Next-generation sequencing revealed divergence in deletions of the preS region in the HBV genome between different HBV-related liver diseases

Jian’an Jia,1,2 Xiaotao Liang,3,4 Shipeng Chen,1 Hui Wang,1,5 Huiming Li,1 Meng Fang,1 Xin Bai,4 Ziyi Wang,6 Mengmeng Wang,1 Shanfeng Zhu,3,4 Fengzhu Sun,4,7 and Chunfang Gao1,*

Abstract
In order to investigate if deletion patterns of the preS region can predict liver disease advancement, the preS region of the hepatitis B virus (HBV) genome in 45 chronic hepatitis B (CHB) and 94 HBV-related hepatocellular carcinoma (HCC) patients was sequenced by next-generation sequencing (NGS) and the percentages of nucleotide deletion in the preS region were analysed. Hierarchical clustering and heatmaps based on deletion percentages of preS revealed different deletion patterns between CHB and HCC patients. Intergenotype comparison also indicated divergence in preS deletions between HBV genotype B and C. No significant difference was found in preS deletion patterns between sera and matched adjacent non-tumour tissues. Based on hierarchical clustering, HCC patients were classed into two groups with different preS deletion patterns and different clinical features. Finally, the support vector machine (SVM) model was trained on preS nucleotide deletion percentages and used to predict HCC versus CHB patients. The prediction performance was assessed with fivefold cross-validation and independent cohort validation. The median area under the curve (AUC) was 0.729 after repeating SVM 500 times with fivefold cross-validations. After parameter optimization, the SVM model was used to predict an independent cohort with 51 CHB patients and 72 HCC patients and the AUC was 0.727. In conclusion, the use of the NGS method revealed a prominent divergence in preS deletion patterns between disease groups and virus genotypes, but not between different tissue types. Quantitative NGS data combined with a machine learning method could be a powerful approach for prediction of the status of different diseases.

INTRODUCTION
The absence of a proofreading function for hepatitis B virus (HBV) polymerase results in a high mutation rate during virus replication, which is thought to play a substantial role in the progression of HBV-associated hepatocellular carcinoma (HCC) [1]. Mutations, especially deletions in the preS region of the HBV genome, occur quite often, and some mutations are associated with HCC development [2, 3]. A meta-analysis demonstrated that infection with preS mutants increased the risk of HCC with an odds ratio of 3.77 [4]. Moreover, numerous clinical and experimental studies have provided strong evidence for the potential role of HBV preS mutants in the carcinogenesis of hepatocytes. Studies have also shown the association between preS deletion accumulation with antiviral treatment on chronic hepatitis B (CHB) patients [5] and the predictive value of the preS deletion mutants for advanced liver disease [6, 7].

Since preS mutants, as well as deletions, have been closely associated with HCC, it is important to investigate if the deletion patterns of this region can predict the progress of liver disease. Recently, with high-throughput and sensitivity, next-generation sequencing (NGS) could present a more precise
genomic profile of the preS deletion pattern, making them applicable as biomarkers of liver disease. Therefore, in the present study, NGS was used to detect the divergence in deletion distribution and abundance of the preS region between CHB and HCC patients and between HBV genotype B and C. Moreover, the deletion patterns of the preS region from matched serum, tumour tissue (TT) and adjacent non-tumour tissue (ANTT) samples were also compared. Clinical features and deletion patterns in the preS region of different HCC patient groups separated by hierarchical clustering were compared. Finally, a canonical machine learning method, known as the support vector machine (SVM) model, was fitted based on preS nucleotide deletion percentages of CHB and HCC patients, and the performance for HCC prediction was evaluated 500 times with fivefold cross-validation and independent cohort validation.

RESULTS

Clinical features and virological characteristics of CHB and HCC patients

Baseline demographics, liver biochemistry tests and virological data of 94 HCC patients and 45 CHB patients are listed in Table S1 (available in the online Supplementary Material). HCC patients have higher serum α-fetoprotein (AFP) levels, and most of the HCC patients are estimated to be at the tumour node metastasis I-II stage. Corresponding data about the independent cohort, including 51 CHB and 72 HCC patients, are presented in Table S2.

Median filtered reads per sample of the preS region of TTs, ANTTs and sera in HCC patients are 4987 (interquartile range (IQR) 2608–16,066), 4896 (IQR 2194–8972) and 5966 (IQR 1205–13,865) respectively, and the counterpart of the serum samples in CHB patients is 3951 (IQR 1205–13,865) and also displayed a non-significant difference. In the independent cohort, median filtered reads per sample of the preS region in HCC and CHB patients were 15,576 (IQR 8350–23,980) and 16,415 (IQR 13,693–20,413) and also displayed a non-significant difference.

PreS region deletion profiles

Owing to the high sensitivity and depth of NGS, deletions in the preS region were quantified and deletions in each nucleotide site were transformed into the deletion percentage of such a site. Subsequently, nucleotide deletion heatmaps in the HBV preS region of the CHB and HCC patients were plotted in Fig. S1. Of the 94 HCC patients, deletions of the preS region in ANTTs were detected in every individual with a median percentage of 8.31 % (IQR 5.50–11.49 %). In contrast, deletions of the preS region in sera of 45 CHB patients had a median percentage of 2.75 % (IQR 0.42–4.22 %), which is remarkably lower than that in the HCC patients (Fig. 1b).

Nucleotides from nt2850 to nt2863 (preS1 aa1-6) in the preS1 region have a similar deletion percentage in both CHB (4.34 % [IQR 2.07–10.86 %]) and HCC groups (6.71 % (IQR 2.70–11.79 %)), and this region had the most common deletions in CHB patients. In the preS2 region, nt2-55 (preS2 aa5-22) of CHB patients showed median nucleotide deletion percentages of 3.24 % (IQR 0.53–6.38 %) (see Fig. 1a). The most frequent nucleotide deletions in HCC patients occurred in nt2997-3197 (preS1 aa50-117, 10.5 % (IQR 5.87–17.65 %)) and nt2-55 (preS2 aa5-22, 11.06 % (IQR 6.24–17.06 %), Fig. 1a). When deletions in several more divergent fragments were compared, several fragments such as nt2997-3197 (preS1 aa50-117), nt3051-3078 (preS1 aa68-77), nt3152-3197 (preS1 aa102-117) in the preS1 region and nt2-55 (preS2 aa5-22) in the preS2 region had a significantly higher frequency of deletions in HCC patients compared to that of CHB patients (Fig. 1b).

Comparison of deletion patterns of the preS region originating from different tissue types

In order to explore whether divergence in preS deletion patterns existed among samples from different tissues, deletion distribution and abundance of preS deletions were compared in matched serum, TT and ANTT sample groups from 40 patients. In general, deletions in the preS region from different tissues displayed coincident distributions and similar frequencies, and preS deletions from sera showed the highest occurrence (Fig. 2a). Matched comparison revealed that prominent divergent deletion percentages were investigated between serum and TT samples (Fig. 2b, upper left, P=0.01), while no divergence was found in preS deletion between serum and ANTT samples or between TT and ANTT samples. Furthermore, no significant difference in deletion abundance was observed among serum, TT and ANTT samples in several fragments of the preS region including nt2850-2863 (preS1 aa1-6), nt2997-3197 (preS1 aa50-117), nt3051-3078 (preS1 aa68-77), nt3152-3197 (preS1 aa102-117) and nt2-55 (preS2 aa5-22) (Fig. 2b). Hierarchical clustering based on nucleotide deletion percentages from matched serum, TT and ANTT samples from 40 patients was performed, and the dendrogram in Fig. S2 shows four groups of serum, TT and ANTT samples from the same patient clustering in the same clade including patients A74, F50, F9 and F61. Twelve pairs of serum and ANTT samples from the same patient are in the same clade such as patient F39 and so on. In addition, four pairs of TT and ANTT samples from the same patient were presented in the same clade such as patient A106 and so on, while only one pair of TT and serum samples from patient F53 was presented together in the dendrogram. Taken together, higher similarity was found in preS deletion between ANTT and serum samples, but not between TT and serum samples. We also used principal coordinate analysis (PCoA) to project the nucleotide deletion percentages onto two-dimensional space; none of the three tissue type samples cluster together (Fig. S3).

Comparison of deletion patterns in the preS region between different HBV genotypes

The deletion landscapes in preS of HBV genotype B and C were compared, and deletion abundance of the preS region in HBV genotype C was remarkably higher than that in
HBV genotype B (Figs 3a and S4 upper left). Furthermore, average nucleotide deletion percentages in nt2997-3197 (preS1 aa50-117) of the preS1 region and nt2-55 (preS2 aa5-22) of the preS2 region were significantly higher in HBV genotype C (Fig. S4). Both distribution and abundance of deletions in the preS region showed divergent features between genotype B and C virus in HCC patients (Fig. 3c), while in CHB patients, no significant difference was observed between the two genotypes (Fig. 3b). Concretely speaking, nt3152-3191 (preS1 aa102-115), nt3205-1-15 (preS2 aa1-9) and nt15-55 (preS2 aa9-22) harboured the most frequent deletions in genotype B virus of HCC patients and nt2850-2863 (preS1 aa1-6), nt2997-3197 (preS1 aa50-117) and nt2-55 (preS2 aa5-22) had the most profuse deletions in genotype C virus of HCC patients. However, in HCC patients, only average nucleotide deletion percentages in nt3058-3078 (preS1 aa71-77) of the preS1 region were prominently higher in the HBV genotype C versus genotype B (Fig. S5) and as such, no statistically significant difference was observed in average nucleotide deletion percentages of the total preS region or fragments of nt2850-2863 (preS1 aa1-6), nt2997-3197 (preS1 aa50-117), nt3205-1-15 (preS2 aa1-9) and nt2-55 (preS2 aa5-22) between HBV genotype B and C, although the genotype C preS region did show higher deletion percentages in these fragments (Fig. S5).

Hierarchical clustering analysis of deletions in the preS region

Hierarchical clustering analysis based on a cosine distance matrix between sample nucleotide deletion percentages in the preS region was carried out, and all of the whole patients, including CHB and HCC patients were classified...
into two groups, most of which properly corresponded to the patients’ diagnosis (Fig. 4a). Individuals in cluster A mostly belonged to the HCC group and cluster B aggregated patients mostly belonged to the CHB group.

Furthermore, 87 HCC patients with intact clinical characteristics were separated into two groups by hierarchical clustering (Fig. 4b). Nucleotide deletions in the cluster C group seemed to reside in the preS1 region, while deletions in the cluster D group mostly aggregated in the preS2 region. Clinical features of the two cluster groups were compared (Table 1), and patients in the cluster D group showed significantly higher lectin-reactive α-fetoprotein (AFP-L3) levels ($P=0.01$), lower albumin levels ($P=0.03$), relatively fewer tumours ($P=0.09$) and large tumour size ($P=0.17$).

**Support vector machine model fitting and validation**

Previous studies have applied viral factors as predictors of HCC [8, 9], therefore, in this study, we attempted to employ quantitative high-dimensional deletion percentages of the preS region to predict the incidence of HCC. Comparison of nucleotide deletions among different tissue types showed no significant difference in preS nucleotide deletion percentages between serum and ANTT samples. Consequently, the radial SVM model was trained on preS nucleotide deletion...
percentages from CHB serum samples and HCC ANTT samples. As a canonical machine learning technique, SVM is a binary numerical classifier designed to judge whether a query data point belongs to an arbitrarily defined class [10, 11]. SVM has been broadly used in bioinformatics studies in the past [10, 12]. Two of its tuning parameters involved in training procedure, cost and gamma, were optimized. After parameter optimizing and model training, fivefold cross-

Fig. 3. Comparison of deletion patterns in the preS region between different HBV genotypes. Comparison of deletion pattern and abundance (average nucleotide deletion percentage) in the preS region between HBV genotype B and genotype C in total patients (a), CHB patients (b) and HCC patients (c). B/C: HBV B/C mixed genotype. The y-axis represents the average deletion percentage per sample in an individual nucleotide site (x-axis).
validation was used to evaluate prediction performance for HCC and the process was repeated 500 times. SVM performance is summarized in Table 2 and the median area under the curve (AUC) is 0.729. In addition, 51 CHB patients and 72 HCC patients were used as an independent validation cohort. After parameter optimization with fivefold cross-validation on preS nucleotide deletion percentages of 45 CHB and 94 HCC patients, the best cost and gamma of the SVM model were 8 and 0.0625 respectively. The SVM was used to predict the HCC patients versus CHB patients in the independent cohort and the performance was summarized in Table 2. The AUC and accuracy of independent validation were 0.727 and 0.764 respectively, which were similar to the cross-validation. The specificity of HCC prediction in the independent cohort was only 0.451, and a serious false-positive error was observed in HCC prediction.

**DISCUSSION**

In the present study, a precise deletion profile of the preS region was depicted thanks to the high sensitivity of NGS. In the individual HBV carrier, viruses with or without nucleotide deletions in the preS region exist together (Fig. 4), just like variant quasispecies with mutations in other regions of the HBV genome [13]. It should be noted that quasispecies may be barely detectable with canonical Sanger sequencing. The coexistence of deleted preS variants and intact ones in a single sample was consistent with previous studies [14–16]. The coexistence of wild-type and partially deleted preS sequences suggested that deletion mutants may replicate with the help of a wild-type preS gene and this combination of virus variants may be more adaptable in hepatocytes. [16]. In addition, although the...
deletion locations and abundance could be demonstrated in the heatmap (Fig. 4), the real nucleotide deletion landscape may be more complex since data processing compressed the whole complex quasispecies into an average single-species level, accompanying signal attenuation. The same as numerous studies [5, 6, 17], deletion in HCC patients occurred far more frequently than it did in CHB patients, and some preS fragments in the HCC group had two–three times more nucleotide deletion percentages compared to the CHB group (Fig. 1), such as nt2997-3197 (preS1 aa50-117), nt3051-3078 (preS1 aa68-77), nt3152-3197 (preS1 aa102-117) in the preS1 region and nt2-55 (preS2 aa5-22) in the preS2 region. All of the divergent deletion regions were located in the C-half of the preS1 region and overlapped with B cell epitopes (preS1 aa72-78, preS1 94–105, preS1 aa106-117, preS2 aa1-26), or the T cell epitope (preS1 94–117) [18]. Additionally, the divergent regions overlapped with some functional domains, including the heat shock protein 70 binding site (preS1 aa74-118), cytosolic anchor-age determinant (preS1 aa81-105), nucleocapsid binding site (preS aa103-127), polymerized human serum albumin binding site (preS2 aa3-16), S-promoter (nt3045-3180), CCAAT binding factor binding site (nt3137-3147) and transaction domain (preS2 aa1-53) [18]. Since HCC patients harboured more deletions in immune epitopes and functional domains within the preS region than CHB patients, preS deletions may play important roles in the advancement of HBV infection via affecting immune identification, intermolecular interactions and gene expression.

Since numerous and substantial divergence occurs between hepatocytes in TTs and ANTTs, such as cell differentiation and proliferation or immune microenvironment, viruses living in TTs are assumed to be dissimilar to their counterparts in ANTTs, which was seldomly studied in the past. The lower sensitivity of Sanger sequencing makes it ill-equipped to deeply probe the subtle diversity among different tissue

| Table 1. Comparison of the features of two groups of HCC patients classified by hierarchical clustering based on deletion patterns of the preS region in the HBV genome |
|---------------------------------|----------------|----------------|
| Gender                         | Cluster C (n=51) | Cluster D (n=36) |
| Female                         | 8              | 6              |
| Male                           | 43             | 30             |
| Age (years)                    | 51 (42–60)     | 53 (45–61)     |
| Tumour size (cm)               | 5.2 (3.15–8)   | 6.4 (4–10.4)   |
| Tumour no.                     |                |                |
| Single                         | 11             | 14             |
| Multiple                       | 40             | 22             |
| TNM stage                      |                |                |
| I                              | 25             | 15             |
| II                             | 16             | 12             |
| III                            | 10             | 9              |
| TBIL (μmol l⁻¹)                | 15.1 (11.4–19.6) | 14.0 (11.0–17.1) |
| ALB (g l⁻¹)                    | 41.2 (39.5–43.7) | 39.7 (37.075–42.4) |
| ALT (U/L)                      | 42.3 (25.9–67.7) | 46.6 (34.5–67.5) |
| GGT (U/L)                      | 80 (46–159)    | 100 (58–153)   |
| AFP (ng ml⁻¹)                  | 158.3 (9.25–1653.5) | 409.6 (41.2–5684.5) |
| AFP-L3 (%)                     | 3.75 (0.00–13.30) | 11.4 (5.63–22.83) |
| Serum HBV DNA (log₁₀ IU ml⁻¹)  | 4.56 (3.89–5.53) | 5.09 (4.33–5.84) |
| HBV genotype                   |                |                |
| Negative                       | 33             | 17             |
| Positive                       | 18             | 16             |
| B                              | 5              | 5              |
| C                              | 43             | 30             |
| B/C                            | 3              | 1              |

*P<0.05.
ANTT, adjacent non-tumour tissue; AFP, α-fetoprotein; AFP-L3, lectin-reactive α-fetoprotein; ALB, albumin; ALT, alanine aminosferase; GGT, γ-glutamyltransferase; HBeAg, hepatitis B e antigen; TBIL, total bilirubin; TNM, tumour node metastasis; TT, tumour tissue.

| Table 2. Performance of the SVM training model on preS nucleotide deletion percentages to predict HCC patients versus CHB patients |
|---------------------------------|----------------|----------------|
| Fivefold cross-validation       | Independent cohort validation |
| AUC                             | Median         | Interquartile range |
| AUC                             | 0.729          | 0.698–0.755       |
| Accuracy                        | 0.755          | 0.741–0.770       |
| Sensitivity                     | 0.830          | 0.787–0.872       |
| Specificity                     | 0.622          | 0.556–0.667       |

AUC, area under the curve; CHB, chronic hepatitis B; HCC, hepatocellular carcinoma; SVM, support vector machine.
types, so the divergence of deletions among different tissues in the preS region was investigated with the more sensitive NGS method in the current study. In general, comparison of matched tissue samples disclosed significant divergence between serum and TT samples, while no significant divergence was found between serum and ANTT samples or between TT and ANTT samples (Fig. 2). These phenomena may be explained in one of three ways. First, virus in serum mostly originated from normal hepatocytes because of their predominance in liver in most HCC patients, providing optimal suitability for virus growth. Second, the different immune selective pressures of each sample type may also play a role. As the primary immune component involved in host anti-tumour reaction to solid tumours [19], CD8+ T-cell function is always suppressed in the tumour microenvironment, producing weak immune selective pressure [20] which, in turn, may result in tolerance to highly antigenic preS antigens with fewer deletions. In contrast, unimpaired immune function in adjacent benign tissue and serum with strong immune selective pressure may contribute to more preS deletion variants, thus facilitating virus escape from immune response. Finally, discrepancy of response to antiviral drugs between virus living in TTs and ANTTs may also contribute to the divergence in preS deletions among sample types.

Several studies have implied that carriers infected by HBV genotype C have a higher incidence of preS deletions than those infected by HBV genotype B (Fig. 3) [18, 21], and the results of our study also concurred with those findings. Two studies [18, 22] found higher incidence of preS deletions of HBV genotype C regardless of different clinical stages, while significant intergenotype divergence of preS deletions was only observed in HCC patients in our study, which is similar to a recent study in New Taipei City [21]. Thus, the association among preS deletions, HBV genotypes and clinical stages should be further studied. On the other hand, highly frequent deletions in nt3150-3197 (preS1 aa68-117) of preS1 and nt5-55 (preS2 aa5-22) of preS2 which were observed (Fig. 3), irrespective of different HBV genotypes, also suggested that such regions may be involved in the development of HCC.

Hierarchical clustering based on a distance matrix is commonly used for the aggregation of samples with similar properties [23, 24]. In this study, several types of distance matrices were computed based on nucleotide deletion percentages in the preS region for hierarchical clustering, and the cosine distance matrix finally showed the best performance (Fig. 4a), better than Manhattan, Euclidean, maximum distance matrix or correlation matrix. The cosine distance matrix maps nucleotide deletion percentages to vectors in a high-dimensional space and computes the distance between two vectors as the cosine of the angle between them. Cosine similarity is the most commonly used method to compute the similarity between directional data in the vector-space mode [25]. The better performance of the cosine distance matrix in the present paper may imply that the preS region of HCC patients may share similar deletion fragments. Such deletion features, even partially, may be associated with higher carcinogenic risk. In addition, k-means clustering was also carried out on distance matrices, but the clustering method could not outstrip the results of the cosine distance matrix.

Hierarchical clustering also classified HCC patients into two clusters, and deletions of preS in the cluster C group seemed to be located in the preS1 region, while deletions in the cluster D group mostly distributed in the preS2 region (Fig. 4b). As one of the histological hallmarks of chronic HBV infection, ground-glass hepatocytes (GGHs), which are characterized by remarkable aggregation of viral surface proteins in ER, could be grouped into two types: type I GGHs harbouring mutated L protein with deletions within the preS1 region, and type II GGHs harbouring mutants with deletions or abolition over the preS2 region [26, 27]. Of interest, cluster C and cluster D in our study are consistent with type I and type II GGHs and type II GGHs are distributed in large cell clusters because of their higher proliferation, which seemed consistent with the relatively large tumour size in the cluster D group of HCC patients (Table 1).

For better disease classification and prediction based on preS deletions, an SVM model was trained, and the most robust radial basis function kernel was employed. After training, the median AUC of SVM reached 0.729 in fivefold cross-validation with 500 iterations and was close to the diagnostic power of AFP (Table 2), as previously reported [28, 29]. Because of the lack of intact serum AFP levels of CHB patients, it is regrettable that the diagnostic efficacy of the present SVM model could not be compared with AFP directly.

In addition, an independent cohort was used to evaluate the SVM model and the AUC reached 0.727 (Table 2). The SVM model gave superior sensitivity, but inferior specificity implying that more CHB patients were misdiagnosed as HCC patients. The high rate of false-positive results may be attributed to the heterogeneity between HBV quasispecies in different cohorts or the limited sample size in our cohort. Furthermore, besides viral factors, other etiological factors may also affect the development of HCC, and the combination of different probable factors may increase prediction efficiency. Nevertheless, owing to the high sensitivity (0.986) of the SVM model, preS deletion detection could be used in the monitoring of HBV infection advancement as well as early biomarkers of HCC. The characteristic of high sensitivity and low specificity also makes this SVM model complementary to the high specificity but low sensitivity of AFP [30]. The superior performance of the SVM model trained purely on preS deletions implied that deletions in the preS region play essential roles in hepatocarcinogenesis, as described in numerous studies [31–33].

We should point out some limitations of our study. First, comparison of preS nucleotide deletion percentages were based on a single summary value of each sample and during
the data processing, lots of information may be lost. So the comparison between groups may not be comprehensive and more precise and adaptable methods should be developed to address such a question. Second, the SVM model was trained on preS nucleotide deletion percentages from tissues of HCC patients and sera of CHB patients. Although no significant divergence was found in preS deletion, when different tissue types from the same patient were compared, nuances in preS deletion patterns between ANTT and serum samples may still be present owing to their substantial histological difference. Third, the performance of the SVM model in the present study is still unsatisfactory, but the prediction performance may be further improved if other viral or clinical features were to be combined. Finally, the sample size is still small and more samples should be included for the evaluation of SVM in the future.

Conclusions
In summary, based on quantitative NGS data, deletion occurrence in the preS region was prominently higher in HCC patients versus CHB patients and higher in HBV genotype B patients than in genotype C patients. Meanwhile, no significant divergence in preS deletion patterns was observed between viruses from serum and matched ANTT samples. Furthermore, SVM trained on preS deletions showed superior performance on HCC prediction and such machine learning methods that show excellent performance in the treatment of high-dimensional NGS data could be further applied in related fields.

METHODS
Patients and samples
A total of 94 HBV-related HCC patients who received complete surgical resection between March 2011 and November 2012 at Eastern Hepatobiliary Surgery Hospital, Shanghai, China were recruited. Serum, TT and paired ANTT samples were collected respectively. As controls, 45 CHB patients were also included, and their corresponding serum samples were collected. Furthermore, an independent cohort of 51 CHB and 72 HCC patients was recruited from January 2013 to January 2015 according to inclusion criteria described above at Eastern Hepatobiliary Surgery Hospital, and serum samples were collected. The study was approved by the Ethics Committee of the Eastern Hepatobiliary Hospital, and written informed consent was obtained from all patients.

HBV DNA extraction and preS region amplification
HBV genomes were extracted from frozen TTs, ANTTs or 200 µl of serum using a QIAamp DNA Mini kit (QIAGEN GmbH, Hilden, Germany), according to the manufacturer’s instructions. HBV DNA levels were determined with the COBAS Amplicor HBV Monitor Test (Roche Diagnostics, Pleasanton, CA, USA). PreS regions of 23 HBV (genotype A–H) references on GenBank (www.ncbi.nlm.nih.gov/projects/genotyping/view.cgi?db=2) were aligned for primer design and the preS region (nt2801-3215-1-79, 495 bp) was amplified with a pair of primers: 5’-CGCCTCATTYTKYGGGTCA-3’ (forward, nucleotides2801-2819), and 5’-TCCKGAACTGGAGCCACC-3’ (reverse, nucleotides 62 to 79). According to the HBV genotype C genome (accession number: X04615), the expectant amplicon length of the preS region used for analysis included preS1 (nt2820-3204) and preS2 (nt3205-3215-1-61) with 457 base pairs after excluding primer regions.

Illumina sequencing of the preS region
Overall, preS regions from 94 pairs of TTs and ANTTs of HCC patients and 45 sera of CHB patients were used for sequencing, and preS regions from 40 matched sera of HCC patients were additionally sequenced. PCR products of the preS region were purified with Agencourt AMPure XP beads (Beckman Coulter, Beverly, MA, USA) and the concentrations of the PCR products were measured using the Qubit dsDNA HS assay kit (Invitrogen, Carlsbad, CA, USA). A library of PCR products of the preS region was prepared using the TruSeq DNA PCR-Free sample preparation kit (Illumina, San Diego, CA, USA), according to the manufacturer’s instructions. The PCR product libraries were mixed with 1 % PhiX control libraries and run on a MiSeq sequencer (Illumina) for paired-end sequencing, according to Illumina’s protocol. Finally, imaging-based detection of the fluorescent signals was analysed using the MiSeq control software (Illumina) and was transferred to sequence data in the FASTQ format. All the raw sequences were submitted to NCBI Sequence Read Archive, and the accession number is SRP096021.

Sequence read mapping and deletion analysis
Quality estimation of raw reads was performed with the fastqc tool (www.bioinformatics.babraham.ac.uk/projects/fastqc/), and reads with a quality score under 20 were discarded. After quality filtration and adapter removal, paired-end reads were joined with FLASH v1.2.10 (http://ccb.jhu.edu/software/FLASH/). Merged preS region sequences were aligned against the reference HBV genome (AF100309) with local BLASTn (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) and e value cutoff was set as 1e-5 to reduce fragmentation result when running local BLAST. Nucleotide deletion localization and frequency in the preS region were attained with compiled Perl script. The nucleotide deletion percentage is defined as 100×(counts of reads with deletion in single nucleotide site)/(total number of reads including such a nucleotide site). Furthermore, HBV Star software[34] was used for virus genotyping and the preS region of 23 HBV (genotype A–H) on GenBank (www.ncbi.nlm.nih.gov/projects/genotyping/view.cgi?db=2) was used as a reference. Each sample had a set of genotype fractions, and the genotype of the most predominant fraction was set as the sample’s genotype.

Hierarchical clustering and support vector machine training
The cosine distance matrix between the samples originated from nucleotide deletion percentages of all samples, was calculated using the R software package (proxy, https://CRAN.
R-project.org/package=proxy), and classification of all patients was carried out using the hierarchical clustering method. Hierarchical clustering based on nucleotide deletion percentages from matched serum, TT and ANTT samples from 40 patients was performed, and the dendrogram was plotted using the R software package (ape, https://CRAN.R-project.org/package=ape). PCoA was also used to project the nucleotide deletion percentages of the three tissue types onto two-dimensional space and the PCoA plot was plotted using R software package (ggtreeplot2, https://CRAN.R-project.org/package=ggtreeplot2). The 94 HCC patients were further separated with hierarchical clustering, and preS deletion pattern and clinical features between the separated groups were compared. Moreover, nucleotide deletion percentages of preS regions from ANTTs of 94 HCC patients and sera of 45 CHB patients were used to train the SVM model. The radial SVM model was trained using the R software package (e1071, https://CRAN.R-project.org/package=e1071), and fivefold cross-validation was performed as follows. All samples, including 45 CHB and 94 HCC, patients were randomly separated into five groups. For each iteration, each group of patients were predicted by the SVM model trained on the other four groups of patients, and finally, all 139 patients were given probabilities of HCC occurrence. Hence, the AUC of the SVM model for predicting HCC occurrence could be calculated, and the iterations were repeated 500 times. Furthermore, a cohort of 72 HCC and 51 CHB patients was used for independent validation. Parameters gamma and cost were optimized and SVM training was performed based on the training cohort (45 CHB and 94 HCC patients). The probability of HCC was predicted by the SVM model and was given to each patient in the validation cohort. The AUC of the independent validation was calculated. The AUC of receiver operating characteristic curves was assessed using the R software package (Daim, https://CRAN.R-project.org/package=Daim).

**Statistical analysis**

Statistical analysis was performed using the R software program, version 3.2.3. The results of continuous variables were expressed as the median with IQR, and preS deletion pattern comparison between different patient types, tissue types and virus genotypes was performed with Student’s t-test, the Mann–Whitney test or the paired Wilcoxon signed-rank test according to data distribution. Deletion abundance of the defined fragment of the preS region was compared between groups according to the average deletion percentage per nucleotide site within the region. In addition, nucleotide deletion percentage heatmaps of the preS region were displayed with the R software package (pheatmapmap, https://CRAN.R-project.org/package=pheatmap).

**Funding information**

This study was supported by the China National Key Projects for Infectious Disease (Nos. 20122×10002-016); National Natural Science Foundation of China (Nos. 81572072, 81171664); Key Projects of Science and Technology Commission of Shanghai Municipality (11JC1416400) to C.F. Gao. We thank the Shanghai Amplicongene Corp. for assistance in this study.

**Conflicts of interest**

The authors declare that there are no conflicts of interest.

**References**