frontiers in Genetics

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fgene.2019.01306

Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data

Harpreet Kaur[1,2], *Anjali Dhall*[2], *Rajesh Kumar*[1,2] *and Gajendra P. S. Raghava*[2*]

[1] Bioinformatics Center, CSIR-Institute of Microbial Technology, Chandigarh, India, [2] Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Shi Ming, Sun Yat-sen University Cancer Center (SYSUCC), China
Yun Hak Kim, Pusan National University, South Korea

***Correspondence:**
Gajendra P. S. Raghava
raghava@iiitd.ac.in

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

Received: *04 September 2019*
Accepted: *26 November 2019*
Published: *10 January 2020*

Citation:
Kaur H, Dhall A, Kumar R and Raghava GPS (2020) Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data. Front. Genet. 10:1306. doi: 10.3389/fgene.2019.01306

The high mortality rate of hepatocellular carcinoma (HCC) is primarily due to its late diagnosis. In the past, numerous attempts have been made to design genetic biomarkers for the identification of HCC; unfortunately, most of the studies are based on small datasets obtained from a specific platform or lack reasonable validation performance on the external datasets. In order to identify a universal expression-based diagnostic biomarker panel for HCC that can be applicable across multiple platforms, we have employed large-scale transcriptomic profiling datasets containing a total of 2,316 HCC and 1,665 non-tumorous tissue samples. These samples were obtained from 30 studies generated by mainly four types of profiling techniques (Affymetrix, Illumina, Agilent, and High-throughput sequencing), which are implemented in a wide range of platforms. Firstly, we scrutinized overlapping 26 genes that are differentially expressed in numerous datasets. Subsequently, we identified a panel of three genes (*FCN3, CLEC1B*, and *PRC1)* as HCC biomarker using different feature selection techniques. Three-genes-based HCC biomarker identified HCC samples in training/validation datasets with an accuracy between 93 and 98%, Area Under Receiver Operating Characteristic curve (AUROC) in a range of 0.97 to 1.0. A reasonable performance, i.e., AUROC 0.91–0.96 achieved on validation dataset containing peripheral blood mononuclear cells, concurred their non-invasive utility. Furthermore, the prognostic potential of these genes was evaluated on TCGA-LIHC and GSE14520 cohorts using univariate survival analysis. This analysis revealed that these genes are prognostic indicators for various types of the survivals of HCC patients (e.g., Overall Survival, Progression-Free Survival, Disease-Free Survival). These genes significantly stratified high-risk and low-risk HCC patients (p-value <0.05). In conclusion, we identified a universal platform-independent three-genes-based biomarker that can predict HCC patients with high precision and also possess significant prognostic potential. Eventually, we developed a web server HCCpred based on the above study to facilitate scientific community (http://webs.iiitd.edu.in/raghava/hccpred/).

Keywords: liver cancer, hepatocellular carcinoma, biomarker, expression, diagnosis, survival, machine learning, classification

INTRODUCTION

Cancer is a heterogeneous disease driven by genomic and epigenomic changes within the cell (Sharma et al., 2010; Dawson and Kouzarides, 2012; Nagpal et al., 2015; Flavahan et al., 2017; Kamel and Al-Amodi, 2017; Chatterjee et al., 2018; Kagohara et al., 2018; Narrandes and Xu, 2018; Nebbioso et al., 2018; Kumar et al., 2019). Gene dysregulation is considered a hallmark of cancer. Among the 22 common cancer type, hepatocellular carcinoma (HCC) ranks at sixth in terms of frequency of occurrence and fourth at cancer-related mortality (Siegel et al., 2019). The etiology of HCC can be induced by multiple factors, especially hepatitis viral infection, alcoholic cirrhosis, and consumption of aflatoxin-contaminated foods (Ho et al., 2016). Although various traditional and locoregional treatment strategies such as hepatic resection (RES), percutaneous ethanol injection (PEI), radiofrequency ablation (RFA), microwave ablation (MWA), and trans-arterial chemotherapy infusion (TACI) have improved the survival rate, patients with HCC still have a late diagnosis and poor prognosis (Tian et al., 2018).

In the past, several studies focus on the identification of biomarkers by comparing the global gene expression changes between cancer tissue and non-tumorous tissues (Shirota et al., 2001; Jia et al., 2007; Marshall et al., 2013; Gao et al., 2015; Kang et al., 2015; Liu et al., 2015; Emma et al., 2016; Komatsu et al., 2016; Cai et al., 2017; Li et al., 2017; Zhang et al., 2017; Li et al., 2018b; Liao et al., 2018; Meng et al., 2018; Wang et al., 2018; Xu et al., 2018; Zheng et al., 2018; Cai et al., 2019; Jiao et al., 2019; Xia et al., 2019; Zhang et al., 2019). Such analyses yield hundreds or thousands of gene signature that are differentially expressed in cancer tissue compared to normal tissue, thus making it difficult to identify a universal subset of genes that play a crucial role in neoplastic transformation and progression (Rhodes et al., 2004). The lack of concordance of signature genes among different studies and extensive molecular variation between the patient's samples restrains the establishment of the robust biomarkers, promising targets and their experimental validation in clinical trials (Vasudevan et al., 2018). The transcriptome signatures have yet to be translated into a clinically useful biomarker, which may be due to a lack of their satisfactory validation performance on independent patient's cohort.

In this regard, treatment of HCC remains unsatisfying as only diagnostic and prognostic biomarkers alpha-fetoprotein (AFP) has been established so far. Several other biomarkers AFP-L3, osteopontin, and glypican-3 are currently being under investigation for the early diagnosis of HCC patients (Ocker, 2018). Advancement in the genomics has created rich public repositories of microarray and high throughput datasets from numerous studies such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013), Genomic Data Common (GDC), and Gene Expression Omnibus (Grossman et al., 2016), (Barrett et al., 2013), which provide the opportunity to study the various aspects of cancer. Thus, novel methods exploring the computational approach by merging multiple datasets from different platforms could provide a new way to establish a robust and universal biomarker for disease diagnosis and prognosis with increased precision and reproducibility. Recently, this approach has been used for biomarker identification of pancreatic adenocarcinoma (PDAC) (Bhasin et al., 2016; Klett et al., 2018). However, various studies employed large-scale data or meta-analysis approaches to identify protein and miRNA expression-based biomarker for HCC diagnosis (Ji et al., 2016; Ding et al., 2017; Chen et al., 2018b; Ji et al., 2018). But, to the best of our knowledge, RNA-expression data are not explored in this regard for identification of the robust biomarker for HCC diagnosis and prognosis.

In order to overcome the limitations of existing methods, we made a systematic attempt to identify genetic biomarkers for HCC diagnosis that apply to a wide range of platforms and profiling techniques. One of the objectives of this study is to identify robust gene expression signatures for discrimination of HCC samples by the integration of multiple transcriptomic datasets from various platforms. Here, we have collected and analyzed a total of 3,981 samples from published datasets, out of which 2,316 and 1,665 are of HCC and normal or non-tumorous tissue samples, respectively. From this, we identified 26 genes, which are commonly differentially expressed in uniform patterns among most of the datasets, which provides a universally activated transcriptional signatures of HCC cancer type. Further, we have established a robust "three-genes-based HCC biomarker" implementing different machine learning techniques to distinguish HCC and non-tumorous samples with high precision. Additionally, the survival analysis of HCC patient's cohorts using these genes revealed their significant prognostic potential in the stratification of high-risk and low-risk patient's groups. To the best of our knowledge, this is the first study regarding HCC cancer type for the identification of universal platform-independent diagnostic biomarkers by integrating data from multiple platforms implementing machine learning approaches.

MATERIALS AND METHODS

Dataset Collection

Collection of Gene Expression Datasets of HCC

In this study, we extract raw expression data of 30 datasets, where 29 transcriptome datasets were obtained from GEO and one is from TCGA; each dataset contains at least 10 samples. The following is the list of datasets obtained from GEO: GSE102079 (Chiyonobu et al., 2018), GSE22405, GSE98383 (Diaz et al., 2018), GSE84402 (Wang et al., 2017), GSE64041 (Makowska et al., 2016), GSE69715 (Sekhar et al., 2018), GSE51401, GSE62232 (Schulze et al., 2015), GSE45267 (Chen et al., 2018a), GSE32879 (Oishi et al., 2012), GSE19665 (Deng et al., 2010), GSE107170 (Diaz et al., 2018), GSE76427 (Grinchuk et al.,

Abbreviations: AUROC, Area under the Receiver Operating Characteristic curve; ETREES, Extra Trees Classifier; SVC-RBF, Support Vector Machine with RBF kernel; TCGA, The Cancer Genome Atlas; KNN, K Neighbors Classifier; HCC, Hepatocellular Carcinoma; MCC, Matthew's correlation coefficient; LR, Logistic Regression; NB, Naive Bayes; RF, Random Forest; PBMCs, Peripheral Blood Mononuclear Cells

2018), GSE39791 (Kim et al., 2014), GSE57957 (Mah et al., 2014), GSE87630 (Woo et al., 2017), GSE46408, GSE57555 (Murakami et al., 2015), GSE54236 (Villa et al., 2016; Zubiete-Franco et al., 2019), GSE65484 (Dong et al., 2015), GSE31370 (Seok et al., 2012), GSE84598, GSE89377, GSE29721 (Stefanska et al., 2011), GSE14323 (Mas et al., 2009), GSE25097 (Lamb et al., 2011; Tung et al., 2011; Wong et al., 2016), GSE14520 (Roessler et al., 2010; Zhao et al., 2015), GSE36376 (Lim et al., 2013), GSE36076). All GEO datasets were obtained using GEOquery package of Bioconductor in R-3.5.3. The TCGA RNA-seq dataset of TCGA-LIHC was downloaded using gdc-client from the GDC data portal. All datasets were curated manually to remove all non-human samples and ensured that only human tissue samples remain in the dataset. Besides, Probe ID mapped to gene symbols extracted from respective platform file and incorporated in the dataset matrix for each dataset. It has been observed that two datasets, i.e., GSE102079 and GSE64041, have three types of samples (HCC, adjacent non-tumor, and normal healthy). Thus, we derived two datasets from GSE102079, called GSE102079_D1 (contains HCC and adjacent non-tumor samples) and GSE102079_D2 (contains HCC and healthy normal samples). Similarly, we derived GSE64041_D1 and GSE64041_D2 datasets from GSE64041. Finally, we derived 32 datasets from original 30 datasets as we derived four datasets corresponding to GSE102079 and GSE64041. Notably, we used one non-invasive dataset (GSE36076), which contains 20 blood samples of peripheral blood mononuclear cells (PBMCs) to evaluate our models.

Pre-Processing of Datasets

Each retrieved raw dataset (**Supplementary Data**) was subjected to a detailed curation process. We have pre-processed dataset matrix individually from each profiling technique for different platforms in a standardized manner. In case of Affymetrix datasets, raw data files were pre-processed with background correction; RMA values were calculated using the Oligo package (Carvalho and Irizarry, 2010). In case of Illumina datasets, raw files were processed using Limma and Lumi packages (Du et al., 2008; Ritchie et al., 2015) and finally log2 values calculated using in-house R scripts. Similarly, raw Agilent-1-color and Agilent-2-color files were pre-processed using Limma package individually, then A-values were generated, which were further transformed to log2 values. Eventually, the average of multiple probes computed that correspond to a single gene for each dataset individually employing in-house R scripts. TCGA-LIHC dataset contains FPKM values, which were further converted to log2 values. Entrez transcript IDs were mapped to the gene symbols using GENCODE v22.

Datasets for the Identification of Differentially Expressed Genes

We divide our datasets into two parts: i) datasets for features extraction and ii) datasets for the development of the prediction models. Twenty-seven out of 32 datasets were selected for identification of differentially expressed genes (DEGs); each dataset contains more than 10 samples (**Figure 1A**). These 27 datasets were derived from 25 original GEO datasets. Out of them, 20 datasets contain HCC v/s adjacent non-tumor samples and 7 datasets contain HCC v/s healthy samples. These datasets encompass a total of 1,199 HCC and 949 normal or adjacent non-tumor samples.

Training and Validation Datasets

In this study, the GSE25097 dataset was used as a training dataset to develop prediction models; it contains 268 HCC and 243 non-tumor samples (**Figure 1B**). The performance of these models was evaluated on the following three datasets: GSE14520, GSE36376, and TCGA-LIHC, and called them as external validation datasets. As shown in **Figure 1B**, each dataset has a minimum of 400 samples. The distribution of all cohorts used in the current study based on sample size is shown in **Figure 1C**. To validate the performance of models on the non-invasive specimen, we also evaluated the performance on the GSE36076 dataset. This dataset contains 20 blood samples of PBMCs; it contains 10 HCC and 10 healthy individuals. In order to reduce the cross-platform artifacts, we performed quantile normalization using the PreprocessCore library of Bioconductor (Grossman, et al., 2016) package, for each dataset as well as for each profiling technique. This approach is well-adapted in the literature (Huang and Qin, 2018; Klett et al., 2018; Pedersen et al., 2018). These datasets contain a total of 1,117 HCC and 716 adjacent non-tumor samples.

Identification of Differentially Expressed Genes

Each gene in 32 datasets was analyzed for differential expression using Student's *t*-test (Welch *t*-test and Wilcoxon *t*-test). It is implemented using in-house R scripts after the assignment of samples to the respective class, i.e., cancer or normal. These tests have been applied previously in different studies for the identification of DEGs (WELCH, 1947; Akaiwa et al., 1999; Carvalho and Irizarry, 2010; Aino et al., 2014; Schulze et al., 2015; Best et al., 2016; Bhasin et al., 2016; Bhalla et al., 2017; Cai et al., 2017; Bhalla et al., 2019; Cai et al., 2019; Kaur et al., 2019). Wilcoxon T-test is used for paired samples and Welch T-test is used for unpaired samples. Only those sets of genes chosen to define DEGs that are statistically differentially expressed between two classes of samples with Bonferroni adjusted p-value less than 0.01. In order to identify a set of differential expression signatures or "core DEGs of hepatocellular carcinoma," DEGs in all 27 datasets were compared. Finally, only those overlapping genes were considered as "core DEGs of hepatocellular carcinoma," which have significant differential expression in at least 80% of cohorts. A similar type of approach was previously implemented in various studies (Bhasin et al., 2016; Klett et al., 2018; Li et al., 2018a).

Identification of Robust Biomarkers for HCC Diagnosis

Ranking and Selection of Features

To reduce the number of genes from the selected set of signature, i.e., "the core genes of hepatocellular carcinoma," genes were

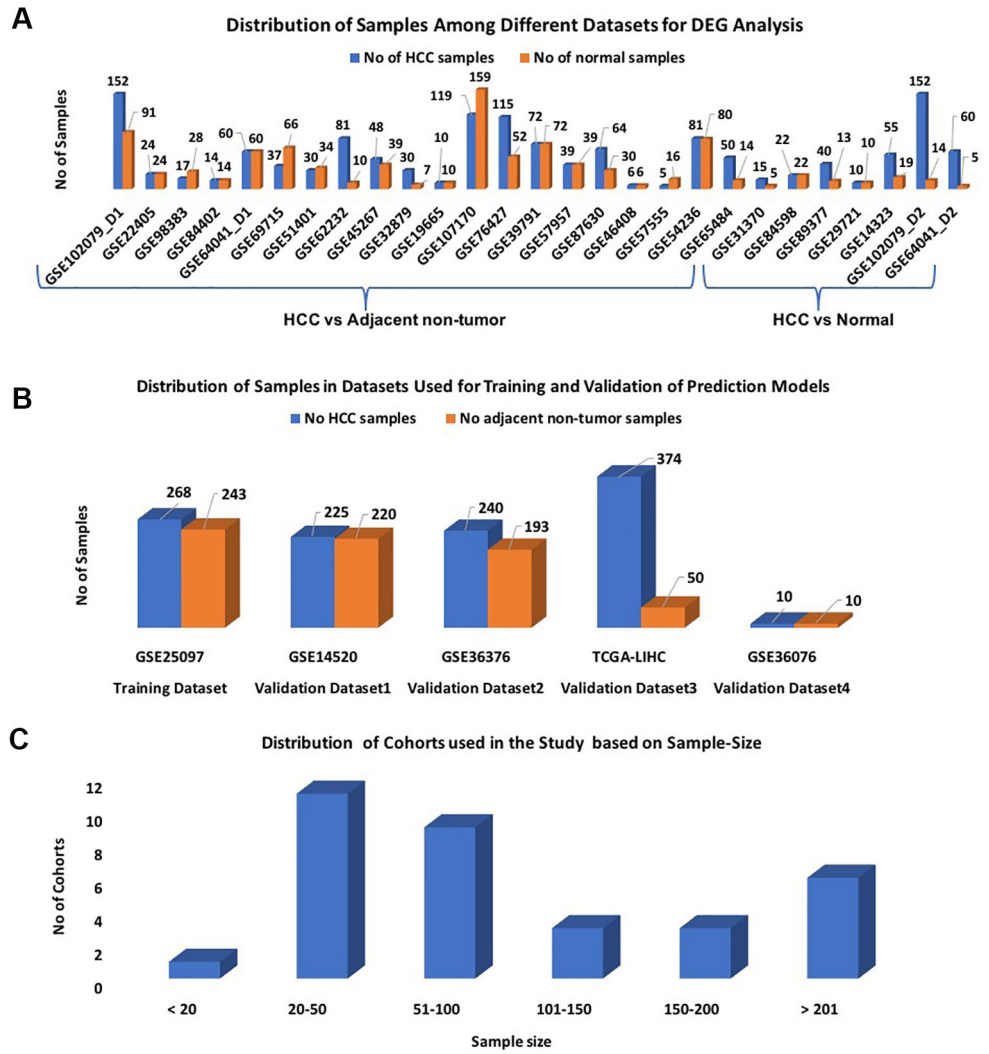

FIGURE 1 | Distribution of samples among datasets used in the study: **(A)** Datasets used for DEG analysis; **(B)** Datasets used for Development of Prediction models; **(C)** Sample-wise distribution of the datasets.

ranked on training dataset (GSE25097) using a simple threshold-based approach (Bhalla et al., 2017; Bhalla et al., 2019; Kaur et al., 2019). In the threshold-based approach, genes with a score above the threshold are assigned to cancer if it is found to be upregulated in cancer and otherwise normal; whereas sample is assigned to normal if the gene is downregulated in cancerous condition. We compute the performance of each gene based on a given threshold and identify the top 10 features having the highest performance. We further identified the top 5 features, which give the best performance when evaluated on the training dataset using a 10-fold cross-validation technique. Features were further reduced from five to four and then four to three using a wrapper-based approach. In this technique, one-by-one each feature is removed, and the prediction model is developed using the remaining features. Finally, a combination of features that give the best performance is selected. This technique is also known as the feature-reduction technique.

Development of Prediction Models

Here, we have developed the prediction models to distinguish HCC and non-tumorous samples using selected features. These models were implemented using Python package Scikit-learn (Pedregosa et al., 2011). A wide range of machine learning techniques have been used for developing these prediction models that include ExtraTrees (ETREES), Naive Bayes, K-nearest neighbor (KNN), Random Forest, Logistic Regression (LR), and SVC-RBF (radial basis function). The optimization of the parameters for the various classifiers was done by using a grid search with AUROC curve as scoring performance measure for selecting the best parameter.

Performance Evaluation of the Prediction Models

In the current study, both internal and external validation techniques were employed to evaluate the performance of models. First, the training dataset is used to develop prediction models and standard 10-fold cross-validation is used for performing internal validation, which is commonly employed in the literature (Burton et al., 2012; Bastani et al., 2013; Kourou et al., 2015; Bhalla et al., 2017; Jiang et al., 2018; Bhalla et al., 2019; Kaur et al., 2019). It is important to evaluate the realistic performance of the model on the external validation dataset, which should not be used for training and testing during model development. Therefore, we evaluated the performance of our models on four independent gene-expression cohorts that include GSE14520, GSE36376, GSE36076, and TCGA-LIHC obtained from GEO and The Cancer Genome Atlas (TCGA) (see **Figure 1B**), which were not used for training. In order to measure the performance of models, we used both threshold-dependent and threshold-independent parameters. In the case of threshold-dependent parameters, we measure sensitivity, specificity, accuracy, and Matthew's correlation coefficient (MCC) using the following equations.

$$Sensitivity\ (Sen) = \frac{TP}{TP + FN} \times 100 \quad (1)$$

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \times 100 \quad (2)$$

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (3)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

where FP, FN, TP, and TN are false positive, false negative, true positive, and true negative predictions, respectively.

In case of threshold-independent measures, we used a standard parameter Area under the Receiver Operating Characteristic (AUROC) curve. The AUROC curve is generated by plotting sensitivity or true positive rate against the false positive rate (1-specificity) at various thresholds. Finally, the area under the curve is calculated to compute a single parameter called AUROC.

Prognostic Potential of Identified HCC Diagnostic Biomarkers

The prognostic potential of the *"three-genes HCC biomarker"* was analyzed using gene-expression data of TCGA-LIHC and GSE14520 cohorts. The TCGA and GSE14520 datasets contain 374 and 219 tumor samples, respectively. Their clinical information was extracted from GEO, GDC, and the literature (Roessler et al., 2010; Liu et al., 2018a). The clinical characteristics of patients are given in **Table S1** (**Supplementary Information File 1**). Univariate survival analyses and risk assessments were performed by survival package in R (Therneau and Grambsch, 2000; Therneau, 2013). The distribution of the survival risk groups is done by using a log-rank test, eventually represented in the form of Kaplan-Meier plots. A p-value < 0.05 was considered the cut-off to describe the statistical significance in all survival analyses. Here, we analyzed four types of survivals, i.e., OS (Overall Survival), DSS (Disease-Specific Survival), DFS (Disease-Free Survival), and PFS (Progression-Free Survival) for TCGA-LIHC cohort, and two types of survivals, i.e., OS and RFS (Recurrence-Free Survival) (also called as DFS) for GSE14520 cohort. Besides, genes from the signature, univariate survival analysis is also performed on clinical characteristics of patients like age, gender, and tumor stage individually. Additionally, multivariate survival analysis was performed to assess the combined effect of clinical characteristics with the signature genes.

Functional Annotation of Signature Genomic Markers

In order to discern the biological relevance of the signature genes, enrichment analysis is performed using Enrichr (Kuleshov et al., 2016). Enrichr executes Fisher exact test to identify enrichment score. It provides Z-score and adjusted p-value, which is derived by applying correction on a Fisher exact test. We have considered only those Gene Ontology (GO) terms that are significantly enriched with adjusted p-value less than 0.05.

RESULTS

Overview

The pipeline of our analysis is illustrated in **Figure 2**. The detail of each step is described below.

Transcriptomic Cores for Hepatocellular Carcinoma

Identification of the Transcriptomic Cores

The individual statistical differential expression analyses of 27 gene-expression datasets resulted in the identification of hundreds of DEGs (**Supplementary Figure 1**). The 9,954 genes are present among each of the 27 datasets (**Supplementary Information File 1**, **Table S2**). Further, the comparative analysis among all 27 datasets scrutinized 26 overlapping genes that are differentially expressed in 80% or more datasets, i.e., 22 datasets. We called these genes as "core genes for hepatocellular carcinoma." Among these 26 genes, 12 are downregulated and 14 are upregulated in HCC in comparison to normal samples. The regulatory patterns of the core genes were consistent among most of the datasets (**Table 1**). Additionally, the expression pattern of these genes in training and three external validation datasets is shown in **Figure S2** (**Supplementary Information File 2**).

Gene Enrichment Analysis of the Transcriptomic Cores

Gene enrichment analysis of these "core genes of HCC" revealed their biological significance. The proteins encoded by the downregulated genes mainly enriched in complement activation and lectin pathways related processes. These genes negatively regulate cellular extravasation. They are also enriched in GO

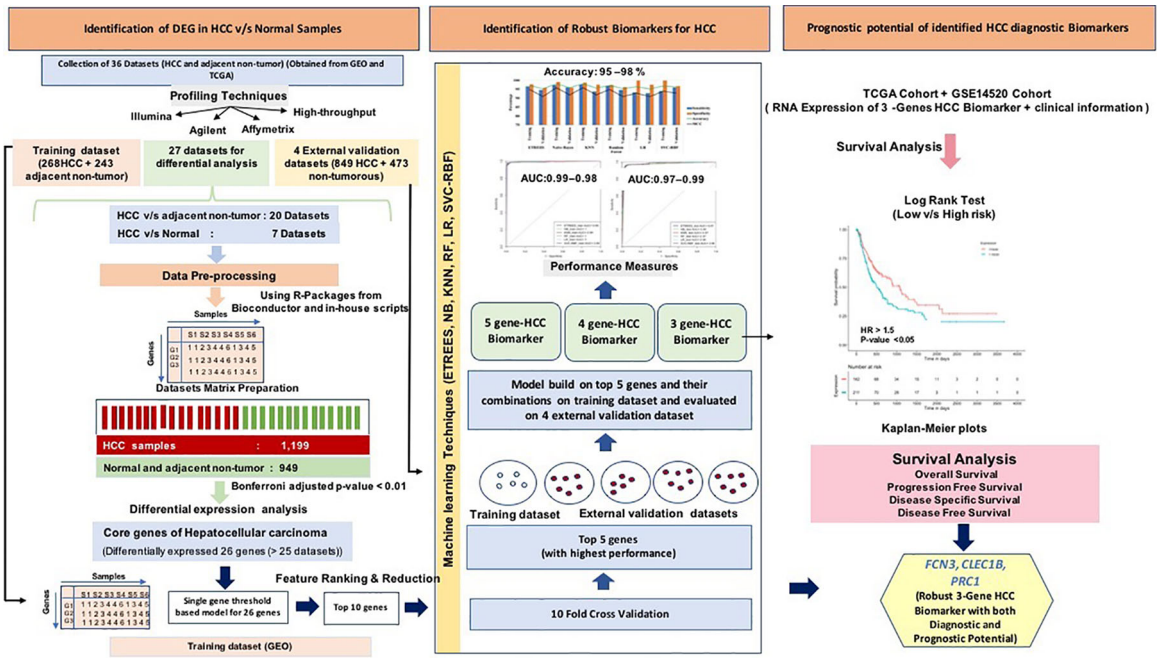

FIGURE 2 | Overview for the analysis implemented in the study.

TABLE1 | List of overlapping 26 genes that are differentially expressed (Core DEGs for HCC) between HCC and adjacent normal or adjacent non-tumor samples with Bonferroni p-values < 0.01.

Gene	#Up	#Down	#Sig	#Non-sig	Up (%)	Down (%)	Sig (%)	Regulation
FCN3	1	26	24	3	3.70	96.30	88.89	Down
CLEC4M	2	25	24	3	7.41	92.59	88.89	Down
FCN2	2	25	24	3	7.41	92.59	88.89	Down
MARCO	3	24	22	5	11.11	88.89	81.48	Down
CRHBP	2	25	22	5	7.41	92.59	81.48	Down
CFP	2	25	22	5	7.41	92.59	81.48	Down
STEAP3	2	25	25	2	7.41	92.59	92.59	Down
HGFAC	4	23	22	5	14.81	85.19	81.48	Down
CLEC1B	2	25	23	4	7.41	92.59	85.19	Down
CXCL12	3	24	24	3	11.11	88.89	88.89	Down
MT1E	3	24	24	3	11.11	88.89	88.89	Down
NSUN5	25	2	24	3	92.59	7.41	88.89	Down
MCM7	24	3	24	3	88.89	11.11	88.89	Up
MCM3	24	3	24	3	88.89	11.11	88.89	Up
ITGA6	24	3	24	3	88.89	11.11	88.89	Up
SSR2	24	3	23	4	88.89	11.11	85.19	Up
STMN1	23	4	24	3	85.19	14.81	88.89	Up
PRC1	24	3	23	4	88.89	11.11	85.19	Up
POLD1	24	3	23	4	88.89	11.11	85.19	Up
PBK	24	3	24	3	88.89	11.11	88.89	Up
IGSF3	22	5	23	4	81.48	18.52	85.19	Up
DTL	24	3	22	5	88.89	11.11	81.48	Up
ZWINT	24	3	22	5	88.89	11.11	81.48	Up
SPATS2	24	3	24	3	88.89	11.11	88.89	Up
GPSM2	23	4	23	4	85.19	14.81	85.19	Up
COL15A1	24	3	22	5	88.89	11.11	81.48	Up

Up, Upregulated in cancer or HCC; Down, Downregulated in cancer or HCC; #Up: No. of datasets in which gene is overexpressed; #Down: No. of datasets in which gene is underexpressed; #Sig: No. of datasets in which gene is significantly differentially expressed; #Non-Sig: No. of datasets in which gene is not significantly differentially expressed; Up (%): Percentage of datasets in which gene is overexpressed; Down (%): Percentage of datasets in which gene is underexpressed; Sig (%): Percentage of datasets in which gene is significantly differentially expressed.

molecular functions like serine-type endopeptidase, oxidoreductase, RNA methyltransferase activity, etc. (**Supplementary Information File 2**, **Figure S3**). Whereas, upregulated core genes are enriched in cell cycle GO biological processes like mitotic spindle organization and mitotic sister chromatid segregation, DNA synthesis and DNA replication, post-replication repair and cellular response to DNA damage stimulus, etc. They are also enriched in GO molecular functions such as exodeoxyribonuclease activity, GDP-dissociation inhibitor activity, DNA polymerase activity and insulin-like growth factor binding, etc. (**Supplementary Information File 2**, **Figure S3**).

Identification of HCC Biomarkers and Development of Prediction Models

Single-Gene Based Prediction Models

All 26 DEGs were ranked on the training dataset using threshold-based approach; ranking is based on their discriminatory power to distinguish HCC from non-tumorous samples (Bhalla et al., 2017; Kaur et al., 2019). The performance of the top 10 genes having maximum discriminatory power is shown in **Table 2**; see **Supplementary Information File 1**, **Table S3** for detail. These top 10 genes showed highest performance with an accuracy > 85%, MCC > 0.75, and AUROC > 0.85. We also evaluate the performance of these top 10 genes using 10-fold cross-validation to understand their robustness as shown in **Table S4** (**Supplementary Information File 1**). We further selected 5 genes out of 10 genes, which exhibit the maximum performance. These genes are *FCN3, CLEC1B, CLEC4M, PRC1*, and *PBK*; models based on these genes have accuracy more than 90% with AUROC > 0.95. In addition, the performance is also evaluated on the external validation datasets. The performance of the method was same on the training dataset but decreases on the external validation for few genes/features (see **Table S5**, **Supplementary Information File 1**).

Multiple-Genes Based Prediction Models

We identified the top five genes based on single gene-based prediction models, as described above. Further, we developed machine learning techniques-based classification models using these top five genes. We called these models as multiple-genes based prediction models as they take multiple genes as input. These models were evaluated on the training as well as validation datasets using internal and external cross-validation. The performance of these models on training as well as on three validation datasets is shown in **Table 3**. As shown in **Table 3**, we got AUROC approximately 0.98 on training as well as on the validation datasets. We further reduced one gene from selected set of five genes using feature reduction technique as described in *Materials and Methods* and obtained a set of four genes (*FCN3, CLEC1B, PRC1, PBK*). Subsequently, machine learning prediction models developed based on them classified HCC and non-tumor samples with accuracy more than 95% with AUROC in the range of 0.97–0.99 on both training and three independent validation datasets as shown in **Table S6** (**Supplementary Information File 1**). Results from this analysis show that we got nearly same performance using four genes-based biomarkers as we got in case of five genes-based biomarkers. Thus, reduction of one feature (five to four) does not affect the performance of our multiple-gene based prediction method. We further reduced features using feature reduction technique and got a set of three genes that contains *FCN3, CLEC1B,* and *PRC1*. Prediction models based on three genes-biomarker got accuracy 95–98% with AUROC in the range of 0.96–0.99 on training as well as independent validation datasets as shown in **Table 4**. The expression pattern of these three genes among samples of training dataset and three external validation datasets is depicted in **Figure 3**. We also tried two gene biomarkers, but there is substantial reduction in the performance on validation datasets. Thus, our final model is developed using a biomarker panel of three genes that include *FCN3, CLEC1B,* and *PRC1*. We considered three-genes based biomarker as the final model because the number of genes is limited. Hence, it is easy to implement in real life as well as economical.

Validation of Models on Blood Samples

In this study, models have been developed on tissue samples, which is complex and difficult to implement for routine testing. The question arises whether this model can also be used to discriminate the samples achieved from non-invasive techniques. Thus, we assessed the performance of our final model on PBMCs/blood samples of GSE36076. These signature genes

TABLE 2 | Top 10 genes based on the simple threshold-based approach.

Gene symbol	Thresh	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC	Mean in HCC	Mean in normal	Mean diff
FCN2	9.78	97.76	99.59	98.63	0.97	0.98	5.76	10.89	–5.13
CLEC4M	7.59	97.01	98.77	97.85	0.96	0.98	4.32	9.37	–5.06
FCN3	10.76	95.15	99.18	97.06	0.94	0.97	7.87	12.32	–4.45
CLEC1B	9.46	95.52	97.94	96.67	0.93	0.97	5.96	11.38	–5.42
CFP	8.14	96.64	94.24	95.50	0.91	0.96	6.15	8.63	–2.48
CRHBP	8.69	92.54	96.71	94.52	0.89	0.95	6.35	10.30	–3.95
PRC1	7.76	91.42	97.12	94.13	0.88	0.94	10.03	6.35	3.68
PBK	6.03	91.04	93.42	92.17	0.84	0.93	8.65	4.41	4.24
DTL	6.71	85.82	94.65	90.02	0.80	0.91	8.72	5.20	3.52
IGSF3	6.93	81.34	91.77	86.30	0.73	0.88	8.10	6.08	2.01

Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve; Thresh, Threshold; Mean diff, Mean in HCC–Mean in normal.

TABLE 3 | Performance of five genes (*FCN3*, *CLEC4M*, *CLEC1B*, *PRC1*, *PBK*) based models on training and validation datasets implementing various machine learning techniques.

Classifier	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI
	Training Dataset					Validation Dataset1				
ETREES	97.39	98.35	97.85	0.96	0.99 (0.99-1)	97.78	94.09	95.96	0.92	0.98 (0.97-0.99)
NB	97.76	99.18	98.43	0.97	0.99 (0.99-1)	97.33	95.45	96.40	0.93	0.98 (0.97-0.99)
KNN	97.39	98.77	98.04	0.96	0.99 (0.99-1)	96.89	96.82	96.85	0.94	0.98 (0.97-0.99)
RF	97.01	97.94	97.46	0.95	0.99 (0.99-1)	97.33	94.55	95.96	0.92	0.98 (0.97-0.99)
LR	97.76	99.59	98.63	0.97	0.99 (0.99-1)	95.56	97.27	96.40	0.93	0.99 (0.98-0.99)
SVC	97.01	100	98.43	0.97	0.99 (0.99-1)	96.89	95.00	95.96	0.92	0.99 (0.98-0.99)
	Validation Dataset2					Validation Dataset3				
ETREES	95	97.41	96.07	0.92	0.98 (0.97-0.99)	97.86	96	97.64	0.89	0.99 (0.98-0.99)
NB	94.58	98.45	96.3	0.93	0.98 (0.96-0.99)	98.13	92	97.41	0.88	0.98 (0.98-0.99)
KNN	92.92	98.45	95.38	0.91	0.97 (0.96-0.99)	97.86	94	97.41	0.88	0.99 (0.98-0.99)
RF	96.67	93.26	95.15	0.9	0.98 (0.97-0.99)	98.4	90	97.41	0.88	0.99 (0.98-0.99)
LR	93.75	98.45	95.84	0.92	0.98 (0.97-0.99)	97.59	98	97.64	0.90	0.99 (0.98-0.99)
SVC-RBF	93.33	98.45	95.61	0.91	0.98 (0.97-0.99)	97.33	98	97.41	0.89	0.99 (0.98-0.99)

ETREES, Extra Trees Classifier; NB, Naive Bayes; KNN, K Neighbors Classifier; RF, Random Forest; LR, Logistic Regression; SVC-RBF, Support Vector Machine with RBF-kernel; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve.

TABLE 4 | Performance of three-genes HCC biomarker-A (*FCN3*, *CLEC1B*, *PRC1*) based models on training and validation datasets implementing various machine learning techniques.

Classifier	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI
	Training Dataset					Validation Dataset1				
ETREES	96.64	97.94	97.26	0.95	0.99 (0.98-0.99)	94.67	95.91	95.28	0.91	0.97 (0.96-0.99)
NB	97.39	99.18	98.24	0.96	0.99 (0.99-1.0)	96.00	95.91	95.96	0.92	0.98 (0.97-0.99)
KNN	97.76	98.77	98.24	0.96	0.99 (0.99-1.0)	93.78	97.73	95.73	0.92	0.97 (0.96-0.99)
RF	97.01	97.53	97.26	0.95	0.99 (0.99-1.0)	94.67	96.36	95.51	0.91	0.97 (0.96-0.99)
LR	93.28	100	96.48	0.93	0.99 (0.99-1.0)	92.89	97.73	95.28	0.91	0.98 (0.97-0.99)
SVC-RBF	94.03	100	96.87	0.94	0.99 (0.98-0.99)	96.00	96.82	96.40	0.93	0.98 (0.97-0.99)
	Validation Dataset2					Validation Dataset3				
ETREES	93.75	96.37	94.92	0.90	0.98 (0.97-0.99)	95.72	98	95.99	0.84	0.99 (0.98-0.99)
NB	94.58	98.45	96.3	0.93	0.98 (0.97-0.99)	98.13	82	96.23	0.82	0.96 (0.95-0.98)
KNN	95.83	97.93	96.77	0.94	0.98 (0.97-0.99)	97.59	96	97.41	0.88	0.99 (0.98-0.99)
RF	95.42	94.3	94.92	0.90	0.98 (0.97-0.99)	95.45	96	95.52	0.82	0.98 (0.97-0.99)
LR	95.42	98.45	96.77	0.94	0.99 (0.98-0.99)	97.33	98	97.41	0.89	0.99 (0.98-0.99)
SVC-RBF	93.33	97.93	95.38	0.91	0.98 (0.97-0.99)	96.79	98	96.93	0.87	0.99 (0.98-0.99)

ETREES, Extra Trees Classifier; NB, Naive Bayes; KNN, K Neighbors Classifier; RF, Random Forest; LR, Logistic Regression; SVC-RBF, Support Vector Machine with RBF kernel; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve.

correctly predicted 90% of both HCC and healthy samples with ROC in the range of 0.91–0.96 and MCC 0.80–0.82. Complete results of prediction models are tabulated in **Table 5**. This demonstrates that our three genes-based models have the ability to discriminate HCC and healthy blood samples with reasonably high accuracy.

Protein-Based Biomarkers

In the past, proteins have been identified as diagnostic biomarkers for HCC. These protein biomarkers are AFP +GPC3 and AFP+GPC3+CK19 (*KRT19*) (Lou et al., 2017; Ocker, 2018). As we do not have their protein expression for these patients' samples, we employed only their gene expression values. Models based on the gene expression of *AFP+GPC3 +KRT19* classified HCC and normal samples of training dataset with an accuracy 67–75%. While this model attained accuracy of 69–77%, 51–87%, and 50–74% on external validation dataset1, dataset2 and dataset3, respectively, as shown in **Table S7** (**Supplementary Information File 1**). Further, the prediction models based on the gene expression of *AFP+GPC3* have improved performance on training dataset with an accuracy of 70–77%, but lower performance on all three validation datasets as given in **Table S8** (**Supplementary Information File 1**).

Survival Analysis to Determine the Prognostic Potential of "Three-Genes HCC Biomarker"

Univariate Survival Analysis for Three-Genes HCC Biomarker

To examine the prognostic potential of the "three-genes HCC biomarker," the univariate survival analysis was performed on TCGA-LIHC and GSE14520 cohorts. The samples were partitioned into low-risk and high-risk groups. Interestingly, all three genes of "three-genes HCC biomarker-A" are significantly

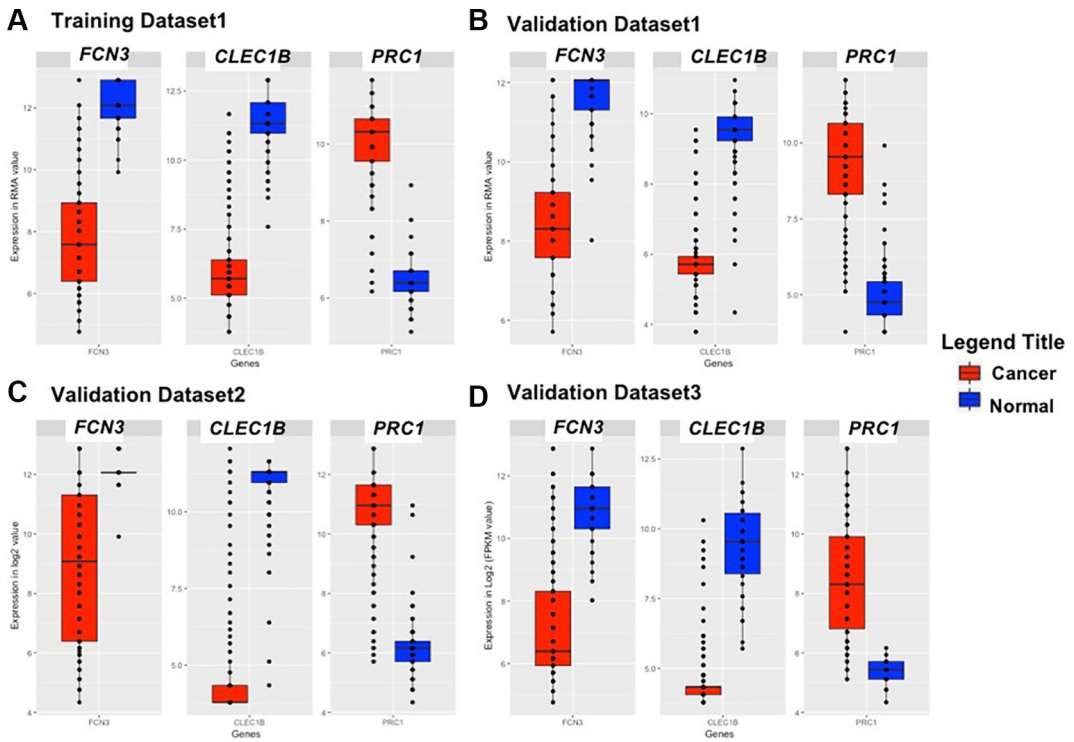

FIGURE 3 | Boxplot representing the expression pattern of three-genes panel-based HCC biomarker in the **(A)** Training Dataset, **(B)** Validation Dataset 1, **(C)** Validation Dataset 2, **(D)** Validation Dataset 3.

TABLE 5 | Performance of three-genes HCC biomarker-A (*FCN3*, *CLEC1B*, *PRC1*) based models on training and validation datasets 4 (containing blood samples, i.e., PBMCs) implementing various machine learning techniques.

Classifier	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI	Sens (%)	Spec (%)	Acc (%)	MCC	AUROC with 95% CI
	Training Dataset					Validation Dataset4				
ETREES	94.78	99.18	96.87	0.94	0.99 (0.979-0.998)	100	80	90	0.82	0.93 (0.854-1.0)
NB	97.39	99.18	98.24	0.96	0.99 (0.989-1.0)	90	90	90	0.80	0.95 (0.81-1.0)
KNN	97.01	99.59	98.24	0.97	0.99 (0.986-1.0)	90	90	90	0.80	0.96 (0.878-1.0)
RF	95.52	99.59	97.46	0.95	0.99 (0.991-1.0)	100	80	90	0.82	0.93 (0.81-1.0)
LR	96.64	100	98.24	0.97	0.99 (0.992-1.0)	90	90	90	0.80	0.96 (0.877-1.0)
SVC	95.15	99.18	97.06	0.94	0.99 (0.988-0.999)	90	90	90	0.80	0.91 (0.744-1.0)

ETREES, Extra Trees Classifier; NB, Naive Bayes; KNN, K Neighbors Classifier; RF, Random Forest; LR, Logistic Regression; SVC-RBF, Support Vector Machine with RBF kernel; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; MCC, Mathews Correlation Coefficient; AUROC, Area under Receiver operator curve.

associated with the survival of HCC patients. For instance, higher expression (greater than mean) of *CLEC1B* and *FCN3* is significantly associated with good outcome of the patients, i.e. OS, DSS, DFS, and PFS; while the overexpression of *PRC1* is significantly associated with poor survival including DSS, DFS, or RFS and PFS of HCC patients for TCGA-LIHC dataset as shown in **Figure 4**. In the GSE14520 dataset, higher expression of *PRC1* is significantly associated with the poor outcome of patients, i.e., OS and DFS or RFS, while the higher expression of *FCN3* is significantly associated with the better outcome of HCC patients as depicted in **Figure 5**. Complete results of survival analysis with HR (Hazard Ratio), with 95% CI and p-value, are presented in **Table S9** (**Supplementary Information File 1**).

Univariate Survival Analysis for Clinical Features

The clinical characteristics of the patients like age, gender, tumor size, and stage are considered as important prognostic indicators for the survival of the patients in different malignancies including HCC (Best et al., 2016; Liu et al., 2018a; Wu et al., 2018; Yang et al., 2019). As the tumor size information is not present in one of the cohorts, therefore, we performed univariate survival analysis using only age, gender, and tumor stage of

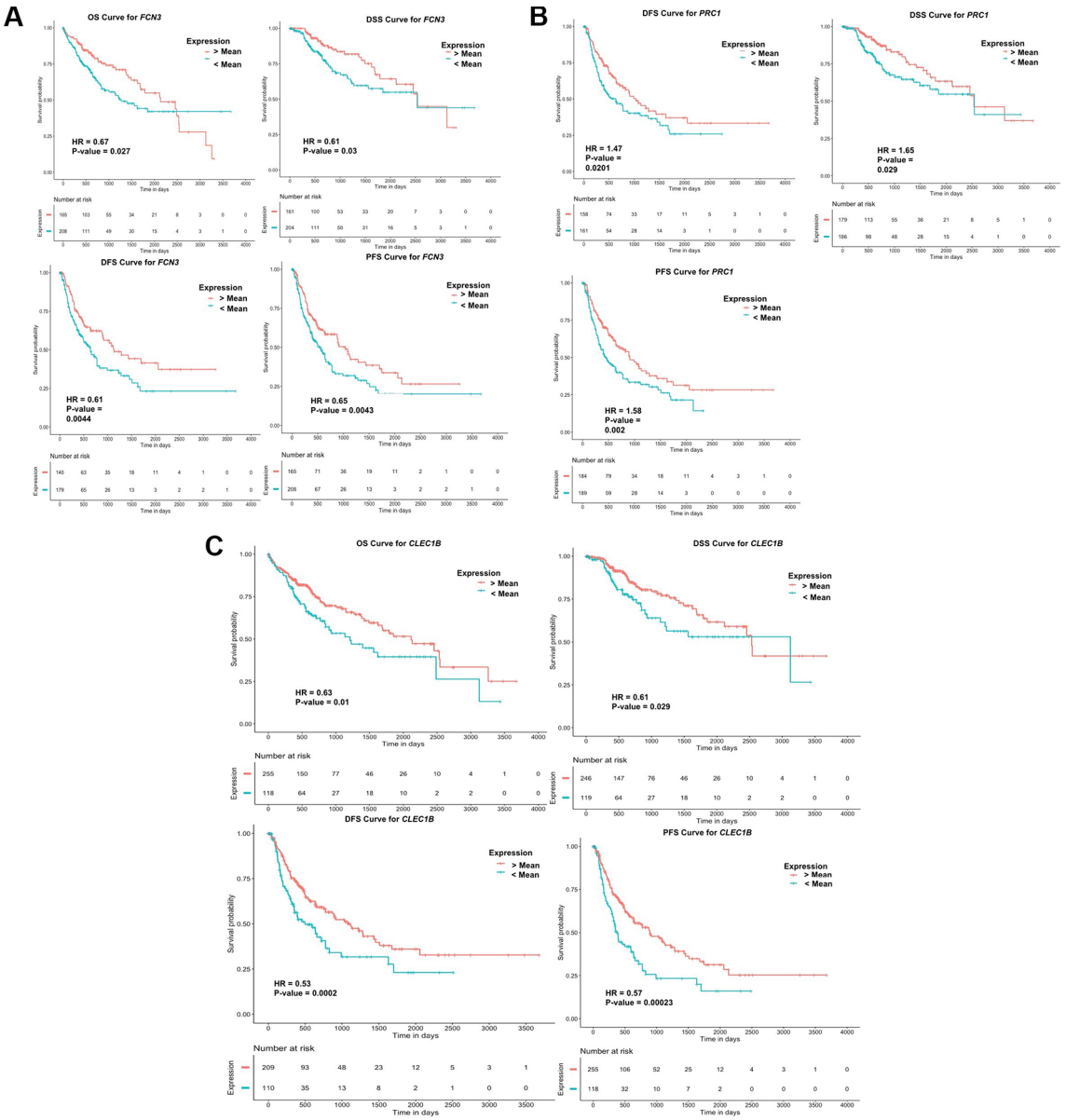

FIGURE 4 | Kaplan Meier survival curves for the risk estimation of HCC patient in TCGA cohort based on the RNA expression of **(A)** *FCN3*, **(B)** *PRC1*, and **(C)** *CLEC1B*.

the patients. This analysis shows that tumor stage is an important clinical factor with prognostic potential that significantly stratified high-risk and low-risk groups of patients in both cohorts, i.e., TCGA-LIHC and GSE14520. For instance, stage individually significantly (p-value <0.0001) stratified risk groups for OS, RFS with HR = 1.73 and HR = 1.65 of TCGA cohorts and with HR = 2.29 and HR = 1.79 of GSE14520 cohort, respectively (**Table S10**, **Supplementary Information File 1**). While the gender and age of patients do not possess high prognostic potential, as shown in **Table S10** (**Supplementary Information File 1**).

Multivariate Survival Analysis

Eventually, the multivariate analysis is performed to assess the independent impact of clinical characteristics and three genes of our signature biomarker that are determined as significant prognostic variables by univariate analysis. From this analysis, tumor stage is identified as the sole independent prognostic factor associated with the survival of HCC patients that significantly (with p-value <0.01) stratified high-risk and low-risk groups of both TCGA-LIHC and GSE1450 cohorts as presented in **Figures S4–S6** (**Supplementary Information File 2**).

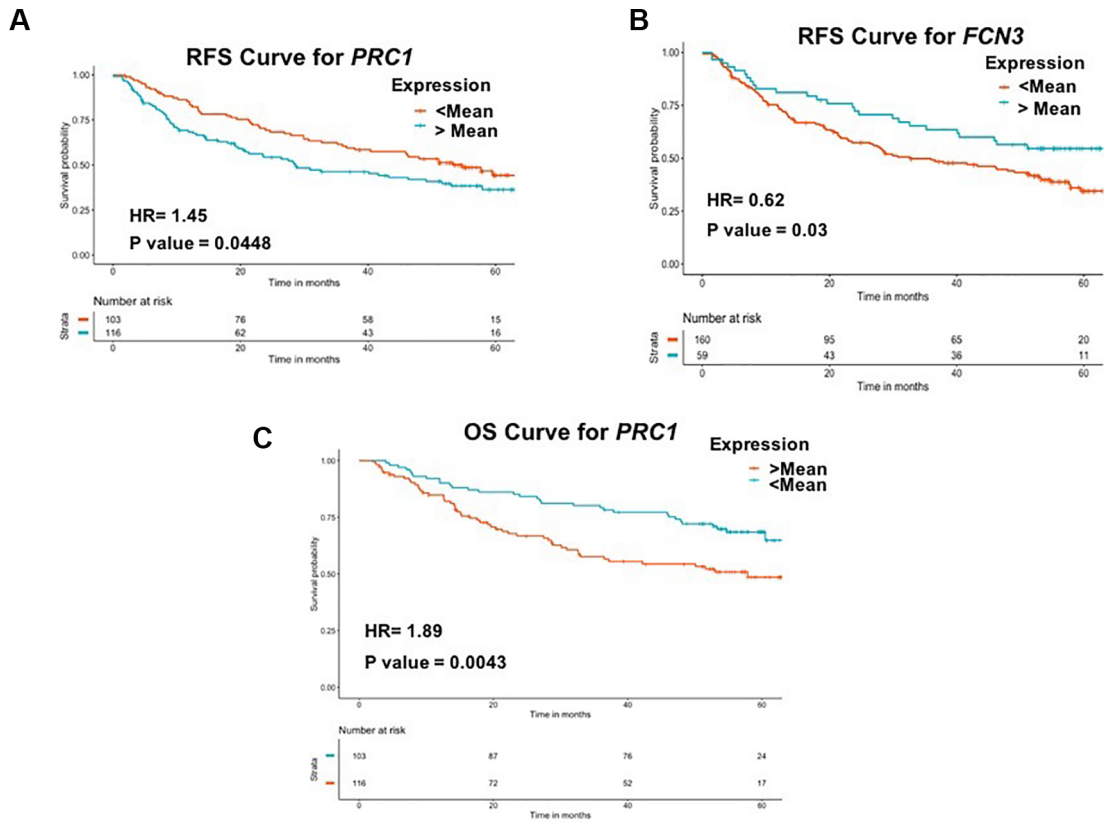

FIGURE 5 | Kaplan Meier survival curves for the risk estimation of HCC patient in GSE14520 cohort based on the RNA expression of **(A)** RFS for *PRC1*, **(B)** RFS for *FCN3*, and **(C)** OS for *PRC1*.

Web Server

To facilitates the scientific community working in the area of liver cancer research, we developed "HCCpred" (Prediction Server for Hepatocellular Carcinoma). In HCCpred, we execute mainly two modules: Prediction Module and Analysis Module based on robust five-genes, four-genes. and three-genes HCC biomarkers and 26 Core genes of HCC identified in the present study for the prediction and analysis of samples from the RNA-expression data. The prediction module permits the users to predict the disease status, i.e., cancerous or normal using RNA expression values of a subset of genes using *in silico* prediction models based on robust five-genes, four-genes, and three-genes HCC biomarkers identified in the present study. Here, the user is required to submit RMA (for Affymetrix), A-value (for Agilent), Log2 value (for Illumina), or FPKM (High throughput RNA-seq data) for a subset of genes or biomarkers. The output result displays a list for patient samples and corresponding predicted status of samples. Moreover, the user can select among the models, i.e., ETREES-based or SVC-RBF based model. Further, the Analysis Module permits the user to analyze the expression pattern of any of the top 10 ranked genes to check whether it is upregulated or downregulated in comparison to HCC samples based on the samples of the current study. This webserver is freely accessible at http://webs.iiitd.edu.in/raghava/hccpred/.

DISCUSSION

HCC is a type of tumor that is associated with the poor prognosis and a high mortality rate among the most common cancer types (Siegel et al., 2019). High recurrence rate and low rate of early detection results in poor prognosis. Accurate diagnosis of HCC may provide the opportunity for appropriate treatment, including traditional available treatment like liver transplantation resection, etc. Although the AFP and DCP proteins are well-established markers for the diagnosis of HCC, their sensitivity and specificity are not optimum (Sauzay et al., 2016). Therefore, the development of a novel robust diagnostic and prognostic biomarker for HCC is needed as it can assist in the existing clinical management of tumor. Towards this, our current report is an attempt to scrutinize a robust transcriptomic biomarker for HCC diagnosis. Briefly, in this study, we provide a novel large-scale analysis-based approach to identify a robust gene expression-based candidate diagnostic biomarker for HCC

derived from multiple transcriptomic profiles/datasets across a variety of platforms obtained from GEO and TCGA. This metadata integration approach employed to elucidate "core HCC DEGs" subset followed by a class prediction by implementing various machine learning algorithms. Eventually, validation on external independent datasets led us to the identification of multiple-genes based robust biomarkers for HCC.

Here, firstly, we have identified 26 genes named as "Core DEGs for HCC" that are uniformly differentially expressed among 80% of datasets. We have considered only these genes for downstream machine learning analysis. In an urge to identify a manageable subset with the minimum number of genes from this list that have a high discriminatory power, we further identified three genes signature-set containing *CLEC1B, FCN3*, and *PRC1*. This "three-genes based HCC biomarker" has predictive accuracy of 95–98% and AUROC 0.96–0.99 on the training and all three independent validation datasets. We further hypothesized that this biomarker gene set might be proved as quite an effective non-invasive diagnostic biomarker for HCC. Therefore, eventually, we validated their discriminatory performance on 20 PBMCs samples (GSE36076) extracted from 10 HCC and 10 healthy individuals. As anticipated, this biomarker set correctly classified 90% of the samples with AUROC in the range of 0.91–0.96. Besides, we also developed the prediction models based on the gene expression of already well-established protein biomarkers of HCC in the literature, i.e., *AFP+GPC3* and *AFP+GPC3+KRT19* (Lou et al., 2017). The prediction models based on *AFP+GPC3 +KRT19* discriminate samples of training dataset with an accuracy of 67–75% and 69–77% of validation dataset1, 55–87% of validation dataset2, and 50–74% of validation dataset3, while the models based on *AFP+GPC3* have quite lower performance on validation datasets. Further, we speculate that "three-genes HCC biomarker" can be explored as an effective novel protein based non-invasive biomarker as they have very good predictive power to distinguish HCC and non-tumor samples at gene expression level from the tissue and PBMC samples. Moreover, the product of *FCN3* gene is released in the serum and bile (Akaiwa et al., 1999; Brown et al., 2015; Pan et al., 2015; Tizzot et al., 2018); thus, this may serve as non-invasive biomarkers for diagnosis of HCC. Furthermore, recently, it has been reported that the protein product of two of the three genes from three-genes HCC biomarker, i.e., *PRC1* and *FCN3*, is also associated with HCC diagnosis and prognosis independently (Liu et al., 2018b; Shen et al., 2018). Hence, we anticipate that the three-genes signature might prove to be a good diagnostic and prognostic marker for HCC at the protein level as well. There is still a need for the validation of the protein product of these genes on a large scale of samples to confirm this hypothesis and their clinical utility.

Interestingly, the robust "three-genes HCC biomarker" contains *FCN3, PRC1,* and *CLEC1B,* has very high diagnostic ability, and also possesses prognostic potential, i.e., they are significantly associated with survival of HCC patients as determined by univariate analysis. For instance, higher expression of *CLEC1B* and *FCN3* significantly associated with the good outcome of HCC patients in TCGA-LIHC cohort; while higher expression of *PRC1* is significantly associated with the poor outcome of HCC patients in both TCGA-LIHC and GSE14520 cohorts. Besides, the role of *CLEC1B* and *PRC1* was previously also revealed in the diagnosis and prognosis of HCC (Chen et al., 2016; Chan et al., 2018; Hu et al., 2018; Kaur et al., 2019). Further, univariate analysis employing clinical factors of patients found that tumor stage of patients can act as a strong prognostic factor in the various types of survival, i.e., OS, RFS/DFS, PFS, and DSS of patients. Eventually, the multivariate survival analysis revealed the tumor stage as a sole independent prognostic factor, which was also corroborated with the previous literature (Aino et al., 2014; Wang and Li, 2019). The correct tumor stage identification is quite a tedious and challenging task in comparison to the quantification of the expression of genes.

In the past, a concern raised by Kaplan et al. is that despite the number of advantages of big studies, large sample size can also magnify the bias associated with an error resulting from sampling or study design (Kaplan et al., 2014). Thus, to reduce the overestimation of inferences from the results of large cohorts, we have included both types of cohorts, i.e., large cohort (sample size >50) and small cohort (sample size <50). We hypothesized that these results might be more reliable and applicable. Additionally, it might be practically more useful in real life, where, usually, small cohorts are available with maximum clinical parameters. Therefore, to ensure that cohort's size does not affect the results derived from the overall study, results should be validated on a small cohort as well. Towards this, we have also validated models built on the training dataset on three large cohorts of external validation dataset and one small cohort (contains 20 blood samples). Thus, these results indicate that there is no overestimation of inferences from the results of cohorts used in the study.

Taken together, we have established a robust three-gene HCC diagnostic biomarker with reasonable performance and possesses both diagnostic and prognostic potential. A meta-data integration pipeline is employed for the identification of a robust biomarker using machine learning techniques, which can work across different platforms. Further, this pipeline can also be used for the analysis of any other cancer type. Although more and more research is under the development of novel biomarkers, further work will be required to implement the clinical utilization of identified biomarker to meet real-world demand. We are anticipating that identifying novel cost-efficient biomarker using predictive technology for the detection of HCC will be promising.

CONCLUSIONS

This study identified and validated a highly accurate three-genes HCC biomarker for discriminating HCC and non-tumorous samples; it also possesses a significant prognostic potential that may facilitate more accurate early diagnosis and risk stratification upon validation in prospective clinical trials. Reasonable performance on the validation dataset of PBMCs samples indicates their non-invasive utility. Moreover, the protein product of *FCN3* is released in the serum and bile. Thus, this may serve as non-invasive protein diagnostic biomarkers. Large-scale non-invasive cohorts are required to confirm their non-invasive

clinical utility. Additionally, the uniform overexpression pattern of *PRC1* among numerous HCC samples suggests it as a novel potential therapeutic target for HCC.

DATA AVAILABILITY STATEMENT

We have taken the Gene-expression data from the public repositories, i.e., GEO (https://www.ncbi.nlm.nih.gov/geo/) and GDC data portal (https://portal.gdc.cancer.gov/).

AUTHOR CONTRIBUTIONS

HK collected the data and created the datasets. HK developed classification algorithms. HK and AD implemented algorithms. HK and AD performed the survival analysis. HK and AD created the back-end server and front-end user interface. HK and GR analyzed the results. HK, RK, and AD wrote the manuscript. GR conceived and coordinated the project, helped in the interpretation and analysis of data, refined the drafted manuscript, and gave complete supervision to the project. All of the authors have read and approved the final manuscript.

FUNDING

This research was funded by J. C. Bose National Fellowship (with Grant No. SRP076), Department of Science and Technology (DST), India.

ACKNOWLEDGMENTS

All the authors acknowledge funding agencies J. C. Bose National Fellowship DST. HK and RK are thankful to Council of Scientific and Industrial Research (CSIR) and AD is thankful to DST INSPIRE for providing fellowships, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01306/full#supplementary-material

SUPPLEMENTARY FIGURE S1 | Distribution of Significantly DEG (Differentially Expressed Genes) among various datasets with Bonferroni adjusted p-value < 0.01.

SUPPLEMENTARY FIGURE S2 | Heatmap representing the expression pattern of "Core genes of HCC" in different datasets.

SUPPLEMENTARY FIGURE S3 | Gene Enrichment analysis of 26 genes or "Core genes of HCC".

SUPPLEMENTARY FIGURE S4 | Multivariate analysis of clinical characteristics and three genes of HCC Biomarker on TCGA cohort for **(A)** OS, **(B)** RFS/DFS.

SUPPLEMENTARY FIGURE S5 | Multivariate analysis of clinical characteristics and three genes of HCC Biomarker on TCGA cohort for **(A)** DSS, **(B)** PFS.

SUPPLEMENTARY FIGURE S6 | Multivariate analysis of clinical characteristics and three genes of HCC Biomarker on GSE14520 cohort for **(A)** OS, **(B)** RFS/DFS.

REFERENCES

Aino, H., Sumie, S., Niizeki, T., Kuromatsu, R., Tajiri, N., Nakano, M., et al. (2014). Clinical characteristics and prognostic factors for advanced hepatocellular carcinoma with extrahepatic metastasis. *Mol. Clin. Oncol.* 2, 393–398. doi: 10.3892/mco.2014.259

Akaiwa, M., Yae, Y., Sugimoto, R., Suzuki, S. O., Iwaki, T., Izuhara, K., et al. (1999). Hakata Antigen, a New Member of the Ficolin/Opsonin p35 Family, Is a Novel Human Lectin Secreted into Bronchus/Alveolus and Bile. *J. Histochem. Cytochem.* 47, 777–785. doi: 10.1177/002215549904700607

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bastani, M., Vos, L., Asgarian, N., Deschenes, J., Graham, K., Mackey, J., et al. (2013). A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. *PloS One* 8, e82144. doi: 10.1371/journal.pone.0082144

Best, J., Bilgi, H., Heider, D., Schotten, C., Manka, P., Bedreli, S., et al. (2016). The GALAD scoring algorithm based on AFP, AFP-L3, and DCP significantly improves detection of BCLC early stage hepatocellular carcinoma. *Z. Gastroenterol.* 54, 1296–1305. doi: 10.1055/s-0042-119529

Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., et al. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci. Rep.* 7, 44997. doi: 10.1038/srep44997

Bhalla, S., Kaur, H., Dhall, A., and Raghava, G. P. S. (2019). Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Sci. Rep.* 9, 15790. doi: 10.1038/s41598-019-52134-4

Bhasin, M. K., Ndebele, K., Bucur, O., Yee, E. U., Otu, H. H., Plati, J., et al. (2016). Meta-analysis of transcriptome data identifies a novel 5-gene pancreatic adenocarcinoma classifier. *Oncotarget* 7, 23263–23281. doi: 10.18632/oncotarget8139

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43, D36–D42. doi: 10.1093/nar/gku1055

Burton, M., Thomassen, M., Tan, Q., and Kruse, T. A. (2012). Gene expression profiles for predicting metastasis in breast cancer: a cross-study comparison of classification methods. *Sci. World J.* 2012, 380495. doi: 10.1100/2012/380495

Cai, J., Li, B., Zhu, Y., Fang, X., Zhu, M., Wang, M., et al. (2017). Prognostic Biomarker Identification Through Integrating the Gene Signatures of Hepatocellular Carcinoma Properties. *EbioMed.* 19, 18–30. doi: 10.1016/j.ebiom.2017.04.014

Cai, C., Wang, W., and Tu, Z. (2019). Aberrantly DNA Methylated-Differentially Expressed Genes and Pathways in Hepatocellular Carcinoma. *J. Cancer* 10, 355–366. doi: 10.7150/jca.27832

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Mills Shaw, K. R., Ozenberger, B. A., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 45 (10), 1113–1120. doi: 10.1038/ng.2764

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26, 2363–2367. doi: 10.1093/bioinformatics/btq431

Chan, H. L., Beckedorff, F., Zhang, Y., Garcia-Huidobro, J., Jiang, H., Colaprico, A., et al. (2018). Polycomb complexes associate with enhancers and promote oncogenic transcriptional programs in cancer through multiple mechanisms. *Nat. Commun.* 9, 3377. doi: 10.1038/s41467-018-05728-x

Chatterjee, A., Rodger, E. J., and Eccles, M. R. (2018). Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin. Cancer Biol.* 51, 149–159. doi: 10.1016/J.SEMCANCER.2017.08.004

Chen, J., Rajasekaran, M., Xia, H., Zhang, X., Kong, S. N., Sekar, K., et al. (2016). The microtubule-associated protein PRC1 promotes early recurrence of hepatocellular carcinoma in association with the Wnt/β-catenin signalling pathway. *Gut* 65, 1522–1534. doi: 10.1136/gutjnl-2015-310625

Chen, C.-L., Tsai, Y.-S., Huang, Y.-H., Liang, Y.-J., Sun, Y.-Y., Su, C.-W., et al. (2018a). Lymphoid Enhancer Factor 1 Contributes to Hepatocellular Carcinoma Progression Through Transcriptional Regulation of Epithelial-Mesenchymal Transition Regulators and Stemness Genes. *Hepatol. Commun.* 2, 1392–1407. doi: 10.1002/hep41229

Chen, H., Zhang, Y., Li, S., Li, N., Chen, Y., Zhang, B., et al. (2018b). Direct comparison of five serum biomarkers in early diagnosis of hepatocellular carcinoma. *Cancer Manage. Res.* 10, 1947–1958. doi: 10.2147/CMAR.S167036

Chiyonobu, N., Shimada, S., Akiyama, Y., Mogushi, K., Itoh, M., Akahoshi, K., et al. (2018). Fatty Acid Binding Protein 4 (FABP4) Overexpression in Intratumoral Hepatic Stellate Cells within Hepatocellular Carcinoma with Metabolic Risk Factors. *Am. J. Pathol.* 188, 1213–1224. doi: 10.1016/j.ajpath.2018.01.012

Dawson, M. A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. *Cell* 150, 12–27. doi: 10.1016/j.cell.2012.06.013

Deng, Y.-B., Nagae, G., Midorikawa, Y., Yagi, K., Tsutsumi, S., Yamamoto, S., et al. (2010). Identification of genes preferentially methylated in hepatitis C virus-related hepatocellular carcinoma. *Cancer Sci.* 101, 1501–1510. doi: 10.1111/j.1349-7006.2010.01549.x

Diaz, G., Engle, R. E., Tice, A., Melis, M., Montenegro, S., Rodriguez-Canales, J., et al. (2018). Molecular signature and mechanisms of hepatitis D virus–associated hepatocellular carcinoma. *Mol. Cancer Res.* 16, 1406–1419. doi: 10.1158/1541-7786.MCR-18-0012

Ding, Y., Yan, J.-L., Fang, A.-N., Zhou, W.-F., Huang, L., Ding, Y., et al. (2017). Circulating miRNAs as novel diagnostic biomarkers in hepatocellular carcinoma detection: a meta-analysis based on 24 articles. *Oncotarget* 8, 66402–66413. doi: 10.18632/oncotarget.18949

Dong, H., Zhang, L., Qian, Z., Zhu, X., Zhu, G., Chen, Y., et al. (2015). Identification of HBV-MLL4 integration and its molecular basis in chinese hepatocellular carcinoma. *PloS One* 10, e0123175. doi: 10.1371/journal.pone.0123175

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547–1548. doi: 10.1093/bioinformatics/btn224

Emma, M. R., Iovanna, J. L., Bachvarov, D., Puleio, R., Loria, G. R., Augello, G., et al. (2016). NUPR1, a new target in liver cancer: implication in controlling cell growth, migration, invasion and sorafenib resistance. *Cell Death Dis.* 7, e2269. doi: 10.1038/cddis.2016.175

Flavahan, W. A., Gaskell, E., and Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science* 357, eaal2380. doi: 10.1126/science.aal2380

Gao, B., Ning, S., Li, J., Liu, H., Wei, W., Wu, F., et al. (2015). Integrated analysis of differentially expressed mRNAs and miRNAs between hepatocellular carcinoma and their matched adjacent normal liver tissues. *Oncol. Rep.* 34, 325–333. doi: 10.3892/or.20153968

Grinchuk, O. V., Yenamandra, S. P., Iyer, R., Singh, M., Lee, H. K., Lim, K. H., et al. (2018). Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. *Mol. Oncol.* 12, 89–113. doi: 10.1002/1878-0261.12153

Grossman, R. L., Heath, A. P., Ferretti, V. V. H. E., Lowy, D. R., Kibbe, W. A., and Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375 (12), 1109–1112. doi: 10.1056/NEJMp1607591

Ho, D. W.-H., Lo, R. C.-L., Chan, L.-K., and Ng, I. O.-L. (2016). Molecular Pathogenesis of Hepatocellular Carcinoma. *Liver Cancer* 5, 290–302. doi: 10.1159/000449340

Hu, K., Wang, Z.-M., Li, J.-N., Zhang, S., Xiao, Z.-F., and Tao, Y.-M. (2018). CLEC1B Expression and PD-L1 expression predict clinical outcome in hepatocellular carcinoma with tumor hemorrhage. *Transl. Oncol.* 11, 552–558. doi: 10.1016/j.tranon.2018.02.010

Huang, H.-C., and Qin, L.-X. (2018). Empirical evaluation of data normalization methods for molecular classification. *PeerJ* 6, e4584. doi: 10.7717/peerj4584

Ji, J., Wang, H., Li, Y., Zheng, L., Yin, Y., Zou, Z., et al. (2016). Diagnostic evaluation of des-gamma-carboxy prothrombin versus α-Fetoprotein for hepatitis B virus-related hepatocellular carcinoma in China: a large-scale, multicentre study. *PloS One* 11, e0153227. doi: 10.1371/journal.pone.0153227

Ji, J., Chen, H., Liu, X.-P., Wang, Y.-H., Luo, C.-L., Zhang, W.-W., et al. (2018). A miRNA combination as promising biomarker for hepatocellular carcinoma diagnosis: a study based on bioinformatics analysis. *J. Cancer* 9, 3435–3446. doi: 10.7150/jca.26101

Jia, H.-L., Ye, Q.-H., Qin, L.-X., Budhu, A., Forgues, M., Chen, Y., et al. (2007). Gene expression profiling reveals potential biomarkers of human hepatocellular carcinoma. *Clin. Cancer Res.* 13, 1133–1139. doi: 10.1158/1078-0432.CCR-06-1025

Jiang, Y., Mei, W., Gu, Y., Lin, X., He, L., Zeng, H., et al. (2018). Construction of a set of novel and robust gene expression signatures predicting prostate cancer recurrence. *Mol. Oncol.* 12, 1559–1578. doi: 10.1002/1878-0261.12359

Jiao, Y., Li, Y., Jiang, P., Han, W., and Liu, Y. (2019). PGM5: a novel diagnostic and prognostic biomarker for liver cancer. *PeerJ* 7, e7070. doi: 10.7717/peerj7070

Kagohara, L. T., Stein-O'Brien, G. L., Kelley, D., Flam, E., Wick, H. C., Danilova, L. V., et al. (2018). Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. *Brief. Funct. Genomics* 17, 49–63. doi: 10.1093/bfgp/elx018

Kamel, H. F. M., and Al-Amodi, H. S. A. B. (2017). Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. *Genomics Proteomics Bioinf.* 15, 220–235. doi: 10.1016/j.gpb.2016.11.005

Kang, L., Liu, X., Gong, Z., Zheng, H., Wang, J., Li, Y., et al. (2015). Genome-wide identification of RNA editing in hepatocellular carcinoma. *Genomics* 105, 76–82. doi: 10.1016/j.ygeno.2014.11.005

Kaplan, R. M., Chambers, D. A., and Glasgow, R. E. (2014). Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *Clin. Transl. Sci.* 7, 342–346. doi: 10.1111/cts.12178

Kaur, H., Bhalla, S., and Raghava, G. P. S. (2019). Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PloS One* 14, e0221476. doi: 10.1371/journal.pone.0221476

Kim, J. H., Sohn, B. H., Lee, H.-S., Kim, S.-B., Yoo, J. E., Park, Y.-Y., et al. (2014). Genomic predictors for recurrence patterns of hepatocellular carcinoma: model derivation and validation. *PloS Med.* 11, e1001770. doi: 10.1371/journal.pmed.1001770

Klett, H., Fuellgraf, H., Levit-Zerdoun, E., Hussung, S., Kowar, S., Küsters, S., et al. (2018). Identification and validation of a diagnostic and prognostic multi-gene biomarker panel for pancreatic ductal adenocarcinoma. *Front. Genet.* 9, 108. doi: 10.3389/fgene.2018.00108

Komatsu, H., Iguchi, T., Masuda, T., Ueda, M., Kidogami, S., Ogawa, Y., et al. (2016). HOXB7 expression is a novel biomarker for long-term prognosis after resection of hepatocellular carcinoma. *Anticancer Res.* 36, 2767–2773.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Kumar, R., Patiyal, S., Kumar, V., Nagpal, G., and Raghava, G. P. S. (2019). In Silico Analysis of Gene Expression Change Associated with Copy Number of Enhancers in Pancreatic Adenocarcinoma. *Int. J. Mol. Sci.* 20, 3582. doi: 10.3390/ijms20143582

Lamb, J. R., Zhang, C., Xie, T., Wang, K., Zhang, B., Hao, K., et al. (2011). Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PloS One* 6, e20090. doi: 10.1371/journal.pone.0020090

Li, L., Lei, Q., Zhang, S., Kong, L., and Qin, B. (2017). Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis. *Oncol. Rep.* 38, 2607–2618. doi: 10.3892/or.20175946

Li, J., Tan, W., Peng, L., Zhang, J., Huang, X., Cui, Q., et al. (2018a). Integrative analysis of gene expression profiles reveals specific signaling pathways associated with pancreatic duct adenocarcinoma. *Cancer Commun. (London England)* 38, 13. doi: 10.1186/s40880-018-0289-9

Li, N., Li, L., and Chen, Y. (2018b). The identification of core gene expression signature in hepatocellular carcinoma. *Oxid. Med. Cell. Longev.* 2018, 3478305. doi: 10.1155/2018/3478305

Liao, X., Liu, X., Yang, C., Wang, X., Yu, T., Han, C., et al. (2018). Distinct diagnostic and prognostic values of minichromosome maintenance gene expression in patients with hepatocellular carcinoma. *J. Cancer* 9, 2357–2373. doi: 10.7150/jca.25221

Lim, H.-Y., Sohn, I., Deng, S., Lee, J., Jung, S. H., Mao, M., et al. (2013). Prediction of disease-free survival in hepatocellular carcinoma by gene expression profiling. *Ann. Surg. Oncol.* 20, 3747–3753. doi: 10.1245/s10434-013-3070-y

Liu, F., Li, H., Chang, H., Wang, J., and Lu, J. (2015). Identification of hepatocellular carcinoma-associated hub genes and pathways by integrated microarray analysis. *Tumori* 101, 206–214. doi: 10.5301/tj.5000241

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018a). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.e11. doi: 10.1016/j.cell.2018.02.052

Liu, X., Li, Y., Meng, L., Liu, X.-Y., Peng, A., Chen, Y., et al. (2018c). Reducing protein regulator of cytokinesis 1 as a prospective therapy for hepatocellular carcinoma. *Cell Death Dis.* 9, 534. doi: 10.1038/s41419-018-0555-4

Lou, J., Zhang, L., Lv, S., Zhang, C., and Jiang, S. (2017). Biomarkers for Hepatocellular Carcinoma. *Biomark. Cancer* 9, 1–9. doi: 10.1177/1179299X16684640

Mah, W.-C., Thurnherr, T., Chow, P. K. H., Chung, A. Y. F., Ooi, L. L. P. J., Toh, H. C., et al. (2014). Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. *PloS One* 9, e104158. doi: 10.1371/journal.pone.0104158

Makowska, Z., Boldanova, T., Adametz, D., Quagliata, L., Vogt, J. E., Dill, M. T., et al. (2016). Gene expression analysis of biopsy samples reveals critical limitations of transcriptome-based molecular classifications of hepatocellular carcinoma. *J. Pathol. Clin. Res.* 2, 80–92. doi: 10.1002/cjp2.37

Marshall, A., Lukk, M., Kutter, C., Davies, S., Alexander, G., and Odom, D. T. (2013). Global gene expression profiling reveals SPINK1 as a potential hepatocellular carcinoma marker. *PloS One* 8, e59459. doi: 10.1371/journal.pone.0059459

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Kong, X., Kulik, L., et al. (2009). Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol. Med.* 15, 85–94. doi: 10.2119/molmed.2008.00110

Meng, C., Shen, X., and Jiang, W. (2018). Potential biomarkers of HCC based on gene expression and DNA methylation profiles. *Oncol. Lett.* 16 (3), 3183–3192. doi: 10.3892/ol.20189020

Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., et al. (2015). Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. *Sci. Rep.* 5, 16294. doi: 10.1038/srep16294

Nagpal, G., Sharma, M., Kumar, S., Chaudhary, K., Gupta, S., Gautam, A., et al. (2015). PCMdb: Pancreatic Cancer Methylation Database. *Sci. Rep.* 4, 4197. doi: 10.1038/srep04197

Narrandes, S., and Xu, W. (2018). Gene Expression Detection Assay for Cancer Clinical Use. *J. Cancer* 9, 2249. doi: 10.7150/JCA.24744

Nebbioso, A., Tambaro, F. P., Dell'Aversana, C., and Altucci, L. (2018). Cancer epigenetics: Moving forward. *PloS Genet.* 14, e1007362. doi: 10.1371/journal.pgen.1007362

Ocker, M. (2018). Biomarkers for hepatocellular carcinoma: What's new on the horizon? *World J. Gastroenterol.* 24, 3974–3979. doi: 10.3748/wjg.v24.i353974

Oishi, N., Kumar, M. R., Roessler, S., Ji, J., Forgues, M., Budhu, A., et al. (2012). Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. *Hepatology* 56, 1792–1803. doi: 10.1002/hep.25890

Pan, J.-W., Gao, X.-W., Jiang, H., Li, Y.-F., Xiao, F., and Zhan, R.-Y. (2015). Low serum ficolin-3 levels are associated with severity and poor outcome in traumatic brain injury. *J. Neuroinflammation* 12, 226. doi: 10.1186/s12974-015-0444-z

Pedersen, C. B., Nielsen, F. C., Rossing, M., and Olsen, L. R. (2018). Using microarray-based subtyping methods for breast cancer in the era of high-throughput RNA sequencing. *Mol. Oncol.* 12, 2136–2146. doi: 10.1002/1878-0261.12389

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *JMLR* 12, 2825–2830.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9309–9314. doi: 10.1073/pnas.0401994101

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007

Roessler, S., Jia, H.-L., Budhu, A., Forgues, M., Ye, Q.-H., Lee, J.-S., et al. (2010). A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 70, 10202–10212. doi: 10.1158/0008-5472.CAN-10-2607

Sauzay, C., Petit, A., Bourgeois, A.-M., Barbare, J.-C., Chauffert, B., Galmiche, A., et al. (2016). Alpha-foetoprotein (AFP): A multi-purpose marker in hepatocellular carcinoma. *Clin. Chim. Acta* 463, 39–44. doi: 10.1016/j.cca.2016.10.006

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* 47, 505–511. doi: 10.1038/ng3252

Sekhar, V., Pollicino, T., Diaz, G., Engle, R. E., Alayli, F., Melis, M., et al. (2018). Infection with hepatitis C virus depends on TACSTD2, a regulator of claudin-1 and occludin highly downregulated in hepatocellular carcinoma. *PloS Pathog.* 14, e1006916. doi: 10.1371/journal.ppat.1006916

Seok, J. Y., Na, D. C., Woo, H. G., Roncalli, M., Kwon, S. M., Yoo, J. E., et al. (2012). A fibrous stromal component in hepatocellular carcinoma reveals a cholangiocarcinoma-like gene expression trait and epithelial-mesenchymal transition. *Hepatology* 55, 1776–1786. doi: 10.1002/hep.25570

Sharma, S., Kelly, T. K., and Jones, P. A. (2010). Epigenetics in cancer. *Carcinogenesis* 31, 27–36. doi: 10.1093/carcin/bgp220

Shen, S., Peng, H., Wang, Y., Xu, M., Lin, M., Xie, X., et al. (2018). Screening for immune-potentiating antigens from hepatocellular carcinoma patients after radiofrequency ablation by serum proteomic analysis. *BMC Cancer* 18, 117. doi: 10.1186/s12885-018-4011-8

Shirota, Y., Kaneko, S., Honda, M., Kawai, H. F., and Kobayashi, K. (2001). Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. *Hepatology* 33, 832–840. doi: 10.1053/jhep.2001.23003

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA. Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Stefanska, B., Huang, J., Bhattacharyya, B., Suderman, M., Hallett, M., Han, Z.-G., et al. (2011). Definition of the Landscape of Promoter DNA Hypomethylation in Liver Cancer. *Cancer Res.* 71, 5891–5903. doi: 10.1158/0008-5472.CAN-10-3823

Therneau, T. (2013). A Package for Survival Analysis in S. R package version 2, 37–4, Available at: http://CRAN.R-project.org/package=survival

Therneau, T., and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tian, G., Yang, S., Yuan, J., Threapleton, D., Zhao, Q., Chen, F., et al. (2018). Comparative efficacy of treatment strategies for hepatocellular carcinoma: systematic review and network meta-analysis. *BMJ Open* 8, e021269. doi: 10.1136/bmjopen-2017-021269

Tizzot, M. R., Lidani, K. C. F., Andrade, F. A., Mendes, H. W., Beltrame, M. H., Reiche, E., et al. (2018). Ficolin-1 and Ficolin-3 Plasma Levels are altered in HIV and HIV/HCV coinfected patients from Southern Brazil. *Front. Immunol.* 9, 2292. doi: 10.3389/fimmu.2018.02292

Tung, E. K.-K., Mak, C. K.-M., Fatima, S., Lo, R. C.-L., Zhao, H., Zhang, C., et al. (2011). Clinicopathological and prognostic significance of serum and tissue Dickkopf-1 levels in human hepatocellular carcinoma. *Liver Int.* 31, 1494–1504. doi: 10.1111/j.1478-3231.2011.02597.x

Vasudevan, S., Flashner-Abramson, E., Remacle, F., Levine, R. D., and Kravchenko-Balasha, N. (2018). Personalized disease signatures through information-theoretic compaction of big cancer data. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7694–7699. doi: 10.1073/pnas.1804214115

Villa, E., Critelli, R., Lei, B., Marzocchi, G., Cammà, C., Giannelli, G., et al. (2016). Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. *Gut* 65, 861–869. doi: 10.1136/gutjnl-2014-308483

Wang, C.-Y., and Li, S. (2019). Clinical characteristics and prognosis of 2887 patients with hepatocellular carcinoma: a single center 14 years experience

from China. *Med. (Baltimore).* 98, e14070. doi: 10.1097/MD.0000000000014070

Wang, H., Huo, X., Yang, X.-R., He, J., Cheng, L., Wang, N., et al. (2017). STAT3-mediated upregulation of lncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOX4. *Mol. Cancer* 16, 136. doi: 10.1186/s12943-017-0680-1

Wang, Z., Teng, D., Li, Y., Hu, Z., Liu, L., and Zheng, H. (2018). A six-gene-based prognostic signature for hepatocellular carcinoma overall survival prediction. *Life Sci.* 203, 83–91. doi: 10.1016/j.lfs.2018.04.025

WELCH, B. L. (1947). The generalisation of student's problems when several different population variances are involved. *Biometrika* 34, 28–35. doi: 10.1093/biomet/34.1-2.28

Wong, K.-F., Liu, A. M., Hong, W., Xu, Z., and Luk, J. M. (2016). Integrin α2β1 inhibits MST1 kinase phosphorylation and activates Yes-associated protein oncogenic signaling in hepatocellular carcinoma. *Oncotarget* 7, 77683–77695. doi: 10.18632/oncotarget.12760

Woo, H. G., Choi, J.-H., Yoon, S., Jee, B. A., Cho, E. J., Lee, J.-H., et al. (2017). Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat. Commun.* 8, 839. doi: 10.1038/s41467-017-00991-w

Wu, G., Wu, J., Wang, B., Zhu, X., Shi, X., and Ding, Y. (2018). Importance of tumor size at diagnosis as a prognostic factor for hepatocellular carcinoma survival: a population-based study. *Cancer Manage. Res.* 10, 4401–4410. doi: 10.2147/CMAR.S177663

Xia, Q., Li, Z., Zheng, J., Zhang, X., Di, Y., Ding, J., et al. (2019). Identification of novel biomarkers for hepatocellular carcinoma using transcriptome analysis. *J. Cell. Physiol.* 234, 4851–4863. doi: 10.1002/jcp.27283

Xu, W., Rao, Q., An, Y., Li, M., and Zhang, Z. (2018). Identification of biomarkers for Barcelona Clinic Liver Cancer staging and overall survival of patients with hepatocellular carcinoma. *PloS One* 13, e0202763. doi: 10.1371/journal.pone.0202763

Yang, J. D., Addissie, B. D., Mara, K. C., Harmsen, W. S., Dai, J., Zhang, N., et al. (2019). Galad score for hepatocellular carcinoma detection in comparison with liver ultrasound and proposal of galadus score. *Cancer Epidemiol. Biomarkers Prev.* 28, 531–538. doi: 10.1158/1055-9965.EPI-18-0281

Zhang, C., Peng, L., Zhang, Y., Liu, Z., Li, W., Chen, S., et al. (2017). The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. *Med. Oncol.* 34, 101. doi: 10.1007/s12032-017-0963-9

Zhang, Y.-L., Ding, C., and Sun, L. (2019). High expression B3GAT3 is related with poor prognosis of liver cancer. *Open Med. (Warsaw Poland)* 14, 251–258. doi: 10.1515/med-2019-0020

Zhao, X., Parpart, S., Takai, A., Roessler, S., Budhu, A., Yu, Z., et al. (2015). Integrative genomics identifies YY1AP1 as an oncogenic driver in EpCAM+ AFP+ hepatocellular carcinoma. *Oncogene* 34, 5095–5104. doi: 10.1038/onc.2014.438

Zheng, Y., Liu, Y., Zhao, S., Zheng, Z., Shen, C., An, L., et al. (2018). Large-scale analysis reveals a novel risk score to predict overall survival in hepatocellular carcinoma. *Cancer Manage. Res.* 10, 6079–6096. doi: 10.2147/CMAR.S181396

Zubiete-Franco, I., García-Rodríguez, J. L., Lopitz-Otsoa, F., Serrano-Macia, M., Simon, J., Fernández-Tussy, P., et al. (2019). Sumoylation regulates LKB1 localization and its oncogenic activity in liver cancer. *EBioMedicine* 40, 406–421. doi: 10.1016/j.ebiom.2018.12.031

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.