frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 27 March 2020
doi: 10.3389/fbioe.2020.00254

Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method

*Zi-Mei Zhang, Jiu-Xin Tan, Fang Wang, Fu-Ying Dao, Zhao-Yue Zhang and Hao Lin**

Key Laboratory for Neuro-Information of Ministry of Education, School of Life Sciences and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

Hepatocellular carcinoma (HCC) is a serious cancer which ranked the fourth in cancer-related death worldwide. Hence, more accurate diagnostic models are urgently needed to aid the early HCC diagnosis under clinical scenarios and thus improve HCC treatment and survival. Several conventional methods have been used for discriminating HCC from cirrhosis tissues in patients without HCC (CwoHCC). However, the recognition successful rates are still far from satisfactory. In this study, we applied a computational approach that based on machine learning method to a set of microarray data generated from 1091 HCC samples and 242 CwoHCC samples. The within-sample relative expression orderings (REOs) method was used to extract numerical descriptors from gene expression profiles datasets. After removing the unrelated features by using maximum redundancy minimum relevance (mRMR) with incremental feature selection, we achieved "11-gene-pair" which could produce outstanding results. We further investigated the discriminate capability of the "11-gene-pair" for HCC recognition on several independent datasets. The wonderful results were obtained, demonstrating that the selected gene pairs can be signature for HCC. The proposed computational model can discriminate HCC and adjacent non-cancerous tissues from CwoHCC even for minimum biopsy specimens and inaccurately sampled specimens, which can be practical and effective for aiding the early HCC diagnosis at individual level.

Keywords: hepatocellular carcinoma, early diagnosis, cirrhosis, REOs, mRMR, support vector machine

OPEN ACCESS

Edited by:
Meng Zhou,
Wenzhou Medical University, China

Reviewed by:
Lei Deng,
Central South University, China
Shaherin Basith,
Ewha Womans University,
South Korea

****Correspondence:***
Hao Lin
hlin@uestc.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *02 January 2020*
Accepted: *12 March 2020*
Published: *27 March 2020*

Citation:
Zhang Z-M, Tan J-X, Wang F, Dao F-Y, Zhang Z-Y and Lin H (2020) Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. Front. Bioeng. Biotechnol. 8:254. doi: 10.3389/fbioe.2020.00254

INTRODUCTION

Liver cancer is the fourth leading cause of death in patients with malignant cancerous (Indhumathy et al., 2018; Villanueva, 2019). Hepatocellular carcinoma (HCC), which accounts for approximately 90% of all liver cancer cases, is frequently diagnosed at a late stage and has a poor prognosis. Thus, the early HCC diagnosis is significant to improve the prognosis and survival of patients (Asia-Pacific Working Party on Prevention of Hepatocellular Carcinoma, 2010). At present, diagnosis of HCC is based on laboratory investigations and imaging techniques (El-Serag, 2011; Hartke et al., 2017). Nevertheless, for HCC, especially for early HCC, current serum biomarkers and tools, such as α-fetoprotein (AFP) and imaging techniques, displayed poor diagnostic sensitivity and specificity (Sun et al., 2015). Liver biopsy is regarded as a good diagnostic choice in clinical practice only when

Abbreviations: CwHCC, cirrhosis tissues in patients with HCC; CwoHCC, cirrhosis tissues in patients without HCC; HCC, hepatocellular carcinoma; IFS, incremental feature selection; mRMR, maximum redundancy minimum relevance; REOs, relative expression orderings; SVM, support vector machine.

imaging techniques cannot provide accurate identification of HCC (Russo et al., 2018). However, the biopsy location is usually inaccurate, which might result in inaccurately sampling and thus decrease the diagnosis successful rate (Forner et al., 2008). Therefore, it is necessary to design new methods or discovery new diagnostic signatures to assist the pathologists in the identification of early HCC using biopsy specimens, even inaccurately sampled biopsy specimens. It is likely that the adjacent non-cancerous tissues (cirrhosis tissues in patients with HCC or normal tissues in patients with HCC) can be affected by cancerous tissues, so that they may obtain some similar molecular characteristics of cancerous tissues (Budhu et al., 2006; Wei et al., 2014).

The existed diagnostic signatures are mainly on the basis of risk scores obtained from signature genes' expression (Wurmbach et al., 2007; Archer et al., 2009; Zhou et al., 2015, 2017; Qu et al., 2019), which are highly sensitive to measurement batch effects (Guan et al., 2018) and are hardly applied in clinical settings. Luckily the relative expression orderings (REO)-based strategy (Zhang et al., 2013; Zhou et al., 2013; Wang et al., 2015; Li et al., 2016), which was firstly proposed by Eddy et al. (2010), is highly robust against experimental batch effects (Cai et al., 2015; Ao et al., 2016; Zhao et al., 2016) and platform differences (Guan et al., 2016), partial RNA degradation (Chen et al., 2017; Liao et al., 2017, 2018; Tang et al., 2018) and uncertain sampling sites within the same cancer tissue (Cheng et al., 2017). And thus the REOs have been used in the early diagnosis of HCC (Ao et al., 2018), gastric cancer (Yan et al., 2019) and colorectal cancer (Guan et al., 2019). In 2018, Ao et al. (2018) obtained 19 gene pairs by using the within-sample REOs. These genes could improve early HCC diagnosis using biopsy specimens, even inaccurately sampled biopsy specimens. However, the rule to identify HCC based on REOs is so simply that some intrinsic relationships among these genes are not revealed. Moreover, the accuracy for HCC diagnosis should still be improved.

Machine learning method is a good choice to uncover underlying patterns (Stephenson et al., 2019). It has been widely employed in bioinformatics (Cao et al., 2017; Bao et al., 2019; Conover et al., 2019; Moritz et al., 2019; Stephenson et al., 2019; Zou and Ma, 2019; Sun et al., 2020). The current work aims to develop a machine learning based method to diagnose HCC within-sample REOs. By removing redundant REOs using minimum redundancy maximum relevance (mRMR), a diagnostic signature consisting of 11 gene pairs was obtained. These signatures were also applied in some independent datasets for examining the performance of these gene pairs for HCC identification. High accuracies were obtained, suggesting that the obtained 11-gene-pair signature based on mRMR is better than the existed 19-gene-pair signature gained by Ao et al. (Ao et al., 2018).

MATERIALS AND METHODS

Data Collection and Preprocessing

The gene expression profiles datasets were freely gained from GEO (Barrett et al., 2005) and TCGA (Tomczak et al., 2015) database. Firstly, according to the type and sampling method of samples, the training datasets were derived from biopsy samples of HCC (D1), surgery samples of HCC (D2), biopsy samples of CwoHCC (D3), and surgery samples of CwoHCC (D4), respectively. To objectively evaluate the model, we separated the samples of each type (D1, D2, D3, and D4) mentioned above into two data subsets: training (80% samples of each type) and testing datasets (20% samples of each type). Finally, the training datasets contained 1091 HCC samples (112 biopsy samples of HCC and 979 surgery samples of HCC) and 242 CwoHCC samples (70 biopsy samples of CwoHCC and 172 surgery samples of CwoHCC). The testing datasets contained 73 biopsy samples (29 HCC samples and 44 CwoHCC samples) and 263 surgery samples (245 HCC samples and 18 CwoHCC samples). The independent datasets, which was comprised of surgical resection samples and biopsy samples, was used to evaluate the performance signature. We used the R package of TCGAbiolinks (Colaprico et al., 2016) to download the gene expression data which including 371 HCC and 50 normal tissues in patients from TCGA data resource[1] (up to October 19, 2019). The details have been listed in **Supplementary Table S1**.

For the raw data (.CEL files) detected by the Affymetrix platform, the RMA (Robust Multi-array Average) algorithm was used for background adjustment. If a gene was matched to multiple probes, the arithmetic mean expression value was used as the gene expression level. For the data sets detected by the Illumina platforms, we directly used the processed expression data.

The Within−Sample Relative Expression Orderings

Within a sample, the REOs of two genes (*a* and *b*) is expressed as $Ea > Eb$ (or $Ea < Eb$) if gene *a* has higher (or lower) expression level than gene *b*. The REOs pattern of a gene pair is regarded as stable if the REOs kept in at least 95% of the samples. A reversal gene pair is a gene pair with stable REOs in both cirrhosis tissues in patients without HCC (CwoHCC) samples and HCC samples, but the REOs patterns are reversed in the second group ($Ea < Eb$ or $Ea > Eb$ in CwoHCC samples but $Ea > Eb$ or $Ea < Eb$ in HCC samples). Here, the reversal gene pairs are selected as the candidate REOs signature for the identification of HCC. Then we obtained the common genes between training datasets and validation datasets and its corresponding gene expression profile. Subsequently, based on the gene expression profiles and reversal gene pairs, we generate a new profile by using 1, 0, and −1 to represent $Ea > Eb$, $Ea < Eb$, and other cases (*Ea* or *Eb* do not exist), respectively.

Feature Selection Through mRMR and IFS Methods

Based on the new profiles, mRMR (minimum Redundancy Maximum Relevance) (Peng et al., 2005) was applied to ranking the gene pairs based on the conditions of maximum relevance with the disease type along with minimum redundancy with other gene pairs.

[1] https://portal.gdc.cancer.gov/repository

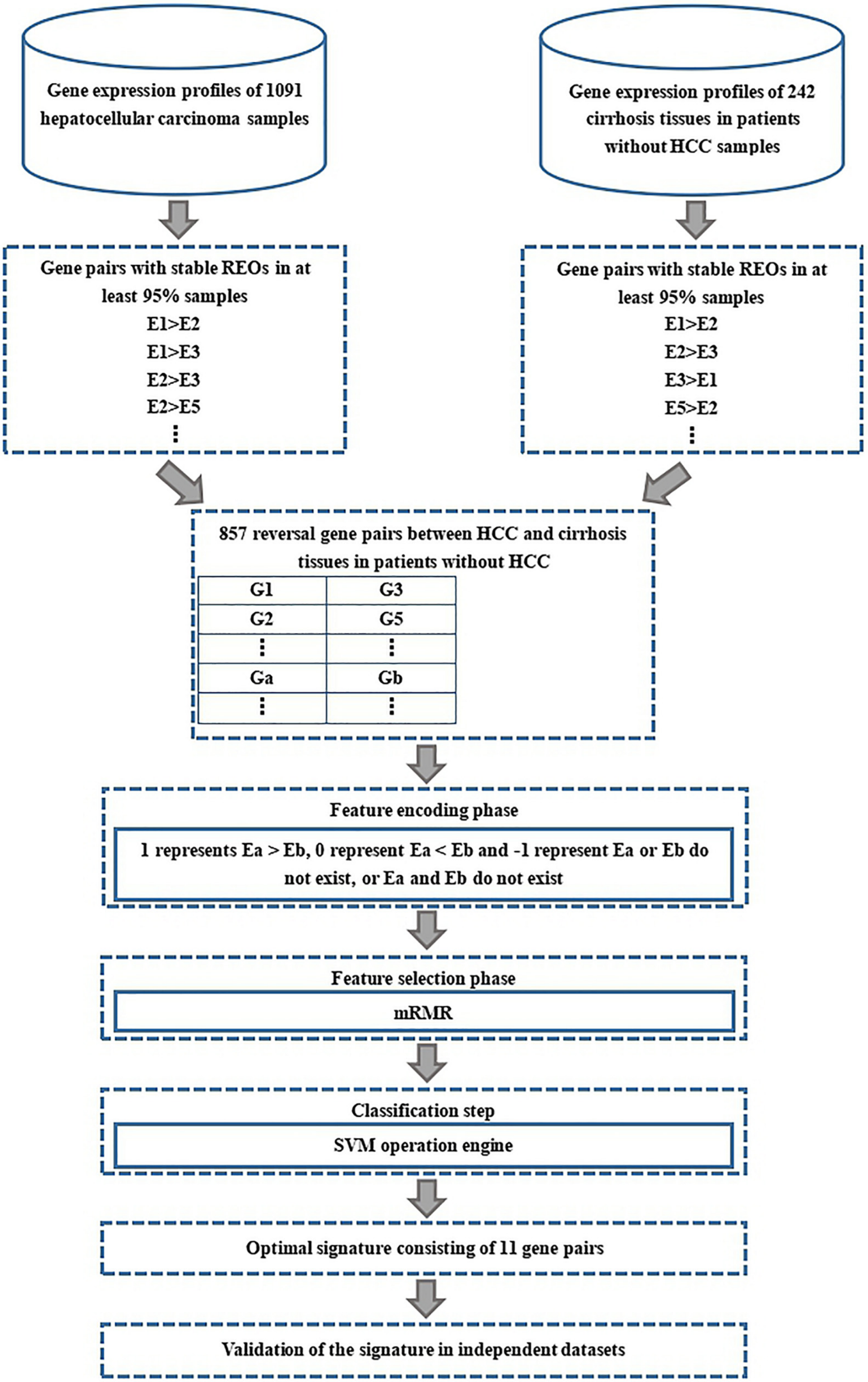
Gene expression profiles of 1091 hepatocellular carcinoma samples
Gene expression profiles of 242 cirrhosis tissues in patients without HCC samples
Gene pairs with stable REOs in at least 95% samples
E1>E2
E1>E3
E2>E3
E2>E5
⋮
Gene pairs with stable REOs in at least 95% samples
E1>E2
E2>E3
E3>E1
E5>E2
⋮
857 reversal gene pairs between HCC and cirrhosis tissues in patients without HCC
G1 G3
G2 G5
⋮ ⋮
Ga Gb
⋮ ⋮
Feature encoding phase
1 represents Ea > Eb, 0 represent Ea < Eb and -1 represent Ea or Eb do not exist, or Ea and Eb do not exist
Feature selection phase
mRMR
Classification step
SVM operation engine
Optimal signature consisting of 11 gene pairs
Validation of the signature in independent datasets

FIGURE 1 | Flowchart presenting the process of developing and validating the HCC diagnostic signature.

Here, Ω represents all 857 gene pairs, *gi* is a gene pair from the 857 gene pairs and *T* is the disease type. The mutual information (*I*) can be formulated as:

$$I(g_i, T) = \int p(g_i, T) \ln\left(\frac{p(g_i, T)}{p(g_i)p(T)}\right) dg_i dT. \quad (1)$$

The mRMR function:

$$mRMR = \frac{1}{|\Omega|} \sum_{g_i \in \Omega} I(g_i, T) - \frac{1}{|\Omega|^2} \sum_{g_i g_j \in \Omega} I(g_i, g_j) \quad (2)$$

where *I(gi, T)* is mutual information between the *gi* gene pair and disease type *T*, *I(gi, gj)* is mutual information between *gi* and *gj*. Then we used incremental feature selection (IFS) (Tan et al., 2019; Yang et al., 2019) method to select the optimal gene pairs from 857 mRMR gene pairs as diagnostic signature. The details about IFS can be found in (Dao et al., 2019).

Classification Through SVM

Support Vector Machine (SVM) is a powerful classification method which has been used extensively in the fields of biological data mining (Cao et al., 2014; Manavalan and Lee, 2017; Manavalan et al., 2017, 2018b, 2019c,d; Tang et al., 2017; Bu et al., 2018; Zhang et al., 2018; Chao et al., 2019a,b; Wang et al., 2019). Here, the free package LibSVM (version 3.23) (Chang and Lin, 2011) was downloaded to implement SVM. Due to its good performance on non-linear problem, RBF (radial basis function) was utilized. The values of two parameters *C* and γ for SVM are determined by the use of grid search with fivefold cross-validation. In present work, the optimal values are *C* = 0.125 and γ = 0.5, respectively.

Performance Metrics

The sensitivity, specificity and accuracy (Basith et al., 2019; Manavalan et al., 2018a,c, 2019a,b) was applied to evaluating the performance of prediction methods. Here, HCC samples were regarded as positive samples; CwoHCC samples were negative samples. Mathematical representation of the above mentioned measures are calculated as:

$$\begin{cases} \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \end{cases} \quad (3)$$

where TP, FN, TN, and FP denotes the number of true positives, false negatives, true negatives, and false positives, respectively. Additionally, the ROC curve and AUC are commonly used to test the balance between true positive rate and false positive rate.

RESULTS

Identification of the Diagnostic Signature

The flow diagram for identifying and validating the diagnostic signature is shown in **Figure 1**. Firstly, total of 13,586,043 stable gene pairs which have an identical REOs in at least 95% of the 1091 HCC samples were recognized. Similarly, we also identified 14,475,509 stable gene pairs which have an identical REOs in at least 95% of the 242 CwoHCC samples. Then, we obtained 857 reversal gene pairs between the HCC samples and CwoHCC samples in the training data (see section "Materials and Methods"). Based on the new profiles (see section "Materials and Methods"), 11 gene pairs shown in **Table 1** were picked out by using mRMR with SVM and regarded as the diagnostic signature. The 11-gene-pair could produce the accuracy of 100% on training data for HCC identification. **Figure 2** showed the IFS process (blue curve).

Examination of the Diagnostic Signature on Independent Datasets

Subsequently, we used biopsy and surgically resected samples to estimate the performance of the 11-gene-pair (see **Table 2**). For 73 biopsy samples in the testing datasets, it yielded accuracy of 100%, sensitivity of 100%, specificity of 100%. For 263 surgically resected samples in the testing datasets, its accuracy is 100%, sensitivity 100%, specificity 100%. In the data set GSE121248, all (100.0%) of the 70 HCC samples were correctly recognized as HCC. For surgically resected samples, 79.79% of the 475 HCC samples from 3 datasets (GSE109211, GSE112790, and GSE102079) were correctly classified. Moreover, the 11-gene-pair based model could correctly identify the 371 HCC and the 50 normal tissues in patients with HCC (NwHCC) samples measured by RNA-seq, in which no RNA-seq information was included (**Table 2**). These results demonstrated that the 11-gene-pair signature could distinguish HCC from non-cancerous liver tissues and the signature was robust to clinicopathological variations. For the 1190 HCC samples and 62 CwoHCC samples, the sensitivity, specificity, and AUC are 91.93%, 100%, and 0.9597 [95% CI (confidence intervals) is 0.9519–0.9674; see in **Figure 3**], respectively.

For biopsy samples, all of 80 cirrhosis tissues in patients with HCC (CwHCC) samples in GSE54236 and all of 97 NwHCC biopsy tissues from 2 datasets (GSE64041 and GSE121248) were

TABLE 1 | The 11−gene−pair signature for early diagnosis of HCC.

Signature	Gene *a*	Gene *b*
pair1	TRMT112	SF3B1
pair2	MFSD5	COLEC10
pair3	FDXR	APC2
pair4	LAMC1	CHST4
pair5	UBE4B	HGF
pair6	NCAPH2	APC2
pair7	HSPH1	MTHFD2
pair8	TMEM38B	AGO3
pair9	PLGRKT	COLEC10
pair10	HNF1A	APC2
pair11	ARPC2	SF3B1

Gene a has a higher expression level than Gene b in HCC patients compared with CwoHCC patients.

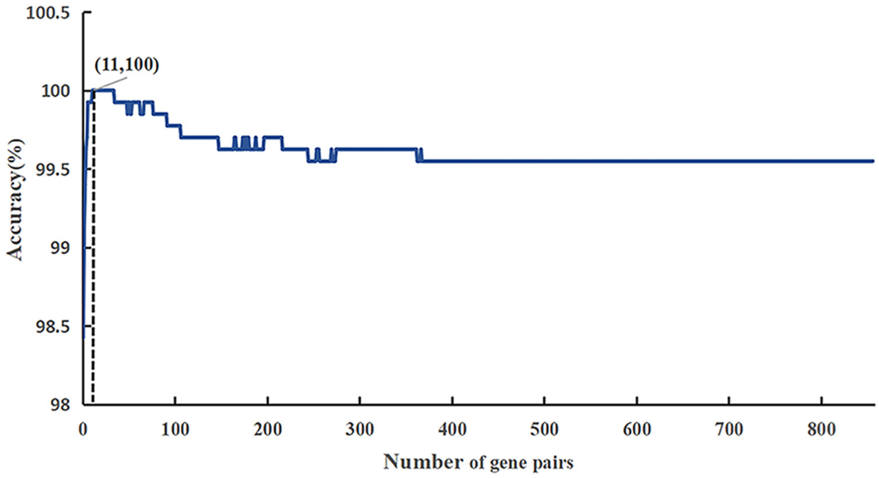

FIGURE 2 | A plot showing the IFS procedure for identifying HCC. When the top 857 features optimized by mRMR were used to perform prediction, the overall success rate reaches an IFS peak of 100% in fivefold cross validation. The solid line represents the ROC curve. The dotted line represents the strategy of randomly guess.

TABLE 2 | The performance of the signature in the validation datasets.

Datasets	NSnHCC	NSpCwoHCC
Testing datasets (biopsy)	100% (29/29)	100% (44/44)
Testing datasets (surgery)	100% (245/245)	100% (18/18)
GSE109211	31.43% (44/140)	–
GSE112790	100% (183/183)	–
GSE102079	100% (152/152)	–
GSE121248	100% (70/70)	–
TCGA	100% (371/371)	–

NSnHCC, number (sensitivity) of HCC samples; NSpCwoHCC, number (specificity) of CwoHCC samples.

correctly classified to HCC. The results proved again that, the 11-gene-pair still displayed good performance that most of HCC adjacent non-cancerous patients (CwHCC and NwHCC) can be correctly recognized, even for the inaccurate samples from biopsy specimens. For surgically resected samples, 93.7% of the 254 CwHCC samples and 100% of the 644 NwHCC samples can be accurately identified (see in **Table 3**). All above results demonstrated again that the obtained 11-gene-pair could be regarded as key biological signatures to diagnose HCC patients.

Comparison With Existing Methods

To further demonstrate the performance of our proposed signatures, we compared our method with 19-gene-pair-based models and recorded results in **Table 3**. An earlier work done by Ao et al. (2018) found that 19-gene-pair can be regarded as diagnostic signature to discriminate HCC and adjacent non-cancerous tissues (cirrhosis or normal) from CwoHCC. Their model could produce 99.69% of accuracy which is lower than that of our 11-gene-pair based model.

For biopsy samples, our proposed model could correctly identify the 70 HCC samples in GSE121248 and the 97 NwHCC biopsy tissues from 2 datasets (GSE64041 and GSE121248) with the accuracy of 100%. Moreover, all 80 CwHCC samples in GSE54236 can be predicted as HCC. Compared with the accuracy (77.5%) of 19-gene-pair based model, the accuracy of 11-gene-pare model could increase to 100%.

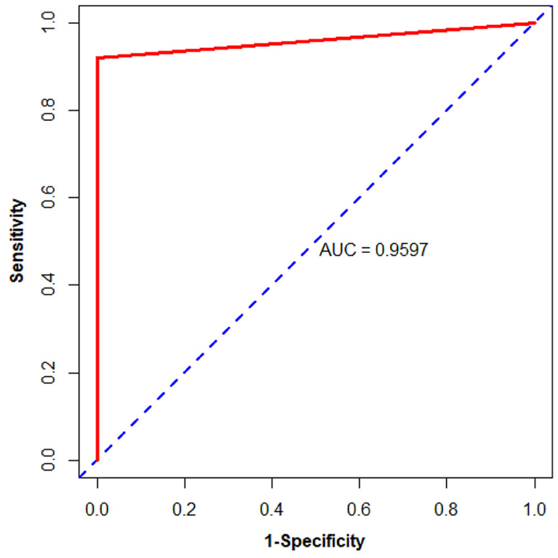

FIGURE 3 | Area under the receiver operating characteristic curve (AUC) of the validation data from public databases of biopsy and surgically resected HCC and CwoHCC samples. The solid line represents the ROC curve. The dotted line represents the strategy of randomly guess.

For surgically resected samples, based on the predictor of 11-gene-pair, 79.8% of the 475 HCC samples from 3 datasets (GSE109211, GSE112790, and GSE102079) and 93.7% of the 254 CwHCC samples from 5 datasets (GSE6764, GSE17548, GSE25097, GSE17967, and GSE63898) can be corrected as HCC. Moreover, the model can accurately predict the 644 NwHCC biopsy tissues integrated from 7 datasets (GSE25097,

TABLE 3 | Comparison of 11 gene pairs with existing methods on independent datasets.

Dataset	**11-gene-pair**			**19-gene-pair**		
	NSnHCC	**NACwHCC**	**NANwHCC**	**NSnHCC**	**NACwHCC**	**NANwHCC**
Datasets from surgical resection						
GSE6764	–	10/10 (100.0%)	–	–	10/10 (100.0%)	–
GSE17548	–	18/20 (90.0%)	–	–	18/20 (90.0%)	–
GSE17967	–	16/16 (100.0%)	–	–	8/16 (50.0%)	–
GSE63898	–	168/168 (100.0%)	–	–	168/168 (100.0%)	–
GSE25097	–	40/40 (100.0%)	243/243 (100.0%)	–	40/40 (100.0%)	243/243 (100.0%)
GSE62232	–	–	10/10 (100.0%)	–	–	10/10 (100.0%)
GSE36376	–	–	193/193 (100.0%)	–	–	172/193 (89.1%)
GSE39791	–	–	72/72 (100.0%)	–	–	71/72 (98.6%)
GSE41804	–	–	20/20 (100.0%)	–	–	20/20 (100.0%)
GSE112790	183/183 (100.0%)	–	15/15 (100.0%)	183/183 (100.0%)	–	15/15 (100.0%)
GSE102079	152/152 (100.0%)	–	91/91 (100.0%)	152/152 (100.0%)	–	91/91 (100.0%)
GSE109211	44/140 (31.4%)	–	–	37/140 (26.4%)	–	–
Total	379/475 (79.8%)	238/254 (93.7%)	644/644 (100.0%)	372/475 (79.3%)	244/254 (96.1%)	622/644 (96.6%)
Datasets from biopsy						
GSE121248	70/70 (100.0%)	–	37/37 (100.0%)	70/70 (100.0%)	–	37/37 (100.0%)
GSE64041	–	–	60/60 (100.0%)	–	–	60/60 (100.0%)
GSE54236	–	80/80 (100.0%)	–	–	62/80 (77.5%)	–
Total	70/70 (100.0%)	80/80 (100.0%)	97/97 (100.0%)	70/70 (100.0%)	62/80 (77.5%)	97/97 (100.0%)

NACwHCC, number (accuracy) of cirrhosis tissues in patients with HCC samples to HCC; NANwHCC, number (accuracy) of normal tissues in patients with HCC samples to HCC.

GSE62232, GSE36376, GSE39791, GSE41804, GSE112790, and GSE102079). Also, the sensitivity of HCC samples increases to 79.8% (19-gene-pair: 79.3%) and the accuracy of NwHCC samples to HCC increases to 100% (19-gene-pair: 96.6%). It can be seen from **Table 3** that in the identification of both HCC and adjacent non-cancerous tissues (CwHCC and NwHCC) from CwoHCC by surgically resected samples, the 11-gene-pair based model displayed better performance than the 19-gene-pair based model, demonstrating that the 11-gene-pair-based model is quite promising in generating reliable results for the early HCC diagnosis.

The above results showed that the proposed 11-gene-pair-based model is powerful on both training datasets and independent datasets. This achievement can be attribute to using within-sample REOs and SVM.

DISCUSSION

Clinical practice has demonstrated that diagnosing the tumors in early stages is key to improve the survival of patient. Although pathology is used as a gold standard for HCC diagnosis, the histological analysis of the HCC biopsy specimen is influenced by the sampling location and tissue amount. In present work, a set of diagnostic signature including 11-gene-pair consisting of 18 genes was identified, which can be used to discriminate HCC and adjacent non-cancerous tissues (CwHCC and NwHCC) from CwoHCC individuals for the early HCC diagnosis.

Ten genes in the signature set, including LAMC1, UBE4B, HSPH1, HNF1A, SF3B1, APC2, CHST4, HGF, MTHFD2, and AGO3, might have a vital role during the hepatocarcinogenesis and are key genes for cancer. For instance, LAMC1 mRNA can promote the development of HCC by competing with miR-124 and supporting the excretion of CD151 (Yang et al., 2017). UBE4B can be used as a potential prognostic marker for HCC treatment due to its carcinogenic effect in human primary HCC (Zhang et al., 2016). Additionally, HNF1A is closely associated with HCC because the number of HNF1A increase when non-cancerous liver develops into high differentiate HCC (Wang et al., 1998). SF3B1 is a highly conserved spliceosomal protein in evolution (Eilbracht and Schmidt-Zachmann, 2001) and its expression increases significantly in liver HCC tissues. Serum anti-SF3B1 autoantibody is a potential diagnostic marker for HCC patients (Hwang et al., 2018). Reportedly, HSPH1 (Yang et al., 2015), APC2 (Ghosh et al., 2016), CHST4 (Gao et al., 2015), HGF (Unic et al., 2018), MTHFD2 (Liu et al., 2016), and AGO3 (Kitagawa et al., 2013) are closely related to HCC.

Subsequently, the 18 genes (11-gene-pair) were used for functional enrichment analysis by using Metascape[2] (Tripathi et al., 2015) on the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and GO (Gene Ontology) terms. In order to determine the significant terms, p-value < 0.05 and the number of enriched genes ≥ 3 were used as the statistical standard. Finally, 18 genes were significantly enriched in the "ribonucleoprotein complex biogenesis," "positive regulation of cellular component biogenesis," "lymphocyte activation," and "chemotaxis" terms based on GO analysis, as well as "Pathways in cancer" according to KEGG analysis. The above analysis showed that the genes of the 11-gene-pair might have vital roles in the development and progression of HCC.

[2] http://metascape.org

In current study, we showed that 11 gen pairs can be applied to accurately diagnose the tumors found in the liver. Further, we shall try to establish a user-friendly web-server for the proposed "11-gene-pair" model. In the future, we will apply other feature selection techniques and algorithms to further improve the diagnosis of cancers.

DATA AVAILABILITY STATEMENT

The datasets used in this study can be freely download from the GEO (https://www.ncbi.nlm.nih.gov/geo/) and TCGA (https://portal.gdc.cancer.gov/repository) repository.

AUTHOR CONTRIBUTIONS

HL designed the study and revised the manuscript. Z-MZ carried out all the data collection and drafted the manuscript. Z-MZ, J-XT, FW, F-YD, and Z-YZ performed the data analysis. All authors approved the final manuscript.

FUNDING

This work was supported by the National Nature Scientific Foundation of China (61772119) and Sichuan Provincial Science Fund for Distinguished Young Scholars (20JCQN0262).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00254/full#supplementary-material

REFERENCES

Ao, L., Song, X., Li, X., Tong, M., Guo, Y., Li, J., et al. (2016). An individualized prognostic signature and multiomics distinction for early stage hepatocellular carcinoma patients with surgical resection. *Oncotarget* 7, 24097–24110. doi: 10.18632/oncotarget.8212

Ao, L., Zhang, Z., Guan, Q., Guo, Y., Guo, Y., Zhang, J., et al. (2018). A qualitative signature for early diagnosis of hepatocellular carcinoma based on relative expression orderings. *Liver Int.* 38, 1812–1819. doi: 10.1111/liv.13864

Archer, K. J., Mas, V. R., David, K., Maluf, D. G., Bornstein, K., and Fisher, R. A. (2009). Identifying genes for establishing a multigenic test for hepatocellular carcinoma surveillance in hepatitis C virus-positive cirrhotic patients. *Cancer Epidemiol. Biomarkers Prev.* 18, 2929–2932. doi: 10.1158/1055-9965.EPI-09-0767

Asia-Pacific Working Party on Prevention of Hepatocellular Carcinoma (2010). Prevention of hepatocellular carcinoma in the Asia-Pacific region: consensus statements. *J. Gastroenterol. Hepatol.* 25, 657–663. doi: 10.1111/j.1440-1746.2009.06167.x

Bao, S., Zhao, H., Yuan, J., Fan, D., Zhang, Z., Su, J., et al. (2019). Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer. *Brief. Bioinform.* doi: 10.1093/bib/bbz118 [Epub ahead of print].

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., et al. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.* 33, D562–D566. doi: 10.1093/nar/gki022

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids* 18, 131–141. doi: 10.1016/j.omtn.2019.08.011

Bu, H. D., Hao, J. Q., Guan, J. H., and Zhou, S. G. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method. *Curr. Bioinform.* 13, 655–660. doi: 10.2174/1574893613666180726163429

Budhu, A., Forgues, M., Ye, Q. H., Jia, H. L., He, P., Zanetti, K. A., et al. (2006). Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell* 10, 99–111. doi: 10.1016/j.ccr.2006.06.016

Cai, H., Li, X., Li, J., Ao, L., Yan, H., Tong, M., et al. (2015). Tamoxifen therapy benefit predictive signature coupled with prognostic signature of post-operative recurrent risk for early stage ER+ breast cancer. *Oncotarget* 6, 44593–44608. doi: 10.18632/oncotarget.6260

Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017). ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22:1732. doi: 10.3390/molecules22101732

Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics* 15:120. doi: 10.1186/1471-2105-15-120

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27.

Chao, L., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019a). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Chao, L., Wei, L., and Zou, Q. (2019b). SecProMTB: a SVM-based classifier for secretory proteins of *Mycobacterium tuberculosis* with imbalanced data set. *Proteomics* 19:e1900007.

Chen, R., Guan, Q., Cheng, J., He, J., Liu, H., Cai, H., et al. (2017). Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. *Oncotarget* 8, 6652–6662. doi: 10.18632/oncotarget.14257

Cheng, J., Guo, Y., Gao, Q., Li, H., Yan, H., Li, M., et al. (2017). Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget* 8, 30265–30275. doi: 10.18632/oncotarget.15754

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

Conover, M., Staples, M., Si, D., Sun, M., and Cao, R. (2019). AngularQA: protein model quality assessment with LSTM networks. *Comput. Math. Biophys.* 7, 1–9. doi: 10.1515/cmb-2019-0001

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Eddy, J. A., Sung, J., Geman, D., and Price, N. D. (2010). Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol. Cancer Res. Treat.* 9, 149–159. doi: 10.1177/153303461000900204

Eilbracht, J., and Schmidt-Zachmann, M. S. (2001). Identification of a sequence element directing a protein to nuclear speckles. *Proc. Natl. Acad. Sci. U.S.A.* 98, 3849–3854. doi: 10.1073/pnas.071042298

El-Serag, H. B. (2011). Hepatocellular carcinoma. *N. Engl. J. Med.* 365, 1118–1127. doi: 10.1056/NEJMra1001683

Forner, A., Vilana, R., Ayuso, C., Bianchi, L., Sole, M., Ayuso, J. R., et al. (2008). Diagnosis of hepatic nodules 20 mm or smaller in cirrhosis: prospective validation of the noninvasive diagnostic criteria for hepatocellular carcinoma. *Hepatology* 47, 97–104. doi: 10.1002/hep.21966

Gao, F., Liang, H., Lu, H., Wang, J., Xia, M., Yuan, Z., et al. (2015). Global analysis of DNA methylation in hepatocellular carcinoma by a liquid hybridization capture-based bisulfite sequencing approach. *Clin. Epigenetics* 7:86. doi: 10.1186/s13148-015-0121-1

Ghosh, A., Ghosh, A., Datta, S., Dasgupta, D., Das, S., Ray, S., et al. (2016). Hepatic miR-126 is a potential plasma biomarker for detection of hepatitis B virus infected hepatocellular carcinoma. *Int. J. Cancer* 138, 2732–2744. doi: 10.1002/ijc.29999

Guan, Q., Chen, R., Yan, H., Cai, H., Guo, Y., Li, M., et al. (2016). Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. *Oncotarget* 7, 68909–68920. doi: 10.18632/oncotarget.11996

Guan, Q., Yan, H., Chen, Y., Zheng, B., Cai, H., He, J., et al. (2018). Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer. *BMC Genomics* 19:99. doi: 10.1186/s12864-018-4446-y

Guan, Q., Zeng, Q., Yan, H., Xie, J., Cheng, J., Ao, L., et al. (2019). A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci.* 110, 3225–3234. doi: 10.1111/cas.14137

Hartke, J., Johnson, M., and Ghabril, M. (2017). The diagnosis and treatment of hepatocellular carcinoma. *Semin. Diagn. Pathol.* 34, 153–159. doi: 10.1053/j.semdp.2016.12.011

Hwang, H. M., Heo, C. K., Lee, H. J., Kwak, S. S., Lim, W. H., Yoo, J. S., et al. (2018). Identification of anti-SF3B1 autoantibody as a diagnostic marker in patients with hepatocellular carcinoma. *J. Transl. Med.* 16:177. doi: 10.1186/s12967-018-1546-z

Indhumathy, M., Nabhan, A. R., and Arumugam, S. (2018). A weighted association rule mining method for predicting HCV-human protein interactions. *Curr. Bioinform.* 13, 73–84. doi: 10.2174/1574893611666161123142425

Kitagawa, N., Ojima, H., Shirakihara, T., Shimizu, H., Kokubu, A., Urushidate, T., et al. (2013). Downregulation of the microRNA biogenesis components and its association with poor prognosis in hepatocellular carcinoma. *Cancer Sci.* 104, 543–551. doi: 10.1111/cas.12126

Li, X., Cai, H., Zheng, W., Tong, M., Li, H., Ao, L., et al. (2016). An individualized prognostic signature for gastric cancer patients treated with 5-Fluorouracil-based chemotherapy and distinct multi-omics characteristics of prognostic groups. *Oncotarget* 7, 8743–8755. doi: 10.18632/oncotarget.7087

Liao, Z., Li, D., Wang, X., Li, L., and Zou, Q. (2018). Cancer diagnosis from isomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Liao, Z., Wang, X., Lin, D., and Zou, Q. (2017). Construction and identification of the RNAi recombinant lentiviral vector targeting human DEPDC7 gene. *Interdiscip. Sci.* 9, 350–356. doi: 10.1007/s12539-016-0162-y

Liu, X., Huang, Y., Jiang, C., Ou, H., Guo, B., Liao, H., et al. (2016). Methylenetetrahydrofolate dehydrogenase 2 overexpression is associated with tumor aggressiveness and poor prognosis in hepatocellular carcinoma. *Dig. Liver Dis.* 48, 953–960. doi: 10.1016/j.dld.2016.04.015

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., and Lee, G. (2019a). 4mCpred-EL: an ensemble learning framework for identification of DNA N^4-methylcytosine sites in the mouse genome. *Cells* 8:1332. doi: 10.3390/cells8111332

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J.* 17, 972–981. doi: 10.1016/j.csbj.2019.06.024

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019c). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019d). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222

Manavalan, B., Shin, T. H., and Lee, G. (2018a). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi: 10.18632/oncotarget.23099

Manavalan, B., Shin, T. H., and Lee, G. (2018b). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O., and Lee, G. (2018c). Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* 17, 2715–2726. doi: 10.1021/acs.jproteome.8b00148

Moritz, S., Pfab, J., Wu, T., Hou, J., Cheng, J., Cao, R., et al. (2019). Cascaded-CNN: deep learning to predict protein backbone structure from high-resolution cryo-EM density maps. *BioRxiv [Preprint]* doi: 10.1038/s41598-020-60598-y

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Qu, K. Y., Gao, F., Guo, F., and Zou, Q. (2019). Taxonomy dimension reduction for colorectal cancer prediction. *Comput. Biol. Chem.* 83:107160. doi: 10.1016/j.compbiolchem.2019.107160

Russo, F. P., Imondi, A., Lynch, E. N., and Farinati, F. (2018). When and how should we perform a biopsy for HCC in patients with liver cirrhosis in 2018? A review. *Dig. Liver Dis.* 50, 640–646. doi: 10.1016/j.dld.2018.03.014

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2019). Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* 20, 185–193. doi: 10.2174/1389200219666180820112457

Sun, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110. doi: 10.1136/jitc-2019-000110

Sun, W., Liu, Y., Shou, D., Sun, Q., Shi, J., Chen, L., et al. (2015). AFP (alpha fetoprotein): who are you in gastrology? *Cancer Lett.* 357, 43–46. doi: 10.1016/j.canlet.2014.11.018

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Tang, H., Cao, R. Z., Wang, W., Liu, T. S., Wang, L. M., and He, C. M. (2017). A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.* 10:1750050. doi: 10.1142/s1793524517500504

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Tripathi, S., Pohl, M. O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D. A., et al. (2015). Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. *Cell Host Microbe* 18, 723–735. doi: 10.1016/j.chom.2015.11.002

Unic, A., Derek, L., Duvnjak, M., Patrlj, L., Rakic, M., Kujundzic, M., et al. (2018). Diagnostic specificity and sensitivity of PIVKAII, GP3, CSTB, SCCA1 and HGF for the diagnosis of hepatocellular carcinoma in patients with alcoholic liver cirrhosis. *Ann. Clin. Biochem.* 55, 355–362. doi: 10.1177/0004563217726808

Villanueva, A. (2019). Hepatocellular carcinoma. *N. Engl. J. Med.* 380, 1450–1462. doi: 10.1056/NEJMra1713263

Wang, H., Sun, Q., Zhao, W., Qi, L., Gu, Y., Li, P., et al. (2015). Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* 31, 62–68. doi: 10.1093/bioinformatics/btu522

Wang, W., Hayashi, Y., Ninomiya, T., Ohta, K., Nakabayashi, H., Tamaoki, T., et al. (1998). Expression of HNF-1 alpha and HNF-1 beta in various histological differentiations of hepatocellular carcinoma. *J. Pathol.* 184, 272–278. doi: 10.1002/(sici)1096-9896(199803)184:3<272::aid-path4>3.0.co;2-k

Wang, Y., Shi, F. Q., Cao, L. Y., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wei, L., Lian, B., Zhang, Y., Li, W., Gu, J., He, X., et al. (2014). Application of microRNA and mRNA expression profiling on prognostic biomarker discovery for hepatocellular carcinoma. *BMC Genomics* 15(Suppl. 1):S13. doi: 10.1186/1471-2164-15-S1-S13

Wurmbach, E., Chen, Y. B., Khitrov, G., Zhang, W., Roayaie, S., Schwartz, M., et al. (2007). Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology* 45, 938–947. doi: 10.1002/hep.21622

Yan, H., Li, M., Cao, L., Chen, H., Lai, H., Guan, Q., et al. (2019). A robust qualitative transcriptional signature for the correct pathological diagnosis of gastric cancer. *J. Transl. Med.* 17:63. doi: 10.1186/s12967-019-1816-4

Yang, W., Zhu, X. J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Yang, Z., Zhuang, L., Szatmary, P., Wen, L., Sun, H., Lu, Y., et al. (2015). Upregulation of heat shock proteins (HSPA12A, HSP90B1, HSPA4, HSPA5 and HSPA6) in tumour tissues is associated with poor outcomes from HBV-related early-stage hepatocellular carcinoma. *Int. J. Med. Sci.* 12, 256–263. doi: 10.7150/ijms.10735

Yang, Z. P., Ma, H. S., Wang, S. S., Wang, L., and Liu, T. (2017). LAMC1 mRNA promotes malignancy of hepatocellular carcinoma cells by competing for MicroRNA-124 binding with CD151. *IUBMB Life* 69, 595–605. doi: 10.1002/iub.1642

Zhang, L., Hao, C., Shen, X., Hong, G., Li, H., Zhou, X., et al. (2013). Rank-based predictors for response and prognosis of neoadjuvant taxane-anthracycline-based chemotherapy in breast cancer. *Breast Cancer Res. Treat.* 139, 361–369. doi: 10.1007/s10549-013-2566-2

Zhang, N., Yu, S., Guo, Y., Wang, L., Wang, P., and Feng, Y. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 13, 50–56. doi: 10.2174/1574893611666160608102537

Zhang, X. F., Pan, Q. Z., Pan, K., Weng, D. S., Wang, Q. J., Zhao, J. J., et al. (2016). Expression and prognostic role of ubiquitination factor E4B in primary hepatocellular carcinoma. *Mol. Carcinog.* 55, 64–76. doi: 10.1002/mc.22259

Zhao, W., Chen, B., Guo, X., Wang, R., Chang, Z., Dong, Y., et al. (2016). A rank-based transcriptional signature for predicting relapse risk of stage II colorectal cancer identified with proper data sources. *Oncotarget* 7, 19060–19071. doi: 10.18632/oncotarget.7956

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med.* 13:231. doi: 10.1186/s12967-015-0556-3

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16. doi: 10.1186/s12943-017-0580-4

Zhou, X., Li, B., Zhang, Y., Gu, Y., Chen, B., Shi, T., et al. (2013). A relative ordering-based predictor for tamoxifen-treated estrogen receptor-positive breast cancer patients: multi-laboratory cohort validation. *Breast Cancer Res. Treat.* 142, 505–514. doi: 10.1007/s10549-013-2767-8

Zou, Q., and Ma, Q. (2019). The application of machine learning to disease diagnosis and treatment. *Math. Biosci.* 320:108305. doi: 10.1016/j.mbs.2019.108305

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.