Original Article

Cytochrome P450 family members are associated with fast-growing hepatocellular carcinoma and patient survival: An integrated analysis of gene expression profiles

Zhao-Zhen Liu[1,2], Li-Na Yan[2,3], Chun-Nan Dong[4], Ning Ma[1,2], Mei-Na Yuan[1,2], Jin Zhou[1,2], Ping Gao[1,2]

[1]Department of Social Medicine and Health Care Management, School of Public Health, [3]Department of Epidemiology and Biostatistics, School of Public Health, [4]Department of Pathogenic Biology, Hebei Medical University, Hebei, [2]Hebei Province Key Laboratory of Environment and Human Health, Hebei, China

Abstract

Background/Aims: The biological heterogeneity of hepatocellular carcinoma (HCC) makes prognosis difficult. Although many molecular tools have been developed to assist in stratification and prediction of patients by using microarray analysis, the classification and prediction are still improvable because the high-through microarray contains a large amount of information. Meanwhile, gene expression patterns and their prognostic value for HCC have not been systematically investigated. In order to explore new molecular diagnostic and prognostic biomarkers, the gene expression profiles between HCCs and adjacent nontumor tissues were systematically analyzed in the present study.

Materials and Methods: In this study, gene expression profiles were obtained by repurposing five Gene Expression Omnibus databases. Differentially expressed genes were identified by using robust rank aggregation method. Three datasets (GSE14520, GSE36376, and GSE54236) were used to validate the associations between cytochrome P450 (CYP) family genes and HCC. GSE14520 was used as the training set. GSE36376 and GSE54236 were considered as the testing sets.

Results: From the training set, a four-CYP gene signature was constructed to discriminate between HCC and nontumor tissues with an area under curve (AUC) of 0.991. Accuracy of this four-gene signature was validated in two testing sets (AUCs for them were 0.973 and 0.852, respectively). Moreover, this gene signature had a good performance to make a distinction between fast-growing HCC and slow-growing HCC (AUC = 0.898), especially for its high sensitivity of 95%. At last, *CYP2C8* was identified as an independent risk factor of recurrence-free survival (hazard ratio [HR] =0.865, 95% confidence interval [CI], 0.754–0.992, $P = 0.038$) and overall survival (HR = 0.849; 95% CI, 0.716–0.995, $P = 0.033$).

Conclusions: In summary, our results confirmed for the first time that a four-CYP gene (*CYP1A2*, *CYP2E1*, *CYP2A7*, and *PTGIS*) signature is associated with fast-growing HCC, and *CYP2C8* is associated with patient survival. Our findings could help to identify HCC patients at high risk of rapid growth and recurrence.

Keywords: Biomarker, fast-growing HCC, hepatocellular carcinoma, prognosis

Address for correspondence: Dr. Ping Gao, Department of Social Medicine and Health Care Management, School of Public Health, Hebei Medical University, No. 361 Zhongshan East Road, Shijiazhuang, 050017, Hebei Province, China. E-mail: 176288913@qq.com

Access this article online	
Quick Response Code:	Website: www.saudijgastro.com
	DOI: 10.4103/sjg.SJG_290_18

How to cite this article: Liu ZZ, Yan LN, Dong CN, Ma N, Yuan MN, Zhou J, *et al.* Cytochrome P450 family members are associated with fast-growing hepatocellular carcinoma and patient survival: Based on integrated analysis of gene expression profiles. Saudi J Gastroenterol 2019;25:167-75.

INTRODUCTION

Liver cancer, predominantly hepatocellular carcinoma (HCC), is the fifth most common cancer in men and the ninth in women. It is the second most common cause of death from cancer worldwide.[1] Although much is known about both the cellular changes that lead to HCC and the etiological agents responsible for the majority of HCC cases (hepatitis B virus, hepatitis C virus, alcohol), the molecular pathogenesis of HCC is still not well understood.[2] Considerable efforts have been devoted to establish staging systems for HCC by using clinical information and pathological classification to provide information at diagnosis on both survival and treatment options.[3-6] However, none of the proposed staging systems encompasses the biological and clinical heterogeneity exhibited by HCCs. One of the important reasons is that these predictive algorithms consider HCCs to be static rather than dynamic entities. They account for the size and number of neoplastic lesions at the time of presentation, yet do not take into account their growth behavior during follow-up, such as tumor doubling time (DT).[7] It therefore appears axiomatic that improving the classification of HCC patients into groups with homogeneous growth pattern will at least improve the application of currently available treatment modalities and at best provide new treatment strategies.

Over the past 20 years, microarray technology has led to the identification of several molecular signatures in HCC. For example, a 164-gene signature has been reported to predict the clinical behavior of metastatic HCC patients.[8] Another study established a five-gene score to predict HCC survival after liver resection.[9] These signatures allow stratification of HCC into several clinically relevant subgroups. Nevertheless, the classification and prediction are still improvable because the high-through microarray contains a large amount of bio-information. Therefore, it is necessary to systematically analyze the expression profiles and explore new molecular signatures.

In order to explore new molecular diagnostic and prognostic biomarkers, we systematically analyzed the gene expression profiles between HCCs and adjacent nontumor tissues in the present study. We demonstrated that 15 cytochrome P450 (CYP) family members could make a distinction between HCCs and nontumor tissues. A four CYP-gene (*CYP1A2*, *CYP2E1*, *CYP2A7*, and *PTGIS*) signature is a useful tool to diagnose HCCs and fast-growing HCCs with high sensitivity and specificity. *CYP2C8* is associated with patient survival in individuals at first diagnosis.

MATERIALS AND METHODS

Study design

Discovery stage: All the HCC Gene Expression Omnibus (GEO) datasets were collected. Then, a published robust rank aggregation (RRA) method was applied to identify the aberrant genes in HCC development.

Training stage: GSE14520 was used as the training set. The diagnostic and prognostic values of aberrant genes were evaluated, and a diagnostic signature was constructed in the training set.

Testing stage: GSE36376 and GSE54236 were used as the testing sets. The diagnostic and prognostic values of aberrant genes were further validated in the testing sets.

HCC patient datasets and patient samples

All the HCC datasets (generated from the Affymetrix Human Genome U133 Plus 2.0 Array) were collected from the publicly available GEO database (http://www.ncbi.nlm.nih.gov/geo/). The selection criteria used in this study are as follows: (1) all specimens classified as tissues; cells, serum, or plasma are not included; (2) all the included datasets must contain paired HCC tumors and adjacent noncancerous tissues; (3) sample size should be greater than three pairs; and (4) if there existed data overlapping, the largest sample size was selected. According to the above screening criteria, five datasets were finally included in this study (GSE62232, GSE55092, GSE17548, GSE33006, and GSE6764).[10-14]

To validate the result from the gene expression profiles above, three other datasets were included. They are Roessler's study, Lim's study, and Villa's study (GEO accession: GSE14520, GSE36376, and GSE54236).[7,15,16] HCC patients and tumor features are detailed in Table 1. It is worth noting that, in Villa's study, HCC patients were grouped into four quartiles according to increasing tumor DT: ≤53 days, 54–82 days, 83–110 days, and ≥111 days, respectively. The detailed procedure is as follows:[7] a new diagnosis of HCC at ultrasound (US) surveillance was eligible if they had a clinical condition that allowed a US-guided liver biopsy of a focal lesion, with the largest lesion biopsied in case of multifocality. To further confirm HCC diagnosis, a CT scan was performed. To measure the growth of lesions, a second CT was performed 6 weeks later. During the 6-week interval, patients did not undergo any specific treatment. This interval is much shorter than the average time to treatment after HCC diagnosis.[17,18] Therefore, no ethical issues were raised. After the second CT, patients were treated according to

Table 1: Clinical characteristics of patients enrolled in this study

Variables	Roessler's study GSE14520 (*n*=242)	Lim's study GSE36376 (*n*=240)	Villa's study GSE54236 (*n*=81)	*P**
Male, *n* (%)	211 (87.2)	199 (82.9)	61 (75.3)	0.04
Median age (years) (range)	50 (22-77)	53 (45-61)	67 (44-88)	NA†
HBV infection, *n* (%)	231 (95.5)	183 (76.3)	10 (12.3)	<0.0001
Tumor characteristics				
Tumor size, median (range) (cm)	7.2 (1.3-17.5)	3.7 (2.5-6.2)	5.8 (3.1-7.4)	NA
Single nodular, *n* (%)	190 (78.5)	183 (76.3)	69 (85.2)	0.239
Vascular invasion, *n* (%)	88 (36.4)	133 (55.4)	9 (11.1)	<0.0001
BCLC‡ stage, 0/A/B/C, *n*	20/152/24/29	0/139/91/10	0/56/14/10	<0.0001
Median follow-up (months)	51.7	NA	25	

*Chi-square test, NA=Not available, BCLC=Barcelona Clinic Liver Cancer

international guidelines. Based on these CT values, tumor growth was classified as fast growth (first quartile) or slow growth (other quartiles).

Acquisition and analysis of gene expression profiles of HCC patients

The raw array data (.CEL files) of five gene expression datasets were retrieved from the GEO database and were uniformly preprocessed using the Robust Multichip Average algorithm for background correction, quantile normalization, and log2-transformation. Then, the differentially expressed genes (DEGs) from each dataset were screened out on the basis of $P \leq 0.05$ and fold change ≥2.

The interlab reproducibility of the results is often problematic due to the small sample size and other factors (such as pathologic staging and surgical outcome) between the studies. To overcome these limitations, a published RRA method was applied to identify the aberrant genes in HCC development.[19,20] The new data frame results were constructed with the standard of adjusted P value < 0.05. The operation process can be performed by the Robust Rank Aggreg package in R software (version 3.3.3).

Then, the DEGs were classified into different functional gene groups by using the DAVID functional classification method.[21] The unsupervised hierarchical clustering of both HCC patients and aberrant genes was performed with R software by using the Euclidean distance and complete linkage method.

Statistical analysis

The continuous variables were analyzed by *t*-test or rank sum test as appropriate. The categorical variables were analyzed by Chi-square test. Binary logistic regression analysis was performed to identify variables independently associated with HCC. To construct a diagnostic model, the candidate genes were fitted in the multivariate logistic regression in the discovery dataset. Odds ratio (OR) and 95% confidence intervals (CI) were estimated by logistic regression model. To visualize the capacity of the risk signature to discriminate between HCC and non-HCC, we summarized the data in a receiver operating characteristic curve.[22]

The Cox proportional method was used to identify risk factors for recurrence-free survival (RFS) and overall survival (OS). The RFS was calculated from the date of tumor resection until the detection of tumor recurrence, or last observation. The OS was defined as the length of time between the surgery and death or the last follow-up. Variables with a P value < 0.05 in univariate analysis were included in the final multivariate model. Then, these variables were applied to build a risk signature. Finally, HCC patients were assigned a risk score according to the risk signature and were divided into high- and low-risk groups using the median of the risk score as the cutoff value. The difference in RFS or OS between high- and low-risk groups was demonstrated by Kaplan–Meier method, and the statistical significance was assessed by two-sided log-rank test. Hazard ratio (HR) and 95% CI were estimated by Cox proportional hazards regression model. Statistical analyses were performed with SPSS version 22.0 software (SPSS Inc., Chicago, IL, USA), GraphPad Prism 7 (GraphPad Software, La Jolla, CA), and MedCalc software version 12.2.1 (MedCalc, Mariakerke, Belgium).

RESULTS

GEO datasets analysis and candidate gene selection

A total of five datasets were included in our study for comprehensive analysis (GSE62232, GSE55092, GSE17548, GSE33006, and GSE6764). After data processing, 2179 DEGs were found in GSE62232; 2627, 2533, 1993, and 1158 DEGs were found in GSE55092, GSE17548, GSE33006, and GSE6764. Then, the method of RRA was used to integrally calculate the DEGs of the five datasets. Finally, 273 genes were identified as the most significantly differential genes. Detailed information is listed in Supplementary Table S1. These 273 genes were

classified into different functional gene groups by using the DAVID functional classification method. Finally, 93 of 273 genes were classified into 13 functional groups, of which 15 CYP family genes formed the largest cluster with the highest enrichment score of 8.47. Then, whether these 15 CYP genes had the ability to classify HCC and predict the outcome of HCC were validated.

Identification of CYP family genes associated with HCC

Three datasets were used to validate the association between CYP family genes and HCC. They are Roessler's study, Lim's study, and Villa's study (GEO accession: GSE14520, GSE36376, and GSE54236). A total of 242 patients were enrolled in the Roessler's study, which is the largest dataset in our study. So, we used this dataset to form a training set. The remaining two were considered as the testing sets. Patients in Villa's study were mostly Caucasians, while patients in Roessler's study and Lim's study were mostly Asians. Table 1 summarized the clinical characteristics of the patients in the training set and testing sets. There was heterogeneity among these three sets in some characteristics, such as sex distribution, hepatitis B virus infection, vascular invasion, and Barcelona Clinic Liver Cancer (BCLC) stage. Such heterogeneity may help to ensure that molecular signatures have real-world applicability across heterogeneous patient populations. Besides, in Villa's study, HCC patients were grouped into fast growing group (n = 20) and slow growing group (n = 61), according to increasing tumor DT. Kaplan–Meier curve analysis of survival showed a significantly lower survival rate for HCC cases in the fast growing group as compared with HCC cases in the slow growing group ($P < 0.0001$).[7]

The expressions of 15 CYP genes in HCC tissues were confirmed in the training set, and the results showed that all of these 15 genes were significantly decreased in HCC tissues as compared to that in the matched nontumor tissues [Figure 1b]. Unsupervised hierarchical clustering of 484 tissues according to the expression patterns of these 15 CYP genes showed two distant clusters, which were highly correlated with HCC (P = 6.53E - 9, Chi-square test; Figure 1a). Indeed, cluster I contained most of the nontumor tissues (n = 236; 97.5%). Conversely, cluster II contained the majority of tumor tissues (n = 241; 99.6%).

Construction of diagnostic signature from the training set and validating this signature in the testing sets

In univariate analysis, all of 15 genes were confirmed to be significantly differentially expressed between HCC tissues and nontumor tissues. In multivariate analysis, 4 of 15 genes reached statistical significance and were used to construct the diagnostic model [Table 2]. The model was as follows: logit (P) = 47.896 − 0.721 × *CYP1A2* − 1.132 × *CYP2E1* − 1.320 × *CYP2A7* − 3.736 × *PTGIS*. The best cutoff point of this model is −0.6513. Possibility above −0.6513 suggested HCC. The area under curve (AUC) for the established four-gene expression signature was 0.991 (95% CI, 0.977–0.997; Figure 2a), higher than the diagnostic value of any of these four genes (the AUCs of *CYP1A2*, *CYP2E1*, *CYP2A7*, and *PTGIS* were 0.973, 0.877, 0.931, and 0.874, respectively).

To confirm our findings, the diagnostic ability of the four-gene expression signature was validated in two testing sets. In Lim's study, with the same cutoff point, the AUC of the four-gene signature was 0.973 (95% CI, 0.953–0.986; Figure 2b). In Villa's study, the four-gene signature could distinguish HCC with an AUC of 0.852 (95% CI, 0.787–0.903; Figure 2c). Moreover, this gene signature had a good performance to make a distinction between fast-growing HCC and slow-growing HCC (AUC = 0.898; 95% CI, 0.810–0.954; Figure 2d), especially for its high sensitivity and specificity (95% and 85.25%, respectively).

Performance of the CYP family genes in HCC outcomes

In the training set, univariate Cox proportional hazard regression was applied to analyze each of the 15 genes. The results showed that 7 of 15 genes were significantly correlated with patient's RFS [Table 3], and another 7 of 15 genes were significantly correlated with patient's OS [Supplementary Table S2]. In multivariate analysis, only *CYP2C8* demonstrated significant correlation between patient's RFS (HR = 0.809; 95% CI, 0.712–0.919) and OS (HR = 0.735; 95% CI, 0.634–0.853).

Each patient in the training set was classified into different prognostic groups (the high- and low-risk

Table 2: Univariate and multivariate logistic regression analysis in the training set

Genes	Univariate analysis			Multivariate analysis		
	OR*	95% CI†	*P*	OR	95% CI	*P*
CYP39A1	0.190	(0.140-0.257)	<0.001			
CYP1A2	0.239	(0.181-0.316)	<0.001	0.486	(0.318-0.744)	0.001
CYP2B6	0.115	(0.079-0.168)	<0.001			
CYP2C8	0.004	(0.001-0.012)	<0.001			
CYP2C9	0.036	(0.018-0.069)	<0.001			
CYP2E1	0.183	(0.122-0.276)	<0.001	0.322	(0.161-0.644)	0.001
CYP2C18	0.222	(0.164-0.301)	<0.001			
CYP4A11	0.058	(0.033-0.101)	<0.001			
CYP2A6	0.255	(0.196-0.334)	<0.001			
CYP2A7	0.186	(0.139-0.250)	<0.001	0.267	(0.151-0.472)	<0.001
CYP26A1	0.174	(0.126-0.240)	<0.001			
CYP3A4	0.130	(0.085-0.197)	<0.001			
CYP2C19	0.153	(0.111-0.212)	<0.001			
CYP4F2	0.124	(0.082-0.186)	<0.001			
PTGIS	0.007	(0.003-0.017)	<0.001	0.024	(0.005-0.111)	<0.001

OR=Odds ratio, CI=confidence intervals

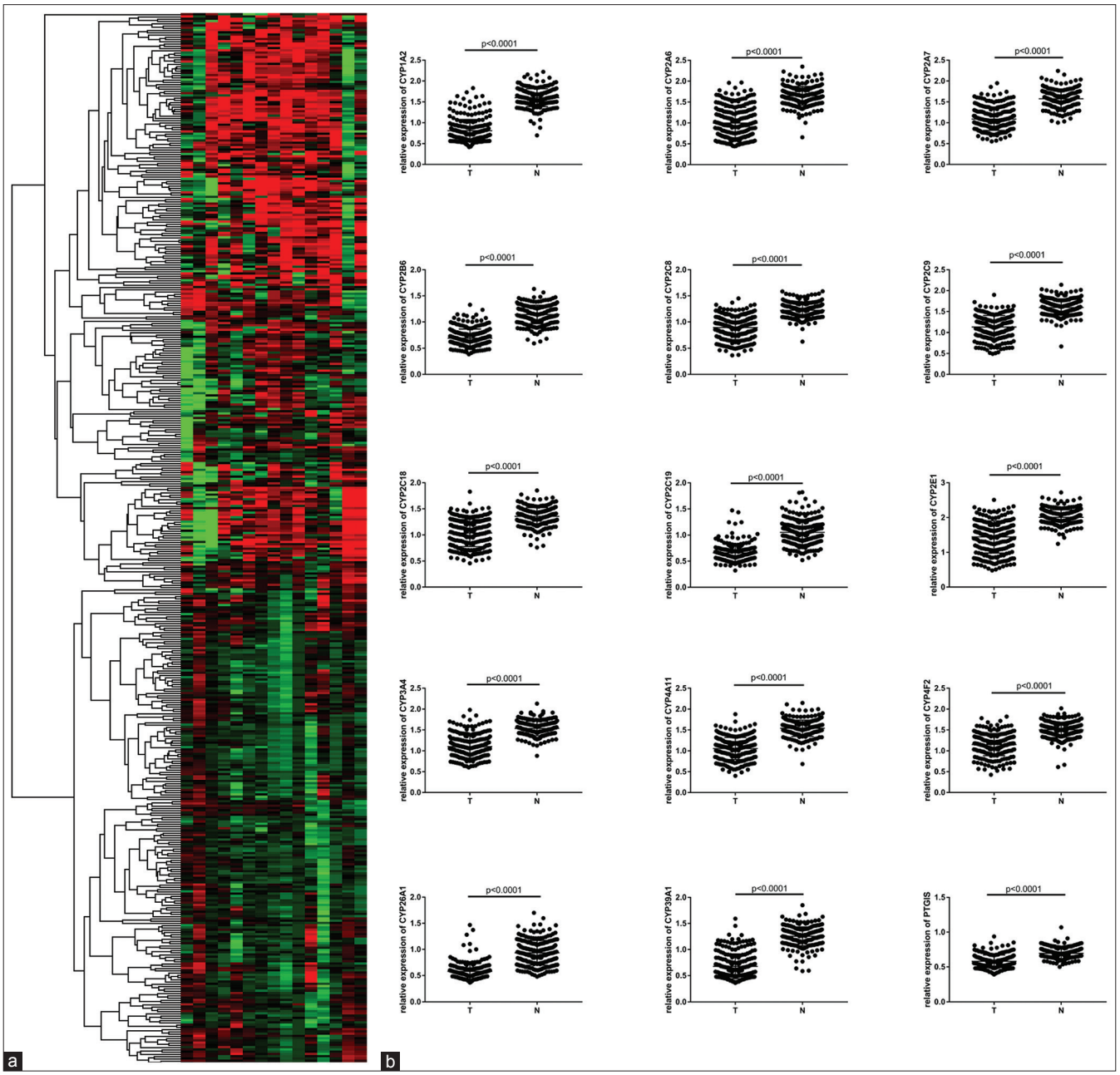

Figure 1: Deregulated cytochrome P450 (CYP) family genes in HCC tumor tissues in the training set. (a) The unsupervised hierarchical clustering heat map of 242 HCC samples and 242 matched adjacent nontumor livers; each row represents an individual tissue sample and each column represents the expression level of an individual CYP gene. (b) Relative expression of the 15 CYP genes in 242 HCC tumor tissues and 242 adjacent nontumor tissues. T = Tumor tissues, N = nontumor tissues

groups) according to the median value of *CYP2C8 (*6.39). Patients in high-risk group had mean and median RFS periods of 32.841 ± 2.468 and 23 months, respectively, whereas the mean and median RFS periods for patients in low-risk group were 44.960 ± 2.298 and 59.5 months, respectively. Correspondingly, the Kaplan–Meier analysis demonstrated a significant difference in RFS between these two groups ($P = 0.0009$; Figure 3a). Meanwhile, patients in high-risk group had significantly shorter OS period than those in low-risk group (mean 41.239 ± 2.461 vs. 54.519 ± 1.947 months; median 51.6 vs. 64.7 months; $P < 0.0001$, log-rank test; Figure 3b). Besides, *CYP2C8* was downregulated in BCLC stage B–C when compared with BCLC stage 0–A ($P = 0.023$).

Cox regression analysis identified *CYP2C8* (HR = 0.865; 95% CI, 0.754–0.992; $P = 0.038$), Tumor Node Metastasis (TNM) stage (HR = 1.641; 95% CI, 1.195–2.254; $P = 0.002$), and BCLC stage (HR = 1.760; 95% CI, 1.312–2.360; $P < 0.0001$) as independent risk factors for RFS.

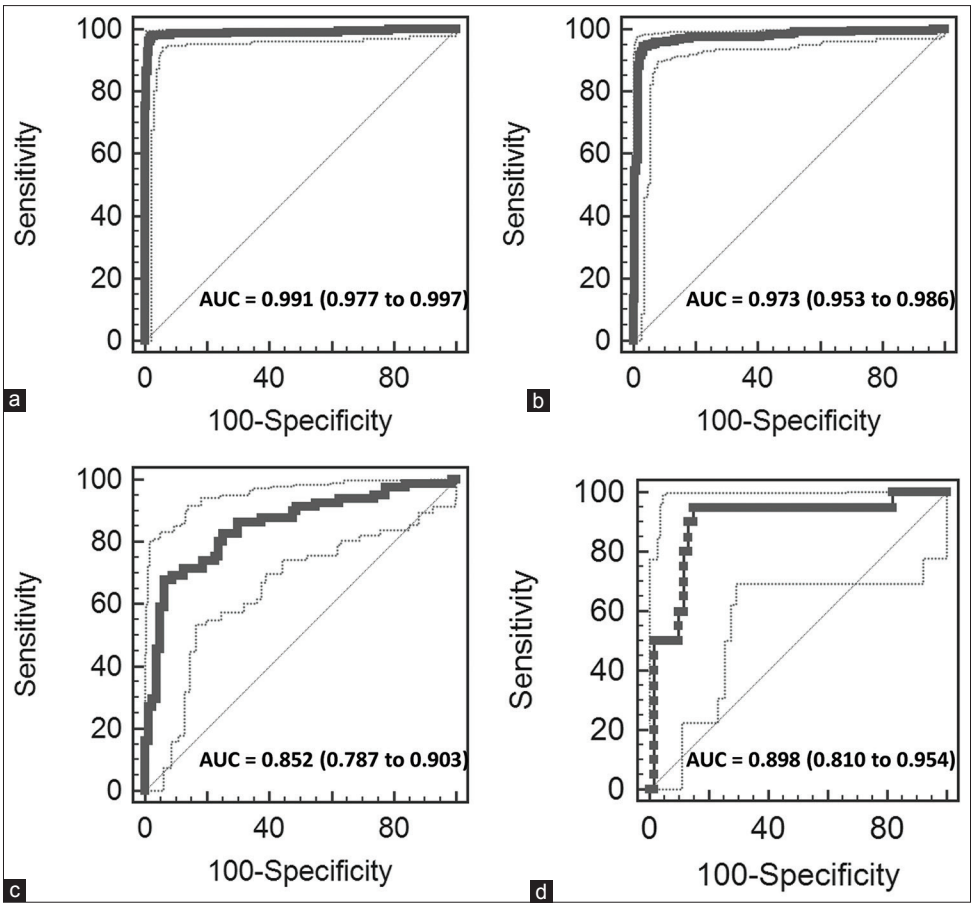

Figure 2: Receiver operating characteristic (ROC) curve analysis of the four-gene (*CYP1A2*, *CYP2E1*, *CYP2A7*, and *PTGIS*) signature in the training and testing sets. In order to compare the predictive value of the four-gene signature, we analyzed the ROC curve of the signature in different datasets. ROC plots for the four-gene panel discriminating HCC in the (a) training set, (b) Lim's dataset, (c) Villa's study; (d) ROC plots for the four-gene panel discriminating between fast-growing HCC and slow-growing HCC. AUC = Area under the curve

Table 3: Cox regression analysis of recurrence-free survival in the training set

Genes	Univariate analysis			Multivariate analysis		
	HR*	95% CI†	*P*	HR	95% CI	*P*
CYP39A1	0.922	(0.823-1.032)	0.158			
CYP1A2	0.909	(0.815-1.014)	0.087			
CYP2B6	0.869	(0.724-1.043)	0.132			
CYP2C8	0.809	(0.712-0.919)	0.001	0.809	(0.712-0.919)	0.001
CYP2C9	0.945	(0.852-10.48)	0.281			
CYP2E1	0.974	(0.920-1.031)	0.357			
CYP2C18	0.990	(0.889-1.103)	0.857			
CYP4A11	0.869	(0.780-0.967)	0.010			
CYP2A6	0.899	(0.834-0.970)	0.006			
CYP2A7	0.864	(0.767-0.973)	0.016			
CYP26A1	0.744	(0.554-1.000)	0.050			
CYP3A4	0.868	(0.774-0.973)	0.015			
CYP2C19	1.023	(0.850-1.231)	0.808			
CYP4F2	0.873	(0.778-0.980)	0.021			
PTGIS	0.994	(0.608-1.625)	0.981			

HR=Hazard ratio, CI=confidence intervals

CYP2C8 (HR = 0.849; 95% CI, 0.716–0.995; *P* = 0.033), TNM stage (HR = 1.723; 95% CI, 1.189–2.498; *P* = 0.004), and BCLC stage (HR = 1.582; 95% CI, 1.120–2.236; *P* = 0.009) were also independent risk factors for OS.

Moreover, Villa's dataset was used to validate the prognostic efficiency of *CYP2C8*. Patients were classified into high- and low-risk groups with the same cutoff point. Patients in high-risk group had mean and median OS periods of 20.450 ± 2.553 and 19 months, respectively, whereas the mean and median OS periods for patients in low-risk group were 39.217 ± 2.440 and 47 months, respectively. Kaplan–Meier curve analysis of survival showed a significantly lower survival rate for patients in high-risk group (*P* = 0.004, Figure 3c).

DISCUSSION

Progression of HCC often leads to vascular invasion and intrahepatic metastasis, which correlate with recurrence after surgical treatment and poor prognosis. Surgical resection and liver transplantation are the only curative treatments for HCC, but eligibility is uncommon. Even if patients underwent surgery, tumor recurrence occurs in more than 70% of cases within 5 years, and the 5-year survival rate is 60–70%.[23,24] In the past years, great efforts have been made to improve our understanding

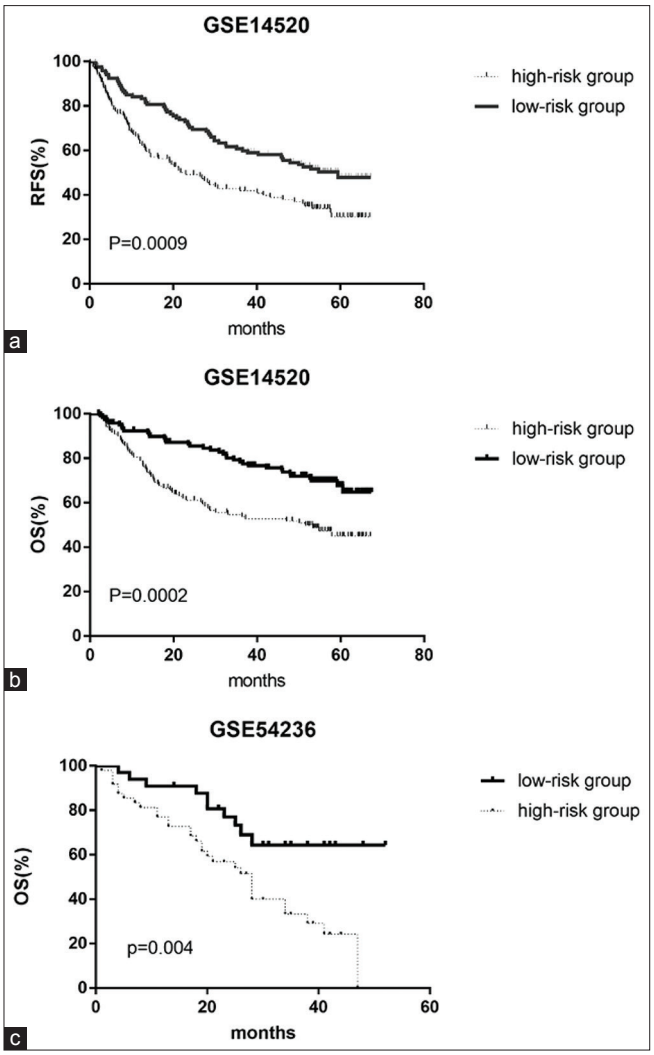

Figure 3: Kaplan–Meier curve for recurrence-free survival (RFS) and overall survival (OS) in patients with HCC with high- or low risk according to the median value of *CYP2C8*. (a, b) RFS and OS in the training set, and (c) OS in Villa's set

of the possible mechanism of progression, metastasis, and recurrence at protein, mRNA, and noncoding RNA levels, which will enable them to benefit from adjuvant therapy.[25,26] More recently, many molecular tools have been developed to assist in patient stratification and prediction with microarray analysis.[8,9,15] Nevertheless, the classification and prediction are still improvable because the high-through microarray contains a large amount of information. Moreover, gene expression patterns and their prognostic value for HCC have not been systematically investigated.

In this study, we have demonstrated that 15 CYP family genes were significantly decreased in HCC tissues as compared to that in the matched nontumor tissues. A four-gene diagnostic signature was constructed for distinguishing between HCC and noncancerous liver, and the results were robust. Besides, *CYP2C8* was identified as an independent risk factor of survival.

Till now, there are several studies about integrated analysis of gene expression profiles in HCC.[27-31] But the integration strategies are different. In a study by Shiraishi *et al*. the authors performed integrated and comparative analyses of whole genomes and transcriptomes of 22 HBV-related HCCs and their matched controls. The results showed that various types of genomic mutations triggered diverse transcriptional changes.[27] In another study, the Wang *et al*. repurposed 7 GEO datasets, which include a total of 267 HCC samples and 67 control samples and then reanalyzed the different genes in these 2 groups.[28] Zheng *et al*. only used one GEO dataset[29] while Chen *et al*. used three GEO datasets and one miRNA dataset to obtain DEGs and miRNAs.[30] In the study by Zhou *et al*. the authors chose four datasets, because they thought these datasets represented different racial populations.[31] In our study, gene expression profiles have been obtained by repurposing five GEO datasets. Although each of the five datasets used Affymetrix Human Genome U133 Plus 2.0 Array to analyze gene expression patterns, the interlab reproducibility of the results is often problematic due to the small sample size and other factors (such as pathologic staging and surgical outcome). To overcome these limitations, RRA approach was applied in this study. It has been specifically designed for comparison of several ranked gene lists and identification of commonly overlapping genes. This method assigns a *P* value to each element in the aggregated list indicating how much better it is ranked compared with a null model expecting random ordering. Finally, 273 genes were identified as the most significantly differential genes. This method has some strength. Most importantly, it is based on a statistical model that naturally allows evaluating the significance of the results. In addition, RRA is easy to compute and robust, not restricting its use to certain subset of problems or requiring all data to be of top quality. This method can also handle variable gene content of different microarray platforms. By defining the rank vector for each gene based only on the datasets where it is present, we do not have to omit the genes that are not present in every platform.[20]

In the four-gene signature, the observation of low *CYP1A2* expression in HCC was also reported by other studies.[32,33] CYP1A2 metabolizes 17β-estradiol to generate the potent antitumor agent 2-methoxyestradiol in HCC. The reduction of CYP1A2 significantly disrupts this metabolic pathway, contributing to the progression and growth of HCC.[33] The results of previous studies also suggest that functional relationship occurs among genes characterizing the

signature identified in this study. Fan *et al.* reported that *CYP2E1* revealed low level of expression in 70% of the tumor tissues, when compared to the adjacent nontumor tissues, at both mRNA and protein levels. The low expression of *CYP2E1* was significantly correlated with the aggressive tumor phenotype, including poor differentiation status, absence of tumor capsule, and younger age of the patients.[34] Moreover, HBx inhibits human *CYP2E1* gene expression via downregulating HNF4α, which contributes to promotion of human hepatoma cell growth.[35] However, *CYP2A7* and *PTGIS* are poorly studied in HCC, and further research may reveal a better understanding of the interaction of HCC and these genes.

Lastly, by using multivariate Cox proportional hazard regression analysis, our study demonstrated that only 1 (*CYP2C8*) of 15 genes was significantly correlated with patient's RFS and OS. Even so, the multivariate analysis with HRs indicated that *CYP2C8* is a significant survival-related risk factor independent of the well-known BCLC staging system. This implies that HCC prognosis could be improved by the combination of *CYP2C8* with the existed staging system. However, the mechanisms of *CYP2C8* in HCC remain unclear.

A major strength of our study is that the samples were derived from three large populations with different races, which ensured that molecular signatures have real-world applicability. Another advantage of our study is that the rigorous data-processing methods and statistical analysis made our results reliable. Nonetheless, there are also limitations in our study. First, in order to measure the growth of lesions, a second CT was performed 6 weeks later. During the 6-week interval, patients did not undergo any specific treatment. This interval is much shorter than the average time to treatment after HCC diagnosis. Therefore, no ethical issues were raised. But this interval could not be compared with average waiting time for treatment as this usually varies due to unintended reasons. Second, our work needs to be re-evaluated in a special cohort of patients with a proper follow-up or in case-control studies. The signatures should be validated in qRT-PCR-based samples; we therefore need to develop the signatures in tissues with qRT-PCR method. Besides, we did not study the mechanisms of the screened CYP family genes; whether these genes can affect the biological functions of HCC cells remains to be studied.

CONCLUSION

A four-gene signature was identified as being able to discriminate between HCC and nontumor tissues as well as identify a subgroup of patients with rapidly growing HCC. Moreover, *CYP2C8* can be used as an independent prognostic risk factor. These results may not only help to identify HCC patients at high risk of rapid growth and recurrence but could also provide insight into the mechanisms of HCC progression, metastasis, and recurrence.

Acknowledgments

This study was supported by Young Fund of Hebei Natural Science Foundation of China (No. H2016206411) and Undergraduate Innovative Training Program of Hebei Medical University (No. USIP2017235).

Financial support and sponsorship

There are no financial supports and sponsorships.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:359-86.
2. Thorgeirsson SS, Grisham JW. Molecular pathogenesis of human hepatocellular carcinoma. Nat Genet 2002;31:339-46.
3. Bruix J, Sherman M, American Association for the Study of Liver D. Management of hepatocellular carcinoma: An update. Hepatology 2011;53:1020-2.
4. European Association For The Study Of The Liver, European Organisation For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: Management of hepatocellular carcinoma. J Hepatol 2012;56:908-43.
5. Marrero JA, Fontana RJ, Barrat A, Askari F, Conjeevaram HS, Su GL, *et al.* Prognosis of hepatocellular carcinoma: Comparison of 7 staging systems in an American cohort. Hepatology 2005;41:707-16.
6. Camma C, Di Marco V, Cabibbo G, Latteri F, Sandonato L, Parisi P, *et al.* Survival of patients with hepatocellular carcinoma in cirrhosis: A comparison of BCLC, CLIP and GRETCH staging systems. Aliment Pharmacol Ther 2008;28:62-75.
7. Villa E, Critelli R, Lei B, Marzocchi G, Camma C, Giannelli G, *et al.* Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut 2016;5:861-9.
8. Ye QH, Qin LX, Forgues M, He P, Kim JW, Peng AC, *et al.* Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 2003;9:416-23.
9. Nault JC, De Reynies A, Villanueva A, Calderaro J, Rebouissou S, Couchy G, *et al.* A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. Gastroenterology 2013;145:176-87.
10. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Gen 2015;47:505-11.
11. Melis M, Diaz G, Kleiner DE, Zamboni F, Kabat J, Lai J, *et al.* Viral expression and molecular profiling in liver tissue versus microdissected hepatocytes in hepatitis B virus-associated hepatocellular carcinoma. J Transl Med 2014;12:230.

12. Yildiz G, Arslan-Ergul A, Bagislar S, Konu O, Yuzugullu H, Gursoy-Yuzugullu O, *et al*. Genome-wide transcriptional reorganization associated with senescence-to-immortality switch during human hepatocellular carcinogenesis. PloS One 2013;8:64016.
13. Huang Y, Chen HC, Chiang CW, Yeh CT, Chen SJ, Chou CK. Identification of a two-layer regulatory network of proliferation-related microRNAs in hepatoma cells. Nucleic Acids Res 2012;40:10478-93.
14. Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, Schwartz M, *et al*. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology 2007;45:938-47.
15. Lim HY, Sohn I, Deng S, Lee J, Jung SH, Mao M, *et al*. Prediction of disease-free survival in hepatocellular carcinoma by gene expression profiling. Ann Surg Oncol 2013;20:3747-53.
16. Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, *et al*. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res 2010;70:10202-12.
17. Yopp AC, Mansour JC, Beg MS, Arenas J, Trimmer C, Reddick M, *et al*. Establishment of a multidisciplinary hepatocellular carcinoma clinic is associated with improved clinical outcome. Ann Surg Oncol 2014;21:1287-95.
18. Singal AG, Waljee AK, Patel N, Chen EY, Tiro JA, Marrero JA, *et al*. Therapeutic delays lead to worse survival among patients with hepatocellular carcinoma. J Natl Compr Canc Netw 2013;11:1101-8.
19. Vosa U, Vooder T, Kolde R, Vilo J, Metspalu A, Annilo T. Meta-analysis of microRNA expression in lung cancer. Int J Cancer 2013;132:2884-93.
20. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 2012;28:573-80.
21. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucl Acids Res 2009;37:1-13.
22. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.
23. Llovet JM, Schwartz M, Mazzaferro V. Resection and liver transplantation for hepatocellular carcinoma. Semin Liver Dis 2005;25:181-200.
24. Poon RT. Prevention of recurrence after resection of hepatocellular carcinoma: A daunting challenge. Hepatology 2011;54:757-9.
25. Giordano S, Columbano A. MicroRNAs: New tools for diagnosis, prognosis, and therapy in hepatocellular carcinoma? Hepatology 2013;57:840-7.
26. Giannelli G, Koudelkova P, Dituri F, Mikulits W. Role of epithelial to mesenchymal transition in hepatocellular carcinoma. J Hepatol 2016;65:798-808.
27. Shiraishi Y, Fujimoto A, Furuta M, Tanaka H, Chiba K, Boroevich KA, *et al*. Integrated analysis of whole genome and transcriptome sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in liver cancers. PLoS One 2014;12:e114263.
28. Wang F, Wang R, Li Q, Qu X, Hao Y, Yang J, *et al*. A transcriptome profile in hepatocellular carcinomas based on integrated analysis of microarray studies. Diagn Pathol 2017;12:4.
29. Wang J, Tian Y, Chen H, Li H, Zheng S. Key signaling pathways, genes and transcription factors associated with hepatocellular carcinoma. Mol Med Rep 2018;17:8153-60.
30. Zhou L, Du Y, Kong L, Zhang X, Chen Q. Identification of molecular target genes and key pathways in hepatocellular carcinoma by bioinformatics analysis. Onco Targets Ther 2018;11:1861-9.
31. Xing T, Yan T, Zhou Q. Identification of key candidate genes and pathways in hepatocellular carcinoma by integrated bioinformatical analysis. Exp Ther Med 2018;15:4932-42.
32. Chen H, Shen ZY, Xu W, Fan TY, Li J, Lu YF, *et al*. Expression of P450 and nuclear receptors in normal and end-stage Chinese livers. World J Gastroenterol 2014;20:8681-90.
33. Ren J, Chen GG, Liu Y, Su X, Hu B, Leung BC, *et al*. Cytochrome P450 1A2 Metabolizes 17 beta-estradiol to suppress hepatocellular carcinoma. PLoS One 2016;11:0153863.
34. Ho JC, Cheung ST, Leung KL, Ng IO, Fan ST. Decreased expression of cytochrome P450 2E1 is associated with poor prognosis of hepatocellular carcinoma. Int J Cancer 2004;111:494-500.
35. Liu H, Lou G, Li C, Wang X, Cederbaum AI, Gan L, *et al*. HBx inhibits CYP2E1 gene expression via downregulating HNF4alpha in human hepatoma cells. PLoS One 2014;9:107913.

Table S1: Significantly differentially expressed genes after integrated calculating through RRA method

Gene symbol	Score	Adjusted *P*	Gene symbol	Score	Adjusted *P*
Upregulated			Downregulated		
SPINK1	5.76E-19	3.15E-14	FCN2	1.19E-17	6.52E-13
AKR1B10	2.75E-17	1.50E-12	CLEC1B	3.03E-16	1.65E-11
HMMR	1.52E-15	8.30E-11	SLC22A1	2.12E-15	1.16E-10
ASPM	5.14E-15	2.80E-10	CXCL14	5.14E-15	2.80E-10
NDC80	1.89E-14	1.03E-09	FCN3	8.58E-15	4.69E-10
ROBO1	2.09E-14	1.14E-09	GLS2	1.89E-14	1.03E-09
CAP2	7.35E-14	4.01E-09	CYP39A1	2.31E-14	1.26E-09
RACGAP1	1.64E-13	8.97E-09	CYP1A2	4.86E-14	2.65E-09
CCNB1	2.14E-13	1.17E-08	CYP2B6	5.76E-14	3.14E-09
TOP2A	2.89E-13	1.58E-08	CNDP1	7.35E-14	4.01E-09
GPC3	3.89E-13	2.12E-08	CLEC4M	1.53E-13	8.38E-09
PRC1	4.13E-13	2.26E-08	C9	2.43E-13	1.32E-08
RRM2	4.88E-13	2.66E-08	CXCL12	4.13E-13	2.26E-08
SPP1	1.19E-12	6.51E-08	APOF	7.04E-13	3.84E-08
DNAJC6	1.42E-12	7.78E-08	ESR1	7.78E-13	4.25E-08
KIF20A	1.55E-12	8.48E-08	CRHBP	1.19E-12	6.51E-08
IGF2BP3	1.62E-12	8.84E-08	TENM1	1.25E-12	6.81E-08
PRR11	2.17E-12	1.18E-07	GHR	1.42E-12	7.78E-08
MAP2	4.74E-12	2.59E-07	DNASE1L3	1.62E-12	8.86E-08
CDKN2C	7.05E-12	3.85E-07	ADRA1A	1.84E-12	1.01E-07
DLGAP5	8.25E-12	4.50E-07	LPA	1.92E-12	1.05E-07
NUSAP1	1.18E-11	6.44E-07	HGF	2.44E-12	1.33E-07
BIRC5	2.03E-11	1.11E-06	HAO2	3.08E-12	1.68E-07
KIF4A	2.74E-11	1.50E-06	IL1RAP	3.08E-12	1.68E-07
NCAPG	3.24E-11	1.77E-06	HAMP	4.62E-12	2.52E-07
CCNB2	3.81E-11	2.08E-06	IGF1	6.20E-12	3.39E-07
MAGEA1	4.26E-11	2.33E-06	LIFR	7.05E-12	3.85E-07
KIF11	5.41E-11	2.95E-06	CYP2C8	7.51E-12	4.10E-07
AURKA	6.52E-11	3.56E-06	NAT2	8.25E-12	4.50E-07
SULT1C2	7.39E-11	4.03E-06	FBP1	1.08E-11	5.90E-07
CCNA2	8.28E-11	4.52E-06	LCAT	1.18E-11	6.44E-07
CDK1	1.14E-10	6.24E-06	GBA3	1.64E-11	8.98E-07
E2F8	1.76E-10	9.63E-06	NNMT	1.69E-11	9.23E-07
CCNE2	2.01E-10	1.10E-05	MARCO	1.88E-11	1.03E-06
BUB1	2.04E-10	1.12E-05	SLCO1B3	1.89E-11	1.03E-06
LCN2	2.42E-10	1.32E-05	ALDOB	2.36E-11	1.29E-06
CENPA	2.80E-10	1.53E-05	RDH16	4.26E-11	2.33E-06
C1orf112	3.06E-10	1.67E-05	SPP2	4.65E-11	2.54E-06
CDKN3	3.45E-10	1.88E-05	CD5L	6.93E-11	3.78E-06
AKR1C3	3.87E-10	2.11E-05	MT1M	7.97E-11	4.35E-06
PBK	6.92E-10	3.78E-05	PLAC8	1.16E-10	6.34E-06
BUB1B	7.55E-10	4.12E-05	COLEC11	1.18E-10	6.45E-06
ECT2	9.69E-10	5.29E-05	CYP2C9	1.51E-10	8.26E-06
EDIL3	1.01E-09	5.51E-05	CYP2E1	1.65E-10	9.00E-06
CDKN2A	1.05E-09	5.71E-05	SRD5A2	1.68E-10	9.15E-06
DEPDC1	1.11E-09	6.05E-05	GLYAT	2.11E-10	1.15E-05
KIAA0101	1.21E-09	6.63E-05	HGFAC	2.11E-10	1.15E-05
TRIM16	1.53E-09	8.36E-05	C8orf4	2.18E-10	1.19E-05
PRKAA2	2.31E-09	0.000126011	AFM	2.32E-10	1.27E-05
CDC20	2.54E-09	0.00013875	ZG16	2.36E-10	1.29E-05
EZH2	2.69E-09	0.00014687	IGFBP3	2.76E-10	1.51E-05
TTK	3.15E-09	0.000172213	ATF5	3.02E-10	1.65E-05
LEF1	3.45E-09	0.000188269	SOCS2	3.35E-10	1.83E-05
ACSL4	4.21E-09	0.000230079	NPY1R	3.71E-10	2.03E-05
CENPF	5.10E-09	0.00027849	KDM8	4.52E-10	2.47E-05
DTL	8.21E-09	0.000448485	COLEC10	4.58E-10	2.50E-05
MELK	9.44E-09	0.000515558	MRC1	5.11E-10	2.79E-05
CDKN2B	9.76E-09	0.000532784	XDH	5.61E-10	3.06E-05
NEK2	1.10E-08	0.000603006	STEAP4	6.15E-10	3.36E-05
FGF13	1.10E-08	0.000603006	MT1F	6.62E-10	3.62E-05
COL15A1	1.26E-08	0.000686421	MBL2	6.99E-10	3.82E-05
CLGN	1.27E-08	0.000692295	SLC7A2	7.17E-10	3.92E-05
STXBP6	1.85E-08	0.00100832	CYP2C18	7.64E-10	4.17E-05
ITGA6	2.15E-08	0.001173191	DCN	8.22E-10	4.49E-05

Contd...

Table S1: Contd...

Gene symbol	Score	Adjusted *P*	Gene symbol	Score	Adjusted *P*
Upregulated			Downregulated		
RRAGD	2.20E-08	0.001199474	STAB2	8.22E-10	4.49E-05
FAM169A	2.44E-08	0.001333329	CIDEB	8.43E-10	4.60E-05
MAGEA6	2.44E-08	0.001333329	CYP4A11	9.28E-10	5.07E-05
PEG10	2.72E-08	0.001483161	CYP2A6	9.30E-10	5.08E-05
KIF14	2.76E-08	0.001509308	RCAN1	1.30E-09	7.09E-05
SLC7A11	2.81E-08	0.0015358	SRPX	1.32E-09	7.22E-05
MAD2L1	3.01E-08	0.001645264	ZGPAT	1.42E-09	7.75E-05
ENAH	3.44E-08	0.001878771	LY6E	1.45E-09	7.92E-05
TKT	3.62E-08	0.001976497	VNN1	1.48E-09	8.06E-05
MAGEA3	4.38E-08	0.002392275	MASP1	1.55E-09	8.46E-05
PTTG1	4.66E-08	0.002544424	CA2	1.69E-09	9.21E-05
ZWINT	5.33E-08	0.002913182	KCNN2	1.76E-09	9.63E-05
CENPU	6.08E-08	0.003320583	CXCL2	2.01E-09	0.000110033
APOBEC3B	1.17E-07	0.006411243	KBTBD11	2.25E-09	0.000123013
DLG5	1.80E-07	0.009808666	KAZN	2.34E-09	0.000128026
TXNRD1	1.94E-07	0.010578678	CETP	3.07E-09	0.000167869
EFCAB2	4.75E-07	0.025921521	GYS2	3.15E-09	0.000172213
CKAP2	4.79E-07	0.02614349	MT1G	3.25E-09	0.000177446
SLC38A6	5.30E-07	0.028919541	MT1H	3.35E-09	0.000182797
NRCAM	5.90E-07	0.032238372	GPM6A	3.60E-09	0.000196524
DHRS2	6.29E-07	0.034360482	THBS1	5.29E-09	0.000289149
TPX2	6.33E-07	0.034574782	AKR1D1	5.81E-09	0.000317473
FAT1	6.85E-07	0.037402125	MT1E	6.12E-09	0.000334159
SMPX	8.02E-07	0.043778089	MT1X	6.28E-09	0.000342744
HOXA3	8.40E-07	0.045901344	HABP2	6.83E-09	0.000373189
			GREM2	7.28E-09	0.000397764
			PLG	7.64E-09	0.000417031
			GSTZ1	7.76E-09	0.000423617
			AGXT	9.12E-09	0.000497916
			MYO10	9.19E-09	0.000501698
			ACSM3	9.33E-09	0.00050933
			MOGAT2	9.47E-09	0.000517054
			ECM1	9.61E-09	0.000524872
			GNMT	9.88E-09	0.000539501
			ADH1A	1.05E-08	0.000573787
			ADAMTS13	1.18E-08	0.000644603
			ANXA10	1.45E-08	0.000794344
			TMEM45A	1.58E-08	0.000861452
			PDGFRA	1.67E-08	0.000914505
			TDO2	1.74E-08	0.000951209
			ASS1	1.77E-08	0.000968444
			FOS	1.78E-08	0.000969969
			SLC10A1	1.85E-08	0.00100832
			BBOX1	1.99E-08	0.001088418
			AZGP1	2.02E-08	0.00110373
			FGFR2	2.15E-08	0.001173191
			EPB41L4B	2.32E-08	0.001269561
			SH3YL1	2.49E-08	0.001357476
			KMO	2.91E-08	0.001589826
			C7	3.17E-08	0.00173112
			ANGPTL6	3.23E-08	0.001761951
			ADH1C	3.58E-08	0.001956276
			PRG4	3.62E-08	0.001976497
			CD1D	3.73E-08	0.002036314
			SLCO4C1	3.99E-08	0.002176958
			HBB	4.05E-08	0.002211793
			FETUB	4.18E-08	0.002282715
			MT1HL1	4.45E-08	0.002429655
			MCC	4.80E-08	0.002623161
			SHBG	4.87E-08	0.002658396
			MT2A	5.49E-08	0.003000286
			IGFALS	5.68E-08	0.00310222
			RBMS3	6.37E-08	0.003476853
			SLC22A7	6.37E-08	0.003476853

Contd...

Table S1: Contd...

Gene symbol	Score	Adjusted *P*	Gene symbol	Score	Adjusted *P*
Upregulated			Downregulated		
			PCK1	6.71E-08	0.00366583
			CHST4	6.84E-08	0.003733479
			HAO1	7.67E-08	0.004187743
			OLFML3	9.02E-08	0.004927537
			CFP	1.14E-07	0.006235826
			FAM134B	1.16E-07	0.006333836
			C6	1.20E-07	0.00656818
			GRAMD1C	1.47E-07	0.008021078
			TFPI2	1.64E-07	0.008980903
			TAT	1.70E-07	0.009284924
			TRPM8	1.74E-07	0.009491841
			CPEB3	1.76E-07	0.009596585
			BHMT	1.86E-07	0.010133342
			CYP2A7	1.98E-07	0.010806761
			CYP26A1	2.06E-07	0.011273951
			NDRG2	2.17E-07	0.011875047
			SLC1A1	2.36E-07	0.012897255
			KCND3	2.41E-07	0.01316171
			ADH6	2.58E-07	0.014062978
			HAL	2.66E-07	0.014545375
			ASPA	3.05E-07	0.016661935
			ANK3	3.05E-07	0.016661935
			F9	3.08E-07	0.016821456
			CYP3A4	3.17E-07	0.017306885
			ADH1B	3.58E-07	0.019532585
			CYP2C19	3.61E-07	0.019712234
			G6PC	3.81E-07	0.020816234
			FOSB	3.99E-07	0.021771039
			OAT	4.02E-07	0.021965865
			ASPN	4.39E-07	0.023986679
			ART4	4.47E-07	0.024406973
			CYP4F2	4.63E-07	0.025264069
			MASP2	4.71E-07	0.025700965
			FOLH1B	5.08E-07	0.027737347
			ACADL	5.12E-07	0.027970839
			NAMPT	5.38E-07	0.029402819
			GADD45B	5.56E-07	0.030387478
			MFAP3L	5.70E-07	0.031142006
			ITGA9	5.94E-07	0.032430592
			SLC19A3	6.28E-07	0.03430112
			ABCA8	6.80E-07	0.037111871
			FAM13A	7.23E-07	0.039481728
			PTGIS	8.10E-07	0.044228221
			EPB41L4A	8.22E-07	0.044891898
			SERPINA4	8.78E-07	0.047971072
			KLKB1	8.85E-07	0.048322697
			BCHE	9.04E-07	0.049389145

Table S2: Cox regression analysis of overall survival in the training set

Genes	Univariate analysis			Multivariate analysis		
	HR	95% CI	*P*	HR	95% CI	*P*
CYP39A1	0.944	(0.826-1.080)	0.403			
CYP1A2	0.906	(0.792-1.035)	0.146			
CYP2B6	0.890	(0.714-1.108)	0.295			
CYP2C8	0.735	(0.634-0.853)	0.000	0.735	(0.634-0.853)	0.000
CYP2C9	0.868	(0.770-0.979)	0.021			
CYP2E1	0.948	(0.888-1.013)	0.116			
CYP2C18	0.943	(0.829-1.072)	0.370			
CYP4A11	0.776	(0.683-0.882)	0.000			
CYP2A6	0.854	(0.778-0.938)	0.001			
CYP2A7	0.800	(0.690-0.926)	0.003			
CYP26A1	0.779	(0.553-1.096)	0.152			
CYP3A4	0.797	(0.692-0.918)	0.002			
CYP2C19	1.004	(0.796-1.267)	0.970			
CYP4F2	0.832	(0.724-0.956)	0.010			
PTGIS	1.087	(0.605-1.952)	0.780			