Observational Study

Medicine®

OPEN

WGCNA combined with machine learning to find potential biomarkers of liver cancer

Jia-Hao Lv, MD[a], A-Jiao Hou, PhD[a], Shi-Hao Zhang, MD[a], Jiao-Jiao Dong, MD[a], Hai-Xue Kuang, PhD[a], Liu Yang, PhD[a], Hai Jiang, PhD[a,*]

Abstract

The incidence of hepatocellular carcinoma (HCC) has been increasing in recent years. With the development of various detection technologies, machine learning is an effective method to screen disease characteristic genes. In this study, weighted gene co-expression network analysis (WGCNA) and machine learning are combined to find potential biomarkers of liver cancer, which provides a new idea for future prediction, prevention, and personalized treatment. In this study, the "limma" software package was used. $P < .05$ and log2 |fold-change| > 1 is the standard screening differential genes, and then the module genes obtained by WGCNA analysis are crossed to obtain the key module genes. Gene Ontology and Kyoto Gene and Genome Encyclopedia analysis was performed on key module genes, and 3 machine learning methods including lasso, support vector machine-recursive feature elimination, and RandomForest were used to screen feature genes. Finally, the validation set was used to verify the feature genes, the GeneMANIA (http://www.genemania.org) database was used to perform protein–protein interaction networks analysis on the feature genes, and the SPIED3 database was used to find potential small molecule drugs. In this study, 187 genes associated with HCC were screened by using the "limma" software package and WGCNA. After that, 6 feature genes (AADAT, APOF, GPC3, LPA, MASP1, and NAT2) were selected by RandomForest, Absolute Shrinkage and Selection Operator, and support vector machine-recursive feature elimination machine learning algorithms. These genes are also significantly different on the external dataset and follow the same trend as the training set. Finally, our findings may provide new insights into targets for diagnosis, prevention, and treatment of HCC. AADAT, APOF, GPC3, LPA, MASP1, and NAT2 may be potential genes for the prediction, prevention, and treatment of liver cancer in the future.

Abbreviations: DEGs = differentially expressed genes, FC = fold change, GO = Gene Ontology, GS = gene significance, HCC = hepatocellular carcinoma, KEGG = Kyoto Gene and Genome Encyclopedia, LASSO = Least Absolute Shrinkage and Selection Operator, MM = membership correlation, PPI = protein-protein interaction, RFE = recursive feature elimination, SVM = support vector machine, WGCNA = weighted gene co-expression network analysis.

Keywords: bioinformatics, biomarkers, liver cancer, machine learning, WGCNA

1. Introduction

Hepatocellular carcinoma (HCC) accounts for about 90% of primary liver cancer and is one of the deadliest and most sick malignancies in humans.[1,2] At now, early stage liver cancer can be efficiently treated by drastic therapy such as surgical resection, liver transplantation, and local ablation. However, due to the subtle beginning and fast development of the illness, most HCC patients are detected at middle to late stages, and the overall treatment success is mediocre.[3] In recent years, large-scale genome-wide association studies and meta-analyses[4–10] have discovered common disease-associated variations in the population. Such genomic profiling might help uncover prospective biomarkers to enhance HCC screening and management by enabling risk-stratified customization. This is a potential new avenue for early detection and therapeutic therapy of liver cancer.

In recent years, gene chip technology has received unprecedented attention and rapid development, with the establishment

This work was supported by the Natural Science Foundation of Heilongjiang Province (No. LH2021H097), Heilongjiang Touyan Innovation Team Program, the National Famous Old Traditional Chinese Medicine Experts Inheritance Studio Construction Program of National Administration of TCM (NO. [2022] No. 75),Traditional Chinese medicine Processing technology inheritance base project, The Seventh Batch of National Old Traditional Chinese Medicine Experts Academic Experience Inheritance Program (NO. [2022] No. 76), the National Natural Science Foundation Matching Project (No. 2018PT02) and National Training Program for Traditional Chinese Medicine Characteristic Technology Inheritance Talents(NO. [2023]).

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are publicly available.

The data in this paper are from public databases and do not involve ethical review issues.

Supplemental Digital Content is available for this article.

[a] Key Laboratory of Basic and Application Research of Beiyao, Heilongjiang University of Chinese Medicine, Ministry of Education, Harbin, China.

* Correspondence: Hai Jiang, Key Laboratory of Basic and Application Research of Beiyao, Heilongjiang University of Chinese Medicine, No. 24 Heping Road, Xiangfang District, Harbin 150040, China (e-mail: jianghai_777@126.com).

How to cite this article: Lv J-H, Hou A-J, Zhang S-H, Dong J-J, Kuang H-X, Yang L, Jiang H. WGCNA combined with machine learning to find potential biomarkers of liver cancer. Medicine 2023;102:50(e36536).

Received: 20 October 2023 / Received in final form: 16 November 2023 / Accepted: 17 November 2023

http://dx.doi.org/10.1097/MD.0000000000036536

of numerous biological information databases in quick succession.[11,12] The advancement of these new technologies for studying various diseases provides a useful tool and opens up new ideas and directions. In recent years, there has been growing interest in applying machine learning and bioinformatics to research.[13–15] By integrating multilayered biological data, including genomes, transcriptomics, proteomics, and metabolomics, these techniques may enable a more comprehensive and systematic understanding of the molecular mechanisms and pathophysiology underlying HCC. The weighted gene co-expression network analysis (WGCNA) approach can analyze large-scale gene expression profile data, identify genes associated with liver cancer, and extract potential biomarkers and therapeutic targets from them.[16] The proper screening of differentially expressed genes for bioinformatics analysis is a worldwide priority. Investigating discrepancies in gene expression between case and control groups may prove highly beneficial for elucidating the origins, diagnosis, and treatment of diseases.

The focus of this project is to elucidate how bioinformatics, machine learning, and WGCNA approaches can be utilized in the study of liver cancer. By contrasting and analyzing gene expression data from malignant and normal tissues of liver cancer patients, genes and pathways associated with the pathogenesis and progression of liver cancer were identified. The genes discovered through these processes were then narrowed down using machine learning techniques. To identify potential small molecule compounds that may contribute to liver cancer biomarkers, the SPIED3 database was also analyzed. With the use of these findings, we aim to improve the diagnosis and management of liver cancer, as well as the care and prognosis for liver cancer patients.

GPC3, AADAT, APOF, LPA, MASP1, NAT2, and its related molecular activities all have a substantial impact in HCC. GPC3 (Fang et al, 2022; Li et al, 2023; Zheng et al, 2022) is abundant in cancer tissues, whereas AADAT, APOF, LPA, MASP1, and NAT2 are abundant in healthy tissues. According to various studies, GPC3 is expressed in a variety of cancers. AADAT, APOF, GPC3, LPA, MASP1, and NAT2 have the potential to be therapeutic targets for hepatocellular carcinoma.

2. Methods and materials

2.1. Datasets information and data processing

The GSE101685 and GSE136247 datasets were obtained from the GEO database (http://www.ncbi.nlm.nih.gov/geo). The GSE101685 (GPL570) dataset includes 24 cancerous and 8 normal liver tissues from liver cancer patients. The GSE136247 (GPL17586) dataset contains 69 samples, including 39 tissue samples from patients with liver cancer and 30 controls of normal liver tissue next to cancer. In the microarray dataset, probes were replaced by matching gene symbols in subsequent analyses (if multiple genes matched, the first gene symbol was selected), while probes that did not match any gene were discarded. For all datasets, multiple rows of the same gene symbol were further analyzed by expression value and mean. The GSE101685, GSE136247 data sets were merged as the training set, and in data processing, batch effects were removed using "removeBatchEffect" in the "limma" package for R, and background correction and normalization were performed using the "limma" package. P value < .05 and log2 |fold change (FC) > 1| were used as thresholds for screening differentially expressed genes (DEGs) in the liver cancer group versus healthy controls.

2.2. Construction of a weighted gene coexpression network to identify key genes with HCC

WGCNA identifies potential gene interactions and correlations with phenotypes by identifying gene co-expression relationships in samples and is used to explore the complex relationships between gene expression profiles and phenotypes.[17] The WGCNA package (version 1.71) of R software was used to analyze the genes in the training set and construct a weighted gene co-expression network. The genes were ranked in descending order by median absolute deviation to take the top 10,000 genes, and this was used to construct the WGCNA network for subsequent analysis. A scale-free network of gene expression profiles was constructed using a pick soft threshold function based on a correlation coefficient of $R^2 > 0.90$. Finally, the dynamic tree-cutting algorithm was used for module identification, and 30 genes were selected as the minimum number for each module. Correlations between different modules and diseases were calculated separately, correlations between modules and diseases were assessed, and heat maps of module-disease correlations were drawn by R software. The module with the highest correlation with both phenotypes was used as the key module, with genes with gene significance (GS) > 0.2, membership correlation (MM) > 0.8, and a P value of .05 in the key module strongly correlated with both disease and module. These genes were used as candidate genes for subsequent analyses.

2.3. Functional enrichment analyses

Gene Ontology (GO) and Kyoto Gene and Genome Encyclopedia (KEGG) pathway enrichment analyses were performed on DEGs screened from liver cancer tissues of liver cancer patients and normal liver tissues using the R package "clusterProfiler" and the DAVID database[18] (https://david.ncifcrf.gov/tools.jsp) to investigate the biological functions and related pathways of DEGs. At P .05, the KEGG enrichment analysis was statistically significant. The 3 components of the GO analysis were biological process, cellular component, and molecular function.

2.4. Candidate gene biomarker identification

Three machine learning algorithms, namely the Least Absolute Shrinkage and Selection Operator (LASSO), support vector machine (SVM), and RandomForest, were used in this study to identify significant HCC diagnostic gene biomarkers. LASSO is a regression analysis algorithm with variable selection and regularization features that help avoid overfitting and improve prediction accuracy. SVM is a widely used supervised machine learning technique for classification and regression, while recursive feature elimination (RFE) algorithms can obtain optimal combinations of variables to maximize model performance, so this study uses the SVM-RFE algorithm to identify characteristic biomarkers with better discriminatory power.[19] RandomForest can build a classification model, and while building this model, the importance of genes to the model can be discriminated and classified, the criticality of individual genes to this classification can be obtained, and important genes can be filtered out in the form of a scoring ranking. Moreover, the RandomForest algorithm has a fairly good adaptability to complex data and has promising prospects for noisy, non-linear, and other high-dimensional genomic data. The overlapping genes calculated by the above 3 algorithms were used as candidate gene biomarkers, and their expression levels in liver cancer were further validated using different datasets.

2.5. PPI (protein–protein interaction) network construction

GeneMANIA (http://www.genemania.org) is a website that allows users to build PPI networks that may be used to predict gene function and discover genes with comparable outcomes.[20] Network ensemble algorithms employ bioinformatics approaches such as physical interaction, co-expression, co-localization, gene enrichment analysis, genetic interaction, and locus prediction. GeneMANIA was utilized in this study to evaluate the PPI networks of feature genes.

2.6. Potential small molecule drugs for the treatment of liver cancer

The SPIED3 database (http://www.spied.org.uk/cgi-bin/HGNC-SPIED3.1.cgi) is a web-based tool designed to facilitate fast and simple quantitative queries on publicly available gene expression data.[21] The query distribution is associated with a given SPIED3 entry based on probability scores from a Pearson regression analysis of continuous expression profile queries or a cumulative binomial distribution ranked according to gene probabilities estimated from the database frequency of discrete expression queries. The output lists the SPIED3 entries that are significantly correlated based on the score ranking. The genes and their expression trends obtained from the screening are imported, and the individual small-molecule compounds that can reverse the expression of uploaded genes and promote down-regulated gene expression are screened by ranking. These compounds have the potential to treat liver cancer.

3. Result

3.1. DEG screening and data preprocessing

Figure 1A depicts the principal component analysis (PCA) results before eliminating the batch effect for numerous datasets, with different colors indicating distinct datasets. Both datasets are split as illustrated, with no crossing. Figure 1B depicts the PCA findings after batch elimination. As demonstrated, the intersection of the 2 datasets may be utilized as a batch for further analysis. At *P* value < .05 and log2FC > 1, 621 genes were identified as DEGs, with 145 genes up-regulated and 476 genes down-regulated. The heatmap in this paper depicts the log2FC of the DEGs in the control and tumor group, as shown in Figure 1C, where each column represents the expression of a different gene in the same sample, and each row represents the expression of the same gene in different samples. Warm colors represent up-regulated genes and cold colors reflect down-regulated genes, with darker colors representing more pronounced up- or down-regulation. The volcano plot can be used to show the distribution of the difference in gene expression levels between 2 groups of samples, with the X-axis log2FC, and the genes with greater differences are distributed with the X-axis at the 2 ends. The

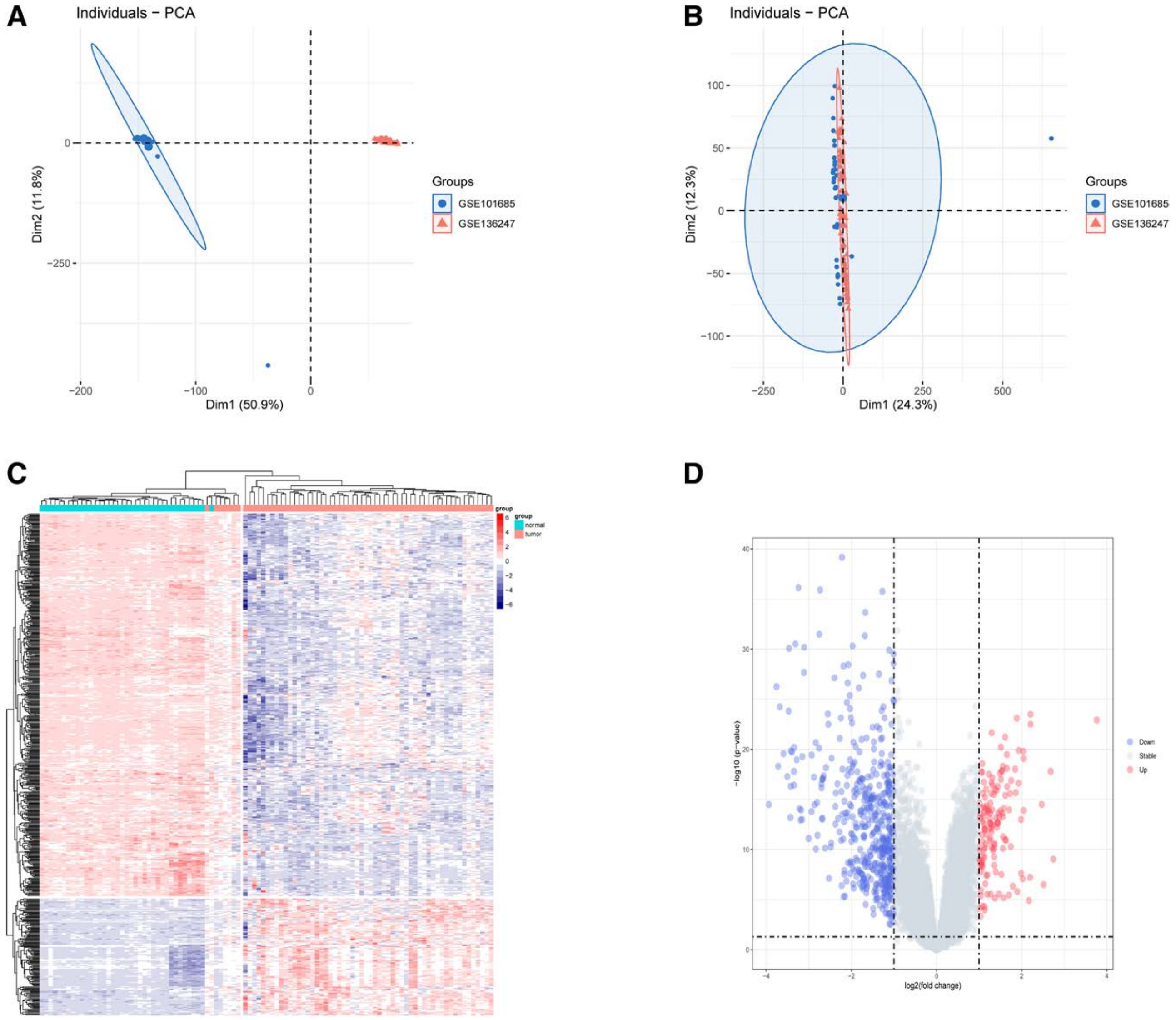

Figure 1. (A and B) PCA of HCC and control samples. (C) Heat map of DEG. (D) Volcano plot of DEGs. DEGs = differentially expressed genes, HCC = hepatocellular carcinoma, PCA = principal component analysis.

Y-axis is represented by –log10 *P* value, the result is shown in Figure 1D, the red dots represent up-regulated genes, the blue dots are down-regulated genes, and the gray dots are stable genes.

3.2. Weighted gene co-expression network construction

The GSE101685 and GSE136247 datasets were retrieved from the GEO data, and 38 normal samples and 63 liver cancer samples were selected to cluster the samples preferentially and to exclude the obviously abnormal samples by setting the threshold, as shown in Figure 2A. Then, as shown in Figure 2B, we set the soft threshold to 10 when $R^2 > 0.9$ and the average connectivity was high. After merging strongly associated modules using the 0.5 clustering height restriction (Fig. 2C), 9 modules were identified for further study. The initialized and merged modules are finally shown under the clustering tree (Fig. 2D). Next, correlations between modules were examined, and the results showed no significant associations between them (Fig. 2E). Transcriptional correlation analysis within modules proved the reliability of the module delineation, showing no substantial association between modules (Fig. 2F). Positive correlations between ME values and clinical features were applied to explore the association between modules and clinical symptoms. The blue module was positively correlated with normal ($R = 0.79$, $P = 5e–08$) and negatively correlated with HCC ($r = –0.79$, $P = 5e–0.8$), whereas the turquoise module was negatively correlated with normal ($R = 0.8$, $P = 3e–08$) and positively correlated with HCC ($r = –0.8$, $P = 3e–08$) (Fig. 2G). Identification of clinically meaningful modules. The results showed that the blue and blue-green modules were highly correlated with HCC in the scatter plot of control MM versus GS (Fig. 2H) and HCC MM versus GS (Fig. 2I). Further testing was performed for key genes in both modules.

3.3. DEGs and functional analysis of key module genes

We obtained 51 key up-regulated genes by taking the intersection of the key genes from the greenyellow module of WGCNA with the up-regulated differential genes (Fig. 3A), and then 136 key down-regulated genes by intersecting the key genes from the blue module with the down-regulated differential genes (Fig. 3B), resulting in a total of 187 key genes associated with the disease. We performed a functional analysis to better understand the biological activities of the key genes in the module (Table S1, Supplemental Digital Content, http://links.lww.com/MD/K992). The findings of GO enrichment analysis indicated that these genes were connected to "carboxylic acid biosynthetic process," "carboxylic acid catabolic process," "chromosomal region," "chromosome, centromeric region," "heme binding," "iron ion binding," and so on, as illustrated in Figure 3D (The X-axis is the GO term, the Y-axis is the number of genes, and the brown, orange and purple colors represent the 3 gene functions biological process, cellular component and, molecular function, respectively). The KEGG results point to a probable connection to "metabolic pathways," "xenobiotic metabolism by cytochrome P450," "retinol metabolism," "chemical carcinogenesis – DNA adducts" and "drug metabolism—cytochrome P450" (Fig. 3C, The X-axis is the *P* value and the Y-axis is the hsa ID, the size of the circle represents the number of genes enriched in this pathway, and the color from blue to yellow represents the size of the *P* value).

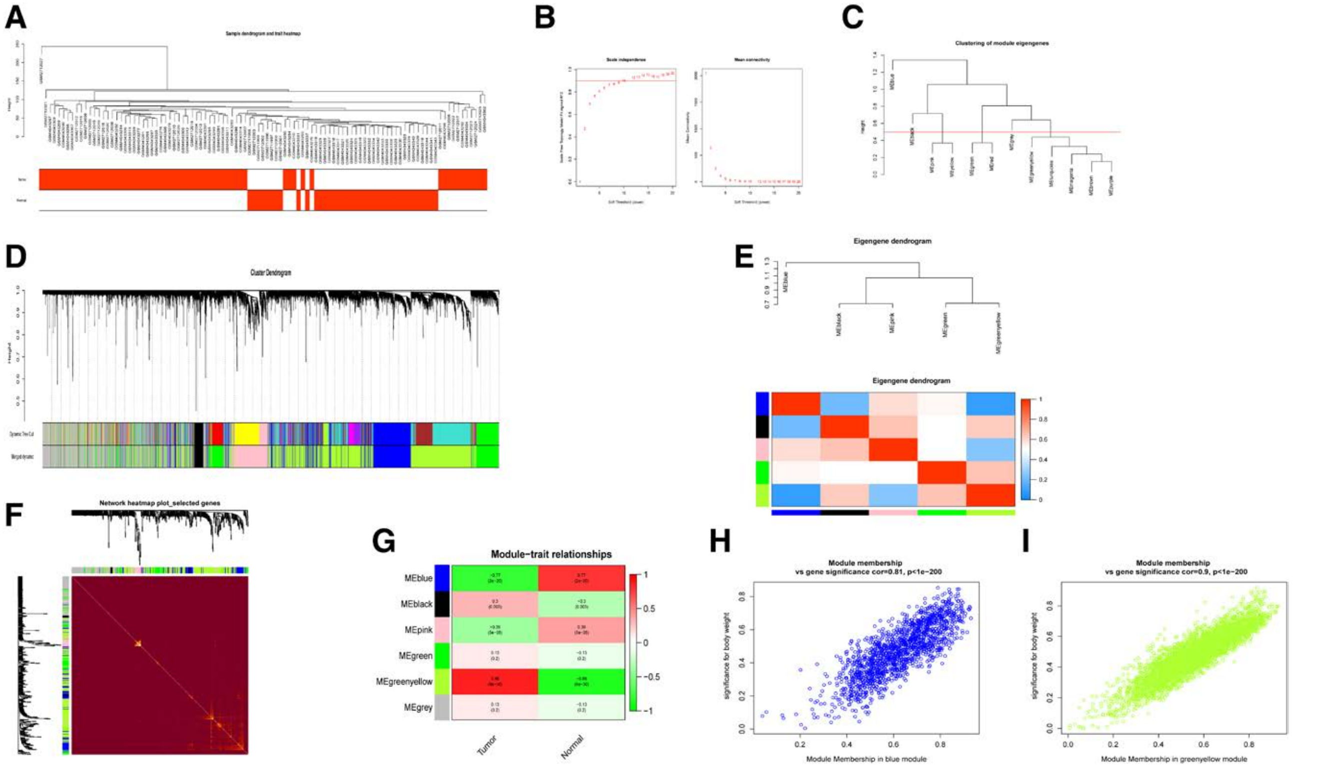

Figure 2. Construction of WGCNA co-expression network. (A) Sample clustering dendrogram with tree leaves corresponding to individual samples. (B) Soft threshold *b* = 10 and scale-free topological fit index (R^2). (C) Clustered dendrograms were cut at a height of 0.5 to detect and combine similar modules. (D) The original and combined modules under the clustering tree. (E) Collinear heat map of module feature genes. Red color indicates a high correlation and blue color indicates opposite results. (F) Clustering dendrogram of module feature genes. (G) Heat map of module–trait correlations. Red represents positive correlations and blue represent negative correlations. (H) MM versus GS scatter plot of HCC. (I) MM versus GS scatter plot of Normal. GS = gene significance, HCC = hepatocellular carcinoma, MM = membership correlation, WGCNA = weighted gene co-expression network analysis.

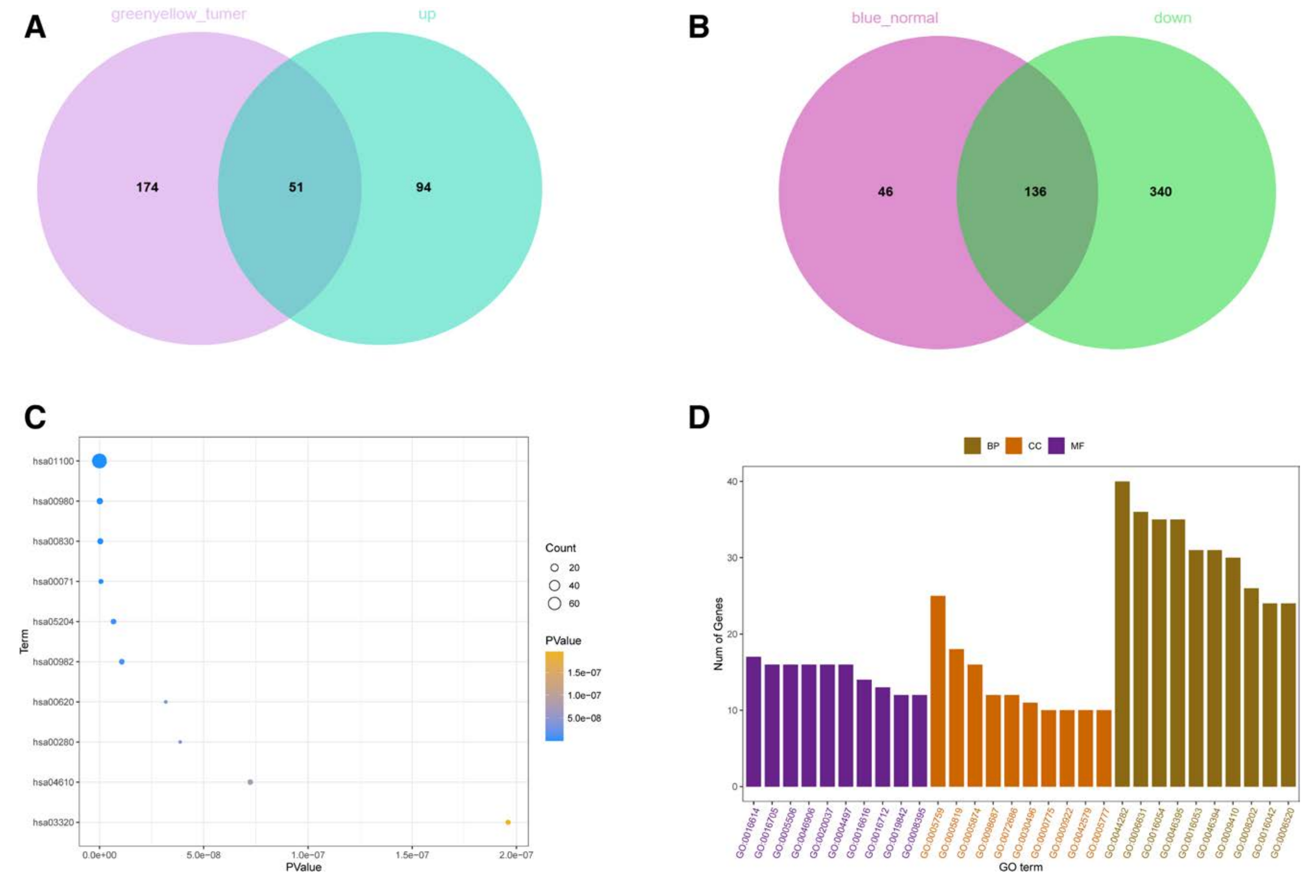

Figure 3. Functional analysis of key module genes merged with DEGs. (A and B) Venn diagram of key module genes versus DEGs. (C) GO analysis. (D) KEGG analysis. DEGs = differentially expressed genes, GO = Gene Ontology, KEGG = Kyoto Gene and Genome Encyclopedia.

3.4. Selection of feature genes

The above genes were used to isolate the feature genes used to diagnose HCC. SVM is a regression or classification-supervised machine learning technique that requires a training set with labels. SVM-RFE is a machine learning technique that trains a subset of features from different classes to narrow down the feature set and find the most predictive features (Table S2, Supplemental Digital Content, http://links.lww.com/MD/K993, Fig. 4A). To compute and select linear models and retain valuable variables, a LASSO regression was performed using the "glmnet" package in R. Binomially distributed variables were then used for LASSO classification, combined with a standard error λ value as a minimum criterion (1-SE criterion) to construct the model, which had good performance (Table S3, Supplemental Digital Content, http://links.lww.com/MD/K994, Fig. 4B and C). Genes were ranked using RandomForest, and their relative values above 0.25 were considered to be typical causes of chance (Table S4, Supplemental Digital Content, http://links.lww.com/MD/K995, Fig. 4D and E). We used the Venn diagram to find the 6 genes that overlapped using the intersection of the 3 methods described above (Fig. 4F).

3.5. Validation of gene expression

Using the combined training set data from GSE101685 and GSE136247, we validated the expression of these 6 genes in HCC and discovered that AADAT, APOF, LPA, MASP1, and NAT2 were dramatically decreased in HCC except for GPC3, which was significantly upregulated in HCC (Figure S1A, Supplemental Digital Content, http://links.lww.com/MD/K997). Furthermore, the validation dataset GSE84402 revealed that GPC3 was dramatically increased in HCC, whereas AADAT, APOF, LPA, MASP1, and NAT2 were all abundantly expressed in normal tissues (Figure S1B, Supplemental Digital Content, http://links.lww.com/MD/K997). As shown, 6 genes were significantly different in both control and tumor groups in both datasets with a consistent trend. It proves the reliability of this study. Gene correlation is also studied in this paper. As shown in Figure 5, the lower left corner represents the correlation value, the size of the circle in the upper right corner represents the correlation, red represents negative correlation, and blue is positive correlation. These 6 genes have strong correlation, which proves that they may have some kind of intrinsic connection with each other. In addition, AADAT, APOF, LPA, MASP1 and NAT2 are positively correlated, which suggests that 5 genes other than GPC3 may have similar effects in disease expression.

3.6. Trait gene interaction analysis

We created a PPI network of feature genes using the GeneMANIA database (Fig. 6A, the 6 circles in the middle represent the 6 genes, and the outer circles represent the genes associated with the 6 genes, with the larger the circle the higher the correlation). GO/KEGG analysis was done on 6 genes to further study the activities of the distinctive genes. The core genes were related with "organic cyclic compound catabolic process," "steroid metabolic process," "blood microparticle," "regulation of hormone levels," and "small molecule catabolic process" according to GO analysis (Fig. 6B). The KEGG analysis revealed that these genes were associated with "metabolic pathways," "complement and coagulation cascades," "coronavirus disease – COVID-19," "retinol metabolism," and "steroid hormone biosynthesis" (Fig. 6C).

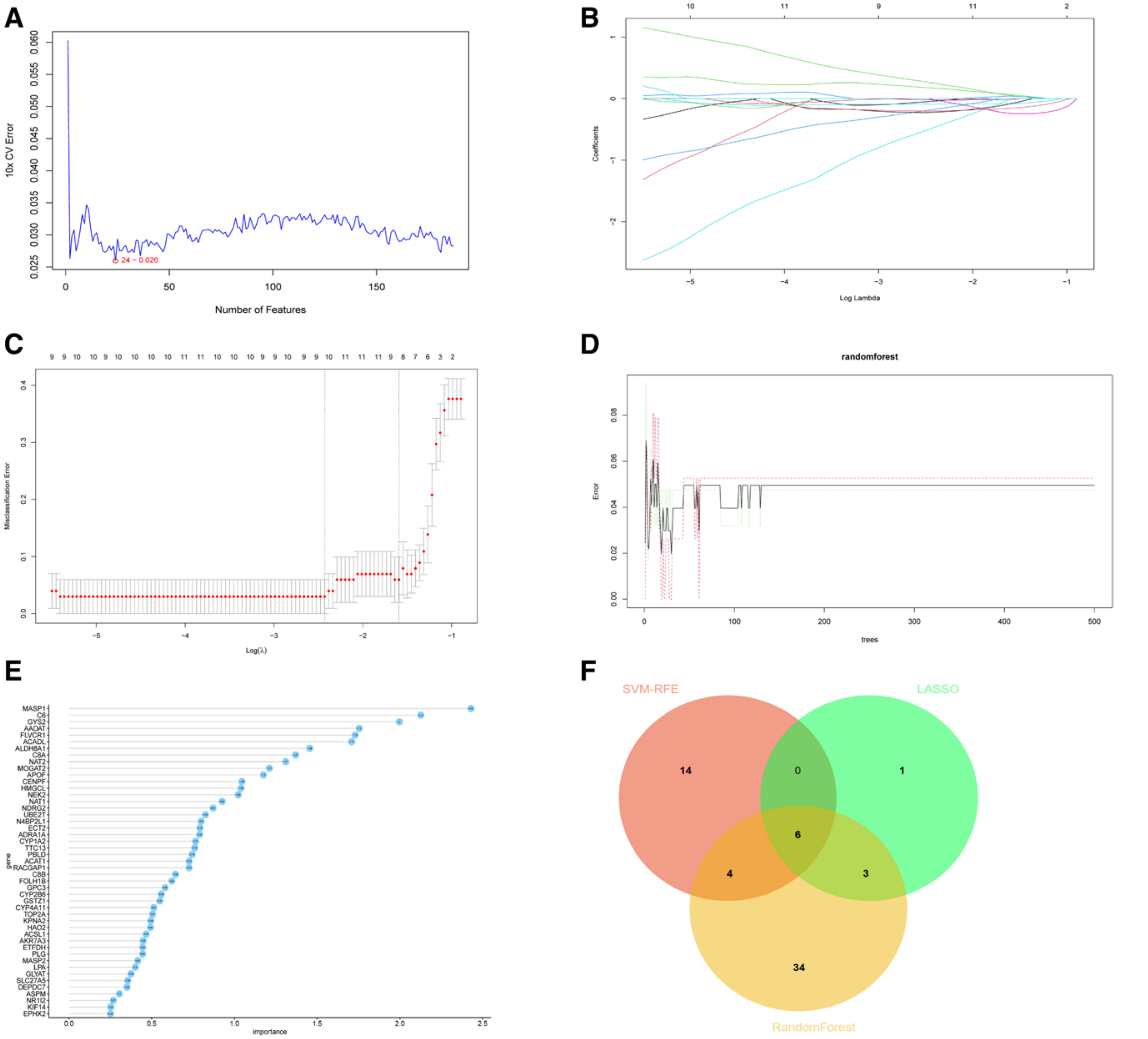

Figure 4. Feature gene selection. (A) Biomarker signature gene expression validation by support vector machine recursive feature elimination (SVM-RFE) algorithm selection. (B and C) Adjustment of feature selection in the minimum absolute shrinkage and selection operator model (lasso). (D) RandomForest error rate versus the number of classification trees. (E) The top 47 relatively important genes. (F) Three algorithmic Venn diagram screening genes.

3.7. Small molecule drug candidates

The identified important genes and their expression patterns were entered into the SPIED3 database to forecast prospective medicines that might reverse critical expression changes. Furthermore, SPIED3 database revealed that metamizole sodium was the co-occurring medicine with the highest overall score (Table S5, Supplemental Digital Content, http://links.lww.com/MD/K996). These chemicals have the potential to cure HCC.

4. Discussion

Globally, liver cancer is the most common fatal malignancy. As the liver involves extensive physiological functions, strong compensation, and no distribution of peripheral nerves, early symptoms of liver cancer are not obvious and can easily be overlooked, thus missing the best time for treatment and being detected only at an advanced stage, leading to increased difficulty in treatment and a poor prognosis. Cancer cells of advanced hepatocellular carcinoma are extremely active in growth, highly invasive, prone to invade the periphery and blood vessels, local spread and bloodstream metastasis, affecting prognosis.[3,22] As a result, new therapeutic approaches[23–26] are urgently needed. Thus far, the exploration of new genetic targets will provide new ideas for liver cancer treatment and therapeutic strategies.[27,28] Molecular research and bioinformatics techniques have been rapidly developed in recent decades.[29–33] Through the enrichment analysis of molecular functions, biological processes, and cellular components, molecular biology can provide clues for a comprehensive and further study of how gene variants and co-expression affect protein function and disease progression. At the same time, the emerging WGCNA is increasingly being used for associations between diseases and associated phenotypes and highly corrected gene module.[34–36] Several studies have elucidated the role of hub genes and their underlying molecular mechanisms in HCC patients through WGCNA analysis. With the development of various detection technologies, machine learning is an effective method to screen disease characteristic genes. With the advancement of artificial intelligence, machine learning is a way to screen for genes that characterize diseases has also come on strong.

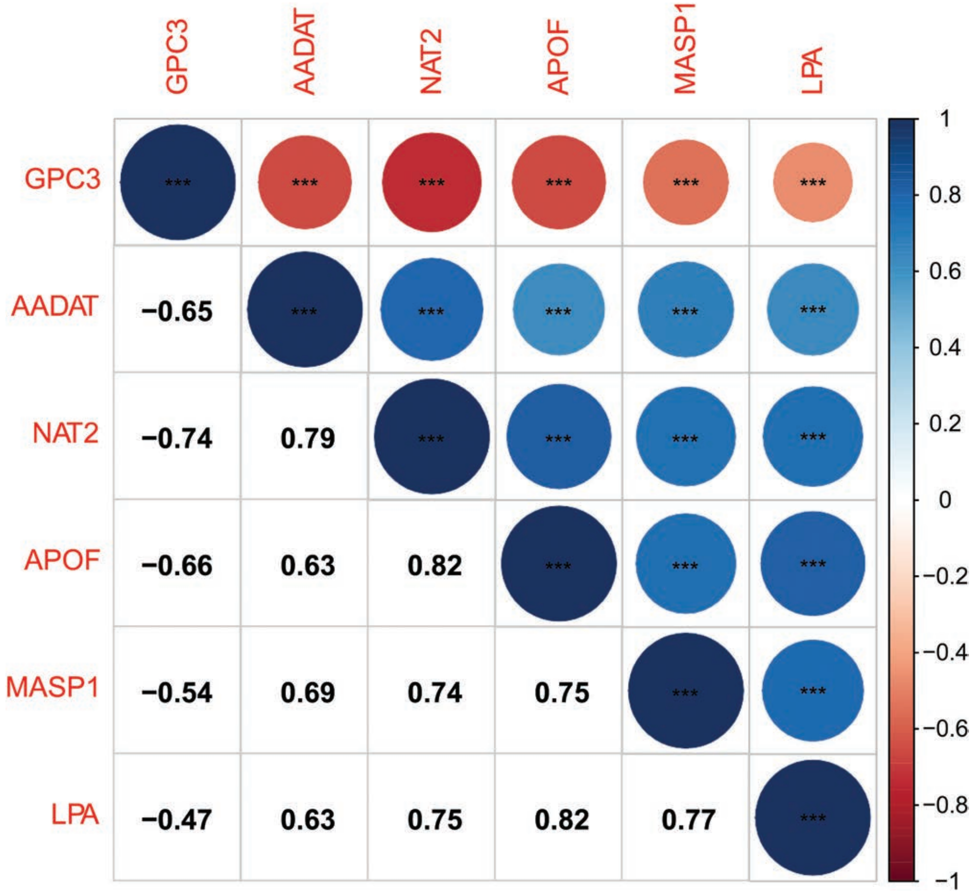

Figure 5. Validation of gene expression correlation between key genes.

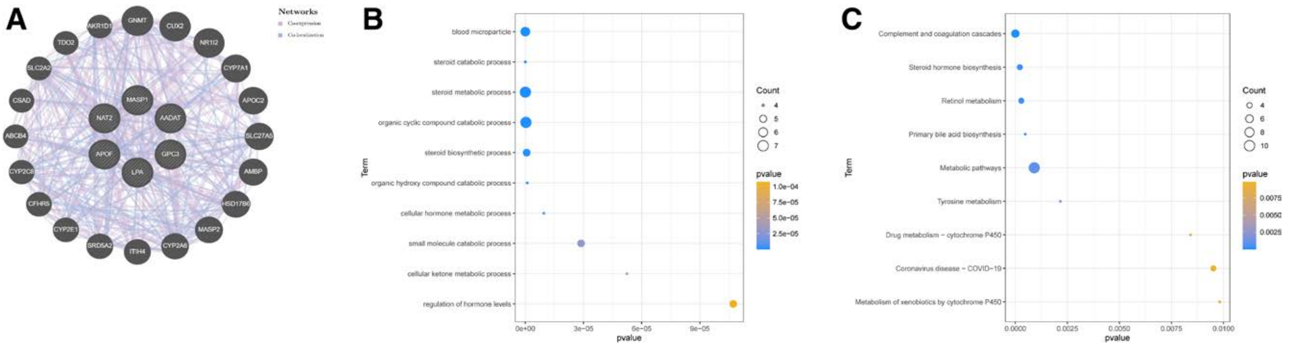

Figure 6. Interaction analysis of key genes. (A) Characterized gene co-expression network. (B) GO analysis of co-expressed genes. (C) Co-expressed gene KEGG analysis. GO = Gene Ontology, KEGG = Kyoto Gene and Genome Encyclopedia.

Machine learning algorithms can learn from large amounts of medical data and automatically identify patterns and regularities to assist doctors in making diagnoses.[37,38] In addition, machine learning algorithms can analyze data such as historical cases and biomarkers to predict the type of disease[39,40] and the level of risk that a patient may be suffering from, and take early intervention measures to reduce the incidence of medical errors. After we initially obtained the DEGs, we applied WGCNA combined with machine learning to further screen the biomarkers. The biomarkers found provide a reference for the early prevention, timely diagnosis and treatment of liver cancer, and the expression of these biomarkers in the patient's body can be used to set up a personalized treatment plan.

In view of this, we aim to broaden the horizons of studying the physiological, pathological, and molecular mechanisms of HCC through bioinformatics and provide new therapeutic targets for clinical treatment. In this study, 621 DEGs were screened, and 145 genes were found to be up-regulated and 476 genes were found to be down-regulated. 187 genes were obtained from WGCNA analysis by screening the corresponding modules most associated with disease and most associated with normal and intersecting the differential genes. Subsequent GO enrichment analysis was obtained and showed that all genes were mainly associated with carboxylic acid biosynthetic process, carboxylic acid catabolic process, chromosomal region, chromosome, centromeric region, heme binding, iron ion binding, while KEGG enrichment analysis showed association with Metabolic pathways, Metabolism of xenobiotics by cytochrome P450, retinal metabolism, Chemical carcinogenesis - DNA adducts, Drug metabolism – cytochrome P450. The results of RandomForest, SVM-RFE, and LAASO regression analyses were taken from the intersection set to identify 6 hub genes. The validation dataset confirmed that GPC3 was highly expressed in HCC, and AADAT, APOF, LPA, MASP1, and NAT2 were highly expressed in normal samples, consistent with the results of the training set analysis. The latter 5 genes were highly similar in biological function, suggesting that the signature genes are associated with disease progression.

Unlike the latter 5 genes mentioned above, GPC3 has been well documented in HCC studies. is a tumor fetal proteoglycan anchored to the cell membrane and is commonly detected in fetal liver but not in healthy adult liver.[41,42] To predict the diagnostic value of serum GPC3 in patients with HCC, a meta-analysis by Yang et al showed an association between GPC3 expression and HCC patient.[43] Another study pointed to GPC3 as a potential novel therapeutic target for osteosarcoma[44]; Yu et al[45] found that the expression of GPC3 protein was significantly higher in lung squamous cell carcinoma than in lung adenocarcinoma. GPC3 may be a candidate marker for detecting lung squamous cell carcinoma. These diverse studies imply that GPC3 is a very diagnostic gene and also remind us that further, more comprehensive, and in-depth studies are needed to determine the link between GPC3 and HCC patients.

AADAT, APOF,[46,47] LPA,[48] MASP1, and NAT2 are all reduced in HCC and are also neuroprotective factors. These 5 genes' expression was likewise lower in cancer samples than in normal tissues. And the results of the GeneMANIA database revealed that these 5 genes had close co-expression interactions. At the moment, no literature has clearly identified the precise mechanism of action of these 5 genes on cancer, and further study is required to determine the unique process of the 5 genes that may diagnose or treat liver cancer.

Among the small molecule compounds screened, there was an association between nonselective β-blockers and reduced risk of HCC in cirrhosis.[49] The small molecule compound nadolol has been reported to be associated with a reduced risk of HCC in patients with cirrhosis,[50] but there are no reports demonstrating that biperiden, metamizole sodium, and H-7 are associated with HCC treatment, and their potential for treating HCC may be related to the modulation of key genes, which needs to be investigated further.

Our study also has some limitations. We use data from public databases, which come from different platforms and are not directly comparable. There are some differences in the inclusion criteria of data sets, generally a lack of corresponding clinical data, and some data sets have fewer clinical samples. In addition, our study is limited to the transcriptome level, and the significance of the findings needs to be further verified by prospective clinical and basic experiments. In conclusion, our study identified GPC3, AADAT, APOF, LPA, MASP1, and NAT2 as potential biomarkers for HCC by applying WGCNA and analyzing HCC transcriptome data. It provides a new perspective for exploring the pathogenesis of liver cancer and a new research clue for preventing the occurrence and development of liver cancer.

5. Conclusion

We conducted a thorough, in-depth study of linked genes and pathways in order to investigate the specific diagnosis of the relationship with HCC as well as possible therapeutic genes. Our identification of 6 hub genes (GPC3, AADAT, APOF, LPA, MASP1, and NAT2) will widen our understanding of molecular pathways and provide more possible therapeutic targets for clinical therapy, which will also need more research to confirm and develop. GPC3 was identified as the most likely target in HCC and several other cancers in subsequent investigations, providing hope for the therapy of human immune-related illnesses and even cancer.

Author contributions

Formal analysis: Jia-Hao Lv.
Investigation: Jiao-Jiao Dong, Hai Jiang.
Methodology: Jia-Hao Lv, A-Jiao Hou, Shi-Hao Zhang.
Project administration: Hai-Xue Kuang, Liu Yang, Hai Jiang.
Software: Jia-Hao Lv.
Supervision: Hai-Xue Kuang, Liu Yang, Hai Jiang.
Validation: Jia-Hao Lv.
Visualization: Jia-Hao Lv.
Writing – original draft: Jia-Hao Lv.

References

[1] Forner A, Reig M, Bruix J. Hepatocellular carcinoma. Lancet. 2018;391:1301–14.
[2] Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. CA Cancer J Clin. 2023;73:17–48.
[3] Anwanwan D, Singh SK, Singh S, et al. Challenges in liver cancer and possible treatment approaches. Biochim Biophys Acta Rev Cancer. 2020;1873:188314.
[4] Pan GQ, Yang CC, Shang XL, et al. The causal relationship between white blood cell counts and hepatocellular carcinoma: a Mendelian randomization study. Eur J Med Res. 2022;27:278.
[5] Caliskan M, Brown CD, Maranville JC. A catalog of GWAS fine-mapping efforts in autoimmune disease. Am J Hum Genet. 2021;108:549–63.
[6] Mountjoy E, Schmidt EM, Carmona M, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet. 2021;53:1527–33.
[7] Gao N, Kong M, Li X, et al. The association between psoriasis and risk of cardiovascular disease: a Mendelian randomization analysis. Front Immunol. 2022;13:918224.
[8] Wu Y, Furuya S, Wang Z, et al. GWAS on birth year infant mortality rates provides evidence of recent natural selection. Proc Natl Acad Sci U S A. 2022;119:e2117312119.
[9] Wu HT, Ji CH, Dai RC, et al. Traditional Chinese medicine treatment for COVID-19: an overview of systematic reviews and meta-analyses. J Integr Med. 2022;20:416–26.
[10] Fernández-Rodríguez R, Álvarez-Bueno C, Cavero-Redondo I, et al. Best exercise options for reducing pain and disability in adults with chronic low back pain: pilates, strength, core-based, and mind-body A network meta-analysis. J Orthop Sports Phys Ther. 2022;52:505–21.
[11] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19:A68–77.
[12] Clough E, Barrett T. The gene expression omnibus database. Methods Mol Biol. 2016;1418:93–110.
[13] Ju JW, Nam K, Sohn JY, et al. Association between intraoperative body temperature and postoperative delirium: a retrospective observational study. J Clin Anesth. 2023;87:111107.
[14] Swanson K, Wu E, Zhang A, et al. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023;186:1772–91.
[15] Chen DL, Cai JH, Wang CCN. Identification of key prognostic genes of triple negative breast cancer by LASSO-based machine learning and bioinformatics analysis. Genes (Basel). 2022;13:902.
[16] Yin Z, Cai B, Zhu C, et al. Identification of key pathways and genes in the dynamic progression of HCC based on WGCNA. Genes (Basel). 2018;9:92.
[17] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 2008;9:559.
[18] Dennis G, Yang J, Gao W, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4:P3.
[19] Huang ML, Hung YH, Lee WM, et al. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. Sci World J. 2014;2014:795624.
[20] Franz M, Rodriguez H, Lopes C, et al. GeneMANIA update 2018. Nucleic Acids Res. 2018;46:W60–4.
[21] Williams G. SPIEDw: a searchable platform-independent expression database web tool. BMC Genomics. 2013;14:765.
[22] Ganesan P, Kulik LM. Hepatocellular carcinoma: new developments. Clin Liver Dis. 2023;27:85–102.
[23] Chen L, Wei X, Gu D, et al. Human liver cancer organoids: biological applications, current challenges, and prospects in hepatoma therapy. Cancer Lett. 2023;555:216048.

[24] Hamaya S, Oura K, Morishita A, et al. Cisplatin in liver cancer therapy. Int J Mol Sci . 2023;24:10858.

[25] Ganesan R, Yoon SJ, Suk KT. Microbiome and metabolomics in liver cancer: scientific technology. Int J Mol Sci. 2022;24:537.

[26] Sidali S, Trépo E, Sutter O, et al. New concepts in the treatment of hepatocellular carcinoma. United Eur Gastroenterol J. 2022;10:765–74.

[27] Huo Y, Zhou Y, Zheng J, et al. GJB3 promotes pancreatic cancer liver metastasis by enhancing the polarization and survival of neutrophil. Front Immunol. 2022;13:983116.

[28] Pan J, Zhang M, Dong L, et al. Genome-Scale CRISPR screen identifies LAPTM5 driving lenvatinib resistance in hepatocellular carcinoma. Autophagy. 2023;19:1184–98.

[29] Wang YC, Tian ZB, Tang XQ. Bioinformatics screening of biomarkers related to liver cancer. BMC Bioinf. 2021;22(Suppl 3):521.

[30] Piñero F, Dirchwolf M, Pessôa MG. Biomarkers in hepatocellular carcinoma: diagnosis, prognosis and treatment response assessment. Cells. 2020;9:1370.

[31] Xu Z, Peng B, Liang Q, et al. Construction of a Ferroptosis-related Nine-lncRNA signature for predicting prognosis and immune response in hepatocellular carcinoma. Front Immunol. 2021;12:719175.

[32] Chen D, Liu J, Zang L, et al. Integrated machine learning and bioinformatic analyses constructed a novel stemness-related classifier to predict prognosis and immunotherapy responses for hepatocellular carcinoma patients. Int J Biol Sci. 2022;18:360–73.

[33] Li L, Lei Q, Zhang S. Screening and identification of key biomarkers in hepatocellular carcinoma: evidence from bioinformatic analysis. Oncol Rep. 2017;38:2607–18.

[34] Nomiri S, Karami H, Baradaran B, et al. Exploiting systems biology to investigate the gene modules and drugs in ovarian cancer: a hypothesis based on the weighted gene co-expression network analysis. Biomed Pharmacother. 2022;146:112537.

[35] Zhao C, Xiong K, Zhao F, et al. Glycosylation-related genes predict the prognosis and immune fraction of ovarian cancer patients based on weighted gene coexpression network analysis (WGCNA) and machine learning. Oxid Med Cell Longev. 2022;2022:3665617.

[36] Wang T, Dai L, Shen S, 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. 2022;13:843408.

[37] Ji Y, Shi B, Li Y. An evolutionary machine learning for multiple myeloma using Runge Kutta Optimizer from multi characteristic indexes. Comput Biol Med. 2022;150:106189.

[38] Huang X, Liu B, Guo S, et al. SERS spectroscopy with machine learning to analyze human plasma derived sEVs for coronary artery disease diagnosis and prognosis. Bioeng Transl Med. 2023;8:e10420.

[39] Li Z, Liu Y, Guo P, et al. Construction and validation of a novel angiogenesis pattern to predict prognosis and immunotherapy efficacy in colorectal cancer. Aging (Albany NY). 2023;15:12413–50.

[40] Wang Q, Huang X, Zeng S, et al. Weighted gene co-expression network analysis and machine learning identified the lipid metabolism-related gene LGMN as a novel biomarker for keloid. Exp Dermatol. 2023:exd.14974.

[41] Zheng X, Liu X, Lei Y, et al. Glypican-3: a novel and promising target for the treatment of hepatocellular carcinoma. Front Oncol. 2022;12:824208.

[42] Schepers EJ, Glaser K, Zwolshen HM, et al. Structural and functional impact of posttranslational modification of Glypican-3 on liver carcinogenesis. Cancer Res. 2023;83:1933–40.

[43] Yang SL, Fang X, Huang ZZ, et al. Can serum glypican-3 be a biomarker for effective diagnosis of hepatocellular carcinoma? A meta-analysis of the literature. Dis Markers. 2014;2014:127831.

[44] Nie JH, Yang T, Li H, et al. Frequently expressed glypican-3 as a promising novel therapeutic target for osteosarcomas. Cancer Sci. 2022;113:3618–32.

[45] Wang J, Jiang KJ, Zhang FY, et al. Characterization and expression analysis of the prophenoloxidase activating factor from the mud crab Scylla paramamosain. Genet Mol Res. 2015;14:8847–60.

[46] Wang YB, Zhou BX, Ling YB, et al. Decreased expression of ApoF associates with poor prognosis in human hepatocellular carcinoma. Gastroenterol Rep (Oxf). 2019;7:354–60.

[47] Shen XB, Huang L, Zhang SH, et al. Transcriptional regulation of the apolipoprotein F (ApoF) gene by ETS and C/EBPα in hepatoma cells. Biochimie. 2015;112:1–9.

[48] Jiang JT, Wu CP, Xu N, et al. Mechanisms and significance of lipoprotein(a) in hepatocellular carcinoma. Hepatobiliary Pancreat Dis Int. 2009;8:25–8.

[49] Wijarnpreecha K, Li F, Xiang Y, et al. Nonselective beta-blockers are associated with a lower risk of hepatocellular carcinoma among cirrhotic patients in the United States. Aliment Pharmacol Ther. 2021;54:481–92.

[50] He X, Zhao Z, Jiang X, et al. Non-selective beta-blockers and the incidence of hepatocellular carcinoma in patients with cirrhosis: a meta-analysis. Front Pharmacol. 2023;14:1216059.