereda, D. Donna, S. Oliviero, A comprehensive bioinformatics analysis of UBE2C in cancers, Int. J. Mol. Sci. 20 (9) (2019).

[38] B.J. Bernard, N. Nigam, K. Burkitt, V. Saloura, SMYD3: a regulator of epigenetic and signaling pathways in cancer, Clin. Epigenet. 13 (1) (2021) 45.

[39] A. Giakountis, P. Moulos, M.E. Sarris, P. Hatzis, I. Talianidis, Smyd3-associated regulatory pathways in cancer, Semin. Cancer Biol. 42 (2017) 70–80.

[40] J.M. Suski, M. Braun, V. Strmiska, P. Sicinski, Targeting cell-cycle machinery in cancer, Cancer Cell 39 (6) (2021) 759–778.

[41] D. Sia, A. Villanueva, S.L. Friedman, J.M. Llovet, Liver cancer cell of origin, molecular class, and effects on patient prognosis, Gastroenterology 152 (4) (2017) 745–761.