Research Paper

Nucleotide variants in hepatitis B virus preS region predict the recurrence of hepatocellular carcinoma

Xi Chen[1,*], Minfeng Zhang[2,*], Nan Li[2], Rui Pu[1], Ting Wu[1], Yibo Ding[1], Peng Cai[1], Hongwei Zhang[1], Jun Zhao[2], Jianhua Yin[1,&], Guangwen Cao[1]

[1]Department of Epidemiology, Faculty of Navy Medicine, Second Military Medical University, Shanghai, China
[2]Department of Surgery, Eastern Hepatobiliary Surgery Hospital, Second Military Medical University, Shanghai, China
*Equal contribution

Correspondence to: Jianhua Yin, Guangwen Cao; **email:** hawkyjh163@163.com, https://orcid.org/0000-0003-1047-7059; gcao@smmu.edu.cn
Keywords: hepatitis B virus, viral variant, prediction model, hepatocellular carcinoma, prognosis
Received: March 31, 2021 **Accepted:** September 3, 2021 **Published:** September 17, 2021

ABSTRACT

Background: Hepatitis B virus (HBV) variants in the preS region have been associated with hepatocellular carcinoma (HCC). However, the effect of the preS variants on HCC prognosis remains largely unknown. We aimed to identify the preS variants that reliably predict postoperative prognosis in HCC.
Methods: RNA-seq data of 203 HCC patients retrieved from public database were screened for the preS variants related to HCC prognosis. The variants in the sera and tumors were then validated in our prospective cohort enrolling 103 HBV-associated HCC patients.
Results: By analyzing prognosis-related gene sets in the RNA-seq data, 12 HBV preS variants were associated with HCC recurrence. Of those, G40C and C147T in the sera predicted an unfavorable recurrence-free survival in our cohort (hazard ratio [HR]=2.18, 95% confidence interval [CI]=1.37-3.47, *p*=0.001 for G40C; HR=1.84, 95% CI=1.15-2.92, *p*=0.012 for C147T). G40C and C147T were significantly associated with microscopic vascular invasion, larger tumor size, and abnormal liver function. Multivariate Cox regression analysis showed that G40C significantly increased the risk of HCC recurrence in patients with postoperative antiviral treatment. The HCC prognosis-prediction model consisting of α-fetoprotein and G40C in the sera achieved the best performance (sensitivity=0.80, specificity=0.70, and area under the curve=0.79). Functional analysis indicated that these two variants were associated with cell proliferation, chromosome instability, carcinogenesis, metastasis, and anticancer drug resistance.
Conclusions: G40C and C147T are serological biomarkers for HCC prognosis and the prognostic model consisting of serological α-fetoprotein and G40C achieved the best performance in predicting postoperative prognosis.

INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the deadliest malignant diseases. Chronic hepatitis B virus (HBV) infection is the most common cause of HCC worldwide [1]. During HBV-induced carcinogenesis, HBV keeps evolving. Some HBV mutants can promote the development of HCC [2, 3]. Due to the absence of a proofreading function for HBV polymerase, HBV has a relatively higher mutation rate during virus replication [4]. Mutations, especially mutations in the preS region of the HBV genome, are associated with advanced liver diseases, including HCC [5]. Both preS1 and preS2 deletions can cause unbalanced production of HBV envelope proteins, with consequent accumulation of the mutated large HBV surface antigen (LHBS) in the endoplasmic reticulum (ER) of hepatocytes, causing ER stress and, ultimately, HCC development [6–8]. A meta-

analysis including 5563 HBV-infected patients has demonstrated that HBV preS deletion is significantly associated with an increased risk of HCC, with a summary odds ratio of 3.0 [9]. Importantly, HBV preS mutations especially deletions and some HCC-associated preS point mutations are present at least 10 years before the development of HCC [10]. These studies indicate that HBV preS mutations can predict the development of HCC in HBV-infected subjects.

The recurrence rate of HCC is high after curative resection. It is important to predict the prognosis of HCC before surgical treatment. However, reliable biomarkers for predicting HCC prognosis are lacking. Previous studies have demonstrated that higher viral load, preS deletion mutations, and higher expression of LHBS with partial pre-S2 deletion in the tumors may significantly predict the postoperative prognosis of HBV-caused HCC (HBV-HCC) cases [11–14]. However, there is no study reporting whether nucleotide variants in the preS region of HBV genome are prognostic for HCC. In this study, HBV variants in the preS region associated with postoperative prognosis of HBV-HCC were first examined by analyzing RNA-seq datasets of HCC tissues. The variants were independently validated in the sera and paired tumor tissues of HBV-HCC patients who received radical hepatectomy in a prospective cohort.

RESULTS

Screening of prognostic HBV preS variants in RNA-seq data

We combined the RNA-seq datasets of tumor samples of 203 HBV-HCC patients from a total of twelve studies. HBV reads could not be detected in 37 of them, while the median depth of HBV was 477.81 (IQR, 71.05–1639.56) in the remaining 166 samples. The variants at the preS region (nt. 2848 to nt. 154) covered by more than 100 reads were extracted by a procedure taking sequencing error into account. The frequencies of those variants without sufficient coverage were considered not available. To estimate the associations between HBV variants and postoperative prognosis of HCC patients, HCC-specific prognosis-related gene sets were retrieved from MsigDB and compiled into an in-house database, which contained six overall survival (OS)-related and seven recurrence-free survival (RFS)-related gene sets (Supplementary Table 1). After adjusting for any potential batch effect, sample-level enrichment scores were calculated for every gene set and correlation tests were performed. It was found that 12 HBV variants were significantly associated with RFS of HCC patients, while none was found to be associated with OS (Supplementary Table 2).

Validation of prognostic HBV preS variants in a prospective cohort

To validate these variant-prognosis associations predicted by RNA-seq data, 103 HBV-HCC patients were enrolled in our prospective cohort (Table 1). Their sera and tumor tissues were collected and subjected to clone-based Sanger sequencing for HBV preS region. Multiple clones were sequenced for each sample, with a median clone number in the serum sample of 9 (IQR, 7–10) and in the tumor sample of 7 (IQR, 6–9). To inspect the inter-subject contamination, a heat map was plotted and indicated that there was no between-subject contamination (Supplementary Figure 1). The presence/absence of the HBV variants was summarized for each sample. Our survival analysis indicated that G40C and C147T in the tumors were significantly associated with RFS (Supplementary Table 3).

In the correlation test of RNA-seq, the frequencies of G40C and C147T were all associated with a gene set named "KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP" consisting of genes upregulated in HCC with early recurrence (r = 0.29, false discovery rate (FDR) = 0.025 for G40C; r = 0.24, FDR = 0.082 for C147T) (Supplementary Table 2). In the survival analysis of our 103 patients, the presence of these two variants in the tumors also predicted unfavorable RFS (HR = 1.78, 95% CI = 1.04–3.05, p = 0.045 for G40C; HR = 1.74, 95% CI = 1.03–2.95, p = 0.039 for C147T) (Figure 1A and Supplementary Table 3).

Evaluation of the HBV variants as serological biomarkers for HCC prognosis

To evaluate if these HBV variants serve as serological biomarkers predicting prognosis of HCC patients, we performed survival analyses using these variants in the sera. The presence of G40C and C147T in the sera both predicted unfavorable RFS (HR=2.18, 95% CI=1.37–3.47, p=0.001 for G40C; HR=1.84, 95% CI=1.15–2.92, p=0.012 for C147T). Kaplan-Meier curves and log-rank tests also confirmed these results (Figure 1B). These two variants are polymorphic sites between the HBV genomes of HBV genotypes B2 and C2 compared to the HBV reference sequences and were highly linked. The wild types at nt. 40 and nt. 147 are G and C in genotype C2, and C and T in genotype B2, respectively. Besides, G40C does not alter amino acid, while C147T results in amino acid change from alanine to valine. By correlation tests, we found that the two variants tended to occur together, and their correlation coefficient was 0.96 in the sera. If the patients with these two variants in the sera were treated as a group, survival analysis and Kaplan-Meier curve could confirm the prediction power of two individual variants (HR=2.18, 95% CI=1.37–

Table 1. Baseline characteristics of patients enrolled in our cohorts[†].

Variable	Level	Cohort in the study (n = 103)
Age - yr		50.09±8.71
Gender	Male	93 (90.3)
	Female	10 (9.7)
BMI -kg/m^2		23.50±3.75
HBV genotype (sera)	B	9 (8.7)
	C	67 (65.1)
	Mixture	27 (26.2)
Ascites	No	89 (86.4)
	Yes	14 (13.6)
Tumor rupture	No	100 (97.1)
	Yes	3 (2.9)
Portal vein tumor thrombi	No	84 (81.6)
	Yes	19 (18.4)
Tumor number	Single	85 (82.5)
	Multiple	18 (17.5)
Tumor size	<3cm	16 (15.5)
	≥3cm	87 (84.5)
Cirrhosis	No	7 (6.8)
	Mild Cirrhosis	71 (68.9)
	Cirrhosis	25 (24.3)
Tumor capsule	Complete	16 (15.5)
	Incomplete	73 (70.9)
	Absence	14 (13.6)
Microsatellite	No	74 (71.8)
	Yes	29 (28.2)
Microscopic vascular invasion	No	65 (63.1)
	Yes	38 (36.9)
Tumor differentiation	I	14 (13.6)
	II	7 (6.8)
	III	82 (79.6)
BCLC staging	0	0 (0.0)
	A	36 (35.0)
	B	48 (46.6)
	C	19 (18.4)
Postoperative antiviral treatment	No	58 (56.3)
	Yes	45 (43.7)
HBeAg	Negative	83 (70.6)
	Positive	20 (19.4)
HBV DNA - $\log_{10}$ copies/mL		3.94±1.36
Total bilirubin (μmol/L)	≤20	82 (79.6)
	>20	21 (20.4)
Direct bilirubin (μmol/L)	≤7	75 (72.8)
	>7	28 (27.2)
Albumin (g/L)	35-55	93 (90.3)
	<35 OR >55	10 (9.7)
AFP (ng/mL)	≤20	42 (40.8)
	>20	61 (59.2)
ALT (U/L)	≤42	55 (53.4)
	>42	48 (46.6)
AST (U/L)	≤37	48 (46.6)
	>37	55 (53.4)
GGT (U/L)	≤61	50 (48.5)
	>61	53 (51.5)
ALP (U/L)	≤129	84 (81.6)

	>129	19 (18.4)
Follow-up time (month)	Median	30.97
	IQR	11.67–59.66
HCC-related death	No	41 (39.8)
	Yes	62 (60.2)
Recurrence	No	30 (29.1)
	Yes	73 (70.9)

†Plus/minus values are means ± SD; Data are number (%), unless otherwise indicated.

ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; GGT, γ –glutamyltranspeptidase; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.

3.47, p=0.001) (Supplementary Figure 2). Next, we investigated the associations of the combined HBV variant with clinical variables. It was found that the combined variant in the sera was significantly associated with microscopic vascular invasion (p<0.001), larger tumor size (p=0.019), and higher levels of γ-glutamyltranspeptidase (GGT) and alkaline phosphatase (ALP) (p=1.15×10^{-3} for GGT and p<0.001 for ALP) (Table 2). We failed to perform stratified analysis for these two variants in the sera of HBV

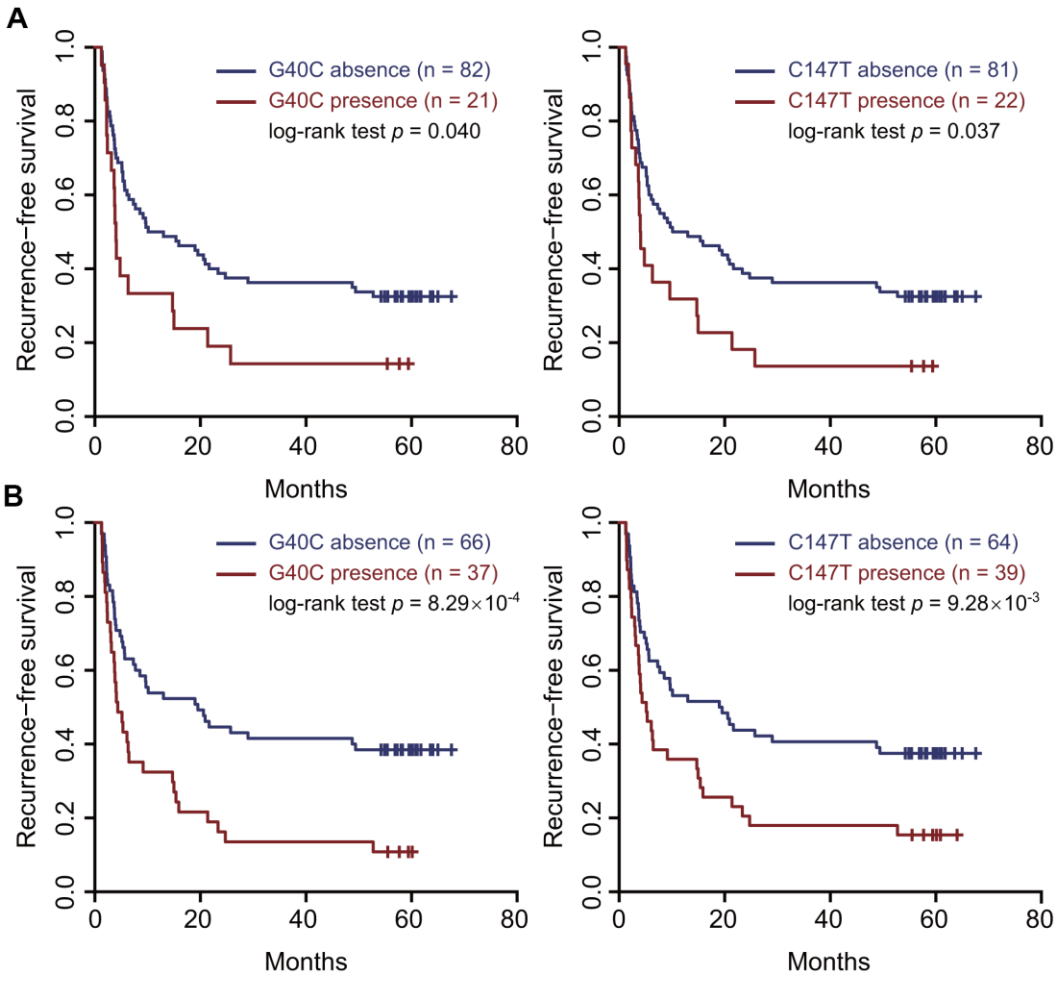


Figure 1. Two HBV preS variants in the tumor tissues and sera predicted an unfavorable recurrence-free survival. (**A**) the tumor tissues. (**B**) the sera. Patients were split into two groups according to the presence (or absence) of the variant. Kaplan–Meier curves were plotted to visualize the difference.

Table 2. Association of the combined HBV variant with clinical variables†.

Variable	Level	The presence of the two HBV variants		
		No	Yes	*P* value
Microscopic vascular invasion	No	50 (48.5)	15 (14.6)	$8.35x10^{-4}$
	Yes	16 (15.5)	22 (21.4)	
Tumor size (cm)		5.89 ± 3.09	7.78 ± 3.88	0.019
GGT (U/L)		86.3 ± 99.0	148.9 ± 122.2	$1.52x10^{-4}$
ALP (U/L)		95.0 ± 48.9	146.9± 184.5	$1.15x10^{-3}$

†Plus/minus values are means ± SD; Data are number (%), unless otherwise indicated. Wilcoxon rank sum test was performed for continuous variables. Chi-square test was applied for count data. The two HBV variants (G40C and C147T) were combined. GGT, γ-glutamyltranspeptidase; ALP, alkaline phosphatase.

genotype C, as these variants were barely present in the samples (1/67 for G40C and 3/67 for C147T). We also scanned all the preS2 deletion sites in our data. It was found that pre-S2 deletion mutants were present in 85 sera and 87 tumor samples, respectively. However, pre-S2 deletion mutants were not associated with prognosis of HCC patients in our data (Supplementary Figure 3).

All virological factors including HBV genotype and HBV variants and clinicopathological factors were subjected to the Cox proportional hazard model analysis to estimate postoperative survival. Significant variables in the univariate Cox regression analysis were included in the multivariate Cox model. The results show that tumor rupture (HR=3.91, 95% CI=1.14–13.35, p=0.03), microscopic vascular invasion (HR=3.03, 95% CI=1.72–5.34, p<0.001), α-fetoprotein (AFP) (HR=1.88, 95% CI=1.04–3.38, p=0.036), and ALP (HR=2.5, 95% CI=1.39–4.49, p=0.002) increased the risk of HCC recurrence, while age (HR=0.97, 95% CI=0.94–0.99, p=0.024) and antiviral treatment (HR=0.15, 95% CI=0.08–0.28, p<0.001) significantly decreased the risk of HCC recurrence (Supplementary Table 4). Because antiviral therapy is a very strong protective factor to prevent the recurrence of HCC, the multivariate Cox proportional hazard models were further stratified by antiviral treatment. The results show that G40C (HR=3.89, 95% CI=1.39–10.87, p=0.01), advanced BCLC staging (HR=4.63, 95% CI=1.53–14.02, p=0.007), and high level of AFP (HR=5.88, 95% CI=1.88–18.39, p=0.002) significantly increased the risk of HCC recurrence in the group with postoperative antiviral treatment; however, G40C was not associated with postoperative recurrence in HCC patients without postoperative antiviral treatment (Supplementary Table 5).

Next, each clinical variable and the two HBV variants were introduced into the Cox proportional hazards model to build a model that could predict postoperative recurrence of HCC. It was found that the model consisting of serum AFP and G40C achieved the best performance (area under curve (AUC) = 0.79, sensitivity= 0.80, and specificity= 0.70). In this model, AFP was further discretized to achieve a better performance (Table 3). If the model was built by dichotomized AFP (≤20 or >20 ng/mL) alone, the prediction power was less optimal (AUC= 0.73, sensitivity=0.75, specificity=0.70) (Figure 2). The statistical test suggested that the AUC of the model containing AFP and G40C was significantly larger than that of the model containing AFP alone (p = 0.029).

Gene dysregulation and biological functions related to the two HBV variants

In the RNA-seq data, the correlation of these two HBV variants in the tumors was 0.99 (Figure 3A), which was quite consistent with the Sanger sequencing data. Furthermore, each of the variants could be identified in almost each read covering this HBV nucleotide in any positive sample. Based on this observation, the tumors in RNA-seq dataset were split into two groups: tumors with high and low frequencies of these variants. In total, 109 tumor samples with sufficient coverage on these variants were retained for differential expression analysis. It was found that 169 genes were differentially expressed between the two groups. Among these genes, 114 were upregulated, while 55 downregulated in the tumors with high frequencies of the HBV variants (Figure 3B and Supplementary Table 6). We observed that the HBV S gene was upregulated in the tumors with high frequency of the HBV variants (fold change = 6.82, FDR = 0.028, Figure 3B, 3C). G40C and C147T were located at the promoter region of the HBV S gene. Therefore, transcription factor binding sites (TFBSs) were predicted for the promoter region (nt. 1 to nt. 154 of the HBV genome). Multiple possible TFBSs were discovered at this region. These data suggest that the variants might be involved in regulating the transcription of HBV S gene (Supplementary Table 7).

Next, gene set enrichment analysis (GSEA) was performed to investigate the gene sets enriched between

Table 3. Risk scores of HCC recurrence based on the presence of G40C and discretized AFP levels†.

AFP (ng/mL)	Discretized AFP	G40C	Score
<20	0	0	0
≥20 and <200	1	0	1.365
≥200 and <400	2	0	2.730
≥400	3	0	4.095
<20	0	1	2.000
≥20 and <200	1	1	3.361
≥200 and <400	2	1	4.727
≥400	3	1	6.092

†AFP, α-fetoprotein; HCC, hepatocellular carcinoma.

these two groups of tumors. In total, 47 positively and 5 negatively enriched gene sets were identified in the data (Supplementary Table 8). In the five negatively enriched gene sets, the top ranked gene set suggested that these HBV variants were associated with proliferation and chromosome instability (CHIANG_LIVER_CANCER_SUBCLASS_PROLIFERATION_DN, normalized enrichment score (NES) = -2.73, familywise-error rate (FWER) < 0.001) (Figure 3D). In addition, its complementary gene set "CHIANG_LIVER_CANCER_SUBCLASS_ PROLIFERATION_UP" was enriched positively in the data (NES = 2.62, FWER < 0.001) (Supplementary Table 8). Another top ranked negatively enriched gene set was "CHIANG_LIVER_CANCER_SUBCLASS_CTNNB1_UP" (NES = -2.66, FWER < 0.001), which confirmed the results of our survival analyses (Figure 3E). Besides these gene sets, other prognosis-predicting gene sets were significantly enriched, including LEE_LIVER_CANCER_SURVIVAL_UP (NES = -2.32, FWER = 0.005) and LEE_LIVER_CANCER_SURVIVAL_DN (NES = 2.39, FWER = 0.001), indicating that the HBV variants predicted unfavorable OS. In addition, SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP (NES = 2.41; FWER < 0.001) and VILLANUEVA_LIVER_CANCER_KRT19_UP (NES = 2.16, FWER = 0.01) were significantly enriched, indicating that the two HBV variants facilitate metastasis. Our results also suggest that these HBV variants are associated with tumorigenesis, anticancer drug resistance, and interferon

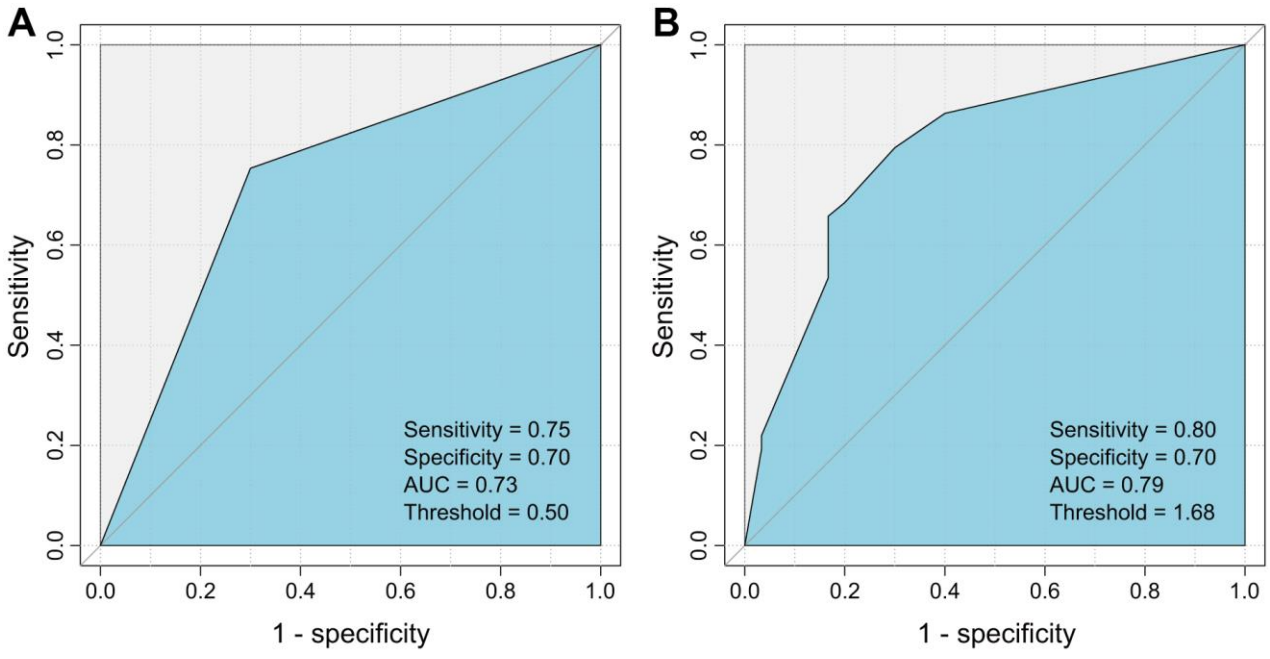


Figure 2. ROC curves for the HCC recurrence prediction models using AFP alone and AFP plus G40C. (**A**) Model using AFP alone. (**B**) Model using AFP and G40C. ROC, receiver operating characteristic; AFP, α-fetoprotein; HCC, hepatocellular carcinoma.

response (BOYAULT_LIVER_CANCER_SUBCLASS_G3_UP, NES = 2.41, FWER = 0.001; KOBAYASHI_EGFR_SIGNALING_24HR_DN, NES = 2.32, FWER = 0.001; and FARMER_BREAST_CANCER_CLUSTER_1, NES = -2.15, FWER = 0.051) (Figure 3F and Supplementary Table 8).

DISCUSSION

In this study, RNA-seq datasets of 203 HCC samples were firstly screened for HCC prognosis-related HBV variants in the preS region. A total of 12 variants related to RFS were initially identified. Of those, G40C and C147T were successfully validated both in the sera and in the tumors of 103 HBV-HCC patients in our prospective cohort. Interestingly, the two variants were actually polymorphic sites between the genomes of HBV genotypes B2 and C2 and highly linked with each other, either in the RNA-seq data of the tumor tissues or in the Sanger sequencing data of both the sera and the tumor tissues. G40C is also a representative of HBV genotype B2 or genotype mixture with genotype B2. Compared to genotype C2, HBV genotype B2 or genotype mixture increases the risk of HCC recurrence, which is concordant with our previous study [15]. The results of multivariate Cox proportional hazard models

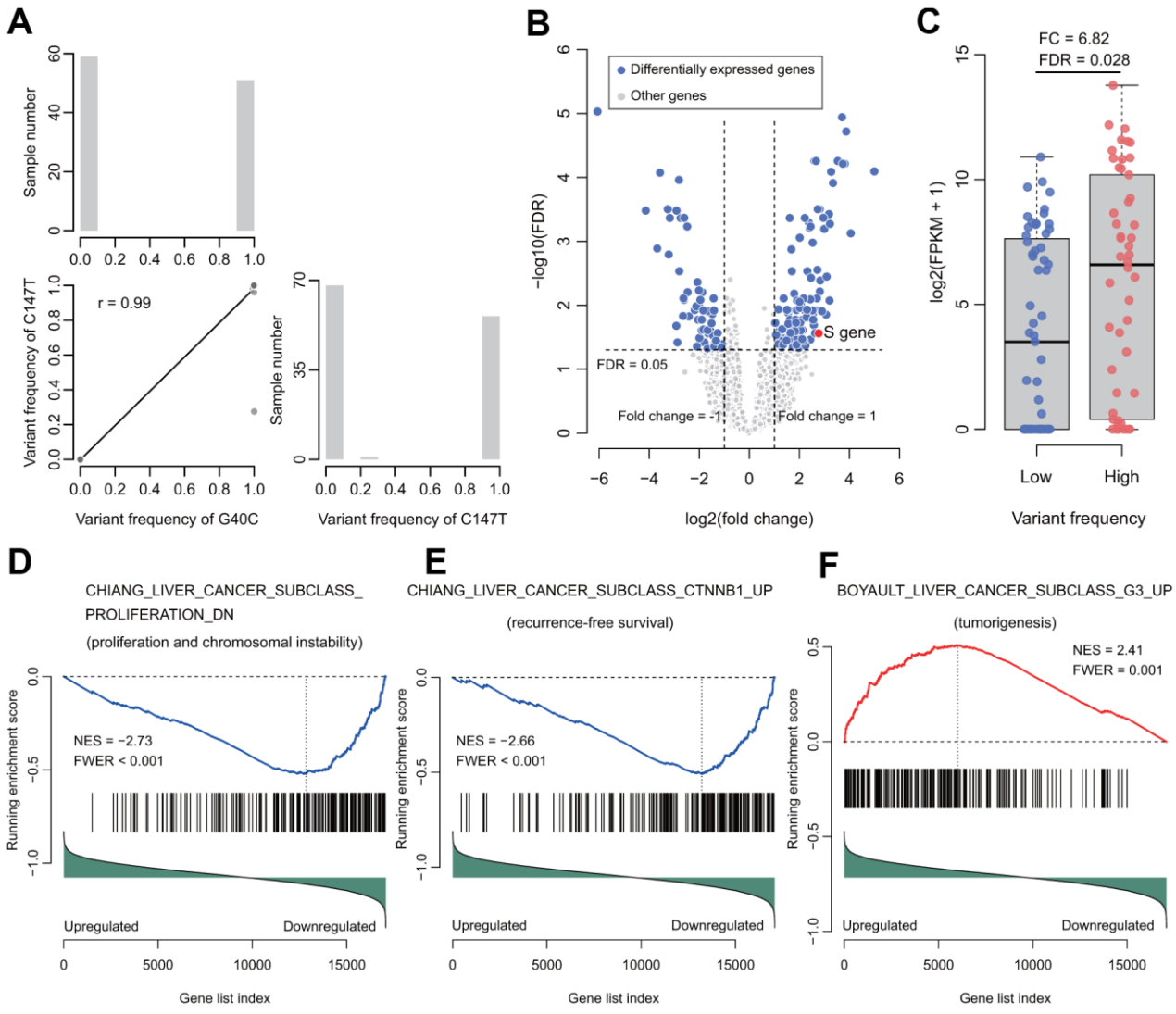


Figure 3. Correlation of the two variants and their functional analysis in RNA-seq data of HBV-HCC tumors. (**A**) The correlation (lower triangle) and frequency distributions (diagonal) of these two variants. The two variants tended to occur simultaneously. (**B**) Volcano plot for differentially expressed genes between groups with high or low frequency of the two variants. (**C**) The HBV S gene was significantly upregulated. (**D**) The gene set represents the gene signature of proliferation and chromosome instability. (**E**) The gene set represents the gene signature of recurrence-free survival. (**F**) The gene set represents the gene signature of tumorigenesis. NES, normalized enrichment score. FWER, familywise-error rate.

show that G40C, advanced BCLC staging, and high level of AFP significantly increased the risk of HCC recurrence in the group with postoperative antiviral treatment, while there was no effect in the group without postoperative antiviral treatment (Supplementary Table 5). Antiviral treatment can decrease the occurrence and recurrence of HCC in high-risk HBV-infected subjects [14, 16, 17]. This result indicates that postoperative antiviral treatment could not decrease the risk of HCC recurrence in patients with HBV genotype B carrying G40C.

Our results show that G40C as a represent of C147T and HBV genotype B2 was a significant serological biomarker for HCC recurrence. AUC of the optimized model consisting of HBV G40C and AFP in the sera was 0.79. The prediction power of this prediction model is better than the one we previously developed using clinical variables composed of the levels of HBV DNA load, the presence of liver cirrhosis, the level of AFP, and BCLC stage [18]. Furthermore, the Cox prediction model based on the AJCC tumor stage and ratios of serum preS2 deletion as deleted by a preS gene chip was also developed for the prediction of postoperative prognosis in HBV-HCC [12, 19]. It has been confirmed that the AUC of this model is 0.741 in the main cohort and 0.704 in the validation cohort [12]. Apparently, our model established in this study should be more powerful than the currently published ones in predicting postoperative prognosis in HBV-HCC. This model is worth translating into clinical practice.

Some studies have confirmed the carcinogenic potential of the preS2 mutated proteins in both transgenic mice and cell culture [7, 20, 21]. The rtM204I/sW196* preS/S truncation induce the cell transformation and tumorigenesis ability via altered host gene expressions, including MGST2, HIF1A, and TGFbi. Downregulated TGFbi may be a common mechanism for oncogenicity in HBV surface truncation mutants [22]. PreS2 deletions modulate cellular processes with a potential impact on liver disease. The accumulation of mutated envelope proteins in the ER leads to ER stress, DNA damage, centrosome overduplication, and genomic instability [23–25]. HBV preS2 interacted with the preS2-responsible region and activated the hTERT promoter, resulting in the upregulation of telomerase activity and the promotion of HCC development [22]. However, the mechanism by which the nucleotide variants in the preS2 of HBV promote the recurrence of HCC remains unknown. G40C and C147T are located in the HBV preS2 region and as well as the promoter region of the HBV S gene. The expression of the HBV S gene was upregulated 6.82 times in the tumors with high frequency of the two variants, compared to those with low frequency of the two variants (Figure 3C). The HBV variants may alter the binding of the TFBSs to the promoter, which regulates the transcription of HBV S gene (Supplementary Table 7). In this study, we also provided evidence showing that the two HBV variants facilitated cell proliferation, chromosome instability, tumorigenesis, metastasis, and anticancer drug resistance. It may explain the reason that HBV Genotype B2 increases the risk of HCC recurrence. Further experimental studies using cell lines and animal models are suggested to validate the cancer promoting function of the HBV with the two variants.

In summary, the present study indicates that HBV preS variants G40C and C147T as representatives of HBV Genotype B2 are highly linked with each other, and may serve as prognostic biomarkers in both sera and tumor tissue samples of HBV-HCC patients. The AUC of the optimized model combining G40C with AFP was 0.79. HBV preS G40C variant and serological AFP are easily examined in HBV-HCC patients and helpful for making therapeutic decision before surgery. Thus, the model is worth translating into clinical practice.

MATERIALS AND METHODS

RNA-seq data analysis to screen for HBV variants related to HCC prognosis

The original RNA-seq data of 203 HCC patients from 12 studies (SRP062885, SRP069212, SRP099053, SRP174991, SRP256409, SRA074279, SRP039694, SRP108560, SRP118972, SRP120360, SRP188371, and SRP220071) were retrieved from the Sequence Read Archive database [26–34]. HBV variants in the region of preS1/preS2 were extracted as previously described and those with average frequencies ≥ 25% and more than 100 valid values were kept for downstream analyses [35]. The read count matrix was obtained by Salmon [36]. During the process, the annotation of human genes was combined with that of HBV ones. Thus, the abundances of HBV genes were evaluated along with the human genes during the quantification process. Combat-Seq method was applied to adjust the potential batch effect among different studies [37]. The gene set names containing “survival” (or “recurrence”) and “liver cancer” were retrieved from gene sets of chemical and genetic perturbations in MsigDB (http://software.broadinstitute.org/gsea/index.jsp) as liver-specific prognosis-related gene sets. Sample-level prognosis scores were calculated by gene set variation analysis in which classical maximum deviation method was performed to compute the enrichment statistics [38]. The parameter “min.sz” was set to 10 during the process. The function “cor.test” in R language was applied to calculate the Pearson correlations between variant frequencies of HBV loci and prognosis scores.

The Benjamini–Hochberg (BH) method was performed among the prognostic gene sets per variant to calculate FDRs. Any association with FDR < 0.1 was kept for downstream analyses. For differential expression gene analysis, edgeR was applied [39, 40]. *P* values were adjusted by BH method. Genes with fold change ≥ 2 and FDR < 0.05 were collected as differentially expressed genes. For GSEA analysis, the gene's read count was converted into Fragments Per Kilobase of exon model per Million mapped fragments (FPKM). Taking the signal-to-noise ratio as input, the "GSEAPreranked" tool in GSEA software was performed to detect the gene sets enriched in the data [41]. Gene sets with FWER ≤ 0.1 were considered as significantly enriched.

Independent validation of the HBV variants in HBV-HCC patients

In total, 103 consecutive HBV-infected HCC patients who received radical hepatectomy from this research group of the Eastern Hepatobiliary Surgery Hospital (Shanghai, China) were enrolled and confirmed by pathology from February 2011 to March 2012. Resected tumors were subjected to pathological examination for tumor-free resection margin > 1 cm without evidence of cancer metastasis. Preoperative peripheral blood samples and tumor tissues of participants were collected and stored at -80° C immediately after surgery. Routine laboratory tests related to liver function were measured using international standard methods and matched reagents (HITACHI 7600, Hitachi Koki Co. Ltd., Hitachinaka City, Japan; Wako Diagnostics Reagents, Wako Pure Chemical Industries Ltd., Osaka, Japan). Alpha-fetoprotein concentrations were routinely measured on the Cobas e601 immunoassay analyzers and matched reagents (Roche Diagnostics, Manheim, Germany) with electrochemiluminescence technology. Participants were surgically treated and followed-up according to the standard protocols as previously described [14]. The follow-up was finished on October 1st, 2019. All participants were self-reported Han Chinese. This study was approved by the ethics committee of Eastern Hepatobiliary Surgery Hospital. All patients provided written informed consent.

HBV DNA of preoperative sera and tumors was extracted using QIAamp DNA blood mini kit (Qiagen, Hilden, Germany). The HBV genome between nt.2743 and nt.255 (from nt.2743 to nt.3215 and from nt.1 to nt.255) was amplified using nested PCR and sequenced using the cloning-based sequencing method as previously described [42]. Ten clones of each sample were randomly selected for Sanger sequencing. Genotyping was performed by HBV subtype analyzer (STAR) as previously described [35]. For a clone, scores were assigned for Genotype A to H. The genotype of a clone was identified as the one with the largest score. Samples with clones of multiple genotypes were defined as mixture. Variants of each clone and preS2 deletion sites were retrieved from BLAST alignments [43]. Clones that failed to align to the HBV genome were excluded from the subsequent analysis. Sample-level variants were then summarized via collecting the variants of all clones in a sample. In simple terms, in any clone of a sample, if a variant was detected at a nucleotide of the HBV genome, then we considered that the variant was present at that locus in that sample. The pairwise distances of the clones from serum samples were calculated by MEGA X and then visualized to inspect identical clones and therefore inter-subject contamination [44].

Statistical analysis

Clinical and baseline characteristics were summarized by using mean values with standard deviation or median values with interquartile range (IQR, 25th to 75th percentiles) for continuous variables. Proportions were applied for categorical values. Univariate Cox regression analysis and log-rank test were applied to estimate the associations between the presence of the viral variants and patients' OS and RFS. Kaplan–Meier method performed survival analysis and generated a survival plot. The selected HBV variants and each clinical variable were subjected to the Cox regression analyses to compute the risk scores and build HCC recurrence prediction models. Given that x_i is the ith variable and β_i is its coefficient, then the risk score of a patient is calculated as:

$$Risk\ Score = \sum_{i=1}^{n} \beta_i \chi_i$$

The number of variables (*i.e.*, *n* in the formula) was set to 2 according to the Harrell's guidelines [45]. The best one was determined by the AUC. The receiver operating characteristic (ROC) curves were plotted by R package pROC [46]. The statistical significance between two AUCs was determined by the one-sided test applied by the function "roc.test" in the pROC package with default parameters. TFBS were predicted by Find Individual Motif Occurrences [47]. The match *p*-value was set to 0.001 to gain more sensitivity. TFBSs with false discovery rate (FDR) of < 0.25 and overlap with G40C or C147T were collected. For clinical variable association test, differences were determined by Wilcoxon rank sum test or χ^2 tests as appropriate. P<0.05 was considered significant. All analyses were two-side and performed using SPSS, version 21 (Armonk, NY).

Ethics committee approval

The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the ethics committee of Eastern Hepatobiliary Surgery Hospital. All patients provided written informed consent.

Availability of data and materials

The sequences of clone sequencing were deposited in GenBank with accession numbers MW179612 - MW180878.

AUTHOR CONTRIBUTIONS

Yin and Cao had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept, design and supervision: Yin and Cao; Acquisition, analysis, or interpretation of data: Chen, Li, Zhang (MZ), Pu, Wu, Ding, Cai, Zhang (HZ), Zhao and Yin; Drafting of the manuscript: Chen, Yin and Cao; Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: Yin, Chen and Pu; Obtained funding: Cao, Yin and Chen; Material support: Li, Zhang (MZ).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This work was funded by National Natural Science Foundation of China (81520108021, 91529305, 81673250, 81521091, 81373067, 81502882), the State Key Infection Disease Project of China (2017ZX10201201-006-001) and the National Key Basic Research Program (973 program) (2015CB554000).

Editorial note

[&]This corresponding author has a verified history of publications using a personal email address for correspondence.

REFERENCES

1. Villanueva A. Hepatocellular Carcinoma. N Engl J Med. 2019; 380:1450–62. https://doi.org/10.1056/NEJMra1713263 PMID:30970190

2. Yin J, Xie J, Liu S, Zhang H, Han L, Lu W, Shen Q, Xu G, Dong H, Shen J, Zhang J, Han J, Wang L, et al. Association between the various mutations in viral core promoter region to different stages of hepatitis B, ranging of asymptomatic carrier state to hepatocellular carcinoma. Am J Gastroenterol. 2011; 106:81–92. https://doi.org/10.1038/ajg.2010.399 PMID:20959817

3. Huang Y, Tai AW, Tong S, Lok AS. HBV core promoter mutations promote cellular proliferation through E2F1-mediated upregulation of S-phase kinase-associated protein 2 transcription. J Hepatol. 2013; 58:1068–73. https://doi.org/10.1016/j.jhep.2013.01.014 PMID:23348237

4. Revill PA, Tu T, Netter HJ, Yuen LK, Locarnini SA, Littlejohn M. The evolution and clinical impact of hepatitis B virus genome diversity. Nat Rev Gastroenterol Hepatol. 2020; 17:618–34. https://doi.org/10.1038/s41575-020-0296-6 PMID:32467580

5. Yin J, Xie J, Zhang H, Shen Q, Han L, Lu W, Han Y, Li C, Ni W, Wang H, Cao G. Significant association of different preS mutations with hepatitis B-related cirrhosis or hepatocellular carcinoma. J Gastroenterol. 2010; 45:1063–71. https://doi.org/10.1007/s00535-010-0253-1 PMID:20419326

6. Zheng Y, Qian YY, Fan H. Pre-S2 and HBV associated hepatocellular carcinoma. Hepatoma Res. 2018; 4:17. https://doi.org/10.20517/2394-5079.2018.08

7. Wang HC, Huang W, Lai MD, Su IJ. Hepatitis B virus pre-S mutants, endoplasmic reticulum stress and hepatocarcinogenesis. Cancer Sci. 2006; 97:683–88. https://doi.org/10.1111/j.1349-7006.2006.00235.x PMID:16863502

8. Ding H, Tu H, Qu C, Cao G, Zhuang H, Zhao P, Xu X, Yang Y, Lu S, and Committee for Prevention and Control of Hepatobiliary and Pancreatic Diseases of Chinese Preventive Medicine Association, and Committee of Hepatology of Chinese Research Hospital Association, and Hepatology Society of Chinese Medical Association, and Prevention of Infection Related Cancer (PIRCA) Group. Guideline for stratified screening and surveillance in patients with high risk of primary liver cancer (2020). Hepatoma Res. 2021; 7:17. https://doi.org/10.20517/2394-5079.2021.13

9. Liu WC, Wu IC, Lee YC, Lin CP, Cheng JH, Lin YJ, Yen CJ, Cheng PN, Li PF, Cheng YT, Cheng PW, Sun KT, Yan SL, et al. Hepatocellular carcinoma-associated single-nucleotide variants and deletions identified by the use of genome-wide high-throughput analysis of hepatitis B virus. J Pathol. 2017; 243:176–92. https://doi.org/10.1002/path.4938

PMID:28696069

10. Zhang AY, Lai CL, Huang FY, Seto WK, Fung J, Wong DK, Yuen MF. Evolutionary Changes of Hepatitis B Virus Pre-S Mutations Prior to Development of Hepatocellular Carcinoma. PLoS One. 2015; 10:e0139478. https://doi.org/10.1371/journal.pone.0139478 PMID:26421619

11. Su CW, Chiou YW, Tsai YH, Teng RD, Chau GY, Lei HJ, Hung HH, Huo TI, Wu JC. The Influence of Hepatitis B Viral Load and Pre-S Deletion Mutations on Post-Operative Recurrence of Hepatocellular Carcinoma and the Tertiary Preventive Effects by Anti-Viral Therapy. PLoS One. 2013; 8:e66457. https://doi.org/10.1371/journal.pone.0066457 PMID:23805222

12. Yen CJ, Ai YL, Tsai HW, Chan SH, Yen CS, Cheng KH, Lee YP, Kao CW, Wang YC, Chen YL, Lin CH, Liu T, Tsai HP, et al. Hepatitis B virus surface gene pre-S_2 mutant as a high-risk serum marker for hepatoma recurrence after curative hepatic resection. Hepatology. 2018; 68: 815–26. https://doi.org/10.1002/hep.29790 PMID:29350774

13. Tsai HW, Lin YJ, Lin PW, Wu HC, Hsu KH, Yen CJ, Chan SH, Huang W, Su IJ. A clustered ground-glass hepatocyte pattern represents a new prognostic marker for the recurrence of hepatocellular carcinoma after surgery. Cancer. 2011; 117:2951–60. https://doi.org/10.1002/cncr.25837 PMID:21692054

14. Yin J, Li N, Han Y, Xue J, Deng Y, Shi J, Guo W, Zhang H, Wang H, Cheng S, Cao G. Effect of antiviral treatment with nucleotide/nucleoside analogs on postoperative prognosis of hepatitis B virus-related hepatocellular carcinoma: a two-stage longitudinal clinical study. J Clin Oncol. 2013; 31:3647–55. https://doi.org/10.1200/JCO.2012.48.5896 PMID:24002499

15. Yin J, Zhang H, Li C, Gao C, He Y, Zhai Y, Zhang P, Xu L, Tan X, Chen J, Cheng S, Schaefer S, Cao G. Role of hepatitis B virus genotype mixture, subgenotypes C2 and B2 on hepatocellular carcinoma: compared with chronic hepatitis B and asymptomatic carrier state in the same area. Carcinogenesis. 2008; 29:1685–91. https://doi.org/10.1093/carcin/bgm301 PMID:18192693

16. Yin J, Wang J, Pu R, Xin H, Li Z, Han X, Ding Y, Du Y, Liu W, Deng Y, Ji X, Wu M, Yu M, et al. Hepatitis B Virus Combo Mutations Improve the Prediction and Active Prophylaxis of Hepatocellular Carcinoma: A Clinic-Based Cohort Study. Cancer Prev Res (Phila). 2015; 8:978–88. https://doi.org/10.1158/1940-6207.CAPR-15-0160 PMID:26290395

17. Lin CL, Kao JH. Prevention of hepatitis B virus-related hepatocellular carcinoma. Hepatoma Res. 2021; 7:9 https://doi.org/10.20517/2394-5079.2020.125

18. Yang F, Ma L, Yang Y, Liu W, Zhao J, Chen X, Wang M, Zhang H, Cheng S, Shen F, Wang H, Zhou W, Cao G. Contribution of Hepatitis B Virus Infection to the Aggressiveness of Primary Liver Cancer: A Clinical Epidemiological Study in Eastern China. Front Oncol. 2019; 9:370. https://doi.org/10.3389/fonc.2019.00370 PMID:31179237

19. Shen FC, Su IJ, Wu HC, Hsieh YH, Yao WJ, Young KC, Chang TC, Hsieh HC, Tsai HN, Huang W. A pre-S gene chip to detect pre-S deletions in hepatitis B virus large surface antigen as a predictive marker for hepatoma risk in chronic hepatitis B virus carriers. J Biomed Sci. 2009; 16:84. https://doi.org/10.1186/1423-0127-16-84 PMID:19751529

20. Luan F, Liu H, Gao L, Liu J, Sun Z, Ju Y, Hou N, Guo C, Liang X, Zhang L, Sun W, Ma C. Hepatitis B virus protein preS2 potentially promotes HCC development via its transcriptional activation of hTERT. Gut. 2009; 58:1528–37. https://doi.org/10.1136/gut.2008.174029 PMID:19651630

21. Yang JC, Teng CF, Wu HC, Tsai HW, Chuang HC, Tsai TF, Hsu YH, Huang W, Wu LW, Su IJ. Enhanced expression of vascular endothelial growth factor-A in ground glass hepatocytes and its implication in hepatitis B virus hepatocarcinogenesis. Hepatology. 2009; 49:1962–71. https://doi.org/10.1002/hep.22889 PMID:19475690

22. Lai MW, Liang KH, Yeh CT. Hepatitis B Virus preS/S Truncation Mutant rtM204I/sW196* Increases Carcinogenesis through Deregulated HIF1A, MGST2, and TGFbi. Int J Mol Sci. 2020; 21:6366. https://doi.org/10.3390/ijms21176366 PMID:32887289

23. Hsieh YH, Su IJ, Wang HC, Chang WW, Lei HY, Lai MD, Chang WT, Huang W. Pre-S mutant surface antigens in chronic hepatitis B virus infection induce oxidative stress and DNA damage. Carcinogenesis. 2004; 25:2023–32. https://doi.org/10.1093/carcin/bgh207 PMID:15180947

24. Wang LH, Huang W, Lai MD, Su IJ. Aberrant cyclin A expression and centrosome overduplication induced by hepatitis B virus pre-S2 mutants and its implication in hepatocarcinogenesis. Carcinogenesis. 2012; 33:466–72.

https://doi.org/10.1093/carcin/bgr296
PMID:22159224

25. Hsieh YH, Chang YY, Su IJ, Yen CJ, Liu YR, Liu RJ, Hsieh WC, Tsai HW, Wang LH, Huang W. Hepatitis B virus pre-S2 mutant large surface protein inhibits DNA double-strand break repair and leads to genome instability in hepatocarcinogenesis. J Pathol. 2015; 236:337–47. https://doi.org/10.1002/path.4531 PMID:25775999

26. Chiu YT, Wong JK, Choi SW, Sze KM, Ho DW, Chan LK, Lee JM, Man K, Cherny S, Yang WL, Wong CM, Sham PC, Ng IO. Novel pre-mRNA splicing of intronically integrated HBV generates oncogenic chimera in hepatocellular carcinoma. J Hepatol. 2016; 64: 1256–64. https://doi.org/10.1016/j.jhep.2016.02.005 PMID:26867494

27. Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, Lv G, Wang S, Wu Y, Yang YT, Wang D, Liu Y, Tang J, et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. Nat Commun. 2017; 8:14421. https://doi.org/10.1038/ncomms14421 PMID:28194035

28. Yoo S, Wang W, Wang Q, Fiel MI, Lee E, Hiotis SP, Zhu J. A pilot systematic genomic comparison of recurrence risks of hepatitis B virus-associated hepatocellular carcinoma with low- and high-degree liver fibrosis. BMC Med. 2017; 15:214. https://doi.org/10.1186/s12916-017-0973-7 PMID:29212479

29. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, Xing B, Sun W, Ren L, Hu B, Li C, Zhang L, Qin G, et al, and Chinese Human Proteome Project (CNHPP) Consortium. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature. 2019; 567:257–61. https://doi.org/10.1038/s41586-019-0987-8 PMID:30814741

30. Kang L, Liu X, Gong Z, Zheng H, Wang J, Li Y, Yang H, Hardwick J, Dai H, Poon RT, Lee NP, Mao M, Peng Z, Chen R. Genome-wide identification of RNA editing in hepatocellular carcinoma. Genomics. 2015; 105:76–82. https://doi.org/10.1016/j.ygeno.2014.11.005 PMID:25462863

31. Gao F, Liang H, Lu H, Wang J, Xia M, Yuan Z, Yao Y, Wang T, Tan X, Laurence A, Xu H, Yu J, Xiao W, et al. Global analysis of DNA methylation in hepatocellular carcinoma by a liquid hybridization capture-based bisulfite sequencing approach. Clin Epigenetics. 2015; 7:86. https://doi.org/10.1186/s13148-015-0121-1 PMID:26300991

32. Jin Y, Lee WY, Toh ST, Tennakoon C, Toh HC, Chow PK, Chung AY, Chong SS, Ooi LL, Sung WK, Lee CG. Comprehensive analysis of transcriptome profiles in hepatocellular carcinoma. J Transl Med. 2019; 17:273. https://doi.org/10.1186/s12967-019-2025-x PMID:31429776

33. Sheng Z, Wang X, Xu G, Shan G, Chen L. Analyses of a Panel of Transcripts Identified From a Small Sample Size and Construction of RNA Networks in Hepatocellular Carcinoma. Front Genet. 2019; 10:431. https://doi.org/10.3389/fgene.2019.00431 PMID:31156698

34. Shen YC, Hsu CL, Jeng YM, Ho MC, Ho CM, Yeh CP, Yeh CY, Hsu MC, Hu RH, Cheng AL. Reliability of a single-region sample to evaluate tumor immune microenvironment in hepatocellular carcinoma. J Hepatol. 2020; 72:489–97. https://doi.org/10.1016/j.jhep.2019.09.032 PMID:31634533

35. Yin J, Chen X, Li N, Han X, Liu W, Pu R, Wu T, Ding Y, Zhang H, Zhao J, Han X, Wang H, Cheng S, Cao G. Compartmentalized evolution of hepatitis B virus contributes differently to the prognosis of hepatocellular carcinoma. Carcinogenesis. 2021; 42:461–70. https://doi.org/10.1093/carcin/bgaa127 PMID:33247709

36. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017; 14:417–19. https://doi.org/10.1038/nmeth.4197 PMID:28263959

37. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020; 2:lqaa078. https://doi.org/10.1093/nargab/lqaa078 PMID:33015620

38. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013; 14:7. https://doi.org/10.1186/1471-2105-14-7 PMID:23323831

39. Lun AT, Chen Y, Smyth GK. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. Methods Mol Biol. 2016; 1418:391–416. https://doi.org/10.1007/978-1-4939-3578-9_19 PMID:27008025

40. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40:4288–97. https://doi.org/10.1093/nar/gks042 PMID:22287627

41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–50. https://doi.org/10.1073/pnas.0506580102 PMID:16199517

42. Li Z, Xie Z, Ni H, Zhang Q, Lu W, Yin J, Liu W, Ding Y, Zhao Y, Zhu Y, Pu R, Zhang H, Dong H, et al. Mother-to-child transmission of hepatitis B virus: evolution of hepatocellular carcinoma-related viral mutations in the post-immunization era. J Clin Virol. 2014; 61:47–54. https://doi.org/10.1016/j.jcv.2014.06.010 PMID:24973814

43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID:2231712

44. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018; 35: 1547–49. https://doi.org/10.1093/molbev/msy096 PMID:29722887

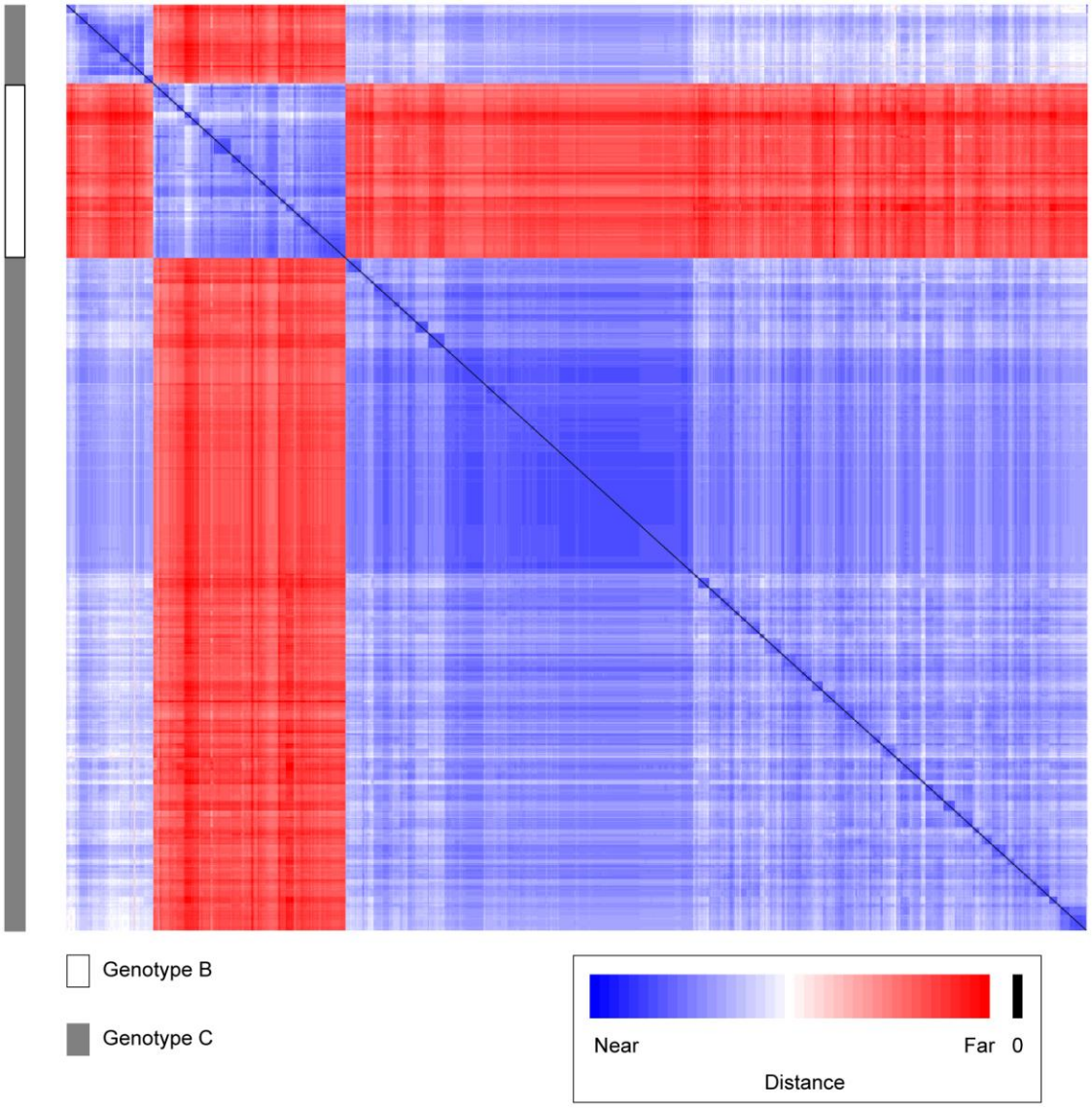
45. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008; 26:1364–70. https://doi.org/10.1200/JCO.2007.12.9791 PMID:18323559

46. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. https://doi.org/10.1186/1471-2105-12-77 PMID:21414208

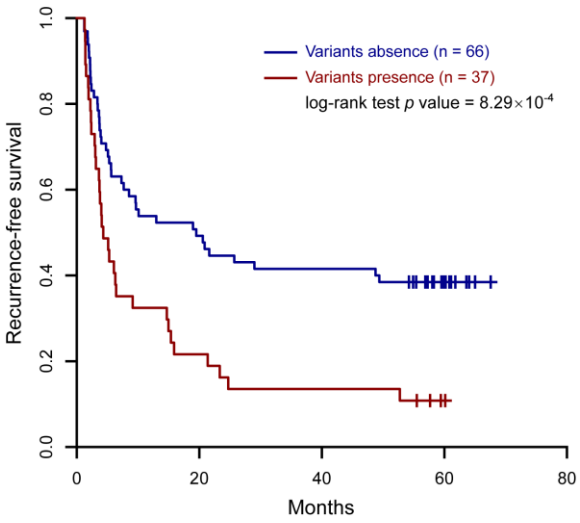
47. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–18. https://doi.org/10.1093/bioinformatics/btr064 PMID:21330290

SUPPLEMENTARY MATERIALS

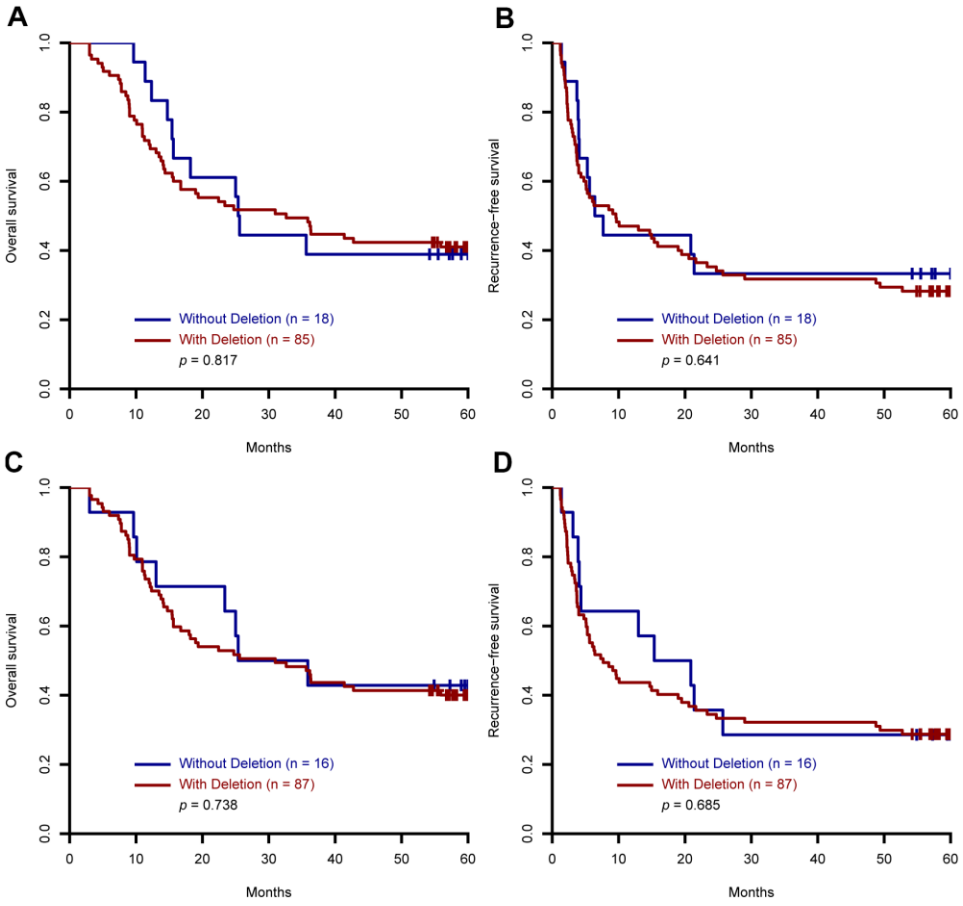
Supplementary Figures



Supplementary Figure 1. Heat map of pairwise distances of the clones from the serum samples. Every row or column represents a clone. The grids on the diagonal line represent the distances between a clone and itself, which are all 0 (marked by black).



Supplementary Figure 2. Combined variants predicted unfavorable recurrence-free survival. G40C and C147T were combined because of their high frequency of concurrence. Kaplan–Meier curve was plotted to visualize the prognosis difference.



Supplementary Figure 3. The prognostic value of the preS2-deletion mutations in the HCC patients from our cohort. (**A**) the serum samples, overall survival; (**B**) the serum samples, recurrence-free survival; (**C**) tumoral samples, overall survival; (**D**) tumoral samples, recurrence-free survival.

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Table 6.

Supplementary Table 1. Liver cancer-specific and prognosis-related gene sets.

Gene set	Pubmed ID
Overall survival-related gene sets	
LEE_LIVER_CANCER_SURVIVAL_DN	15349906
LEE_LIVER_CANCER_SURVIVAL_UP	15349906
HOSHIDA_LIVER_CANCER_SURVIVAL_UP	18701503
HOSHIDA_LIVER_CANCER_SURVIVAL_DN	18701503
KIM_LIVER_CANCER_POOR_SURVIVAL_UP	21320499
KIM_LIVER_CANCER_POOR_SURVIVAL_DN	21320499
Recurrence-free survival-related gene sets	
IIZUKA_LIVER_CANCER_EARLY_RECURRENCE	12648972
HOSHIDA_LIVER_CANCER_LATE_RECURRENCE_UP	18923165
HOSHIDA_LIVER_CANCER_LATE_RECURRENCE_DN	18923165
WOO_LIVER_CANCER_RECURRENCE_UP	18381945
WOO_LIVER_CANCER_RECURRENCE_DN	18381945
KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	15288478
KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_DN	15288478

Supplementary Table 2. Association between HBV variants frequencies and HCC recurrence-related gene sets.

Variant	Gene set	Pearson correlation	*P* value	FDR
G40C	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.29	2.11E-03	2.54E-02
G45C	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.28	2.19E-03	2.63E-02
A85G	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.27	2.21E-03	2.65E-02
A87G	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.25	5.39E-03	6.47E-02
T93C	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.26	4.51E-03	5.42E-02
C96A	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.25	4.77E-03	5.72E-02
C99A	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.26	4.13E-03	4.96E-02
C105T	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.26	3.91E-03	4.69E-02
C110G	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.25	6.11E-03	7.33E-02
C127A	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.26	3.69E-03	4.43E-02
G132A	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.28	1.82E-03	2.19E-02
C147T	KUROKAWA_LIVER_CANCER_EARLY_RECURRENCE_UP	0.24	6.87E-03	8.24E-02

Abbreviations: FDR, false discovery rate.

Supplementary Table 3. Recurrence-related HBV variants validated by sanger sequencing in the tumors.

Variant	Hazard ratio	95% confidence interval	*P* value
G40C	1.78	1.04–3.05	4.49E-02
G45C	1.63	0.95–2.79	8.63E-02
A85G	1.74	1.02–2.98	5.30E-02
A87G	1.72	1.01–2.91	5.45E-02
T93C	1.54	0.90–2.62	1.29E-01
C96A	1.74	1.02–2.98	5.30E-02
C99A	1.60	0.95–2.71	9.00E-02
C105T	0.94	0.55–1.61	8.31E-01
C110G	1.69	0.98–2.92	7.23E-02
C127A	1.43	0.86–2.40	1.83E-01
G132A	1.59	0.95–2.67	8.94E-02
C147T	1.74	1.03–2.95	4.84E-02

Supplementary Table 4. Cox regression analysis for the factors significantly affected the recurrence of postoperative HCC patients.

Variables		No. (%) of patients (n=103)	Univariate analysis		Multivariate analysis	
			HR (95% CI)	*P* value	HR (95% CI)	*P* value
Age (years)		103 (100)	0.96 (0.94–0.99)	0.008	0.97 (0.94–0.99)	0.024
HBV genotype	Genotype C	67 (65.05)	1			
	Not Genotype C	9 (8.74)	1.97 (1.23–3.15)	0.005		
G40C	G	66 (64.1)	1			
	C	37 (35.9)	2.17 (1.36–3.47)	0.001		
C147T	C	64 (62.1)	1			
	T	39 (37.9)	1.83 (1.15–2.92)	0.01		
Tumor rupture	No	100 (97.1)	1		1	
	Yes	3 (2.9)	4.57 (1.40–14.91)	0.012	3.91 (1.14–13.35)	0.03
Portal vein tumor thrombi	No	84 (81.6)	1			
	Yes	19 (18.4)	4.24 (2.44–7.36)	<0.001		
Tumor size	<3cm	16 (15.5)	1			
	≥3cm	87 (84.5)	2.52 (1.16–5.52)	0.02		
Tumor capsule	Complete	16 (15.5)	1			
	Incomplete/Absence	87 (84.5)	2.55 (1.17–5.57)	0.019		
Microscopic vascular invasion	No	65 (63.1)	1		1	
	Yes	38 (36.9)	5.26 (3.22–8.59)	<0.001	3.03 (1.72–5.34)	<0.001
Tumor differentiation	I/II	21 (20.4)	1			
	III	82 (79.6)	2.82 (1.40–5.69)	0.004		
BCLC staging	0/A	36 (35.0)	1			
	B/C	67 (65.0)	3.00 (1.74–5.20)	<0.001		
Postoperative antiviral treatment	No	58 (56.3)	1		1	
	Yes	45 (43.7)	0.17 (0.10–0.29)	<0.001	0.15 (0.08–0.28)	<0.001
AFP	≤20 ng/mL	42 (40.8)	1		1	
	>20 ng/mL	61 (59.2)	2.78 (1.62–4.76)	<0.001	1.88 (1.04–3.38)	0.036
AST (U/L)	≤37	48 (46.6)	1			
	>37	55 (53.4)	1.71 (1.07–2.74)	0.024		
GGT (U/L)	≤61	50 (48.5)	1			
	>61	53 (51.5)	1.99 (1.25–3.18)	0.004		
ALP (U/L)	≤129	84 (81.6)	1		1	
	>129	19 (18.4)	2.05 (1.19–3.54)	0.01	2.50 (1.39–4.49))	0.002

Abbreviations: CI, confidence interval; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; HR, hazard ratio; AST, aspartate aminotransferase; GGT, γ –glutamyltranspeptidase; AFP, α-fetoprotein; ALP, alkaline phosphatase.

Supplementary Table 5. Cox regression analysis for the factors significantly affected the recurrence of postoperative HCC patients stratified by antiviral treatment.

Variables		Multivariate analysis	
		HR (95% CI)	*P* value
With postoperative antiviral treatment (n=45)			
G40C	G	1	
	C	3.89 (1.39–10.87)	0.01
BCLC staging	0/A	1	
	B/C	4.63 (1.53–14.02)	0.007
AFP (ng/mL)	≤20	1	
	>20	5.88 (1.88–18.39)	0.002
Without postoperative antiviral treatment (n=58)			
Age (years)		0.96 (0.93–0.99)	0.024
Tumor rupture	No	1	
	Yes	5.07 (1.43–17.99)	0.012
Microscopic vascular invasion	No	1	
	Yes	4.00 (2.07–7.72)	<0.001
ALP (U/L)	≤129	1	
	>129	2.85 (1.43–5.65)	0.003

Abbreviations: CI, confidence interval; HCC, hepatocellular carcinoma; HR, hazard ratio; AFP, α-fetoprotein; ALP, alkaline phosphatase.

Supplementary Table 6. Differentially expressed genes in the tumors with high/low frequencies of two HBV variants.

Supplementary Table 7. Transcription factor binding sites predicted in the region of preS2.

Motif ID	Transcription factor	Start	Stop	Strand	Score	*P* value	FDR
MA0484.1	HNF4G	138	152	-	8.38	2.25e-4	0.06
MA0505.1	Nr5a2	29	43	+	4.90	4.93e-4	0.07
MA0505.1	Nr5a2	40	54	-	3.34	7.89e-4	0.07
MA0505.1	Nr5a2	35	49	+	3.21	8.21e-4	0.07
MA1101.1	BACH2	30	43	-	6.69	3.69e-4	0.10
MA0114.3	Hnf4a	136	151	-	-4.58	4.23e-4	0.11
MA0501.1	MAF::NFE2	28	42	-	3.16	4.18e-4	0.12
MA1147.1	NR4A2::RXRA	35	49	-	5.77	5.69e-4	0.15
MA0150.2	Nfe2l2	32	46	-	3.61	7.08e-4	0.20
MA0728.1	Nr2f6(var.2)	27	41	+	-18.34	8.13e-4	0.23
MA0138.2	REST	135	155	+	-3.04	9.58e-4	0.23

Abbreviation: FDR, false discovery rate.

Supplementary Table 8. Gene sets enriched in the tumors with high/low frequencies of two HBV variants.

Gene set	Size	ES	NES	P value	FWER
Gene sets enriched in the tumors with low frequencies of two HBV variants					
CHIANG_LIVER_CANCER_SUBCLASS_PROLIFERATION_DN	178	-0.52	-2.73	<1E-3	<1E-3
CHIANG_LIVER_CANCER_SUBCLASS_CTNNB1_UP	170	-0.51	-2.66	<1E-3	<1E-3
CAIRO_HEPATOBLASTOMA_CLASSES_DN	213	-0.44	-2.35	<1E-3	0.003
LEE_LIVER_CANCER_SURVIVAL_UP	177	-0.44	-2.32	<1E-3	0.005
FARMER_BREAST_CANCER_CLUSTER_1	47	-0.52	-2.15	<1E-3	0.051
Gene sets enriched in the tumors with high frequencies of two HBV variants					
NIKOLSKY_BREAST_CANCER_8Q23_Q24_AMPLICON	145	0.62	2.86	<1E-3	<1E-3
CHIANG_LIVER_CANCER_SUBCLASS_PROLIFERATION_UP	174	0.56	2.62	<1E-3	<1E-3
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	152	0.52	2.41	<1E-3	<1E-3
BOYAULT_LIVER_CANCER_SUBCLASS_G3_UP	190	0.51	2.41	<1E-3	0.001
LEE_LIVER_CANCER_SURVIVAL_DN	174	0.51	2.39	<1E-3	0.001
KAMMINGA_EZH2_TARGETS	40	0.65	2.39	<1E-3	0.001
NIKOLSKY_BREAST_CANCER_8Q12_Q22_AMPLICON	122	0.54	2.38	<1E-3	0.001
NIKOLSKY_BREAST_CANCER_17Q11_Q21_AMPLICON	105	0.54	2.32	<1E-3	0.001
KOBAYASHI_EGFR_SIGNALING_24HR_DN	251	0.47	2.32	<1E-3	0.001
BURTON_ADIPOGENESIS_3	93	0.54	2.32	<1E-3	0.001
MITSIADES_RESPONSE_TO_APLIDIN_DN	247	0.47	2.28	<1E-3	0.001
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	140	0.50	2.27	<1E-3	0.001
SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6	445	0.44	2.25	<1E-3	0.002
WONG_EMBRYONIC_STEM_CELL_CORE	332	0.45	2.24	<1E-3	0.002
BORCZUK_MALIGNANT_MESOTHELIOMA_UP	308	0.44	2.20	<1E-3	0.006
AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_UP	294	0.45	2.19	<1E-3	0.007
VILLANUEVA_LIVER_CANCER_KRT19_UP	167	0.46	2.16	<1E-3	0.01
CHIN_BREAST_CANCER_COPY_NUMBER_UP	23	0.68	2.16	<1E-3	0.01
CHIANG_LIVER_CANCER_SUBCLASS_UNANNOTATED_DN	192	0.46	2.15	<1E-3	0.011
RHODES_CANCER_META_SIGNATURE	65	0.54	2.15	<1E-3	0.011
ABRAMSON_INTERACT_WITH_AIRE	43	0.57	2.14	<1E-3	0.014
DAVICIONI_MOLECULAR_ARMS_VS_ERMS_DN	174	0.45	2.11	<1E-3	0.022
LI_WILMS_TUMOR_VS_FETAL_KIDNEY_1_DN	164	0.46	2.10	<1E-3	0.023
LE_EGR2_TARGETS_UP	107	0.48	2.09	<1E-3	0.029
REN_BOUND_BY_E2F	61	0.52	2.07	<1E-3	0.042
WOO_LIVER_CANCER_RECURRENCE_UP	103	0.48	2.07	<1E-3	0.043
PAL_PRMT5_TARGETS_UP	200	0.43	2.07	<1E-3	0.043
WAMUNYOKOLI_OVARIAN_CANCER_GRADES_1_2_UP	139	0.46	2.07	<1E-3	0.044
HEIDENBLAD_AMPLIFIED_IN_PANCREATIC_CANCER	55	0.53	2.07	<1E-3	0.047
BIDUS_METASTASIS_UP	210	0.43	2.06	<1E-3	0.051
JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_DN	56	0.53	2.06	<1E-3	0.052
WHITEFORD_PEDIATRIC_CANCER_MARKERS	113	0.47	2.06	<1E-3	0.052
JIANG_AGING_CEREBRAL_CORTEX_DN	45	0.55	2.06	1.55E-03	0.054
MEINHOLD_OVARIAN_CANCER_LOW_GRADE_DN	19	0.67	2.05	<1E-3	0.058
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP	66	0.51	2.05	<1E-3	0.059
GRAHAM_NORMAL_QUIESCENT_VS_NORMAL_DIVIDING_DN	87	0.49	2.05	<1E-3	0.061
NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPLICON	298	0.41	2.04	<1E-3	0.069
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_14	143	0.45	2.04	<1E-3	0.072
MARKEY_RB1_ACUTE_LOF_UP	229	0.42	2.04	<1E-3	0.074
DING_LUNG_CANCER_EXPRESSION_BY_COPY_NUMBER	100	0.47	2.03	<1E-3	0.076
CROONQUIST_IL6_DEPRIVATION_DN	95	0.47	2.03	<1E-3	0.076
HOSHIDA_LIVER_CANCER_SUBCLASS_S2	114	0.45	2.03	<1E-3	0.082
RHODES_UNDIFFERENTIATED_CANCER	69	0.50	2.03	<1E-3	0.085
VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_UP	169	0.43	2.02	<1E-3	0.088
BLUM_RESPONSE_TO_SALIRASIB_DN	336	0.40	2.02	<1E-3	0.091
FUJII_YBX1_TARGETS_DN	194	0.43	2.02	<1E-3	0.091
MARKEY_RB1_CHRONIC_LOF_UP	107	0.47	2.02	<1E-3	0.097

Abbreviation: ES, enrichment score; NES, normalized ES; FWER, familywise-error rate.