Title: Machine learning algorithms based on proteomic data mining accurately predict the recurrence of hepatitis B-related hepatocellular carcinoma

Short Title: MLA for predicting HCC recurrence

Authors: Gong Feng,[1#] Na He,[2#] Harry Hua-Xiang Xia,[3] Man Mi,[4] Ke Wang,[4] Christopher D. Byrne,[5] Giovanni Targher,[6] Hai-Yang Yuan,[7] Xin-Lei Zhang,[7] Ming-Hua Zheng[7,8,9*] Feng Ye[1*]

Affiliations:

[1] The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China;

[2] The First Affiliated Hospital of Xi'an Medical University, Xi'an, China;

[3] Department of Gastroenterology, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China;

[4] Xi'an Medical University, Xi'an, China;

[5] Southampton National Institute for Health Research Biomedical Research Centre, University Hospital Southampton, Southampton General Hospital, Southampton, UK;

[6] Section of Endocrinology, Diabetes and Metabolism, Department of Medicine, University and Azienda Ospedaliera Universitaria Integrata of Verona, Verona, Italy;

[7]NAFLD Research Center, Department of Hepatology, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China;

[8] Institute of Hepatology, Wenzhou Medical University, Wenzhou, China;

[9] Key Laboratory of Diagnosis and Treatment for The Development of Chronic Liver Disease in Zhejiang Province, Wenzhou, China.

#Co-first author: Gong Feng and Na He

***Corresponding author:**

Feng Ye, MD, PhD

Department of Infectious Disease, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China 325000, China.

E-mail: yefeng.jiaotong@xjtu.edu.cn.

Ming-Hua Zheng, MD, PhD

NAFLD Research Center, Department of Hepatology, the First Affiliated Hospital of Wenzhou Medical University; No. 2 Fuxue Lane, Wenzhou 325000, China.

E-mail: zhengmh@wmu.edu.cn; fax: (86) 577-55578522; tel: (86) 577-55579622.

Number of figures and tables: 6 figures

Abbreviation list

AUROC = area under the ROC curve, CPTAC = Clinical Proteomics Tumor Analysis Consortium, DEPs = differentially expressed proteins, GO = gene ontology, HBV = hepatitis B virus, HCC = hepatocellular carcinoma, HCV = hepatitis C virus, HPA = Human Protein Atlas, MLA = machine learning algorithm, MLP = multi-layer perceptron, SVM = support vector machine, TIMER = Tumor Immune Estimation Resource, TME = tumor microenvironment

Funding sources

Science and Technology Project of Shaanxi Province (2021ZDLSF02-09), Youth Project of Science and Technology Department of Shaanxi Province (2022JQ-986), and Shaanxi Provincial Department of Education 2020 Special Scientific Research Plan for Emergency Public Health Safety (20JG028). GT is supported in part by grants from the School of Medicine, University of Verona, Verona, Italy. CDB is supported in part by the Southampton NIHR Biomedical Research Centre (IS-BRC-20004), UK.

Acknowledgments

Not applicable.

Conflicts of interest

All authors: nothing to declare.

Data sharing statement

No additional data are available.

ABSTRACT

Background and Aims*:* Over 10% of hepatocellular carcinoma (HCC) cases recur each year, even after surgical resection. Currently, there is a lack of knowledge about the causes of recurrence and the effective prevention. Prediction of HCC recurrence requires diagnostic markers endowed with high sensitivity and specificity. This study aims to identify new key proteins for HCC recurrence and to build machine learning algorithms for predicting HCC recurrence.

Methods: The proteomics data for analysis in this study were obtained from the Clinical Proteomics Tumor Analysis Consortium (CPTAC) database. We analyzed different proteins based on cases with or without recurrence of HCC. Survival analysis, Cox regression analysis, and area under the ROC curves (AUROC > 0.7) were used to screen for more significant differential proteins. Predictive models for HCC recurrence were developed using four machine learning algorithms.

Results: A total of 690 differentially expressed proteins between 50 relapsed and 77 non-relapsed hepatitis B-related HCC patients were identified. Seven of these proteins had an AUROC > 0.7 for 5-year survival in HCC, including BAHCC1, ESF1, RAP1GAP, RUFY1, SCAMP3, STK3, and TMEM230. Among the machine learning algorithms, the random forest algorithm showed the highest AUROC values (AUROC: 0.991, 95%CI 0.962-0.999) for identifying HCC recurrence, followed by the support vector machine (AUROC: 0.893, 95%Cl 0.824-0.956), the logistic regression (AUROC: 0.774, 95%Cl 0.672-0.868), and the multi-layer perceptron algorithm (AUROC: 0.571, 95%Cl 0.459-0.682).

Conclusions: Our study identifies seven novel proteins for predicting HCC recurrence and the random forest algorithm as the most suitable predictive model for HCC recurrence.

Keywords: Recurrence of hepatocellular carcinoma, Proteomics, CPTAC database, Machine learning models

Keywords: Recurrence of hepatocellular carcinoma, Proteomics, CPTAC database, Machine learning models

INTRODUCTION

Hepatocellular carcinoma (HCC) is the most common form of liver cancer and accounts for ~90% of cases [1, 2]. The estimated number of new cases of HCC worldwide in 2020 is about 906,000, and the number of deaths was about 830,000[3]. HCC ranks 6th in the total of new cases and 3rd in the number of deaths amongst all cancers [3]. The prognosis of HCC is quite poor, and the data from the Global Cancer Survival Trends Surveillance demonstrates that the 5-year net survival rate of HCC worldwide from 2000-2014 ranged from ~5% to 30%; the 5-year net survival rate of HCC in China from 2010-2014 was about 14% [4]. Hepatitis B virus (HBV) and hepatitis C virus (HCV) remain the most important risk factors for HCC [5]. An essential element for improving the prognosis of patients with HCC is the early identification of recurrence and the implementation of appropriate therapeutic strategies. Circulating levels of alpha-fetoprotein and PIVKA-II are used to detect HCC and are closely associated with HCC prognosis. However, these two biomarkers are sometimes increased in patients with hepatitis [6]. Therefore, exploring new biomarkers with greater sensitivity and specificity for HCC recurrence is an urgent challenge in clinical practice.

Proteomics plays an important role in cancer research, both in terms of biomarkers, antitumor drugs and new therapeutic targets[7]. Proteomics approaches based on mass spectrometry have gained popularity in oncological research. These proteomics approaches have powerful capabilities for protein characterization, quantification, and post-translational modification analysis, and several results have been reported [8-10]. The CPTAC (Clinical Proteomic Tumor Analysis Consortium) database has abundant proteomic data for mining. The advantages of this database are that while most other databases analyze gene expression at the mRNA level, the CPTAC database describes gene expression at the protein level, closer to the most primitive manifestation of the disease. Furthermore, the CPTAC database contains a large amount of clinical data, allowing analysis of the relationship between protein, survival and clinical conditions [11-13]. Hence, the CPTAC database was used for proteomic data mining to find new protein biomarkers for HCC recurrence.

It is known that algorithms built on markers tend to have better diagnostic efficacy than single indicators. Machine learning algorithms (MLA) have several advantages over traditional statistical models. For example, MLA is less likely to overlook unexpected predictor variables, can help identify important influences and more marginal ones, and facilitates continuous updating and optimization of algorithms[14]. MLA is a useful tool in the field of liver disease research, and the random forest algorithm has been used to build a predictive model for significant fibrosis in non-alcoholic fatty liver disease[14]. The support vector machine (SVM), the random forest algorithm, the logistic regression, and other algorithms have also been used to detect HCC [15]. Though some algorithms have been previously established to predict HCC recurrence, they still fall short of meeting clinical requirements. A few models have been developed specifically to detect tumor recurrence after liver surgical resection, including the Singapore liver cancer recurrence score and the Surgery-Specific Cancer of the Liver Italian Program (SS-CLIP). However, none of these has been independently validated and none have excellent area under the ROC curve (AUROC) [16]. Therefore, a more precise prognostic and recurrent prediction model is urgently needed. We tried to use MLA to build predictive models for HCC recurrence to provide insights to improve the prognosis of HCC.

METHODS

Data sources

The data analyzed in this study were obtained from the publicly available CPTAC (Clinical Proteomic Tumor Analysis Consortium) database. In the CPTAC database, genomic and proteomic data were integrated to identify and characterize the full range of proteins found in normal and tumor tissues and to identify potential biomarkers for tumors [11, 17]. We downloaded the data entitled integrated proteogenomic characterization of HBV-related HCC for this analysis[18]. From the CPTAC database, recurrence group (n = 50) and non-recurrence group (n = 77) of hepatitis-B-related HCC samples were obtained after excluding HCC patients with no recurrence information. The Tumor Immune Estimation Resource (TIMER) database was used for the relationship between key differentially expressed proteins (DEPs) and immune infiltrating cells. The Human Protein Atlas (HPA) and TIMER databases were used to explore the relationship between DEPs and HCC [19-21]. Data from immunohistochemistry were extracted in the HPA database. The flow chart shown in **Figure 1** summarizes the study's research idea.

Machine learning algorithms

According to the learning method, machine learning can be divided into supervised, unsupervised, and reinforcement learning[22]. Supervised learning refers to computer training with some known inputs and corresponding correct output data to predict the results of other input data; supervised learning is the most common form of learning in medical research, which is commonly used in classification and regression problems. We developed four supervised learning algorithms for predicting HCC recurrence, including the support vector machine (SVM), the multi-layer perceptron (MLP), the logistic regression, and the random forest algorithms, respectively. The random forests are integrated by decision trees, which emerged to address the relatively weak generalization ability of decision trees[23]. The different decision trees in a random forest are not correlated. Whenever a classification task was conducted, each decision tree in the random forest was assessed separately, and each decision tree yielded its own classification result. The random forest would take the final result of whichever decision trees had the most classifications [14]. The random forest can be highly synchronized for the training process, which has a speed advantage for training large samples in the era of big data. SVM is a sparse and robust classifier using a hinge loss function to compute empirical risk and adds a regularization term to the solution system to optimize structural risk [24]. The core of SVM was proposed between 1992 and 1995 and is the next hot research topic after neural networks. SVM is characterized by its ability to simultaneously minimize empirical error and maximize geometric edge areas and to solve small sample size problems[25]. MLP has a long history of application in medical research, especially in image classification, detection and prediction[26, 27]. MLP is a forward-structured artificial neural network, which can have multiple hidden layers in between, in addition to the input and output layers. It is proposed mainly to solve the nonlinear problems that a single-layer perceptron cannot solve[28]. The MLP does not specify the number of hidden layers; therefore, the number of layers can be chosen according to the individual needs [29]. There is also no limit to the number of neurons in the output layer. Logistic regression is a classical algorithm, which is often used for dichotomous information.

Statistical analysis

This study used R (version 4.0.1), R Bioconductor, and the Perl language for statistical analyses. Fold change (FC) indicates the expression ratio between two samples (groups). We selected differentially expressed proteins based on $|\log2FC|>1$ and a P-value < 0.05 [30]. Survival analysis, Cox regression analysis, and ROC curves were used to further assess differentially expressed proteins. Survival-related proteins were those with significant p-values that were selected based on the Kaplan-Meier analysis. The random forest prediction model was mainly based on the random forest and varSelRF packages[31]. The SVM model used mainly the *svm* function from the e1071 package, and the MLP model was built mainly using the *keras* package [32].

RESULTS

Differentially expressed proteins (DEPs) and functional enrichment analysis

Using |log2FC|>1 and a P-value < 0.05, 690 DEPs were attained between the recurrence and non-recurrence HCC groups **(Supplementary Table 1**). To determine the function of the DEPs, gene ontology (GO) enrichment and KEGG pathway analyses were utilized. GO analysis revealed that DEPs exhibited significant enrichment in three biological processes (BPs): mitochondrial electron transport, mitochondrial respiratory chain complex I assembly, and Cajal body protein localization. Molecular function (MF) was significantly enriched in oxido-reductase activity, Ras GTPase binding, phospholipid binding, and NADH dehydrogenase activity. Cell components (CC) were mainly enriched in the early endosome, oxido-reductase complex, respiratory chain complex, and respiratory chain complex 1 (**Figure 2A**). As per the KEGG pathway analysis, DEPs were enriched in pathways related to neurodegeneration, PD-L1 expression, and PD-1 checkpoint pathways involved in cancer, chemical carcinogenesis, oxidative phosphorylation, nonalcoholic fatty liver disease, and hepatitis B (**Figure 2B).**

Constructing and analyzing protein-protein interaction (PPI) network

A PPI network based on the interactions between DEPs was developed to delve into the link between DEPs at the protein level (**Supplementary Figure 1**). The PPI network was constructed using a total of 1,054 interactions and 297 nodes, with the top ten most contiguous nodes between genes, being UBA52, AKT1, LCK, SHC1, PTGES3, CD4, NDUFB7, NDUFB8, CCT4 and PTPN6.

Survival analysis

Survival information was garnered from the CPTAC database, and we found 39 survival-related proteins by Kaplan-Meier analysis (all P < 0.05) (**Supplementary Table 2**). Based on this result, we conducted univariable and multivariable Cox regression analyses. Subsequently, 32 (**Supplementary Table 3**) and 18 (**Supplementary Table 4**) differential proteins were obtained. Next, 1-year, 3-year, 5-year survival ROC curves were performed from the 18 independent prognostic proteins. According to the criterion of the area under the 5-year survival ROC curves > 0.7, seven important proteins, including BAHCC1, ESF1, RAP1GAP, RUFY1, SCAMP3, STK3, TMEM230, were screened (**Supplementary Figure 2**). The Kaplan-Meier survival curves for seven DEPs are shown in **Figure 3**. In the **Figure 4** are reported the heat map of 7 key differentially expressed proteins between the recurrence and non-recurrence HCC groups.

Performance of machine-learning models for HCC recurrence

Figure 5 (A) illustrates the performance of four machine-learning models based on seven key proteins in predicting HCC recurrence. The AUROC curves for SVM, MLP, logistic regression, and random forest were 0.893 (95%Cl 0.824-0.956), 0.571

(95%Cl 0.459-0.682), 0.774 (95%Cl 0.672-0.868), and 0.991 (95%Cl 0.962-0.999), respectively. Among these four models, the random forest model performed best. **Figure 5 (B)** also shows a feature-importance plot from the random forest model. The seven variables with the highest importance (from high to low) were: ESF1, SCAMP3, RAP1GAP, BAHCC1, STK3, RUFY1, TMEM230.

Immune cell infiltration analysis and immunohistochemistry

We also examined the relationship between key differential proteins and immune cell infiltration. We found that ESF1, SCAMP3, RAP1GAP, BAHCC1, STK3, and RUFY1 were correlated to B Cell, CD8+ T Cell, CD4+ T Cell, Macrophage, Neutrophil, and Dendritic Cell **(Figure 6).** In the HPA database, we used immunohistochemistry to compare the expression of these key differential proteins in the normal liver tissue and HCC tissue. In **Supplementary Figure 3**, RUFY1, TMEM230, and STK3 were absent or only weakly expressed in the normal hepatic tissue, but were moderately to strongly expressed in the HCC tissue. Meanwhile, ESF1 was expressed at a low level in non-tumor tissues but at a high level in HCC tissues. Additionally, the TIMER database revealed that ESF1, SCAMP3, RAP1GAP, BABCC1, STK3, and RUFY1 were highly overexpressed in HCC patients **(Supplementary Figure 4).**

DISCUSSION

To our knowledge, there are no reliable and accurate predictive tools for HCC recurrence so far. Our study has uncovered important proteins closely associated with HCC recurrence from a proteomic perspective and has constructed the most appropriate machine learning prediction model for HCC recurrence.

In the present study, we found that 690 differential proteins were associated with HCC recurrence. To find proteins of more clinical value, Cox regression and ROC curve analyses were performed. The most important seven of these proteins (ESF1, SCAMP3, RAP1GAP, BAHCC1, STK3, RUFY1, TMEM230) were independent influencers of HCC prognosis and had a good predictive value for 5-year survival in HCC.

The key proteins identified have also been confirmed in previous studies. ESF1 was significantly associated with survival in HCC[33]. Kang et al. found that SCAMP3 might become a target for HCC therapy due to its potential role in promoting metastasis in HCC cells through the EGFR-MAPK p38 signaling pathway [34]. Additionally, Zhang also showed that SCAMP3 expression was correlated with several survival-related genes. Therefore, SCAMP3 might be a diagnostic or prognostic biomarker for HCC[35]. Kim et al. reported that when Hippo kinases Mst1 and Mst2 in the liver were abrogated in mammals, they led to the rapid formation of HCC and activated various molecules and associated signaling, including STAT3 [36]. Chen et al. suggested that RUFY1 was involved in the function of Rab14, promoting the metastasis of HCC cells [37].

In this study, we used multiple machine learning algorithms to build predictive models for HCC recurrence, and found that the random forest algorithm had the best diagnostic performance. The random forest is highly accurate due to its use of integrated algorithms, and outperform most individual algorithms. The introduction of randomness makes the random forest algorithm less prone to over-fitting and performs well on the test set. Due to the combination of trees, a random forest algorithm can process non-linear data. Moreover, the random forest algorithm can handle high-dimensional data that is either categorical or continuous data. Moreover, the random forest algorithm does not require normalization of the dataset, and it is quick to train.

To explore whether these seven key DEPs have other values, we analyzed their links to the immune microenvironment and the occurrence of HCC. We found that these key differential proteins were associated with immune infiltrating cells. The tumor microenvironment (TME) is a complex and evolving environment whose composition varies by tumor type and consists mainly of immune cells, stromal cells, blood vessels, and extracellular base (ECM), of which immune cells are key components of TME[38]. Furthermore, an increasing number of investigators have found that infiltrating immune cells in hepatocellular carcinoma TME may be related to prognosis of HCC[38]. Studies have also shown that M1-type macrophages, CD4+T cells, CD8+T cells and B cells are all associated with a good prognosis of HCC[39, 40].

Conversely, M2-type macrophages, regulatory T cells, regulatory B cells are associated with a poor prognosis of HCC [41, 42]. The relationship between these DEPs and immune cells will provide more evidence to further enhance the efficacy of immunotherapy for HCC and find new strategies to effectively curb HCC recurrence and metastasis prevention [43, 44]. In the HAP and TIMER databases, we also found that these key proteins were differentially expressed in both HCC and normal liver tissues, meaning that these proteins are related to both the occurrence of HCC and the recurrence of HCC, and are HCC important markers that merit further investigations.

There are also some limitations to this study. Firstly, the sample size of the study was limited to the training set data, and there was insufficient data to validate the diagnostic performance of the random forest prediction model. Secondly, the findings of this study were only derived from data mining and were not confirmed in clinical specimens or basic research. Thirdly, as machine learning resembles black blindness, the algorithms cannot derive a specific formula. Besides, this study is only a preliminary exploration of the priority of the algorithms, not the application of the algorithms. Fourthly, the association between these key proteins and microvascular infiltration of liver cancer cells has not been clearly illustrated.

In conclusion, we screened key proteins associated with recurrence of HCC by bioinformatics methods and found that the random forest algorithm has an excellent predictive value for recurrence of HCC. These screened proteins may account for new diagnostic biological markers for HCC recurrence or targets for therapies, setting a new direction for future scientific exploration in this field.

REFERENCES

[1] Campbell C, Wang T, McNaughton AL, Barnes E, Matthews PC. Risk factors for the development of hepatocellular carcinoma (HCC) in chronic hepatitis B virus (HBV) infection: a systematic review and meta-analysis. *Journal of viral hepatitis*. 2021; **28**: 493-507.

[2] Alim A, Karatas C. Prognostic Factors of Liver Transplantation for HCC: Comparative Literature Review. *J Gastrointest Cancer*. 2021.

[3] Sung H, Ferlay J, Siegel RL*, et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021; **71**: 209-49.

[4] Allemani C, Matsuda T, Di Carlo V*, et al.* Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet (London, England)*. 2018; **391**: 1023-75.

[5] McGlynn KA, Petrick JL, El-Serag HB. Epidemiology of Hepatocellular Carcinoma. *Hepatology*. 2021; **73 Suppl 1**: 4-13.

[6] von Felden J, Schulze K, Krech T*, et al.* Circulating tumor cells as liquid biomarker for high HCC recurrence risk after curative liver resection. *Oncotarget*. 2017; **8**: 89978-87.

[7] Shruthi BS, Vinodhkumar P, Selvamani. Proteomics: A new perspective for cancer. *Adv Biomed Res*. 2016; **5**: 67.

[8] Zhang X, Yu W, Cao X, Wang Y, Zhu C, Guan J. Identification of Serum Biomarkers in Patients with Alzheimer's Disease by 2D-DIGE Proteomics. *Gerontology*. 2022: 1-13.

[9] Mota FSB, Nascimento KS, Oliveira MV*, et al.* Potential protein markers in children with Autistic Spectrum Disorder (ASD) revealed by salivary proteomics. *International journal of biological macromolecules*. 2022; **199**: 243-51.

[10] Chong L, Zhu Y. Mass spectrometry-based proteomics for abiotic stress studies. *Trends Plant Sci*. 2022.

[11] Wu P, Heins ZJ, Muller JT*, et al.* Integration and Analysis of CPTAC Proteomics Data in the Context of Cancer Genomics in the cBioPortal. *Mol Cell Proteomics*. 2019; **18**: 1893-8.

[12] Tong M, Yu C, Zhan D*, et al.* Molecular subtyping of cancer and nomination of kinase candidates for inhibition with phosphoproteomics: Reanalysis of CPTAC ovarian cancer. *EBioMedicine*. 2019; **40**: 305-17.

[13] Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics (Oxford, England)*. 2018; **34**: 1615-7.

[14] Feng G, Zheng KI, Li YY*, et al.* Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD. *Journal of hepato-biliary-pancreatic sciences*. 2021; **28**: 593-603.

[15] Spann A, Yasodhara A, Kang J*, et al.* Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review. *Hepatology*. 2020; **71**: 1093-105.

[16] Chan EE, Chow PK. A review of prognostic scores after liver resection in hepatocellular carcinoma: the MSKCC, SLICER and SSCLIP scores. *Jpn J Clin Oncol*. 2017; **47**: 287-93.

[17] Edwards NJ, Oberti M, Thangudu RR*, et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res*. 2015; **14**: 2707-13.

[18] Gao Q, Zhu H, Dong L*, et al.* Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell*. 2019; **179**: 561-77 e22.

[19] Li A, Estigoy C, Raftery M*, et al.* Heart research advances using database search engines, Human Protein Atlas and the Sydney Heart Bank. *Heart Lung Circ*. 2013; **22**: 819-26.

[20] Miura K, Ishida K, Fujibuchi W*, et al.* Differentiating rectal carcinoma by an immunohistological analysis of carcinomas of pelvic organs based on the NCBI Literature Survey and the Human Protein Atlas database. *Surg Today*. 2012; **42**: 515-25.

[21] Zhu Y, Liu Z, Lv D*, et al.* Identification of PYGL as a key prognostic gene of glioma by integrated bioinformatics analysis. *Future Oncol*. 2022.

[22] Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. *Anesthesiology*. 2020; **132**: 379-94.

[23] Ahmad MW, Reynolds J, Rezgui Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*. 2018; **203**: 810-21.

[24] Gao B, Wu TC, Lang S*, et al.* Machine Learning Applied to Omics Datasets Predicts Mortality in Patients with Alcoholic Hepatitis. *Metabolites*. 2022; **12**.

[25] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*. 2020; **408**: 189-215.

[26] Fernandez M, Ban F, Woo G*, et al.* Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *Journal of Chemical Information and Modeling*. 2018; **58**: 1533-43.

[27] Albaradei S, Thafar M, Alsaedi A*, et al.* Machine learning and deep learning methods that use omics data for metastasis prediction. *Computational and structural biotechnology journal*. 2021; **19**: 5008-18.

[28] Dehuri S, Roy R, Cho S-B, Ghosh A. An improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification. *Journal of Systems and Software*. 2012; **85**: 1333-45.

[29] Santra S, Jana M. Predicting the evolution of number of native contacts of a small protein by using deep learning approach. *Comput Biol Chem*. 2022; **97**: 107625.

[30] Feng G, Li XP, Niu CY*, et al.* Bioinformatics analysis reveals novel core genes associated with nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Gene*. 2020; **742**: 144549.

[31] Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal*. 2015; **7**: 19-33.

[32] Meyer D, Wien FT. Support vector machines. *The Interface to libsvm in package e1071*. 2015; **28**.

[33] Kang C, Jia X, Liu H. Development and validation of a RNA binding protein gene pair-associated prognostic signature for prediction of overall survival in hepatocellular carcinoma. *Biomed Eng Online*. 2020; **19**: 68.

[34] Kang L, Zhang ZH, Zhao Y. SCAMP3 is regulated by miR-128-3p and promotes the metastasis of hepatocellular carcinoma cells through EGFR-MAPK p38 signaling pathway. *Am J Transl Res*. 2020; **12**: 7870-84.

[35] Zhou A, Liu H, Tang B. Comprehensive Evaluation of Endocytosis-Associated Protein SCAMP3 in Hepatocellular Carcinoma. *Pharmgenomics Pers Med*. 2020; **13**: 415-26.

[36] Kim W, Khan SK, Gvozdenovic-Jeremic J*, et al.* Hippo signaling interactions with Wnt/β-catenin and Notch signaling repress liver tumorigenesis. *The Journal of clinical investigation*. 2017; **127**: 137-52.

[37] Chen TW, Yin FF, Yuan YM*, et al.* CHML promotes liver cancer metastasis by facilitating Rab14 recycle. *Nat Commun*. 2019; **10**: 2510.

[38] Kurebayashi Y, Ojima H, Tsujikawa H*, et al.* Landscape of immune microenvironment in hepatocellular carcinoma and its additional impact on histological and molecular classification. *Hepatology*. 2018; **68**: 1025-41.

[39] Jin Z, Lei L, Lin D*, et al.* IL-33 Released in the Liver Inhibits Tumor Growth via Promotion of CD4(+) and CD8(+) T Cell Responses in Hepatocellular Carcinoma. *J Immunol*. 2018; **201**: 3770-9.

[40] Xu X, Tan Y, Qian Y*, et al.* Clinicopathologic and prognostic significance of tumor-infiltrating CD8+ T cells in patients with hepatocellular carcinoma: A meta-analysis. *Medicine (Baltimore)*. 2019; **98**: e13923.

[41] Chen Y, Wen H, Zhou C*, et al.* TNF-α derived from M2 tumor-associated macrophages promotes epithelial-mesenchymal transition and cancer stemness through the Wnt/β-catenin pathway in SMMC-7721 hepatocellular carcinoma cells. *Experimental cell research*. 2019; **378**: 41-50.

[42] Shao Y, Lo CM, Ling CC*, et al.* Regulatory B cells accelerate hepatocellular carcinoma progression via CD40/CD154 signaling pathway. *Cancer letters*. 2014; **355**: 264-72.

[43] Giraud J, Chalopin D, Blanc JF, Saleh M. Hepatocellular Carcinoma Immune Landscape and the Potential of Immunotherapies. *Front Immunol*. 2021; **12**: 655697.

[44] Budhu A, Forgues M, Ye QH*, et al.* Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell*. 2006; **10**: 99-111.

FIGURE LEGENDS

Figure1. The flow chart summarizing the screening process of important proteins.

Figure 2. Functions of the identified differentially expressed proteins using GO enrichment (A) and KEGG pathway analysis (B).

Figure 3. Kaplan-Meier survival curve analysis for seven differentially expressed proteins.

Figure 4. Heat map of 7 key differentially expressed proteins between the HCC non-recurrence group (marked as “A”) and the HCC recurrence group (marked as “B”)

Figure 5. (A) ROC curve comparisons of the different algorithms.

(B) Ranking of the importance of the seven differentially expressed proteins.

Figure 6. The relationship between key differentially expressed proteins and infiltrating immune cells.

SUPPLEMENTARY MATERIAL

Supplementary Table 1. 690 differentially expressed proteins.

Supplementary Table 2. 39 survival-related proteins by Kaplan-Meier survival curve analysis (P < 0.05).

Supplementary Table 3. Univariable Cox regression analysis of the proteins (P < 0.05).

Supplementary Table 4. Multivariable Cox regression analysis of the proteins (P < 0.05).

Supplementary Figure1. The protein-protein interaction network.

Supplementary Figure2. Survival ROC curves of seven important proteins (area under of 5-years survival ROC curves > 0.7).

Supplementary Figure 3. Representative protein expressions of RUFY1, TMEM230, STK3, and ESF1 explored in the HPA database.

Supplementary Figure 4. ESF1, SCAMP3, RAP1GAP, BABCC1, STK3, and RUFY1 proteins significantly over-expressed in HCC. LIHC: Liver Hepatocellular Carcinoma.