

88835_Auto_Edited.docx

Name of Journal: *World Journal of Clinical Oncology*

Manuscript NO: 88835

Manuscript Type: ORIGINAL ARTICLE

Clinical and Translational Research

Establishment of a prognosis predictive model for liver cancer based on the expression of genes in the ubiquitin-proteasome pathway

Prognosis model for liver cancer

Hua Li, Yi-Po Ma, Hai-Long Wang, Cai-Juan Tian, Yi-Xian Guo, Hong-Bo Zhang, Xiao-Min Liu, Peng-Fei Liu

Hua Li, ² Department of Endoscopy, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy of Tianjin, Tianjin's Clinical Research Center for Cancer, Tianjin 300060, China

Yi-Po Ma, Department of Critical Care Medicine, Dingzhou City People's Hospital, Dingzhou City 073000, Hebei, China

Hai-Long Wang, Peng-Fei Liu, ³ Department of Oncology, Tianjin Academy of Traditional Chinese Medicine Affiliated Hospital, Tianjin 300120, ² China

Cai-Juan Tian, Hong-Bo Zhang, Tianjin Marvel Medical Laboratory, Tianjin Marvelbio Technology Co.,Ltd, Tianjin, China., Tianjin 300180, China

Yi-Xian Guo, Department of Intelligent, ⁵ Tianjin Yunquan Intelligent Technology Co., Ltd, Tianjin 300381, China

Xiao-Min Liu, Department of Oncology, Tianjin Huanhu Hospital, Tianjin 300350, China

Co-first authors: Hua Li and Yi-Po Ma.

Co-corresponding authors: Xiao-Min Liu and Peng-Fei Liu.

Author contributions: Xiaomin Liu and Pengfei Liu conceptualized and designed the research. Hua Li and Yipo Ma collected the data and wrote the manuscript. Hailong Wang conducted the data mining and prepared the figures. Caijuan Tian, Yixian Guo, and Hongbo Zhang conducted the bioinformatics analysis. All authors have read the manuscript and approved the approved the final manuscript.

Supported by the Tianjin Municipal Natural Science Foundation, No. 21JCYBJC01110.

Corresponding author: Peng-Fei Liu, MD, Chief Doctor, Doctor, Surgical Oncologist, Department of Oncology, Tianjin Academy of Traditional Chinese Medicine Affiliated Hospital, No. 354 North Road, Hongqiao District, Tianjin 300120, China. liupengfeitj@163.com

Received: October 11, 2023

Revised: December 27, 2023

Accepted:

Published online:

Abstract

BACKGROUND

Ubiquitin proteasome pathway (UPP) was proven to play important roles in cancer.

AIM

To investigate the prognostic significance of genes and build a predictive model in the UPP in liver cancer.

METHODS

In this study, UPP-related E1, E2, E3, deubiquitylating enzyme, and proteasome gene sets were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, aiming to screen the prognostic genes using univariate and multivariate regression analysis and form a prognosis predictive model in The Cancer Genome Atlas liver cancer cases.

RESULTS

Five genes (including Autophagy Related 10 (ATG10), Proteasome 20S Subunit Alpha 8 (PSMA8), Proteasome 20S Subunit Beta 2 (PSMB2), Ubiquitin Specific Peptidase 17 like family member 2 (USP17L2), and Ubiquitin Specific Peptidase 8 (USP8)) were proven significantly correlated with prognosis to form the predictive model of liver cancer. Among training, validation, and Gene Expression Omnibus (GEO) sets, the overall survival (OS) existed significantly between high-risk and low-risk groups. The expressions of the five genes were significantly associated with immunocyte infiltration, tumor stage, and postoperative recurrence. A total of 111⁸ differentially expressed genes (DEGs) were identified between high-risk and low-risk groups and they were enriched in 20 and five gene ontology (Go) and KEGG pathways. Cell division cycle 20 (CDC20), kelch repeat and BTB domain containing 11 (KBTBD11), and DDB1 and CUL4 associated factor 4 like 2 (DCAF4L2) were the DEGs in the E3 gene set that correlated with survival.

CONCLUSION

In conclusion, we constructed a prognosis-predictive model in patients with liver cancer, which contains five genes that associate with immunocyte infiltration, staging, and postoperative recurrence.

Key Words: Liver cancer; Ubiquitin proteasome pathway; Prognosis prediction; Gene expression; Immune infiltration

Li H, Ma YP, Wang HL, Tian CJ, Guo YX, Zhang HB, Liu XM, Liu PF. Establishment of a prognosis predictive model for liver cancer based on the expression of genes in the ubiquitin-proteasome pathway. *World J Clin Oncol* 2024; In press

Core Tip: This study unveils the crucial role of the ubiquitin proteasome pathway (UPP) in liver cancer prognosis. Five key genes (ATG10, PSMA8, PSMB2, USP17L2, USP8) identified from The Cancer Genome Atlas datasets form a robust prognostic model, accurately predicting liver cancer outcomes. Immunocyte infiltration analysis highlights associations with immune cell abundance, while clinical correlations link these genes to staging and recurrence. Differential gene expression and pathway enrichment elucidate underlying biological processes. E3 Ligase analysis identifies specific ligases (CDC20, KBTBD11, DCAF4L2) with significant expression differences, further emphasizing the UPP's integral role in liver cancer development and providing valuable insights for precision medicine and prognosis prediction.

INTRODUCTION

With the prevalence of liver cancer increasing, the annual growth rate is up to 2%-3% among male malignant tumors^[1] and the survival rate is 18% following the pancreatic cancer (9%) in 2020^[2]. 336400 new liver cancer cases were detected in China in 2016^[3],

and sugar-sweetened food causing the sharply elevating the incident rate (18.0 per 100000) of liver cancer must be given extra attention^[4].

Hepatitis B/C virus (HBV or HCV) infection, addicted to alcohol, liver cirrhosis, fatty hepatitis, and eating aflatoxin contaminated food are the risk factors for liver cancer^[5]. Imaging examinations for liver cancer include Ultrasonography, dynamic contrast-enhanced computed tomography (CT), multimodal MRI scan, 18F-fluorodeoxyglucose positron emission tomography/CT and so on. Based on virtual liver biopsy sampling pipeline for eliminating the biopsy sampling bias may be the potential diagnostic methods to investigate the nature of the lesions and etiology^[6]. In recent years, using statistical models combined with machine learning techniques to elevate the diagnostic precision of serum biomarkers such as α -fetoprotein (AFP) and cell-free DNA or RNA are widely applied to earlier diagnosis for Hepatocellular carcinoma^[7]. Additionally, surgical resection, transplantation, ablation, chemotherapy, and immunotherapy are common treatment options for liver cancer patients^[8]. However, effective surveillance and prediction the prognosis of liver cancer still face multiple challenges due to the heterogeneity.

The ubiquitin proteasome pathway (UPP) is one of the key pathways of protein selective degradation in organisms^[9], 26S proteasome and multi-catalytic protein hydrolytic complex are related to cell cycle, proliferation, differentiation, apoptosis, transcription, signal transduction, immune response, stress response, and extracellular effectors^[10]. The malfunction of UPP is linked to various diseases, such as nerve, carcinogenesis, infection, autoimmunity, and inflammation. Based on The Cancer Genome Atlas (TCGA) datasets and 961 UPSGs (ubiquitin proteasome-system genes), Liu *et al*^[11] found DDB1 and CUL4 associated factor 13 (DCAF13), cell division cycle 20 (CDC20), and proteasome 20S subunit beta 5 (PSMB5) has excellent performance to predict the survival of liver cancer patients. Zhang *et al*^[12] identified seven Ubiquitin proteasome system (UPS)-based prognostic signatures, while autophagy related 10 (ATG10) was correlation with liver cancer development and prognosis through autophagy, immune response and tumor metastasis. Therefore, proteasome inhibitors,

as a potential and effective anti-tumor drug, have attracted a growing body of attention and research. In this study, we examined the correlation the gene expression of UPP pathway with the prognosis of liver cancer, to screen out some key genes and construct a prognosis predictive model, in order to provide a new horizon for the role and potential mechanism of the UPP pathway in the development of liver cancer.

MATERIALS AND METHODS

Gene sets and data collection

The UPP-related gene set included 857 genes from the UPP-related Kyoto Encyclopedia of genes and genomes (KEGG) pathway [13], among which the E1 gene set contained ten genes, the E2 gene set had 38 genes, 651 genes were related to E3, 112 genes were related to deubiquitylating enzyme (DUB), and 46 genes were related with the proteasome.

A total of 424 expression data of samples related to liver cancer were downloaded from The Cancer Genome Atlas (TCGA) database (<https://portal.gdc.cancer.gov/>). Three recurrent samples, 50 normal tissue samples, and one sample without overall survival (OS) were deleted and 370 remaining samples were randomly divided into the training group ($n = 296$) and validation group ($n = 74$) based on the ratio of 4:1. Another validation set is the gse54236 data set downloaded from the Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/geo/>). The data set includes 162 samples, including 81 tumor samples.

Construction and validation of prognosis predictive model

Univariate and multivariate regression analyses were used to screen the prognostic genes in the E1, E2, DUB, and proteasome gene sets using Survival (version 3.2-3) and Glmnet (version 4.0-2) R packages. The threshold of univariate analysis was $P < 0.1$, and stepwise regression of multivariate analysis was used to screen genes associated with OS. The risk score of the screened genes was calculated to construct a prognosis predictive model, and the prognostic ability was judged according to the receiver

operating characteristic (ROC) curve drawn using Proc (version 1.16.2) package. According to the risk score, the patients were divided into high-risk and low-risk groups. The Maxstat (version 0.7-25) R package is used to calculate the optimal segmentation value. The log rank method was used to compare the differences of OS between the two groups, and Survival (version 3.2-3) and Survminer (version 0.4.8) R packages were used to draw the survival curve. In the validation group, the same method was used to verify the model.

Immunocyte infiltration

The abundance of 40 types of immune cells in each sample was analyzed using GVSA (version 1.32.0). The correlation analysis between the screened genes and immune-related indicators was performed using Psych (version 2.0.8) and Corrplot (version 0.84) R packages.

Analysis of correlation between clinical parameters and gene expressions

The clinical parameters were compared between the high-risk and the low-risk groups using an independent sample t-test or two-sample Wilcoxon test, and Spearman correlation analysis was used to determine whether the gene expressions and risk scores were statistically related to clinical parameters. The Psych (version 2.0.8) and Corrplot (version 0.84) R packages were used for plotting. Univariate Cox regression analysis was used to determine the relationship between OS and clinical parameters, as well as the relationship between the gene expressions and the occurrence of postoperative recurrence.

Differentially expressed genes and the enrichment analysis

Between the high-risk and low-risk groups, the Limma package(version 3.40.2) was used to screen DEGs, with the threshold of $P < 0.05$ and $|\log_2FC| > 1$. Then the functional enrichment analysis was carried out using the database for annotation,

visualization, and integrated discovery (DAVID, <https://david.ncicrf.gov/>) to analyze the enriched gene ontology (GO) terms and KEGG pathways of the DEGs.

Core DEGs in E3 gene set

As a specific substrate recognition element, E3 Ligase plays an important role in the ubiquitin-mediated proteolytic cascade^[14]. Because of its specificity, the relationship between the gene expressions in E3 and prognostic risks was analyzed separately. Similar to the DEGs screening method, the Limma package was used to screen the DEGs in the E3 gene set between the high-risk and the low-risk groups, and the screening threshold was $P < 0.05$ and $|\log_2FC| > 1$.

Statistical analysis

IBM SPSS Statistics 21 software and R (version 3.6.2) were used for statistical analysis. The Shapiro-Wilk test was used for normality test, the independent sample t test or the two-sample Wilcoxon test were used to analyze the differences of variables between two groups. The chi-square test or Fisher's test was used for analysis of categorical variables. The log-rank method was used to test the significance of survival data.

RESULTS

Constructed the prognosis-predicted model for liver cancer by using five genes including ATG10, PSMA8, PSMB2, USP17L2, and USP8

The clinical parameters of the whole samples, cases in the training group, and validation group are listed in Table 1. P value referred to the statistical results between the training group and the validation group. A total of five genes significantly related to prognosis were screened to construct the prognosis predictive model, namely ATG10, proteasome 20S subunit alpha 8 (PSMA8), proteasome 20S subunit beta 2 (PSMB2), ubiquitin specific peptidase 17 like family member 2 (USP17L2), and ubiquitin specific peptidase 8 (USP8) (Table 2). In the training group, the area under curve (AUC) values of one, three, five, and ten year survival were 0.724, 0.659, 0.643, and 0.624, respectively

(Figure 1A). High-risk and low-risk groups of all patients were classified according to the score of risk. Besides, there was the significantly difference between the high-risk and low-risk group ($P < 0.001$, Figure 1B). In the validation group, the AUC of one, three, five, and ten year survival were respectively 0.614, 0.66, 0.64, and 0.649, and the low-risk group exhibited higher survival probability than high-risk group ($P = 0.012$, Figure 1C and D), suggesting that the model can well predict the prognosis in liver cancer patients. In the GSE54236 set, the AUC of one, and three year survival were respectively 0.563, 0.678, and high-risk group and low-risk group could be classified through the risk score of the best dividing point method. There were significantly difference in survival time between the high-risk group and low-risk group ($P = 0.014$, Figure 1E and F).

Significance of five gene expressions in immunocyte infiltration

Through the above analysis, we found that the expression level of five genes can predict the prognosis of liver cancer. Because immunocyte infiltration is commonly affected by gene expressions, we then studied the correlation the expression level of the five genes with the 40 types immune cells' abundance. PSMA8 was associated with the abundance of the most immune cells, a total of 28 immune cells have significant differences at different PSMA8 expression levels. Followed by USP17L2, ATG10, USP8, and PSMA8, with 25, 23, 18, and 13 types of cells that related to the expression levels of these genes, respectively (Figure 2). The abundance of most cells is negatively correlated with the expression levels of ATG10, USP17L2, and USP8, while PSMA8 and PSMB2 are positively correlated with the abundance of most cells(Figure 2).

The five genes are mainly associated with stages and postoperative recurrence

PSMA8, PSMB2 expressions, and risk scores were significantly different between males and females (Figure 3A-C). For pathological and clinical staging, ATG10 expressions and risk scores were different between the T2 and T3 stage, expressions of PSMB2 and USP17L2 were significantly different between the T1 and T3 stage (Figure 3D-G); risk

score was statistically different between the N0 and NX stage, which were obviously lower in patients of the N0 stage than that of NX stage (Figure 3H); between patients in stage I and stage II, the levels of ATG10, PSMB2, USP17L2 expression, and risk scores were significantly different (Figure 3I-L). Moreover, PSMA8, PSMB2, USP17L2, and USP8 expressions were all correlated with albumin result upper limit, among which PSMA8 and USP17L2 were positively correlated, and PSMB2 and USP8 were negatively correlated (Figure 3M). There was also a negative correlation between risk score and albumin result upper limit, indicating that as the risk value increased, the albumin levels decreased, leading to an elevated prognosis risk for patients (Figure 3M).

Postoperative recurrence included extrahepatic recurrence, local recurrence, intrahepatic recurrence, and new primary tumor. After Cox univariate analysis, ATG10, PSMA8, and USP8 were found significantly correlated with postoperative recurrence, as well as risk score ($P < 0.05$, Figure 3N).

DEGs and their enriched pathways between high- and low-risk groups

A total of 111 DEGs were screened out between the high-risk group and low-risk group, among them 27 were up-regulated and 84 were down-regulated (Figure 4A). These DEGs were associated with 20 GO terms, comprising 9 biological processes, 6 cellular components, and 5 molecular functions (Figure 4B). Five KEGG pathways enriched were GABAergic synapse, morphine addiction, neuroactive ligand-receptor interaction, retrograde endocannabinoid signaling and cell cycle (Figure 4C).

DEGs in the E3 gene set ⁶ between the high- and low-risk groups

Among the high-risk and low-risk groups, significant differences were observed in three genes within the E3 gene set: CDC20, Kelch Repeat and BTB Domain Containing 11 (KBTBD11), and DDB1 and CUL4 Associated Factor 4 Like 2 (DCAF4L2). In the high-risk group, CDC20 and DCAF4L2 exhibited elevated expression levels, whereas KBTBD11 showed higher expression in the low-risk group. This suggests a negative

correlation between the expression of CDC20 and DCAF4L2 and survival, while KBTBD11 displayed a positive correlation with the prognosis of liver cancer.(Figure 5).

DISCUSSION

The key factor to cell survival lies in the balance of protein synthesis and decomposition. UPP is an ATP-dependent non-lysosomal protein degradation pathway, which is important for the body to regulate the level and function of intracellular proteins, thus efficiently and selectively degrading intracellular proteins. This study showed that UPP genes ATG10, PSMA8, PSMB2, USP17L2, and USP8 expressions were significantly correlated with the prognosis of liver cancer. The prognosis model constructed by these five genes could accurately predict the prognosis of patients ($P < 0.001$ and $P = 0.012$ in training and validation groups, Figure 1). These genes were statistically correlated with different clinical parameters and immune cell abundance (Figures 2 and 3). The model categorized all patients into either a high-risk group or a low-risk group, a total of 111 DEGs were screened between the two groups, which were enriched in GO terms related to protein binding, GABA-A receptor, synapse, *etc.*, and KEGG pathways of Retrograde endocannabinoid signaling, Neuroactive ligand-receptor interaction, Morphine addiction, GABAergic synapse, and Cell cycle (Figure 4).

In many malignant tumors, those five genes were found to promote the development of a variety of tumors, including liver cancer^[15-22]. Our results showed that the increased expression of ATG10, PSMA8, and PSMB2 increased the risk of death ($P = 0.018$, 0.049 , and 0.013); while the expressions of USP17L2 and USP8 increased and the corresponding risk of death decreased ($P = 0.002$ and 0.089). According to previous studies, the overexpression of ATG10 and PSMB2 in tumors promoted the invasion or metastasis of tumor cells^[16,18], and USP8 showed the opposite function^[21,22]. Besides, PSMA8 could affect the progression and prognosis of colorectal cancer due to the strong association with PSMB2^[23]. Interestingly, the higher PSMA8 expression levels were correlation with good prognoses for breast cancer through epigenetic regulation^[24]. In our study, it was found that it is positively correlated with the prognosis of patients

with liver cancer. On the contrary, USP17L2 has been found to be overexpressed in a variety of tumors^[19,20], which is similar to our results. However, recent studies have found that up-regulation of USP17L2 causes chemotherapy resistance in colorectal cancer, knockdown of USP17L2 could overcome bromodomain and extra-terminal domain inhibitor resistance in prostate cancer cells^[25,26]. Hence, the role of USP17L2 in liver cancer still requires further exploration.

The global immune system functions pose great technical challenges to the research of tumor immune interaction^[27,28]. Because immune infiltration plays a key role in the development of liver cancer^[27], we conducted a thorough correlation analysis to identify the immune cells associated with the prognosis model, subsequently. Minor alterations in the distribution of immune cells could potentially exert diverse impacts on the progression of tumors^[29]. In this study, the myeloid dendritic cell was the immune cell with a significant difference only between high and low expression of the PSMB2 gene, as well as neutrophils and Th17 between different expressions of the USP17L2 gene (Figure 2). However, no significant correlation was found between ticks and gene expressions, it is imperative to undergo additional confirmation and validation in an independent cohort. Furthermore, exploring the connection between the expression levels of some checkpoints and immune infiltration, as well as the tumor microenvironment (TME), will be a hotspot for future.

Moreover, one or more of the expressions of the five genes and risk score were different among T, N, and clinical staging (Figure 3). It is widely known that stage is a key prognostic factor for malignant tumors^[30]. At the same time, all genes except ATG10 and risk score were correlated with the upper limit of albumin (Figure 3). The risk score is not only statistically significant in different stages, but also negatively correlated with the upper limit of albumin, and also correlated with the occurrence of postoperative recurrence, which proves that our model has a certain value in clinical prediction of recurrence and prognosis.

E3 Ligase is the key factor in UPP, which can specifically recognize different substrates and show high selectivity of protein degradation. Therefore, we analyzed the

E3 gene set independently of E1, E2, DUB, and proteasome-related genes. Finally, the expression levels of CDC20, KBTBD11, and DCAF4L2 were identified as significantly different between high-risk and low-risk groups, which were also included in the above 111 DEGs. CDC20 plays a vital role for chromosome segregation and mitosis^[31]. It regulates the stability of phosphorylated Mitotic centromere-associated kinesin (MCAK) in metaphase-anaphase transition^[32], which may play a role as a cancer protein to promote the development and progression of liver cancer. In the study of Zheng *et al*^[33], CDC20, ¹proliferating cell nuclear antigen (PCNA), and minichromosome maintenance complex component 6 (MCM6) synergistically affect the regulation of the cell cycle and may be potential prognostic factors of liver cancer. Shi *et al*^[34] found that CDC20 serves as a crucial factor in the development of hepatocellular carcinoma (HCC) by controlling the prolyl-4-hydroxylase domain 3 (PHD3) protein. By analyzing four expression profiles from the GEO database, the upregulation of CDC20 in HCC tissues indicates poor OS and disease-free survival^[35]. Recently, KBTBD11 were identified as newly discovered adipogenesis-related genes^[36]. In diverse cancer types, such as colorectal cancer (CRC), hepatocellular carcinoma (HCC), and head and neck squamous cell carcinoma (HNSCC), the expression of KBTBD11 was significantly decreased in tumor tissues as compared to normal tissues.^[37] That is consistent with our results that patients in the high-risk group had lower KBTBD11 gene expression levels. DCAF4L2 is a member of the E3 Ligase complex, which is usually used as a mediator of protein-protein interaction and negatively regulates NF- κ B signal transduction. Overexpression of DCAF4L2 has been observed in human colon cancer^[38]. In a study of HCC, overexpression of DCAF4L2 is a common feature of nonalcoholic steatohepatitis-associated HCC and viral hepatitis-associated HCC, which can be used as a candidate therapeutic target for HCC^[39]. Our results also found overexpression of DCAF4L2 in high-risk patients, suggesting a poor diagnosis of patients with liver cancer.

One of the main shortcomings of this study is the lack of clinical cases. All the data are from TCGA and GEO, resulting in the lack of clinical data for some patients, it is unable to validate the expressions of the five genes and comprehensively analyze their

correlation with clinical and prognostic indicators. This constitutes a preliminary study, and the results reported are exploratory. We intend to validate these results and the detailed mechanisms in future studies.

CONCLUSION

In conclusion, we used gene expression data in TCGA to screen genes in the UPP pathway that significantly correlated with the prognosis of liver cancer. And combined with clinical characteristics to comprehensively analyze the prognostic relationship, immunocyte infiltration, and related pathways. Our findings indicated that UPP plays an important role in the development of liver cancer, which provides new insights into the early prediction of prognosis and precision medicine in liver cancer.

5%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.frontiersin.org Internet	53 words — 1%
2	referencecitationanalysis.com Internet	51 words — 1%
3	pubmed.ncbi.nlm.nih.gov Internet	34 words — 1%
4	www.sciencegate.app Internet	15 words — < 1%
5	buscador.una.edu.ni Internet	14 words — < 1%
6	www.dovepress.com Internet	14 words — < 1%
7	Yilong Pan, Linlin Wang, Yaofeng Xie, Yuan Tan, Cheng Chang, Xueshan Qiu, Xiaodong Li. "Characterization of differentially expressed plasma proteins in patients with acute myocardial infarction", Journal of Proteomics, 2020 Crossref	12 words — < 1%
8	bmccancer.biomedcentral.com Internet	12 words — < 1%

EXCLUDE QUOTES ON
EXCLUDE BIBLIOGRAPHY ON

EXCLUDE SOURCES OFF
EXCLUDE MATCHES < 12 WORDS