HCC-Associated Single Nucleotide Variants and Deletions Identified Using Genome-Wide High Throughput Analysis of HBV

Wen-Chun Liu\textsuperscript{1,2}, I-Chin Wu\textsuperscript{1,2}, Yen-Chien Lee\textsuperscript{3,4}, Chih-Peng Lin\textsuperscript{5}, Ji-Hong Cheng\textsuperscript{6}, Yih-Jyh Lin\textsuperscript{7}, Chia-Jui Yen\textsuperscript{1,2}, Pin-Nan Cheng\textsuperscript{1,2}, Pei-Fu Li\textsuperscript{8}, Yi-Ting Cheng\textsuperscript{8}, Pei-Wen Cheng\textsuperscript{9}, Koun-Tem Sun\textsuperscript{9}, Shu-Ling Yan\textsuperscript{1}, Jia-Jhen Lin\textsuperscript{1}, Jui-Chu Yang\textsuperscript{10}, Kung-Chao Chang\textsuperscript{10}, Cheng-Hsun Ho\textsuperscript{1,2}, Vincent S. Tseng\textsuperscript{11}, Bill Chia-Han Chang\textsuperscript{5}, Jaw-Ching Wu\textsuperscript{12,13}, Ting-Tsung Chang\textsuperscript{1,2}

\textsuperscript{1} Department of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan, ROC

\textsuperscript{2} Infectious Disease and Signaling Research Center, National Cheng Kung University, Tainan, Taiwan, ROC

\textsuperscript{3} Department of Oncology, Tainan Hospital, Ministry of Health and Welfare, Tainan, Taiwan, ROC

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/path.4938

This article is protected by copyright. All rights reserved.
4 Institute of Clinical Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan, ROC

5 Yourgene Bioscience, Taipei, Taiwan, ROC

6 Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC

7 Department of Surgery, National Cheng Kung University College of Medicine and Hospital, Tainan, Taiwan, ROC

8 Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan, ROC

9 Department of Information and Learning Technology, Science and Engineering College, National University of Tainan, Tainan, Taiwan, ROC

10 Human Biobank, Research Center of Clinical Medicine, National Cheng Kung University Hospital, Tainan, Taiwan, ROC

11 Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, ROC

12 Institute of Clinical Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan, ROC

This article is protected by copyright. All rights reserved.
Correspondence to: Ting-Tsung Chang, Department of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, 138 Sheng-Li Road, Tainan 70403, Taiwan, ROC. E-mail: ttchang@mail.ncku.edu.tw and Jaw-Ching Wu, Translational Research Division, Medical Research Department, Taipei Veterans General Hospital and Institute of Clinical Medicine, National Yang-Ming University 201 Shih-Pai Road, Sec. 2, Taipei 112, Taiwan. E-mail: jcwu@vghtpe.gov.tw

Running title: HBV Variants and Deletions Associated with HCC

Keywords: Deletion index; Hepatocarcinogenesis; Meta-analysis; Next-generation sequencing; U-shaped distribution

Conflict of interest

The authors declare that there are no conflicts of interest.
Abstract

This study investigated hepatitis B virus (HBV) single nucleotide variants (SNVs) and deletion mutations linked with hepatocellular carcinoma (HCC). Ninety-three HCC patients and 108 non-HCC patients were enrolled for HBV genome-wide next-generation sequencing (NGS) analysis. Systemic literature review and meta-analysis were performed to validate NGS-defined HCC-associated SNVs and deletions. The experimental results identified 60 NGS-defined HCC-associated SNVs, including 41 novel SNVs, and their pathogenic frequencies. Each SNV was specific for either genotype B (n=24) or genotype C (n=34), except nt53C for both genotypes. The pathogenic frequencies of these HCC-associated SNVs exhibited a distinct U-shaped distribution pattern. According to meta-analysis and literature review, 167 HBV variants from 109 publications were categorized into 4 levels (A–D) of supporting evidence associated with HCC. The proportion of NGS-defined HCC-associated SNVs among these HBV variants declined significantly from 75% of 12 HCC-associated variants by meta-analysis (Level A) to 0% of 10 HCC-unassociated variants by meta-analysis (Level D) (P<0.0001). PreS deletions were significantly associated with HCC, in terms of deletion index, for both
genotypes B ($P=0.030$) and C ($P=0.049$). For genotype C, preS deletions involving a specific fragment (nt2977–3013) were significantly associated with HCC (HCC vs non-HCC, 6/34 vs 0/32, $P=0.025$). Meta-analysis of preS deletions showed significant association with HCC (summary odds ratio: 3.0, 95% confidence interval, 2.3 to 3.9).

Through transfecting Huh7 cells, all the five novel NGS-defined HCC-associated SNVs in the small surface region influenced hepatocarcinogenesis pathways, including ER-stress and DNA repair systems, by microarray, real-time PCR and western blot analysis. Their carcinogenic mechanisms are worthy of further research.
Introduction

Patients with chronic hepatitis B virus (HBV) are at risk of developing liver cirrhosis and hepatocellular carcinoma (HCC) [1-3]. The HBV genome presents a highly compact organization with four overlapping open reading frames (ORFs) and regulatory elements [4]. Ten HBV genotypes (A–J) have been identified, with genotypes B and C occurring mainly in Asia, while A and D occur mainly in Africa and Europe [5].

Owing to the lack of proofreading activity of HBV reverse transcriptase, the composition of the viral quasispecies evolves over time depending on selective pressure, including host immune response. The characteristics of viral quasispecies have been implicated in the exacerbation of chronic hepatitis B (CHB) and the development of liver cancer [6]. Although numerous HBV variants have been studied as to whether they are genuine risk factors for HCC development, uncertain or controversial results remain [7,8].

Next-generation sequencing (NGS) is a high-throughput analysis for characterizing genetic diversity in viral strains and can characterize minor strains among viral quasispecies in a way that is not possible by direct sequencing or cloning. This study
seeks HCC-associated HBV single nucleotide variants (SNVs) with their pathogenic frequencies using NGS analysis of whole HBV genomes, and compares the findings with systemic literature review and meta-analysis.

**Materials and Methods (see supplementary file for details)**

**Patient enrollment**

Patients were retrospectively selected from databases of 475 chronic HBV infected patients with HCC and 1036 chronic HBV infected patients without HCC in National Cheng Kung University Hospital and Taipei Veterans General Hospital. These patients were enrolled with negativity of anti-hepatitis C virus antibodies, anti-hepatitis D antibodies, and anti-human immunodeficiency virus antibodies. All serum samples analyzed were obtained before any antiviral treatment and had HBV viral loads more than 10,000 IU/mL, successful HBV genome PCR amplification and successful NGS. For HCC patients, the serum samples were collected when HCC was diagnosed for the first time. For genotypes B and C, patients in the HCC and non-HCC groups were matched for age and sex. Finally, 93 HCC patients and 108 non-HCC patients were enrolled in the NGS analysis.
Next-generation sequencing

The HBV full genome was amplified by PCR with nine primer sets (Table S1). NGS was performed using the Illumina HiSeq™ 2500. The sample-specific reference sequence was utilized as a mapping reference sequence for NGS datasets [9-12]. The HBV genotypes were determined by real-time PCR [13] and phylogenetic sequence analysis [14]. PreS deletion mutants were classified into types I-V. PreS deletion mutants involving functional sites, S, T, N, M and P were analyzed [15].

Systemic literature review and meta-analysis

The MEDLINE (1966 to March 10, 2015) and EMBASE (1950 to December 2014) databases were searched using keywords “hepatitis B virus”, “mutation” and “hepatocellular carcinoma”. A random-effects model was applied to obtain summary odds ratios. Publication bias was measured using Begg’s funnel plots and the Egger’s test. Meta-analysis was only performed for HBV variants studied in at least 4 articles.

Construction of plasmid DNA containing HBV DNA and transfection into Huh7
The plasmid pEGFP-N1-LHBS-genoB/C encoding the wild-type large HBV surface protein (LHBS) was used as a template for mutagenesis to create mutations at small surface region. These plasmids were transfected into Huh7 cells.

**Real-time PCR**

Total RNA was extracted from Huh7 cells 72 hours after transfection. Real-time PCR was performed using primers specific for \textit{APE1}, \textit{GRP78}, and \textit{GAPDH} genes with ABI 7500 Fast Real-Time PCR System using a SYBR Green method.

**Microarray analysis and biological interpretation**

Gene expression data from transfected Huh7 cells were analyzed by Agilent SurePrint Microarray. The complete dataset is available on the Gene Expression Omnibus database (Accession Number GSE87804). Normalized Microarray results were imported into Ingenuity Pathway Analysis (IPA) software. The major molecular and cellular functions, physiological system developmental functions, and disease and disorder development by these LHBS variants were investigated [16].
Western Blot Analysis

Proteins were extracted from transfected Huh7 cells and proteins related to ER stress and DNA repair were analyzed by western blotting.

Statistical analysis

Continuous variables were compared by Student’s t test or Mann-Whitney test. Categorical variables were compared by Chi-square test or Fisher’s exact test. The SNV frequency ($f_{m/\text{quasi}}$) is the frequency of a specific substitution within one subject ($f_{m/\text{quasi}} = \text{number of reads with a specific substitution} / \text{the number of total reads covering that site in one sample}$) [17]. The odds ratio (OR) of a specific HBV SNV for HCC is defined as: $\text{OR} = (\text{HCC}_E \times \text{non-HCC}_{\text{NE}}) / (\text{HCC}_{\text{NE}} \times \text{non-HCC}_E)$. The subscript $E$ indicates the number of patients with SNV frequencies not less than the cut-off value, and the subscript $\text{NE}$ indicates the number of patients with SNV frequencies less than the cut-off value. For each specific SNV, the cut-off value of SNV frequency which corresponds to the highest OR was used for statistical analysis. The $P$ values of these OR were calculated with Chi-square test. With respect to each genotype, HCC-associated SNVs discovered by NGS analysis were determined after
adjustment for multiple testing by Benjamini and Hochberg’s false discovery rate (FDR) method [18] for those HCC-associated variants evidenced by meta-analysis and those SNVs with $P$ values of OR<0.05 by NGS analysis.

Results

Clinical characteristics of patients

Table S3 lists the clinical characteristics of 93 HCC patients and 108 non-HCC patients enrolled in NGS analysis. The distributions of gender, age, AST, ALT, creatinine and HBeAg status were similar between the HCC and non-HCC groups for both genotypes B and C.

HBV genome-wide characterization of substitutions

The mapping quality of NGS was good, as indicated in Table S4. Figure 1 and Figure S2 display the genome-wide characterization of all genotypes B and C HBV substitutions identified in each patient. Complex quasi-species compositions were observed in both HCC and non-HCC groups. SNV I is defined as the dominant variant of HBV in the non-HCC group. SNV II is defined as the variant other than
SNV I at the same nucleotide position. The SNV II frequency in each patient is shown in a heatmap [19].

HCC-associated SNVs discovered by NGS analysis

For genotype B HBV, SNVs located at 23 nucleotide sites were significantly associated with HCC. These SNVs included pre-core (preC) mutations (1896A and 1899A). Nucleotide sites 273A and 2227T were protective factors for HCC, whereas 273G and 2227G were risk factors for HCC. All the other 21 SNVs were risk factors for HCC; of these, 6 were SNV I, and 15 were SNV II (Table 1A). Among the 25 SNVs, 17 were missense mutations which changed the amino acids in the polymerase, preS2, surface and preC/core regions (Table S5A.). Seven of the 17 missense mutations, and 4 of 8 silent mutations, involved the regulatory elements including CpG island I/II/III, X promoter, Enhancer (Enh) I, epsilon loop (µ loop) and basal core promoter (BCP) (Figure S2B).

For genotype C HBV, 35 SNVs located at 35 nucleotide sites were significantly associated with HCC. These SNVs included BCP mutation (1764A and 1653T). All the 35 SNVs were risk factors for HCC, while 18 of them were SNV II (Table 1B).
Among the 35 SNVs, 28 were missense mutations located at 4 ORFs, particularly at the preS1 region and the spacer domain of polymerase (Table S5B). Twenty-one of the 28 missense mutations and 3 of the 6 silent mutations were at the regulatory elements, namely CpG island I/II/III, negative regulatory element (NRE)/core upstream regulatory sequence (CURS)/BCP, Enh I/II, core promoter and S2 promoters (Figure S2C).

Validation of NGS-defined HCC-associated SNVs by systemic literature review and meta-analysis

To validate these 60 NGS-defined HCC-associated HBV SNVs, a systemic literature review of 2,270 potentially relevant articles was conducted, of which 109 articles [15,20-127] were enrolled as appropriate for final analysis (Figure S3 and Table S6). The articles discuss 167 HBV variants, which were categorized into 4 levels of supporting evidence associated with HCC. The proportions of NGS-defined HCC-associated SNVs among HBV variants with different levels of evidence associated with HCC declined significantly with decreasing levels of supporting evidence from Level A to Level D (chi-square test for the 2x4 contingency table,
Nine of the 12 HCC-associated variants supported by meta-analysis (Level A) were consistently identified as HCC-associated SNVs by NGS analysis. For the only three variants of Level A group not identified by NGS analysis, 1762T was identified as an HCC-associated SNV by our NGS-based subgroup analysis of HBeAg-positive patients with genotype C HBV infection (OR: 6.3, 95% CI = 1.1 to 34.0), whereas 1766T and 1768A mainly expressed in genotypes A and D, but not in genotypes B and C among our patients (Figure 2B and Figure S4L, S4N, S4O).

Seven (12%) of 60 HCC-associated HBV variants studied by less than 4 reports and supported in at least one report (Level B) were identified by NGS analysis. Only 2 (2%) of 85 HCC-unassociated variants supported by all previous studies with fewer than 4 reports (Level C) were identified by NGS analysis (Table S7.) None of the 10 HCC-unassociated variants by meta-analysis (Level D) were identified by NGS analysis (Figures S4 and S5).

The U-shaped distribution pattern of SNV frequency

A distinctive U-shaped distribution pattern of SNV II frequency was identified, with
most having either high (>80%) or low values (<20%) (Figure S6). The cut-off values of HCC-associated SNV frequency, *i.e.* pathogenic frequency, clearly showed a similar U-shaped distribution pattern (Table 1 and Figure 2C). Among 60 NGS-defined HCC-associated SNVs, 18 of 19 reported previously had cut-off values of SNV frequencies >20% and 16/19 had cut-off values of SNV frequencies >50%. Among the 41 novel HCC-associated SNVs, 25 and 13 had cut-off values of SNV frequencies >80% and <20%, respectively.

**Deletion mutations in HBV genome**

Deletion hotspots located in the preC/C gene, preS region and BCP region in both genotypes B and C HBV. The preS region had the greatest deletion frequencies and the most complex deletion patterns (Figure 3). Tables S8 and S9 list the deletion mutations of HBV in each subject discovered by NGS analysis. No significant difference was observed between HCC and non-HCC patients in the proportion having deletion mutations, either for genotype B (10/40 vs 8/47, *P*=0.36) or genotype C (34/53 vs 32/61, *P*=0.21). The proportion of patients with deletion mutations in the preS region (especially preS1) was much higher in the genotype C group than in the
genotype B group, among both HCC patients and non-HCC patients (Table S10). Intriguingly, the HCC group had more patients with deletion mutations involving nt2977–3013 (amino acid (aa.) 43–56) (6/34 vs 0/32, \( P=0.025 \), Figure 3B and 3D), deletion patterns II or III (18/34 vs 8/32, \( P=0.039 \)), and deletion mutations at functional sites S (17/34 vs 7/32, \( P=0.034 \)) and T (17/34 vs 7/32, \( P=0.034 \)) in the preS region [15] (Figure S7), than the non-HCC group among genotype C infected patients with HBV deletions. Deletion index is calculated as: Deletion index = \( \Sigma[(L_{\text{del}} \times R_{\text{del}})] / L_{\text{whole}} \), where \( L_{\text{del}} \) indicates length of the deletion fragment within a gene or region; \( R_{\text{del}} \) indicates ratio of the deletion fragment in a quasispecies within one subject; \( L_{\text{whole}} \) indicates whole length of a gene or region. HCC patients had higher deletion index at preS region than non-HCC patients, for genotype B (\( P=0.030 \)) and genotype C (\( P=0.045 \)) (Figure S8). Meta-analysis of preS deletions also showed a significant association with HCC (summary OR: 3.0, 95% CI, 2.3 to 3.9) (Figure S4W).

**NGS-defined HCC-associated SNVs in small S region induce ER-stress, regulate DNA repair and affect hepatocarcinogenesis pathways**

The levels of GRP78 mRNA were 1.27-fold (\( P=0.030 \)) and 1.38-fold (\( P=0.033 \))
higher in A273G mutant (genotype B) and C293G mutant (genotype C), respectively, than controls. The levels of $APE1$ mRNA were 1.57-fold ($P=0.026$) and 1.41-fold ($P=0.018$) higher in C446G mutant (genotype C) and C446G/A456G mutant (genotype C), respectively, than controls (Figure S9). From microarray and Ingenuity Pathway Analysis, all the novel NGS-defined HCC-associated SNVs and their combinations in small S region could be involved in hepatocarcinogenesis pathways, by inducing cell apoptosis, regulating cell cycle, death and survival, and affecting DNA replication, recombination and repair. Intriguingly, these double or triple mutants enhanced those effects regarding hepatocarcinogenic pathways (Figure 4). For example, the triple mutant (A293G/C446G/A456G) in genotype C significantly affected 19805 genes compared to wild type construct. Among those affected genes, 311 belonged to the DNA repair-related genes with 206 down-regulated and 105 up-regulated. The up-regulated genes (Mu/Wt, log$_2$ratio $\geq$1) included PML, HMGA2, PARP1, XRCC1, PLK1, SLX1A and RMI2. The down-regulated genes (Mu/Wt, log$_2$ratio $\leq$-1) included PCNA, PMS2, XRCC4, ATM, MDC1 and RAD51C. Consistent with the results of real-time PCR analysis, our microarray analysis showed significantly higher expression of the ER-stress markers $GRP78$ and $APE1$ mRNA in
small S mutants than controls (GRP78: 1.53-fold in A273G and 1.93-fold in C293G; APE1: 1.58-fold in C446G and 1.56-fold in C446G/A456G).

At the protein level, all NGS-defined HCC-associated SNVs in small S region of genotype B influenced ER stress-related factors such as PERK, IRE1±, Ero1-Lα, GRP78 and calnexin, except for T216C LHBS variant with only a trend to increase GRP78 expression (Figure S10A). In genotype C, three LHBS single mutated variants (A293G, C446G, and A456G) influenced PERK, IRE1±, Ero1-Lα and calnexin, except GRP78 up-regulated to some extent. Most of the double and triple LHBS affected the expression of PERK, IRE1±, Ero1-Lα, and calnexin, but not GRP78. Interestingly, the variants of genotype C resulted in accumulation of LHBS proteins more prominently than those of genotype B (Figure S10B). Additionally, all these genotype B LHBS variants and most of genotype C LHBS variants affected DNA repair systems through inducing XRCC1 and reducing APE1 (Figure S10). The C446G/A456G LHBS variants of genotype C induced Ape1 protein, compatible with results at the RNA level (Figures S9 and S10).

Discussion
This is the first large study to investigate HCC-associated HBV SNVs with their pathogenic frequencies using NGS analysis of whole HBV genomes. It is also the first to compare NGS-defined SNVs with HBV variants studied previously using systemic literature review and meta-analysis. This high-throughput sequencing analysis revealed that SNVs exist in 49.9% and 51.3% of nucleotide sites in HBV genomes of genotypes B and C, respectively. The preS region, including S2 promoter, had the highest mutation rate (the proportion of nucleotide sites with SNVs in a gene or region), while DR1, DR2, ε loop and poly-A signal had lower ones (Figures 1, S2 and S11).

Sixty NGS-defined HCC-associated SNVs were found. The proportions of NGS-defined HCC-associated SNVs among HBV variants declined significantly along with the decreasing levels of supporting evidence from Level A to Level D after combining the results from NGS, systemic literature review and meta-analysis (Figure 2A), indicating that these NGS results were valid. All HCC-associated HBV variants supported by meta-analysis (Level A), except 1766T and 1768A (mainly expressed in genotypes A and D), but including the subgroup analysis of 1762T, were identified by NGS analysis [23,45,61]. These 10 HCC-associated HBV variants were located in the
C-terminal transactivation domain of HBx/BCP (1613A, 1653T, 1674C, 1753G, 1762T and 1764A), preC region (1846T, 1896A and 1899A) and preS2 region (53C). This study provides the first meta-analysis to report association with HCC for 53C, 1613A, 1674C and 1846T. No NGS-defined HCC-associated SNVs were discovered among the 10 HCC-unassociated variants by meta-analysis (Level D).

Notably, this HBV whole-genome high-throughput sequencing analysis revealed 41 novel HCC-associated SNVs. These novel SNVs were discovered probably for the following reasons. First, NGS analysis can define HCC-associated SNVs with their specific sites, nucleotides and pathogenic frequencies, instead of presence or absence of specific major stains identified mainly by direct sequencing as in previous studies. Second, consistent with the unique U-shaped distribution pattern of SNV frequency [17], 32% of 41 novel HCC-associated SNVs had cut-off values of SNV frequencies below 20%, which could not be detected by direct sequencing. The HBV SNVs with low frequencies still have high viral content in CHB patients with high viral loads. Therefore, HCC-associated SNVs with low frequencies could have carcinogenic effects [128]. Coexistence of specific HCC-associated SNVs, even some with low frequencies, could increase the risk of HCC occurrence according to microarray
analysis (Figure 4). Third, 40 of 41 novel HCC-associated SNVs did not locate in previously reported hotspot regions, such as X gene, BCP region and preC region of an HBV genome [88,95,129]. For the genotype B group, 17 of 18 novel HCC-associated SNVs were located in the core region, small S region and P gene, while only one located in the preC region. For the genotype C group, all 23 novel HCC-associated SNVs were located in the preS region, small S region, core region and P gene (Figure S2). Point mutations of the preS region has rarely been explored, although the association of preS region deletions and HCC is extensively studied [130-132]. These novel HCC-associated SNVs might play a significant role in HCC development. The SNVs with a characteristic U-shape frequency distributions probably resulted from evolution processes and are consistent with a previous report [17,133]. Our results indicate that HCC-associated SNVs that lead to hepatocarcinoma are not related to the presence or absence of the specific SNVs, but with certain pathogenic frequencies.

Intriguingly, the 60 NGS-defined HCC-associated SNVs were distinct for both genotype B (n=24) and genotype C (n=34), except 53C for both genotypes. Genotype C HBV showed HCC-associated SNVs in the X gene (n=6), BCP region (n=6) and
preS1 region (n=15), while genotype B HBV did not. Genotype B HBV bear more HCC-associated SNVs in preC/core gene than genotype C HBV (n=10 vs n=2) (Figure S2). Previous studies also indicated that baseline BCP 1762T/1764A mutations increase the risk of HCC development exclusively in genotype C and not in genotype B [128]. Moreover, only two of the 12 HCC-associated variants supported by meta-analysis could receive further subgroup analysis of genotypes. All variants showed significant association with HCC in genotype C but not in genotype B (1653T, summary OR: 2.9, 95% CI, 2.1 to 3.9; 1753V, summary OR: 2.0, 95% CI, 1.1 to 3.6), which is consistent with our NGS results. Distinct genotype-specific HCC-associated SNVs possibly cause distinct clinical characteristics in HCC patients infected with different HBV genotypes. For instance, BCP 1762T/1764A mutations were found to be distinct HCC-associated SNVs for genotype C in this study, and independent predictors for postoperative recurrence in HCC in a previous study [134]. This might be the reason that genotype C HBV is associated with increased risk of HCC and an increased tumor recurrence rate [128,135,136]. In addition, the dN/dS ratios of the codons affected by these SNVs were calculated by DiversiTool (http://josephhughes.github.io/btctools/) [137]. Purifying selection (dN/dS<1)
involved about 60% of these codons (Figure S12 and Table S11). These findings are worthy of further studies.

From real-time PCR and microarray analysis, the NGS-defined HCC-associated SNVs and their combinations in small S region may induce ER-stress and influence hepatocarcinogenesis through affecting cell cycle, cell apoptosis, cell death, cell survival, DNA replication, DNA recombination, DNA repair, etc. The correlation of SNV combinations with HCC were supported by the results of SKAT analysis [66,138] (Table S12). For example, 311 DNA repair-related genes were affected by the triple mutant (A293G/C446G/A456G) in genotype C, involving DNA double-strand break repair by homologous recombination (HDR), non-homologous end joining (NHEJ), mismatch repair (MMR) and nucleotide excision repair (NER). Most genes involved in HDR, MMR and NER were down-regulated (Table S13). Among the NHEJ genes, those participating in the canonical NHEJ (C-NHEJ) were reduced but those participating in the alternative NHEJ (A-NHEJ), a major source of genomic instability, were induced. Western blot analysis demonstrated that LHBS variants with HCC-associated HBV SNVs in small S region induced ER stress with LHBS protein accumulation and altered DNA repair systems through XRCC1 and Ape1 (Figure
Our meta-analysis, which included 32 studies and evaluated a total of 5563 HBV-infected patients, demonstrated that infection with preS deletions was associated with a 3.0-fold increased risk of HCC (Figure S4W). For both genotypes B and C HBV, only preS deletion was significantly associated with HCC in terms of deletion index. The deletion index is composed of the deletion length and the deletion frequency by NGS analysis. Our sequencing analysis found that deletions were more often located in the 3’-terminus of the pre-S1 region and the 5’-terminus of the pre-S2 region (Figure 3), as reported previously [15]. Both preS1 and preS2 deletions can cause unbalanced production of envelope proteins, with a consequent accumulation of mutated large surface protein (LHBs) in the ER of hepatocytes, activating ER stress and ultimately HCC development [132,139,140]. Furthermore, among genotype C HBV infected patients with any deletion mutation, the HCC group had more patients with preS deletion mutations involving nt2977–3013 (aa.43–56) (Figure 3), deletion patterns II or III, and functional sites S and T (Figure S7), than the non-HCC group. The deletion involved in the function of envelope protein synthesis via S-promoter and CAD/Hsc70 binding activity. HCC-associated preS deletion
mutations involved a specific fragment (nt2977–3013, aa.43–56), which is essential for efficient hepadnaviral infectivity [141,142]. However, the hepatocarcinogenic mechanism of this fragment requires further elucidation.

According to the preceding systemic literature review and *in vitro* studies, the NGS-defined HCC-associated SNVs were unlikely to be discovered by chance. However, there are some limitations in this study. Because of the limited sample size, the novel HCC-associated SNVs discovered by our NGS analysis need to be confirmed by further cellular and animal model experiments. These NGS-defined HCC-associated HBV SNVs might not be applicable to patients with low viral load because of possible selection bias in this study. Furthermore, HBV variants in liver might not fully contribute to those in serum of CHB patients. Some differences of viral quasispecies at preC region have been found between those in serum and those in liver of CHB patients [143]. Several preS deletion mutants could result in less ability to replicate or produce infectious particles in the liver, contribute to HBV integration and ER stress via the accumulation of S protein, and probably play an important role in hepatocarcinogenesis [144-149]. It will be valuable to investigate HBV quasispecies in liver in the future.
Studies on the properties of HBV quasispecies indicate that distinct viral variants with greatest fitness evolve through immune selection during the natural course of chronic HBV infection [150] and play crucial roles in HCC development [151,152]. The development of HCC is a highly heterogeneous process involving mutations in HBV proteins and regulatory elements. Unique HBV oncoproteins (HBx isoforms and preS mutants) and mutated preC/core proteins can induce hepatocarcinogenesis through activation of ER-independent pathways [152], induction of ER and oxidative stress [153-155], regulation of microRNA expression [156], lipid metabolism disturbance [157] or epigenetic modification through modified genomic methylation status [158]. Mutations of HBV regulatory elements probably induce hepatocarcinogenesis through oncoprotein expression modulation [151,152], HBV DNA integration leading to chromosomal instability (DR1/DR2) [159] or HBV DNA methylation (CpG islands) [160] (Figure 5). We confirmed that all our novel NGS-defined HCC-associated SNVs in the small S region, particularly those with double and triple mutants, could influence hepatocarcinogenesis pathways (Figure 4) and some of them could induce ER-stress (Figures S9 and S10). In conclusion, NGS analysis revealed 41 novel HCC-associated SNVs and preS deletion, which involved HBV ORFs and regulatory
elements and are worthy of further investigation in hepatocarcinogenesis.

Acknowledgements

This work was supported by grants from the Ministry of Science and Technology, R.O.C (MOST 102-2628-B-006-008-MY3, NSC 99-2314-B-006-028-MY3, and NSC 95-2321-B-010-005), the Ministry of Health and Welfare, Executive Yuan, R.O.C (MOHW104-TDU-B-211-124-001 and MOHW106-TDU-B-211-144-003 for the Center of Excellence), and the Infectious Disease and Signaling Research Center of National Cheng Kung University. We thank the Human Biobank of the Research Center of Clinical Medicine and Cancer Data Bank of the Cancer Center of National Cheng Kung University Hospital Cancer Center for providing HCC samples. We are grateful to Dr. Sheng-Hsiang Lin for providing statistical consulting services from the Biostatistics Consulting Center, National Cheng Kung University Hospital. We thank Chun-Pei Cheng, Yi-Cheng Tsai and Chia-Hua Li for assisting NGS analysis. We are grateful to Jui-Chu Yang for providing the bioinformatics consulting services of IPA from the Human Biobank, Research Center of Clinical Medicine, National Cheng Kung University Hospital. We are indebted to Professor Wenya Huang and Professor
Kung-Chia Young for invaluable suggestions, and Ted Kony and I-Chen Li for English editing.

**Author Contributions:** W-CL and I-CW contributed equally to this article. W-CL handled experimental design, experiment performance, clinical and NGS data analyses, biometric consulting and manuscript writing; I-CW was responsible for statistical analysis and manuscript writing; Y-CL, S-LY and J-JL helped in meta-analysis; C-PL and BC-HC executed NGS and analyzed high-throughput data using bio-informatic software; J-HC, P-FL, Y-TC, P-WC, VST and K-TS assisted in analyzing the classification-based association rules; Y-JL, C-JY and P-NC provided clinical samples from HCC patients; J-CY and K-CC provided the bioinformatics consulting services of IPA; C-HH helped in sample preparation, clinical data arrangement and statistical analysis; T-TC and J-CW coordinated the study, gathered clinical data and serum samples from patients with chronic hepatitis B, and composed the manuscript.

**Abbreviations:** amino acid (aa.), basal core promoter (BCP), chronic hepatitis B
(CHB), core upstream regulatory sequence (CURS), epsilon loop (µ loop), Enhancer (Enh) I, hepatitis B virus (HBV), hepatocellular carcinoma (HCC), homologous recombination (HDR), Ingenuity Pathway Analysis (IPA), mismatch repair (MMR), negative regulatory element (NRE), next-generation sequencing (NGS), non-homologous end joining (NHEJ), nucleotide excision repair (NER), odds ratio (OR), open reading frames (ORFs), pre-core (preC), single nucleotide variants (SNVs), small surface (S).
SUPPLEMENTARY MATERIAL ONLINE
Supplementary material file including Material and Methods, Legends for Figures S1-S12, and Tables S1- S13

**Figure S1.** Begg’s funnel plot of 1753V, 1896A and 1899A

**Figure S2.** HBV genome-wide characterization of substitutions with distinct NGS-defined HCC-associated SNVs in genotype B and genotype C

**Figure S3.** Flowchart of article selection

**Figure S4.** Summary odds ratios of hepatocellular carcinoma

**Figure S5.** Odds ratios of NGS analysis in 10 HBV variants with Level D evidence

**Figure S6.** The U-shaped distribution pattern of SNV II frequency

**Figure S7.** The percentage of patients with distinct preS deletion patterns among patients with any deletion mutation of HBV full genome comparing HCC and non-HCC groups

**Figure S8.** Distribution of deletion index for different HBV gene fragments in HCC and non-HCC patients

**Figure S9.** HBV small S mutants induces the ER stress compared with HBV wildtype small S proteins

**Figure S10.** LHBS variants with HCC-associated HBV SNVs in small S region
induced ER stress and involved DNA repair system

**Figure S11.** Mutation rates of HBV open reading frames and regulatory elements

**Figure S12.** The dN/dS ratios of the codons affected by NGS-defined HCC-associated SNVs in each HBV gene

**Table S1.** Primer sets for amplifying full HBV genome in next-generation sequencing (NGS)

**Table S2.** Oligonucleotides used for PCR mutagenesis

**Table S3.** Clinical characteristics of chronic hepatitis B patients with and without HCC

**Table S4.** Mapping statistics of NGS reads

**Table S5A.** Seventeen significant HCC-associated SAVs among genotype B HBV infected patients

**Table S5B.** Twenty-eight significant HCC-associated SAVs among genotype C HBV infected patients

**Table S6.** Characteristics of 109 studies included in the meta-analysis

**Table S7.** NGS-defined HCC-associated HBV SNVs with Level B evidence or Level C evidence

**Table S8.** Characteristics of genotype B HBV deletion patterns among patients with...
and without HCC

Table S9. Characteristics of genotype C HBV deletion patterns among patients with and without HCC

Table S10. Proportions of patients with deletion mutations among HBV genome comparing genotype B and genotype C groups

Table S11. The substitution patterns of NGS-defined HCC-associated SNVs

Table S12. The association of NGS-defined SNVs and their combinations at small S gene with HCC by SKAT

Table S13. Expression profiles of essential genes of the five major DNA repair mechanisms in Huh7 cell with small S triple mutants of genotype C HBV
References

13. Liu WC, Mizokami M, Buti M, et al. Simultaneous quantification and genotyping of hepatitis B virus for genotypes A to G by real-time PCR and


25. Baptista M, Kramvis A, Kew MC. High prevalence of 1762(T) 1764(A) mutations in the basic core promoter of hepatitis B virus isolated from black


46. Guo X, Jin Y, Qian G, et al. Sequential accumulation of the mutations in core promoter of hepatitis B virus is associated with the development of


134. Yeh CT, So M, Ng J, et al. Hepatitis B virus-DNA level and basal core promoter A1762T/G1764A mutation in liver tissue independently predict postoperative survival in hepatocellular carcinoma. *Hepatology* 2010; 52:
1922-1933.


146. Hildt E, Munz B, Saher G, *et al.* The PreS2 activator MHBs(t) of hepatitis B


158. Watanabe Y, Yamamoto H, Oikawa R, *et al.* DNA methylation at hepatitis B
viral integrants is associated with methylation at flanking human genomic sequences. *Genome Res* 2015; **25**: 328-337.


Table 1A. Twenty-five significant HCC-associated SNVs with their cut-off values of SNV frequency ($f_{\text{m/quasi}}$) among genotype B HBV infected patients through NGS analysis.

<table>
<thead>
<tr>
<th>Site NT</th>
<th>Median SNV frequency ($f_{\text{m/quasi}}$, %)</th>
<th>Odds ratio b</th>
<th>Cut-off (95%CI)</th>
<th>Proportion of patients with SNV frequency $\geq$ cut-off value (%)</th>
<th>Corrected threshold</th>
<th>P</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(10th, 25th, 75th, and 90th percentiles)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC group</td>
<td>Non-HCC group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=40)</td>
<td>(N=47)</td>
<td>(N=40)</td>
<td>(N=47)</td>
<td>(N=40)</td>
<td>(N=47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53 C 38 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 23, 100)</td>
<td>4.5 (1.8, 11.5)</td>
<td>32</td>
<td>55</td>
<td>21.3</td>
<td>0.0025</td>
<td>0.008</td>
</tr>
<tr>
<td>216 C 0.0 (0.0, 0.0, 80, 100)</td>
<td>8.8 (0.0, 0.0, 55, 82)</td>
<td>5.6 (1.2, 28.3)</td>
<td>100</td>
<td>20</td>
<td>4.3</td>
<td>0.0391</td>
<td>0.046</td>
</tr>
<tr>
<td>273 G 21 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 5.6, 100)</td>
<td>5.0 (2.0, 12.8)</td>
<td>12.2</td>
<td>57.5</td>
<td>21.3</td>
<td>0.0012</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>273</td>
<td>A</td>
<td>79 (0.0, 0.0, 100, 100)</td>
<td>100 (0.0, 94, 100, 100)</td>
<td>0.3 (0.1, 0.8)</td>
<td>55</td>
<td>42.5</td>
<td>19.1</td>
</tr>
<tr>
<td>529</td>
<td>G</td>
<td>20 (0.0, 0.0, 100, 100)</td>
<td>3.7 (0.0, 0.0, 18, 100)</td>
<td>4.1 (1.4, 12.0)</td>
<td>100</td>
<td>37.5</td>
<td>12.8</td>
</tr>
<tr>
<td>530</td>
<td>A</td>
<td>100 (80, 100, 100, 100)</td>
<td>100 (32, 89, 100, 100)</td>
<td>4.2 (1.3, 14.0)</td>
<td>100</td>
<td>90</td>
<td>68.1</td>
</tr>
<tr>
<td>724</td>
<td>C</td>
<td>100 (87, 95, 100, 100)</td>
<td>100 (81, 87, 100, 100)</td>
<td>4.0 (1.3, 12.0)</td>
<td>91.7</td>
<td>87.5</td>
<td>63.8</td>
</tr>
<tr>
<td>1173</td>
<td>G</td>
<td>100 (97, 100, 100, 100)</td>
<td>100 (85, 100, 100, 100)</td>
<td>9.2 (1.1, 76.5)</td>
<td>94.1</td>
<td>97.5</td>
<td>80.9</td>
</tr>
<tr>
<td>1221</td>
<td>C</td>
<td>0.0 (0.0, 0.0, 4.6, 45)</td>
<td>0.0 (0.0, 0.0, 0.0, 16)</td>
<td>4.1 (1.1, 14.1)</td>
<td>4.3</td>
<td>27.5</td>
<td>8.5</td>
</tr>
<tr>
<td>1242</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 62, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 71)</td>
<td>5.1 (1.8, 14.6)</td>
<td>5.5</td>
<td>42.5</td>
<td>12.8</td>
</tr>
<tr>
<td>1359</td>
<td>A</td>
<td>0.0 (0.0, 0.0, 3.7, 75)</td>
<td>0.0 (0.0, 0.0, 0.0, 4.2)</td>
<td>3.6 (1.0, 12.5)</td>
<td>2.4</td>
<td>25</td>
<td>8.5</td>
</tr>
<tr>
<td>1896</td>
<td>A</td>
<td>100 (0.0, 1.1, 100, 100)</td>
<td>65 (0.0, 0.0, 100, 100)</td>
<td>3.5 (1.1,10.7)</td>
<td>96.5</td>
<td>57.5</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1899</td>
<td>A</td>
<td>0.0 (0.0, 0.0, 99, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 94)</td>
<td>5.8 (1.1, 31.0)</td>
<td>96.1</td>
<td>32.5</td>
<td>6.4</td>
</tr>
<tr>
<td>1913</td>
<td>C</td>
<td>100 (58, 100, 100, 100)</td>
<td>100 (31, 86, 100, 100)</td>
<td>5.5 (1.1, 28.2)</td>
<td>94.3</td>
<td>62.5</td>
<td>97.9</td>
</tr>
<tr>
<td>2095</td>
<td>G</td>
<td>100 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 100, 100)</td>
<td>3.2 (1.3, 8.0)</td>
<td>10.1</td>
<td>72.5</td>
<td>44.7</td>
</tr>
<tr>
<td>2120</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 65, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 67)</td>
<td>3.6 (1.0, 12.5)</td>
<td>67.2</td>
<td>25</td>
<td>8.5</td>
</tr>
<tr>
<td>2213</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 0.0, 85)</td>
<td>0.0 (0.0, 0.0, 0.0, 0.0)</td>
<td>9.8 (1.2, 83.1)</td>
<td>1.7</td>
<td>17.5</td>
<td>2.1</td>
</tr>
<tr>
<td>2226</td>
<td>T</td>
<td>0.0 (0.0, 0.0, 0.0, 91)</td>
<td>0.0 (0.0, 0.0, 0.0, 0.0)</td>
<td>9.8 (1.2, 83.1)</td>
<td>1.7</td>
<td>17.5</td>
<td>2.1</td>
</tr>
<tr>
<td>2227</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 0.0, 91)</td>
<td>0.0 (0.0, 0.0, 0.0, 0.0)</td>
<td>9.8 (1.2, 83.1)</td>
<td>1.8</td>
<td>17.5</td>
<td>2.1</td>
</tr>
<tr>
<td>2227</td>
<td>T</td>
<td>100 (0.0, 93, 100, 100)</td>
<td>100 (100, 100, 100, 100)</td>
<td>0.2 (0.1, 0.8)</td>
<td>95.5</td>
<td>25</td>
<td>6.4</td>
</tr>
<tr>
<td>2441</td>
<td>C</td>
<td>0.0 (0.0, 0.0, 20, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 75)</td>
<td>11.5 (1.4, 96.5)</td>
<td>100</td>
<td>20</td>
<td>2.1</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved.
<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2444</td>
<td>C</td>
<td>100 (68, 100, 100, 100)</td>
<td>100 (42, 80, 100, 100)</td>
<td>4.8 (1.6, 14.3)</td>
<td>92.6</td>
<td>85</td>
<td>57.4</td>
<td>0.0104</td>
<td>0.018 Significant</td>
</tr>
<tr>
<td>2525</td>
<td>T</td>
<td>0.0 (0.0, 0.0, 95, 100)</td>
<td>0.0 (0.0, 0.0, 14, 71)</td>
<td>7.5 (1.5, 36.7)</td>
<td>95.5</td>
<td>25</td>
<td>4.3</td>
<td>0.01</td>
<td>0.014 Significant</td>
</tr>
<tr>
<td>2583</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 91, 100)</td>
<td>0.0 (0.0, 0.0, 14, 85)</td>
<td>6.5 (1.2, 32.3)</td>
<td>100</td>
<td>22.5</td>
<td>4.3</td>
<td>0.0201</td>
<td>0.03 Significant</td>
</tr>
<tr>
<td>2690</td>
<td>A</td>
<td>0.0 (0.0, 0.0, 12, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 8.6)</td>
<td>4.9 (1.2, 19.3)</td>
<td>12</td>
<td>25</td>
<td>6.4</td>
<td>0.0316</td>
<td>0.042 Significant</td>
</tr>
</tbody>
</table>

Note: a SNV frequency (f_{m/quasi}) of a specific HBV SNV in one subject was counted as the frequency of a specific substitution within one subject, i.e. the number of reads with a specific substitution/the number of total reads covering that site in one sample. Data are expressed as median, 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles. b Odds ratio (OR) of a specific HBV SNV for HCC was defined as: OR = (HCC_E x non-HCC_NE) / (HCC_NE x non-HCC_E). The subscript E means the number of patients with SNV frequency no less than the cut-off value; the subscript NE means the number of patients with SNV frequency less than the cut-off value. c For each specific SNV, the cut-off value of SNV frequency which corresponds to the highest odds ratio was used for statistical analysis. P values of odds
ratios were calculated by Chi-square test. With respect to genotype B, HCC-associated SNVs discovered by NGS were determined after adjustment for multiple testing by Benjamini and Hochberg’s false discovery rate (FDR) method for those HCC-associated variants evidenced by meta-analysis and those SNVs with P values of OR<0.05 by NGS analysis.

Table 1B. Thirty-five significant HCC-associated SNVs with their cut-off values of SNV frequency ($f_{m/quasi}$) among genotype C HBV infected patients through NGS analysis.
<table>
<thead>
<tr>
<th>Site NT</th>
<th>Median SNV frequency (f_{\text{nuqasi, } %}) (10^{\text{th}}, 25^{\text{th}}, 75^{\text{th}}, \text{and } 90^{\text{th}} \text{ percentiles})</th>
<th>Odds ratio b</th>
<th>Cut-off value (%)</th>
<th>P value</th>
<th>False discovery rate (FDR) (10^{\text{th}}, 25^{\text{th}}, 75^{\text{th}}, \text{and } 90^{\text{th}} \text{ percentiles})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HCC group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=53)</td>
<td>(N=61)</td>
<td>(N=53)</td>
<td>(N=61)</td>
<td>(N=53)</td>
<td>(N=61)</td>
</tr>
<tr>
<td>53 C</td>
<td>0.0 (0.0, 0.0, 57, 91)</td>
<td>0.0 (0.0, 0.0, 6.2, 29)</td>
<td>5.1 (1.6, 16.7)</td>
<td>52.4</td>
<td>26.4</td>
</tr>
<tr>
<td>293 G</td>
<td>3.6 (0.0, 0.0, 51, 100)</td>
<td>0.0 (0.0, 0.0, 6.1, 87)</td>
<td>2.9 (1.3, 6.4)</td>
<td>1.8</td>
<td>50.9</td>
</tr>
<tr>
<td>446 G</td>
<td>0.0 (0.0, 0.0, 17, 87)</td>
<td>0.0 (0.0, 0.0, 0.0, 46)</td>
<td>5.6 (2.0, 15.2)</td>
<td>6.4</td>
<td>37.7</td>
</tr>
<tr>
<td>456 G</td>
<td>0.0 (0.0, 0.0, 0.0, 21)</td>
<td>0.0 (0.0, 0.0, 0.0, 0.0)</td>
<td>6.0 (1.2, 29.3)</td>
<td>10.2</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>633</td>
<td>A</td>
<td>0.0 (0.0, 0.0, 3.3, 50)</td>
<td>0.0 (0.0, 0.0, 0.0, 0.0)</td>
<td>5.6 (1.1, 27.5)</td>
<td>3.6</td>
</tr>
<tr>
<td>834</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 95, 100)</td>
<td>0.0 (0.0, 0.0, 17, 52)</td>
<td>7.6 (2.1, 28.2)</td>
<td>76.9</td>
</tr>
<tr>
<td>1092</td>
<td>C</td>
<td>100 (95, 100, 100, 100)</td>
<td>100 (0.0, 100, 100, 100)</td>
<td>5.6 (1.2, 26.6)</td>
<td>50.1</td>
</tr>
<tr>
<td>1155</td>
<td>C</td>
<td>100 (98, 100, 100, 100)</td>
<td>100 (0.0, 100, 100, 100)</td>
<td>5.6 (1.2, 26.6)</td>
<td>87.1</td>
</tr>
<tr>
<td>1386</td>
<td>A</td>
<td>3.6 (0.0, 0.0, 87, 100)</td>
<td>0.0 (0.0, 0.0, 13, 83)</td>
<td>5.2 (1.1, 25.9)</td>
<td>96.7</td>
</tr>
<tr>
<td>1613</td>
<td>A</td>
<td>7.0 (0.0, 0.0, 75, 100)</td>
<td>0.0 (0.0, 0.0, 39, 81)</td>
<td>10.1 (1.2, 83.6)</td>
<td>86.7</td>
</tr>
<tr>
<td>1653</td>
<td>T</td>
<td>11 (0.0, 0.0, 90, 100)</td>
<td>0.0 (0.0, 0.0, 6.5, 100)</td>
<td>4.2 (1.3, 14.2)</td>
<td>60.8</td>
</tr>
<tr>
<td>1674</td>
<td>C</td>
<td>0.0 (0.0, 0.0, 38, 100)</td>
<td>0.0 (0.0, 0.0, 4.2, 54)</td>
<td>11.5 (1.4, 94.8)</td>
<td>55.8</td>
</tr>
<tr>
<td>1753</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 0.0, 23)</td>
<td>0.0 (0.0, 0.0, 0.0, 12)</td>
<td>3.1 (1.1, 9.4)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>1764</td>
<td>A</td>
<td>100 (0.0, 100, 100, 100)</td>
<td>97 (0.0, 7.8, 100, 100)</td>
<td>4.6 (1.7, 12.4)</td>
<td>100</td>
</tr>
<tr>
<td>1846</td>
<td>T</td>
<td>8.7 (0.0, 0.0, 52, 100)</td>
<td>0.0 (0.0, 0.0, 21, 94)</td>
<td>3.1 (1.2, 8.2)</td>
<td>21.1</td>
</tr>
<tr>
<td>2201</td>
<td>T</td>
<td>100 (96, 100, 100, 100)</td>
<td>100 (0.0, 100, 100, 100)</td>
<td>3.7 (1.1, 11.9)</td>
<td>93.3</td>
</tr>
<tr>
<td>2573</td>
<td>C</td>
<td>0.0 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 100)</td>
<td>4.2 (1.3, 14.2)</td>
<td>100</td>
</tr>
<tr>
<td>2594</td>
<td>A</td>
<td>0.0 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 3.9, 100)</td>
<td>4.0 (1.3, 12.4)</td>
<td>100</td>
</tr>
<tr>
<td>2708</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 0.0, 97)</td>
<td>5.8 (1.5, 22.1)</td>
<td>100</td>
</tr>
<tr>
<td>2840</td>
<td>T</td>
<td>14 (6.4, 10, 17, 22)</td>
<td>9.6 (0.0, 4.4, 15, 23)</td>
<td>8.3 (1.8, 38.3)</td>
<td>4.4</td>
</tr>
<tr>
<td>2875</td>
<td>A</td>
<td>4.8 (0.0, 0.0, 95, 100)</td>
<td>5.7 (0.0, 0, 50, 93)</td>
<td>5.6 (1.7, 18.3)</td>
<td>94.8</td>
</tr>
<tr>
<td>2889</td>
<td>G</td>
<td>100 (87, 100, 100, 100)</td>
<td>100 (7.8, 79, 100, 100)</td>
<td>5.5 (1.8, 17.6)</td>
<td>85.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2901</td>
<td>T</td>
<td>100 (90, 100, 100, 100)</td>
<td>100 (73, 90, 100, 100)</td>
<td>4.1 (1.1, 15.4)</td>
<td>85.1</td>
</tr>
<tr>
<td>2931</td>
<td>C</td>
<td>100 (92, 100, 100, 100)</td>
<td>100 (76, 86, 100, 100)</td>
<td>9.1 (2.0, 41.6)</td>
<td>87.1</td>
</tr>
<tr>
<td>2988</td>
<td>C</td>
<td>100 (100, 100, 100, 100)</td>
<td>100 (88, 100, 100, 100)</td>
<td>9.0 (1.1, 73.6)</td>
<td>92.1</td>
</tr>
<tr>
<td>2989</td>
<td>A</td>
<td>100 (100, 100, 100, 100)</td>
<td>100 (88, 100, 100, 100)</td>
<td>9.0 (1.1, 73.6)</td>
<td>91.7</td>
</tr>
<tr>
<td>2997</td>
<td>T</td>
<td>100 (98, 100, 100, 100)</td>
<td>100 (89, 100, 100, 100)</td>
<td>9.0 (1.1, 73.6)</td>
<td>90.8</td>
</tr>
<tr>
<td>2998</td>
<td>C</td>
<td>100 (100, 100, 100, 100)</td>
<td>100 (89, 100, 100, 100)</td>
<td>9.0 (1.1, 73.6)</td>
<td>92.2</td>
</tr>
<tr>
<td>3006</td>
<td>A</td>
<td>100 (100, 100, 100, 100)</td>
<td>100 (84, 98, 100, 100)</td>
<td>14.1 (1.8, 111.8)</td>
<td>94.2</td>
</tr>
<tr>
<td>3009</td>
<td>G</td>
<td>100 (97, 100, 100, 100)</td>
<td>100 (88, 100, 100, 100)</td>
<td>5.0 (1.0, 24.0)</td>
<td>92.9</td>
</tr>
<tr>
<td>3016</td>
<td>C</td>
<td>100 (100, 100, 100, 100)</td>
<td>100 (90, 100, 100, 100)</td>
<td>9.0 (1.1, 73.6)</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3021</td>
<td>A</td>
<td>100 (100, 100, 100, 100)</td>
<td>100 (84, 100, 100, 100)</td>
<td>11.4 (1.4, 91.9)</td>
<td>92.6</td>
</tr>
<tr>
<td>3066</td>
<td>T</td>
<td>100 (94, 100, 100, 100)</td>
<td>100 (72, 94, 100, 100)</td>
<td>5.6 (1.2, 26.6)</td>
<td>86.8</td>
</tr>
<tr>
<td>3097</td>
<td>A</td>
<td>100 (95, 100, 100, 100)</td>
<td>100 (84, 100, 100, 100)</td>
<td>4.5 (1.2, 16.8)</td>
<td>93.5</td>
</tr>
<tr>
<td>3120</td>
<td>G</td>
<td>0.0 (0.0, 0.0, 100, 100)</td>
<td>0.0 (0.0, 0.0, 19, 95)</td>
<td>4.8 (1.6, 14.4)</td>
<td>95.9</td>
</tr>
</tbody>
</table>

Note: a SNV frequency (f_{m/quasi}) of a specific HBV SNV in one subject was counted as the frequency of a specific substitution within one subject, i.e. the number of reads with a specific substitution/the number of total reads covering that site in one sample. Data are expressed as median, 10\textsuperscript{th}, 25\textsuperscript{th}, 75\textsuperscript{th}, and 90\textsuperscript{th} percentiles. b Odds ratio (OR) of a specific HBV SNV for HCC was defined as follows:

\[ OR = \frac{\text{HCC}_E \times \text{non-HCC}_N}{\text{HCC}_N \times \text{non-HCC}_E} \]

The subscript \(_E\) means the number of patients with SNV frequency no less than the cut-off value; the subscript \(_N\) means the number of patients with SNV frequency less than the cut-off value. c For each specific SNV, the cut-off value of SNV frequency which corresponds to the highest odds ratio was used for statistical analysis. \( P \) values of
odds ratios were calculated by Chi-square test. With respect to genotype C, HCC-associated SNVs discovered by NGS were determined after adjustment for multiple testing by Benjamini and Hochberg’s false discovery rate (FDR) method for those HCC-associated variants evidenced by meta-analysis and those SNVs with P values of OR<0.05 by NGS analysis.
Figure Legends

Figure 1. HBV genome-wide characterization of substitutions in genotype B (HCC, n=40; non-HCC, n=47). Comparing HCC and non-HCC groups, the distribution of total substitutions along the hepatitis B virus (HBV) genome and their frequencies within quasispecies were demonstrated. HBV genome: four HBV genes and their nucleotide positions; Sequencing coverage: the average sequencing coverage at each site; Mutation rate per 9-bp: proportions of nucleotides with SNVs in each 9-bp window shown as blue-red color scale at one base sliding step; Average SNV II frequency: the average values of SNV II frequency (SNV II \( f_{m/quasi} \)); SNV II frequency in each patient: SNV II and their SNV frequency (\( f_{m/quasi} \)) at each site shown as blue-red color scale in each patient. SNV I was defined as the dominant variant of HBV in non-HCC group. SNV II was the variant other than SNV I at the same nucleotide position. Almost all SNV I have SNV frequency higher than 80% either in HCC or non-HCC group. H: HCC group. N: Non-HCC group.

Figure 2. Validation of NGS-defined HCC-associated SNVs by a systemic literature review and meta-analysis. (A) The proportions of NGS-defined
HCC-associated SNVs among HBV variants with different levels of supporting evidence declined significantly with decreasing levels from Level A to Level D. One hundred and sixty-seven HBV variants, which had been studied previously, were categorized into 4 groups according to supporting evidence associated with HCC. The proportion of NGS-defined HCC-associated SNVs among each group declined significantly along with decreasing levels of supporting evidences from A to D (chi-square test for the 2x4 contingency table, $P<0.0001$). A: HCC-associated HBV variants favored by meta-analysis with at least 4 studies; B: HCC-associated HBV variants favored by at least one study with fewer than 4 relevant studies; C: HBV variants not associated with HCC provided by all studies with fewer than 4 relevant studies; D: HBV variants not associated with HCC provided by meta-analysis with at least 4 studies. (B) Odds ratios of NGS analysis in 12 HCC-associated HBV variants with Level A evidence. Level A included HCC-associated HBV variants evidenced by meta-analysis with at least 4 studies. Nine of them were identified by NGS analysis. * 1766T and 1768A mainly expressed in genotype A and D HBV and were not identified as HCC-associated SNVs by this NGS analysis. † 1762T could be identified as HCC-associated SNVs in a subgroup analysis of HBeAg-positive
patients with genotype C HBV infection by NGS analysis (odds ratio: 6.3, 95% confidence interval [CI] = 1.1 to 34.0). (C) The distribution of cut-off values of HCC-associated SNV frequency \(f_{m/quasi}\) among the genotype B and genotype C groups. Red circles represent SNVs reported to be associated with HCC. Yellow circles represent SNVs investigated in only one study, but with a negative result. Empty circles represent SNVs not investigated previously. The SNV, nt53C, was the risk variant for HCC in both genotypes B and C groups.

Figure 3. Genome-wide deletion distribution of HBV comparing HCC and non-HCC patients. Y axis represents the proportion of patients with deletion for each nucleotide or amino acid among patients with any deletion mutation of HBV full genome. The number of patients with any deletion was 10 in HCC group and 8 in non-HCC group for genotype B, while 34 in HCC group and 32 in non-HCC group for genotype C, respectively. Inverted Y axis represents the average deletion frequency \(f_{del/quasi}\) for each nucleotide or amino acid among patients with any deletion mutation of HBV full genome. (A) Full genome of genotype B HBV; (B) Full genome of genotype C HBV. The upper portions of (A) and (B) show ORFs for
all genes and 4 domains of the P gene. (C) PreS region of genotype B HBV; (D) PreS region of genotype C HBV. The upper portions of (C) and (D) show B and T cell epitopes and 5 functional sites (S, T, N, M, and P) among preS region. \( f_{\text{del/quasi}} \): the deletion frequency within one subject, i.e. the number of reads with deletion/ the number of total reads covering that site in one sample. TP: terminal protein of polymerase; SP: spacer domain of polymerase; RT: reverse transcriptase of polymerase; RH: RNase H of polymerase. S: S-promoter; T: topology (CAD or Hsc70 binding site); N: nucleocapsid binding site; M: the start codon of M protein; P: pHSA site.

Figure 4. Ingenuity Pathway Analysis of microarray data and the proposed model for transformation of the Huh7 cell line with different HBV small S mutant expression. (A) Biological functions in transfected Huh7 cell with different three genotype B HBV small S mutants (T216C, A273G and T216C/A273G double mutants); (B) Biological functions in transfected Huh7 cell with different seven genotype C HBV small S mutants (A293G, C446G, A456G, A293G/C446G, C446G/A456G, A293G/A456G, and A293G/C446G/A456G triple mutants). Upper:
Key functional differences in molecular and cellular functions; Middle: Key functional differences in physiological system development and function; Lower: Key functional differences in disease and disorders. Bars indicate significance. The length of the bar indicates only that the differentially expressed proteins are related to this pathway but does not indicate upregulation or downregulation of the pathway.

Figure 5. Overview of HBV variants in hepatocarcinogenesis.
Figure 1.tif
Figure 2.tif
Figure 4.tif

Title: Hepatocellular carcinoma-associated single-nucleotide variants and deletions identified by the use of genome-wide high-throughput analysis of hepatitis B virus.

Date: 2017-10


Persistent Link: http://hdl.handle.net/11343/293330