RESEARCH ARTICLE

The characteristic of the synonymous codon usage and phylogenetic analysis of hepatitis B virus

Xiaoming Qi[1,2,3] · Chaojun Wei[1,2,3] · Yonghong Li[1,2,3] · Yu Wu[1,2,3] · Hui Xu[1,2,3] · Rui Guo[1,2,3] · Yanjuan Jia[1,2,3] · Zhenhao Li[1,2,3] · Zhenhong Wei[1,2,3] · Wanxia Wang[1,2,3] · Jing Jia[1,2,3] · Yuanting Li[1,2,3] · Anqi Wang[1,2,3] · Xiaoling Gao[1,2,3]

Abstract

Background Hepatitis B virus (HBV) infection is a crucial medical issue worldwide. The dependence of HBV replication on host cell machineries and their co-evolutionary interactions prompt the codon usage pattern of viral genes to translation selection and mutation pressure.

Objective The evolutionary characteristics of HBV and the natural selection effects of the human genome on the codon usage characteristics were analyzed to provide a basis for medication development for HBV infection.

Methods The codon usage pattern of sequences from different HBV genotypes of our isolates and reference HBV genome sequences downloaded from the National Center for Biotechnology Information (NCBI) database were analyzed by computing the relative synonymous codon usage (RSCU), nucleotide content, codon adaptation index (CAI) and the effective number of codons (ENC).

Results The highest ENC values were observed in the C genotypes, followed by the B genotypes. The ENC values indicated a weak codon usage bias (CUB) in HBV genome. The number of codons differentially used between the three genotypes was markedly higher than that of similarly used codons. High CAI values indicated a good adaptability of HBV to its host. The ENC plot indicated the occurrence of mutational pressure in the three genotypes. The mean Ka/Ks ratios in the three genotypes were lower than 1, which indicated a negative selection pressure. The CAI and GC3% plot indicated the existence of CUB in the HBV genome.

Conclusions Nucleotide composition, mutation bias, negative selection and mutational pressure are key factors influencing the CUB and phylogenetic diversity in HBV genotypes. The data provided here could be useful for developing drugs for HBV infection.

Keywords Hepatitis B virus · Codon usage pattern · Phylogenetic evolution · Mutation pressure · Translation selection

✉ Xiaoling Gao
gaoxl008@hotmail.com

1   The Institute of Clinical Research and Translational Medicine, Gansu Provincial Hospital, 204 Donggang West Road, Chengguan District, Lanzhou 730000, China

2   NHC Key Laboratory of Diagnosis and Therapy of Gastrointestinal Tumor, Gansu Provincial Hospital, Lanzhou 730000, China

3   Gansu Provincial Biobank and Bioinformation Engineering Research Center, Lanzhou 730000, China

Introduction

Hepatitis B virus (HBV) infection, the main causal factor for liver diseases such hepatitis, cirrhosis and liver cancer, is a significant global health concern worldwide (Benhenda et al. 2013; Binh et al. 2019; Bonvicino et al. 2014; Kim et al. 2007, 2017; Sarkar and Chakravarty 2015; Shih et al. 2018). The incidence of HBV is annually increasing, and developing suitable therapeutic approaches is extremely difficult (Li et al. 2017; Nelson et al. 2016; Stasi et al. 2016). HBV is a small particle retroid virus characterized by its circular partially double-stranded DNA genome (Gerelsaikhan et al. 1996), with four over-lapping open reading frames including large S region, PreC/C, X, P gene (Kramvis and Kew 1998; Ma et al. 2015).

The large S protein is present in external (Le-HBsAg) and internal (Li-HBsAg) topological conformations (Churin et al. 2015). The Le-HBsAg conformation allows the attachment of HBV to cellular receptors, which is the initial step of viral infection. In Li-HBsAg conformation, the large S protein participates in virion morphogenesis and regulates the contact with the nucleocapsid (Taylor 2013). The surface protein gene S encoding for large, middle and small S proteins (De Maddalena et al. 2007) is important for studying genome evolution (Chen et al. 2013) because of its complete overlapping with the polymerase gene (Pavesi 2015; Torresi 2002). The HBV genome evolution depends on the complex interaction between viral and host factors, which determines the persistence and progression of HBV infection. Phylogenetic studies of HBV genome have revealed at least ten HBV genotypes, designated A–J (Tian and Jia 2016). These genotypes are major phylogenetic variants playing certain roles in the pathogenesis of HBV infection and are correlated with the progression and long-term outcome of HBV and its epidemiology, which indicates their role in clinical practice (Lin and Kao 2017; Tian and Jia 2016). Generally, genotype A is associated with a better response to interferon therapy; genotype C and, to lesser extent, B usually represent a risk factor for perinatal infection and are associated with advanced liver conditions such as cirrhosis and HCC (Lee et al. 2013; Yang et al. 2008); genotype D may be linked with poor response to interferon therapy (Tian and Jia 2016).

HBV genotypes are distributed in an ethnogeographical manner. This is caused by genome evolution secondary to natural selection, mutational pressure and genetic drift (Guan et al. 2018; Kattoor et al. 2015; Li et al. 2019; Tyagi et al. 2017). Codon usage bias, the preferential use of codons encoding a given amino acid, is vital for examining the adaptation of exogenous genes to the hosts and can play a significant role in enhancing the expression of these genes via codon optimization (Goni et al. 2012; Kattoor et al. 2015; Li et al. 2019; Zhou et al. 2013). Codon usage is equally important for studying the evolution and ecological adaptation of diverse organisms (Chakraborty et al. 2017; Ma et al. 2016; Muthabathula et al. 2018; Zhou et al. 2014). The evolutionary characteristics of HBV and the natural selection effects of the human genome on the codon usage characteristics is important for providing a basis for precision medicine. Therefore, studies on the codon usage pattern of HBV would be vital in elucidating the evolution of HBV, its adaptation to the host and the molecular mechanism in hepatitis, and provide useful data on the virulence of each HBV genotype. However, studies on the CUB of HBV are scarce (Li et al. 2015a; Ma et al. 2011; Pavesi 2015), which needs further in-depth analysis.

In our study, genotyping of HBV sequences from patients in Gansu province was performed and the synonymous codon usage pattern and evolutionary dynamics of different HBV genotypes as well as the adaptation of HBV and its host. The findings of this work will be essential in elucidating the mechanisms driving the molecular evolution of HBV and hepatitis B pathogenesis, while providing a theoretical basis for clinical practice.

Methods

Study population

A total of 79 patients (47 men and 32 women; mean age = 39.57 ± 13.78 years old, range 13–75) were included in this study. The patients visited Gansu Provincial Hospital between November 2017 and January 2018. All patients had persistent seropositivity of HBsAg and showed: (1) no evidence of hepatocellular carcinoma (HCC) or other metastatic liver disease and (2) no evidence for concomitant hepatitis HCV/HDV or HIV infection or autoimmune liver disease. Our study was reviewed and approved by the institutional Ethics Committee of Gansu Provincial People Hospital. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee of Gansu Provincial People Hospital and complied with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Demographic, biochemical and serologic data

HBsAg, anti-HBs, HBeAg, anti-HBe, anti-HCV, and anti-HIV were determined by the microparticle enzyme immunoassay method while anti-HBV was detected by the enzyme immunoassay method (Abbott Laboratories, IL). HBV DNA levels were tested by using a commercial liquid hybridization assay (Digene, MD), with a lower limit of detection of 5 pg/ml.

HBV DNA extraction and quantification

HBV DNA was isolated from the patient plasma following the procedure provided by the manufacturer using "Instant Virus DNA kit" (AJ Roboscreen, Analytikajena Biosolutions, GmbH, Germany). The DNA was quantified following the manufacturer's recommendations using the RoboGene® HBV Quantification kit (AJ Roboscreen, analytikajena Biosolutions, GmbH, Germany).

HBV genotyping

The genotype specific primers were used for genotyping by PCR. The regular PCR reaction mixture was as follows: universal primers S1-2 and P1A (1 μl each), 10 μl GoTaq® Green Master mix (Promega, USA), template DNA (6 μl)

and 2 μl of ddH$_2$O. The reaction program was as follows: 95 °C for 10 min, 30 cycles of 20 s at 94 °C, 20 s at 55 °C and 1 min at 72 °C, followed by 7 min at 72 °C on an ABI 9700 PCR platform (USA). For nested PCR, Mix 1 with a common antisense primer and sense primers for A, B and C genotypes and mix two with antisense primers for D, E and F genotypes with a common sense primer were used. After 1st round of PCR, 2 μl of the product were taken and added as DNA template to each mix; the reaction mixture was composed of 1 μl of each primer, 8 μl of ddH2O and 8 μl of GoTaq® Green Master mix. The nested PCR conditions were as follows: 10 min at 95 °C, followed by 40 cycles of 45 s at 94 °C, 20 s at 63 °C, and 60 s at 72 °C, followed by 7 min at 72 °C with the above amplification platform. The PCR products were separated on 2% agarose gel and a solution of ethidium bromide was used for staining. The revelation was done by ultraviolet fluorescence (BioRad Gel Doc-XR, USA). Samples with viral load higher than 100 IU/ml but whose genotypes were not identified or detected were considered as untypable samples.

DNA sequencing data and phylogenetic analysis

Twenty-one HBV nucleotide sequences were downloaded from Genbank in the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/Genbank/) and added to our sequencing data for analysis. The nucleotide composition percentage (A%, U%, G% and C%) and the percentage of nucleotide in the third position of the codon (A3%, U3%, G3% and C3%) of HBV coding sequence were computed using the CAIcal platform (http://genomes.urv.es/CAIcal) (Puigbò et al. 2008). The Mega6 software (Tamura et al. 2013) was used for the construction of the neighbor joining phylogenetic tree by setting the bootstrap value to 1000.

The calculation of the relative synonymous codon usage (RSCU)

The RSCU values for the coding sequences of our 79 HBV and 21 reference HBV sequences were computed using the CAIcal platform (http://genomes.urv.es/CAIcal) (Puigbò et al. 2008). RSCU values equal to 1.0 indicated that the codon was selected equally and randomly whereas the RSCU higher or lower than 1.0 implied higher frequency or lower frequency, respectively. In addition, the codons with RSCU exceeding 1.6 were considered as over-represented synonymous codons while those with RSCU lower than 0.6 under-represented ones.

ENc analysis

ENc analysis, which quantifies the absolute CUB, was used to estimate the CUB of the coding sequences of HBV. ENc = 20 indicates an excessive CUB while $EN_C = 61$ indicates that there is no CUB. Smaller ENc values indicate larger codon preference in a gene. ENC is calculated using the formula: $Nc = 2 + 9/F2 + 1/F3 + 5/F4 + 3/F6$, wherein the Fk (k = 2, 3, 4, or 6) represents the mean of the Fk values for k-fold degenerate amino acids. The F value is the probability to randomly choose two identical codons encoding for an amino acid. Herein, the ENc values were generated using the CodonW software version 1.4.4.

Principal component analysis PCA

Principal component analysis (PCA) was performed to uncover the main tendency of the codon usage pattern between HBV strains. The PCA was done using the R package "ade" based on the RSCU values.

Codon adaptation index (CAI) analysis

To explore the codon usage preferences, we analyzed the codon adaptation index using the online tool CAIcal (CAI; http://genomes.urv.es/CAIcal) (Puigbò et al. 2008) considering *H. sapiens* cells as reference. Human gene datasets were arbitrarily chosen from the Ensembl (http://www.ensembl.org) database. Student's t test was applied to analyze the difference among CAI values from different groups. The expected value of CAI (e-CAI) was computed at the 95% confidence interval online using the CAIcal tool (http://genomes.urv.es/CAIcal/E-CA) based on the Kolmogorov–Smirnov test. The RSCU values of the host *H. sapiens* were downloaded from the codon usage database (https://hive.biochemistry.gwu.edu/dna.cgi?cmd=refseq_processor&id=545408).

Correlation analysis

The Pearson's correlation was performed to unveil the correlation between variables using the Hmisc package (https://cran.r-project.org/web/packages/Hmisc/index.html) in R. The correlation between nucleotide composition in the third position of codon (A$_3$%, C$_3$%, U$_3$% and G$_3$%) and general nucleotide composition (A%, C%, U% G%) in HBV coding sequences was determined. The correlation between the codon usage pattern and A$_3$%, C$_3$%, U$_3$% and G$_3$% of HBV was also computed.

Statistical analysis

The differences between groups were analyzed using the GraphPad Prism software. One-way analysis of variance (one-way ANOVA) was performed and followed by the Bonferroni post-hoc test. p value cutoff lower than 0.05 was considered for statistical differences between groups.

Results and discussion

### The cluster of HBV sequences from patients and reference sequence by Phylogenetic analysis

To identify the genotypes of HBV that infected the patients, PCR genotyping was performed and followed by sequencing. Our data showed that three different genotypes of HBV (B, C and D genotypes) were identified from the HBV patients. Neighbor-Joining phylogenetic tree analysis showed that there were few substitutions in the sequences. They could be distinctly clustered to the reference sequences obtained from the NCBI database (Fig. 1). There was a significant difference in HBV sequences from different genotypes despite their high identity (Fig. 1). Different sequences of the same genotype were clustered together and had good aggregation with the reference sequences of the same genotype (Fig. 1).

Nucleotide composition analysis of HBV

CUB can be greatly influenced by the overall nucleotide content of the genome (Li et al. 2015b; Ma et al. 2015). A previous study suggested that nucleotide bias is an important factor of the virus-specific codon usage that limits the role of codon selection and translational control (van Hemert and Berkhout 2016; van Hemert et al. 2016). Therefore, we first determined the nucleotide compositions of the HBV genome to highlight the potential influence of the nucleotide constraints on codon usage. Our results indicated that the mean compositions of nucleotides A (27.31% ± 0.67), C (27.75%±0.95) and G (27.44% ± 0.80) were significantly higher compared to T (17.50% ± 0.87) (Fig. 2). No significant difference ($p > 0.05$) was found among A, C and G composition (Fig. 2). One-way ANOVA test indicated that the differences in A vs. T, C vs. T, and T vs. G comparisons were statistically significant ($p < 0.05$). The percentages of nucleotides at the third codon position were: 24.26% ± 1.94

Fig. 1 Phylogenetic tree of HBV sequence from our patients and reference sequence. HBV isolated from the serum of chronic HBV infected patients visited Gansu Provincial Hospital. Phylogenetic tree was constructed by Neighbor-Joining method, based on genotype sequences of 79 HBV isolates and 21 reference sequences from GenBank database
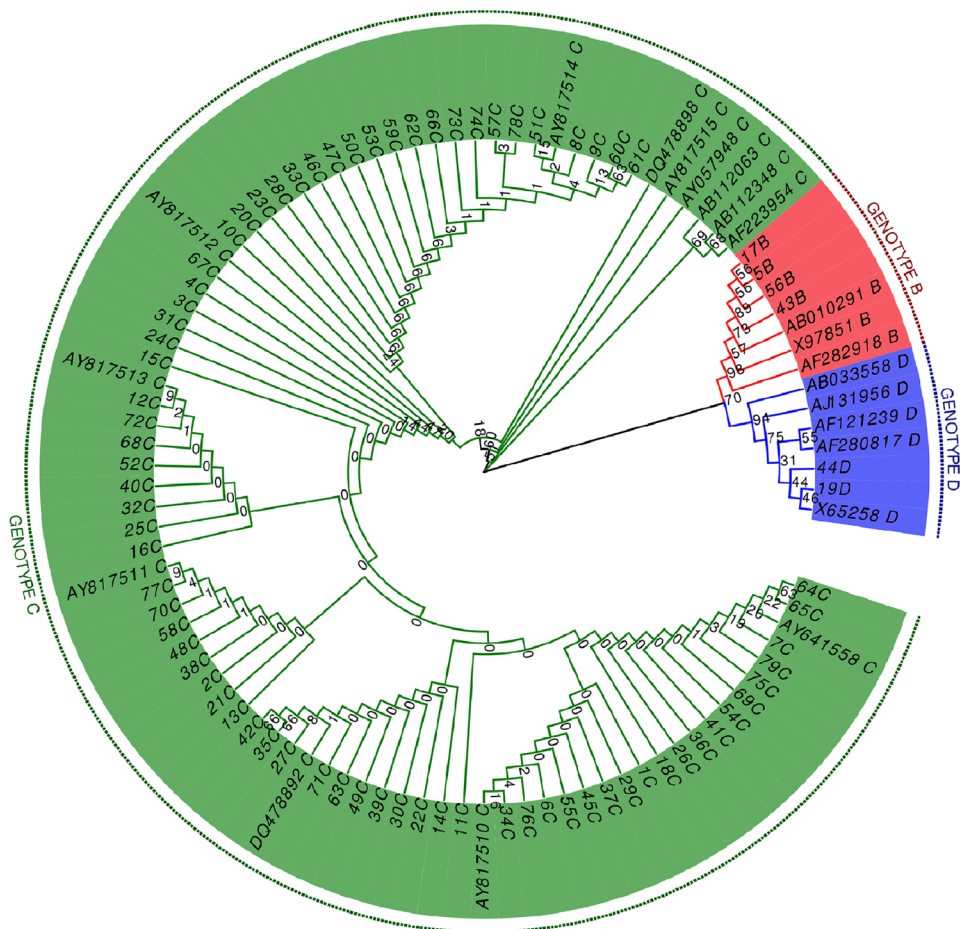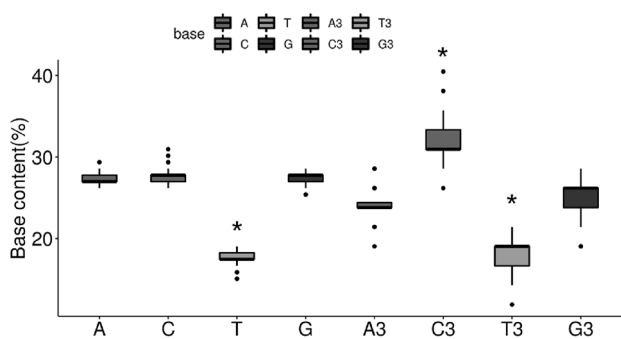
Fig. 2 Base content and composition in bases at third position of codons in HBV sequences. *p < 0.05 compared to A or A3 content as revealed by one-way ANOVA analysis,

for A3; 31.83%±3.00 for C3; 18.48% ± 2.30 for T3; and 25.43% ± 2.37 for G3 (Fig. 2). These values were different from the expected total nucleotide contents (the percentage of a given nucleotide in the analyzed sequences). Specifically, the percentages of A3, C3, T3 and G3 were significantly lower than A, C, T and G. The Pearson correlations among nucleotide content (A%, T%, G%, C%, GC%) and the

percentage of nucleotides at the third position of the codon (A3%, T3%, G3%, C3%, GC3%) were analyzed to clarify whether the effect on CUB was due to translational selection or pressure alone. The overlap scatter-plot of the content of each nucleotide and the content of the nucleotide at the third codon position was depicted in Fig. 3. A positive correlation was generally indicated between the nucleotide composition and the nucleotide composition of the nucleotide at the third codon position in each of the different genotypes of HBV (Fig. 3). The positive correlations between A% and A3% ($r = 0.56$, $p < 0.001$) and C% and C3% ($r = 0.88$, $p < 0.001$) suggested that the constraint of nucleotide content under mutation pressure defines the profile of CUB. In addition, significant positive correlations were found between T% and T3% ($r = 0.81$, $p < 0.001$) (Fig. 3), suggesting that natural selection might not impact on the codon usage pattern.

Codon usage patterns in HBV genotypes

To assess the overall CUB in HBV, the extent of CUB in the HBV genotypes was determined and compared based on the effective number of codons (ENC) and relative synonymous

Fig. 3 The overlap scatter-plot of the content of each nucleotide and the content of each nucleotide at the synonymous third position of codon
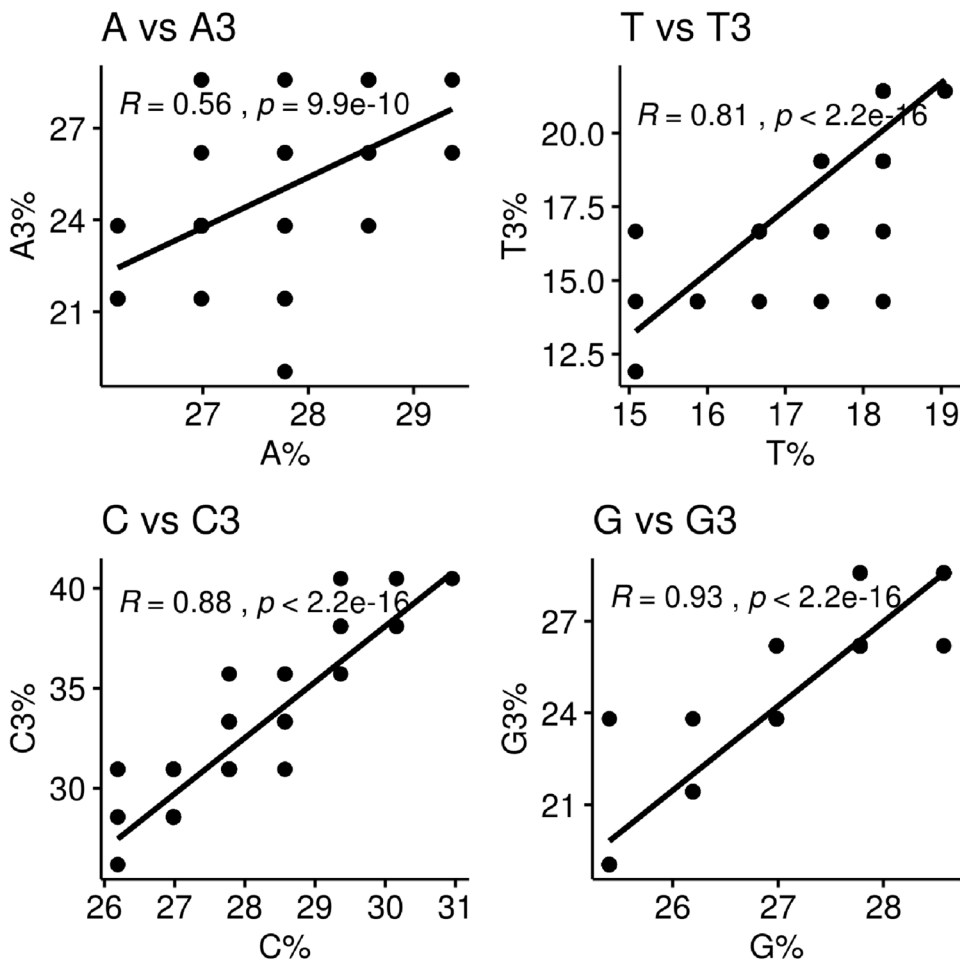
codon usage (RSCU) values of the 79 HBV sequences codon and CAIcal, respectively. According to previous studies, ENC values less than 35 mean high codon preference and ENC values more than 50 reveal general random codon usage (Wang et al. 2018). Our data showed that the ENC mean values (Fig. 4) ranged from 32.5 to 61, with an average ENC value of 50.23 ± 4.25 considering all three genotypes. The ENC value indicated a relatively weak CUB in HBV sequences because only 1% (D genotype) of the HBV sequences had an ENC value < 35. The mean ENC value of HBV genotype C (51.12 ± 3.44) was significantly ($p < 0.05$) higher than those of B (46.39 ± 2.88) and D (42.35 ± 5.86). Significant ($p < 0.05$) difference in ENC was also found between B and D genotypes. Thus, the genotype C tended to use more types of codons to produce proteins, suggesting that their genes might undergo weaker selection constraints with respect to replication speed and transcription efficiency and accuracy compared to B and D genotypes.
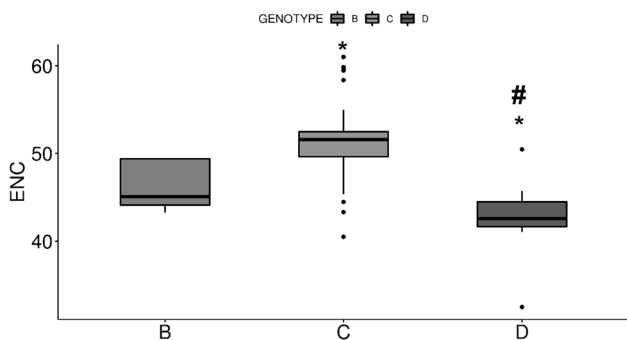

Fig. 4 The ENC values of different codons in the sequences of different HBV sequences. *$p < 0.05$ compared to B, #$p < 0.05$ compared to C

The RSCU values of each codon within the sequences of 79 HBV isolates were summarized in Supplementary Table S1. Of all of the 64 identified codons, 20 were not present in the 79 HBV sequences. These unused codons were CGA (Arg), CGC (Arg), CGG (Arg), CGU (Arg), UGC (Cys), UGU (Cys), AUA (Ile), AUC (Ile), AUU (Ile), CUU (Leu), UUA (Leu), AUG (Met), UCG (Ser),UAA (Ter), UAG (Ter), UGA (Ter), ACG (Thr), ACU (Thr), UAU (Tyr) and GUU (Val). The remaining 44 codons were used by the HBV genome and were unevenly distributed. In HBV genotype B, 38 codons were unused, 8 codons had mean RSCU < 1, 1 codon had mean RSCU = 1 and 17 codons had mean RSCU > 1. The RSCU values of high-frequency codons in the B genotype ranged between 1 and 6. In the genotype C, 26 codons were unused, 21 had mean RSCU of < 1, one codon had mean RSCU value of 1 while 16 codons had mean RSCU > 1, with the maximum RSCU value of 6. For the genotype D, our results indicated that 26 codons were unused while 10 codons had RSCU values of < 1, 1 codon had mean RSCU of 1 and 17 had mean RSCU of > 1 (Fig. 5). The heatmap and clustering based on the RSCU values (Fig. 6) included 3 types of profiles: preferential codons in red, unused in blue and less-preferred codons in white. The used codons in HBV sequences, as well as the reference strains from Genbank had higher RSCU values than that of human cell, indicating the well adaption of HBV to its host. These results implied that translational selection in nature has an effect on the pattern of synonymous codon usage and the evolutionary pattern of HBV.

Moreover, to determine the differential usage preference of codons among genotypes, the mean RSCU values were used for one-ANOVA analysis followed by Bonferroni posttest (Fig. 7). The results showed that 31 codons
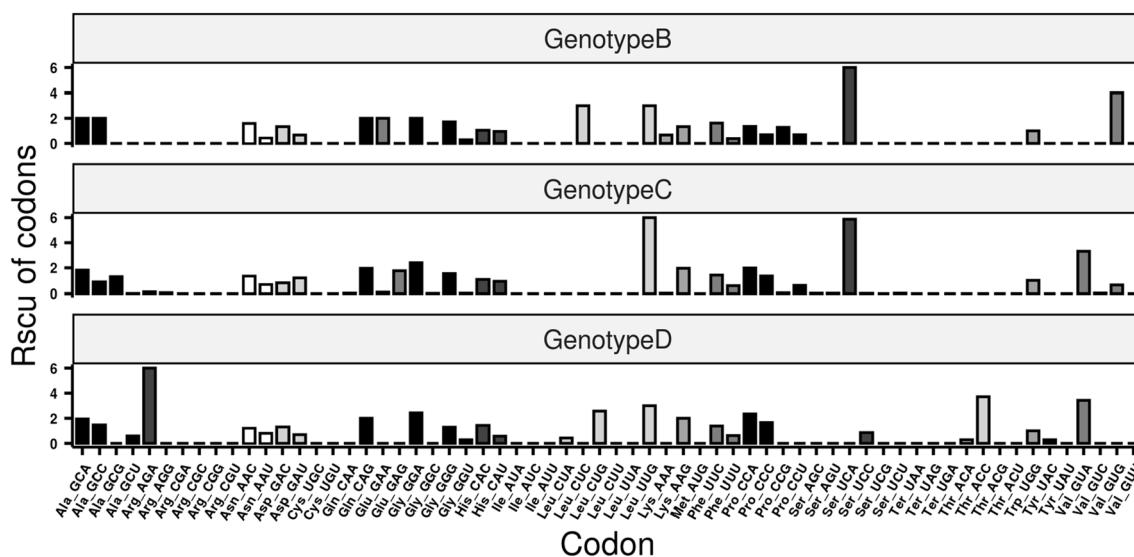

Fig. 5 The relative synonymous codon usage (RSCU) values of different codons from the sequences of different genotypes of HBV
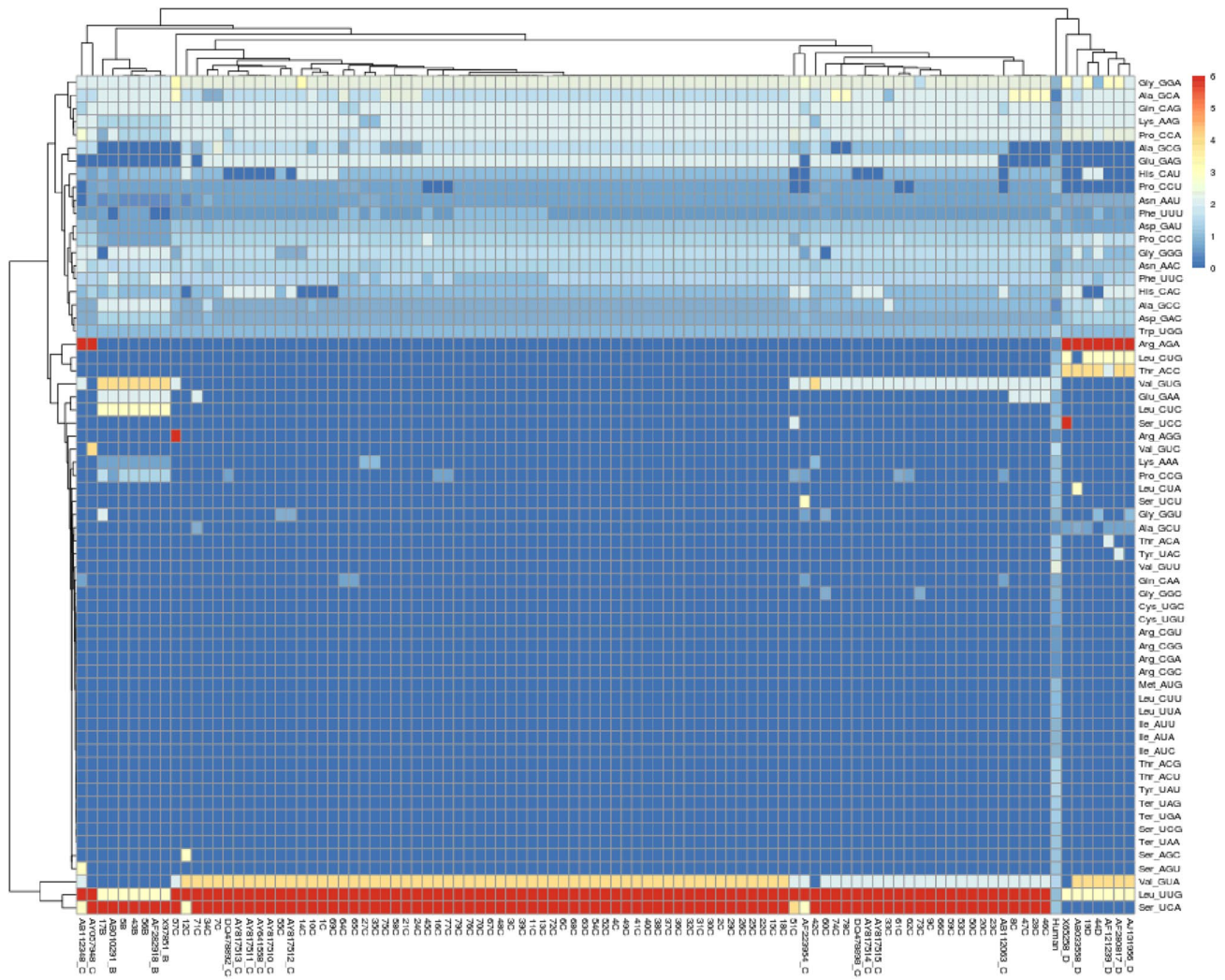
Fig. 6 Hierarchical cluster analysis and heat map of the relative synonymous codon usage (RSCU) values of each codon in HBV. Each square in the heat map represents the log ratio of the RSCU value of each codon (in rows) within the HBV genome (in columns). Colors indicate the magnitude of RSCU values: white, RSCU = 1 (no bias in codon usage); blue RSCU < 1; red, RSCU > 1

were differentially used among the three genotypes, with 29 of these codons showing a p value of < 0.01 while 2 codons showed p value comprised between 0.01 and 0.05. The remaining 13 codons showed no difference regarding to the usage frequency among HBV genotypes (p > 0.05). These results hinted that the number of differentially used codons between the HBV genotypes was higher than similarly used ones.

Genetic correlation based on synonymous codon usage in HBV

To explore whether the synonymous codon usage could influence the genotypes of HBV, the principal component analysis (PCA) was performed based on the RSCU values. The PCA detected the first principal component (PC1)

which accounted for 44.96% of the total synonymous codon usage variation, and the second principal component (PC2) accounting for 28.49% of total variation (Fig. 8). It can be observed that different HBV genotypes were distinctly separated from each other. Moreover, the genotype B and D showed obviously different genetic characteristics but showed obvious aggregation with genotype C. Thus, the codon usage variation might be one of the factors driving HBV evolution.

The effect of mutation pressure on codon usage of HBV

We employed three approaches based on codon usage indices, namely a neutrality analysis, an ENC plot, and the ratio of synonymous to non-synonymous substitutions
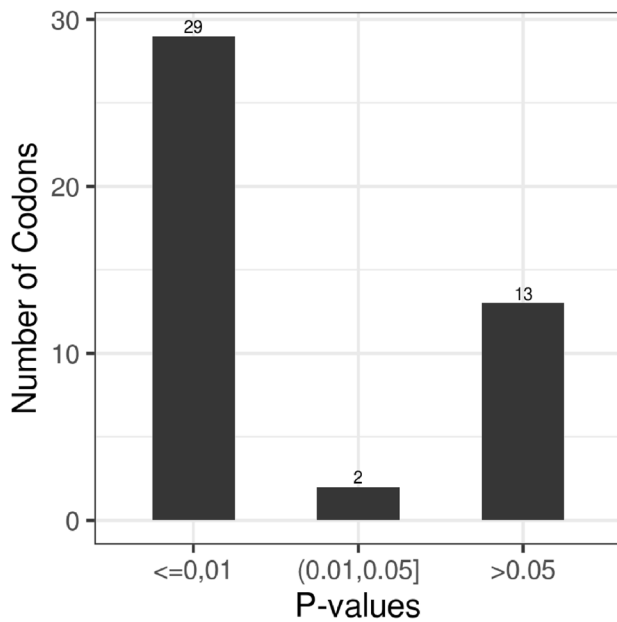
Fig. 7 Significantly different codons. The p values were calculated using average RSCU values of each codon in the comparison of the three genotypes
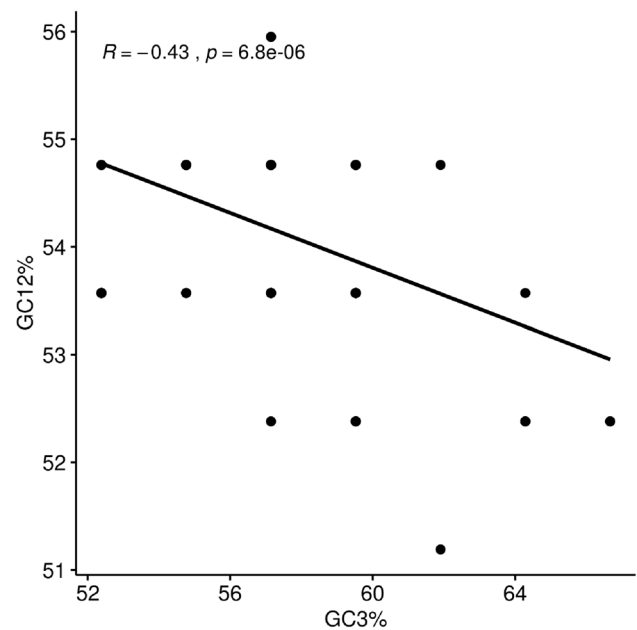

Fig. 9 Correlation between GC content at first and second codon position (GC12%) with that at synonymous third codon positions (GC3%)
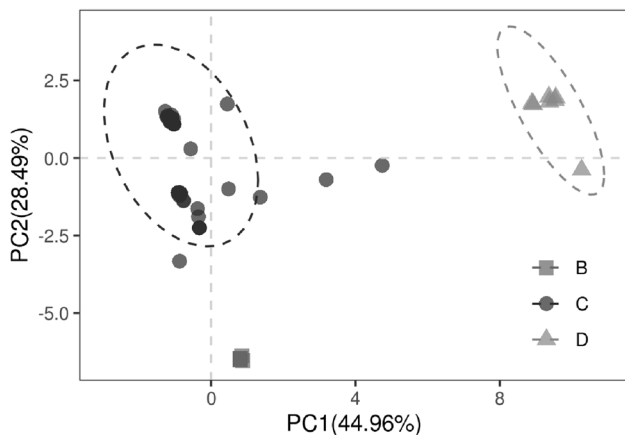

Fig. 8 The genetic characteristic of HBV based on different genotypes. Principal component analysis was used for separating the genotypes based on the RSCU values

(Ka/Ks), to elucidate the different evolutionary mechanisms operating in the HBV genome.

The neutrality analysis was performed to quantify the mutational pressure. The average values of GC content in the first and second positions (GC12) was 54.15% ± 0.89 while that in the third position (GC3) of codons was 57.26% ± 3.02 (Fig. 9). A significant negative correlation was observed between GC12 and GC3 among all clinical isolates (R = − 0.43, p value = 0.001). Separately, no significant correlation between GC12 and GC3 was found in each genotype. These results suggested that directional

mutation bias played a minor role in the evolution of HBV genome.

To examine the factors affecting HBV CUB, the ENC values were plotted against the GC3 percentage. As shown in Fig. 10, the points are the actual ENC values and the curve corresponds to the expected ENC values with the only factor of mutation of HBV coding sequences (Fig. 10). It can be observed that the isolates were distributed under and over the expected ENC curve of the C genotype. This implied mutation pressure and factors such as translational selection impact on the CUB of HBV coding sequences for genotype C. The ENC values of genotypes B and D were all under the curve, which suggested that translational selection may be predominant in these genotypes.

The Ka/Ks ratio is a simple measure of selection pressure on codons, which reveals neutrality (Ka/Ks = 1), negative or purifying selection (Ka/Ks < 1), and positive selection (Ka/Ks > 1) (Ma et al. 2011; Woolley et al. 2003). The closer the ratio reaches to 1, the smaller the selection pressure is (Ma et al. 2011). We calculated the Ka/Ks ratio for each genotype, separately. The Ka/Ks ratio ranged from 0 to 0.25 in HBV B genotype, from 0 to 1.98 in the D genotype with only one isolate showing Ka/Ks ≥ 1, and from 0 to 1.49 in C genotype with 3 sequences showing Ka/Ks ≥ 1 (Fig. 11). The majority of these values were significantly lower than 1, implying that the three genotypes are under intense purifying selection. Further comparison of the Ka/Ks ratio between genotypes by one-way ANOVA indicated no significant difference between the Ka/Ks ratio of the three genotypes

Fig. 10 Distribution of the codon usage index, ENc and GC content at synonymous third codon position (GC3%). The curve shows the expected codon usage of GC compositional constraints alone account for CUB
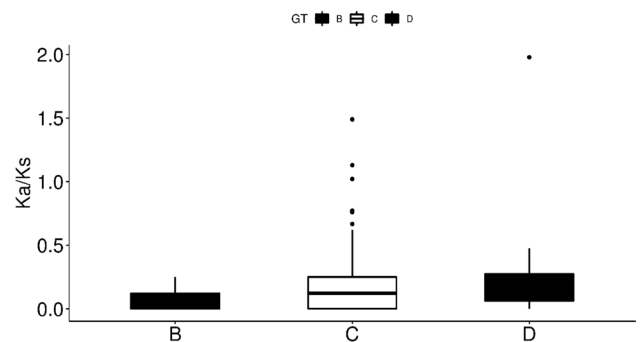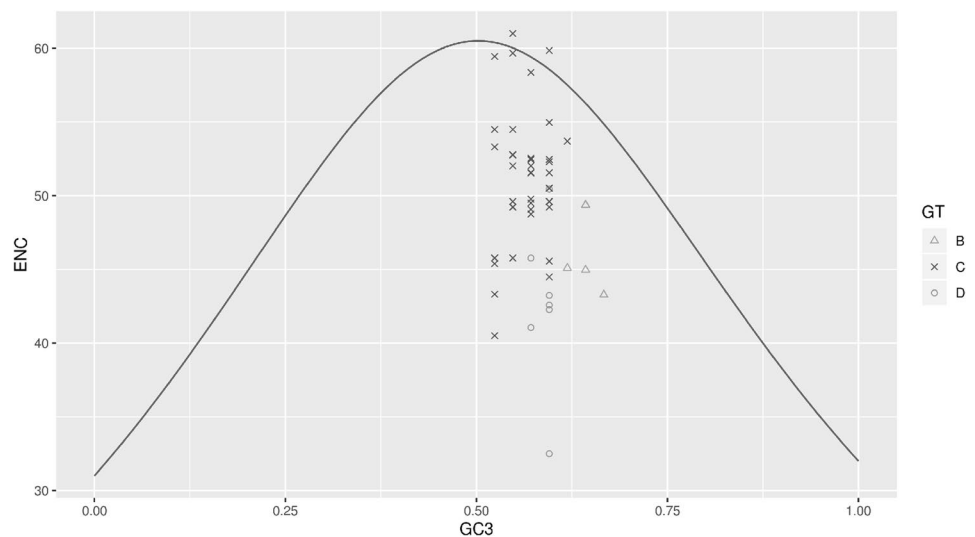

Fig. 11 Box-plots of the ratio of synonymous to non-synonymous substitutions (Ka/Ks) in the three HBV genotypes. No significant difference was found (p > 0.05)

(Fig. 11), suggesting that there was no difference in the purifying selection among the 3 genotypes.

Adaptation of HBV to the human genome

The CAI values range from 0 to 1, and high CAI values indicate higher levels of CUB (Subramanian and Rup Sarkar 2015). Codon adaptation index (CAI) analyses were performed to determine the codon usage optimization and adaptation of HBV in relation to its hosts. CAI values for all codons were calculated by reference to the codon usage of *H. sapiens*. We found that, in relation to *H. sapiens*, the CAI values of HBV S small protein-coding regions were in the range of 0.78–0.90 (Fig. 12). This study found a tendency of relatively high CAI values (> 0.5), which disclosed a good adaptability of HBV to its hosts and a low translation pressure (Subramanian and Rup Sarkar 2015). The tendency of high CAI values for *H. sapiens* suggests that selection pressure from *H. sapiens* can affect the codon usage of HBV and
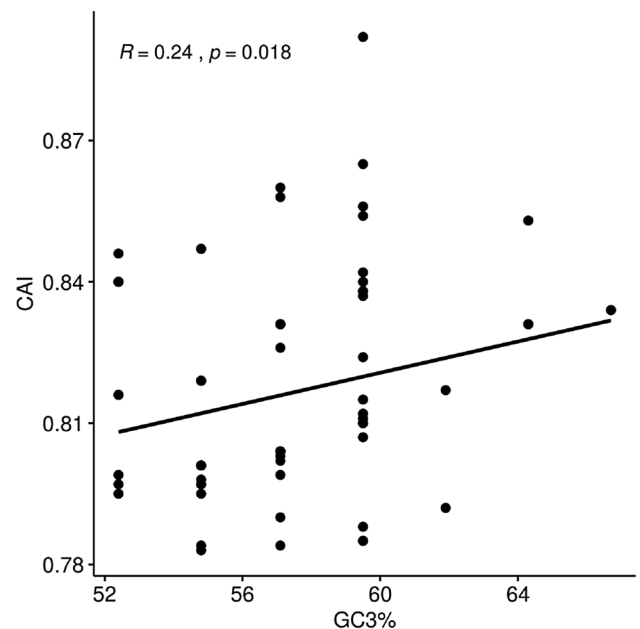

Fig. 12 CAI value vs. GC3%. The different genotype points represent the correlation between gene expression and nucleotide composition of HBV coding sequence

that the evolution of codon usage in HBV allows it to use the translation machinery of *H. sapiens* more efficiently. In addition, the results suggested that these differences are related to codon usage preferences. Our results about codon usage preferences were consistent with published works (Ma et al. 2011). Turning to the expression levels of viral product in host cells, the relationship between CAI and GC3% indicates that the various CUBs among B, C, D genotypes exist in the process of evolution of HBV (Fig. 12). This demonstrated that the synonymous codon usage patterns of HBV might play an important role in the optimized expression level of

viral product of HBV. Some previous reports pointed out that HBV genotypes have been increasingly associated with differences in virologic and clinical features, such as response to antiviral therapies and severity of liver disease (Enomoto et al. 2006; Kao et al. 2000; Palumbo 2007; Wai et al. 2002).

Conclusions

This is the first study revealing the codon usage pattern within HBV genotypes. Our study indicated that the CUB of HBV genotypes is low with good adaptability to the human genome. Our findings hinted that mutation bias and mutational pressure were the prevalent factors in shaping HBV codon usage patterns. The present findings have great prospects for elucidating the molecular evolution and functional mechanisms of HBV. The present data will be of clinical importance, especially for studying the pathogenesis of hepatis B and developing treatment drugs.

Data availability All data generated or analyzed during the present study are included in this published article.

Compliance with ethical standards

Conflict of interest Xiaoming Qi, Chaojun Wei, Yonghong Li, Yu Wu, Hui Xu, Rui Guo, Yanjuan Jia, Zhenhao Li, Zhenhong Wei, Wanxia Wang, Jing Jia, Yuanting Li, Anqi Wang and Xiaoling Gao declare that they have no conflict of interest.

Ethical approval Our study was reviewed and approved by the institutional Ethics Committee of Gansu Provincial People Hospital. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee of Gansu Provincial People Hospital and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Benhenda S, Ducroux A, Rivière L, Sobhian B, Ward MD, Dion S, Hantz O, Protzer U, Michel ML, Benkirane M et al (2013) Methyltransferase PRMT1 is a binding partner of HBx and a negative regulator of hepatitis B virus transcription. J Virol 87:4360–4371. https://doi.org/10.1128/jvi.02574-12

Binh MT, Hoan NX, Tong HV, Sy BT, Trung NT, Bock CT, Toan NL, Song LH, Bang MH, Meyer CG et al (2019) NTCP S267F variant associates with decreased susceptibility to HBV and HDV infection and decelerated progression of related liver diseases. Int J Infect Dis 80:147–152. https://doi.org/10.1016/j.ijid.2019.01.038

Bonvicino CR, Moreira MA, Soares MA (2014) Hepatitis B virus lineages in mammalian hosts: potential for bidirectional cross-species transmission. World J Gastroenterol 20:7665–7674. https://doi.org/10.3748/wjg.v20.i24.7665

Chakraborty S, Uddin A, Choudhury MN (2017) Factors affecting the codon usage bias of SRY gene across mammals. Gene 630:13–20. https://doi.org/10.1016/j.gene.2017.08.003

Chen P, Gan Y, Han N, Fang W, Li J, Zhao F, Hu K, Rayner S (2013) Computational evolutionary analysis of the overlapped surface (S) and polymerase (P) region in hepatitis B virus indicates the spacer domain in P is crucial for survival. PLoS One 8:e60098

Churin Y, Roderfeld M, Roeb E (2015) Hepatitis B virus large surface protein: function and fame. Hepatobiliary Surg Nutr 4:1–10. https://doi.org/10.3978/j.issn.2304-3881.2014.12.08

De Maddalena C, Giambelli C, Tanzi E, Colzani D, Schiavini M, Milazzo L, Bernini F, Ebranati E, Cargnel A, Bruno R et al (2007) High level of genetic heterogeneity in S and P genes of genotype D hepatitis B virus. Virology 365:113–124

Enomoto M, Tamori A, Nishiguchi S (2006) Hepatitis B virus genotypes and response to antiviral therapy. Clin Lab 52:43–47

Gerelsaikhan T, Tavis JE, Bruss V (1996) Hepatitis B virus nucleocapsid envelopment does not occur without genomic DNA synthesis. J Virol 70:4269–4274

Goni N, Iriarte A, Comas V, Soñora M, Moreno P, Moratorio G, Musto H, Cristina J (2012) Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. Virol J 9:263. https://doi.org/10.1186/1743-422x-9-263

Guan DL, Ma LB, Khan MS, Zhang XX, Xu SQ, Xie JY (2018) Analysis of codon usage patterns in *Hirudinaria manillensis* reveals a preference for GC-ending codons caused by dominant selection constraints. BMC Genomics 19:542. https://doi.org/10.1186/s12864-018-4937-x

Kao JH, Wu NH, Chen PJ, Lai MY, Chen DS (2000) Hepatitis B genotypes and the response to interferon therapy. J Hepatol 33:998–1002. https://doi.org/10.1016/s0168-8278(00)80135-x

Kattoor JJ, Malik YS, Sasidharan A, Rajan VM, Dhama K, Ghosh S, Bányai K, Kobayashi N, SinghRK (2015) Analysis of codon usage pattern evolution in avian rotaviruses and their preferred host. Infect Genet Evol 34:17–25. https://doi.org/10.1016/j.meegid.2015.06.018

Kim H, Jee YM, Song BC, Shin JW, Yang SH, Mun HS, Kim HJ, Oh EJ, Yoon JH, Kim YJ et al (2007) Molecular epidemiology of hepatitis B virus (HBV) genotypes and serotypes in patients with chronic HBV infection in Korea. Intervirology 50:52–57. https://doi.org/10.1159/000096313

Kim SY, Kyaw YY, Cheong J (2017) Functional interaction of endoplasmic reticulum stress and hepatitis B virus in the pathogenesis of liver diseases. World J Gastroenterol 23:7657–7665. https://doi.org/10.3748/wjg.v23.i43.7657

Kramvis A, Kew MC (1998) Structure and function of the encapsidation signal of hepadnaviridae. J Viral Hepat 5:357–367. https://doi.org/10.1046/j.1365-2893.1998.00124.x

Lee MH, Yang HI, Liu J, Batrla-Utermann R, Jen CL, Iloeje UH, Lu SN, You SL, Wang LY, Chen CJ et al (2013) Prediction models of long-term cirrhosis and hepatocellular carcinoma risk in chronic hepatitis B patients: risk scores integrating host and virus profiles. Hepatology 58:546–554. https://doi.org/10.1002/hep.26385

Li HM, Wang JQ, Wang R, Zhao Q, Li L, Zhang JP, Shen T (2015a) Hepatitis B virus genotypes and genome characteristics in China. World J Gastroenterol 21:6684–6697. https://doi.org/10.3748/wjg.v21.i21.6684

Li J, Zhou J, Wu Y, Yang S, Tian D (2015b) GC-Content of synonymous codons profoundly influences amino acid usage. G3 (Bethesda) 5:2027–2036. https://doi.org/10.1534/g3.115.019877

Li L, Han T, Zang L, Niu L, Cheng W, Lin H, Li KY, Cao R, Zhao B, Liu Y et al (2017) The current incidence, prevalence, and residual risk of hepatitis B viral infections among voluntary blood donors in China. BMC Infect Dis 17:754. https://doi.org/10.1186/s12879-017-2861-3

Li G, Zhang W, Wang R, Xing G, Wang S, Ji X, Wang N, Su S, Zhou J (2019) Genetic analysis and evolutionary changes of the torque teno sus virus. Int J Mol Sci. https://doi.org/10.3390/ijms20122881

Lin CL, Kao JH (2017) Natural history of acute and chronic hepatitis B: The role of HBV genotypes and mutants. Best Pract Res Clin Gastroenterol 31:249–255. https://doi.org/10.1016/j.bpg.2017.04.010

Ma MR, Ha XQ, Ling H, Wang ML, Zhang FX, Zhang SD, Li G, Yan W (2011) The characteristics of the synonymous codon usage in hepatitis B virus and the effects of host on the virus in codon usage pattern. Virol J 8:544. https://doi.org/10.1186/1743-422x-8-544

Ma MR, Hui L, Wang ML, Tang Y, Chang YW, Jia QH, Yang XP, Wang XH, Ha XQ (2015) Synonymous codon selection in the hepatitis B virus translation initiation region. Genet Mol Res 14:8955–8963. https://doi.org/10.4238/2015.August.7.4

Ma YP, Ke H, Liang ZL, Liu ZX, Hao L, Ma JY, Li YG (2016) Multiple evolutionary selections involved in synonymous codon usages in the Streptococcus agalactiae genome. Int J Mol Sci 17:277. https://doi.org/10.3390/ijms17030277

Muthabathula P, Suneetha S, Grace R (2018) Genome-wide codon usage bias analysis in Beauveria bassiana. Bioinformation 14:580–586. https://doi.org/10.6026/97320630014580

Nelson NP, Easterbrook PJ, McMahon BJ (2016) Epidemiology of hepatitis B virus infection and impact of vaccination on disease. Clin Liver Dis 20:607–628. https://doi.org/10.1016/j.cld.2016.06.006

Palumbo E (2007) Hepatitis B genotypes and response to antiviral therapy: a review. Am J Ther 14:306–309. https://doi.org/10.1097/01.pap.0000249927.67907.eb

Pavesi A (2015) Different patterns of codon usage in the overlapping polymerase and surface genes of hepatitis B virus suggest a de novo origin by modular evolution. J Gen Virol 96:3577–3586. https://doi.org/10.1099/jgv.0.000307

Puigbò P, Bravo IG, Garcia-Vallvé S (2008) CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3:38. https://doi.org/10.1186/1745-6150-3-38

Sarkar N, Chakravarty R (2015) Hepatitis B virus infection, microRNAs and liver disease. Int J Mol Sci 16:17746–17762. https://doi.org/10.3390/ijms160817746

Shih C, Yang CC, Choijilsuren G, Chang CH, Liou AT (2018) Hepatitis B Virus. Trends Microbiol 26:386–387. https://doi.org/10.1016/j.tim.2018.01.009

Stasi C, Silvestri C, Voller F, Cipriani F (2016) The epidemiological changes of HCV and HBV infections in the era of new antiviral therapies and the anti-HBV vaccine. J Infect Public Health 9:389–395. https://doi.org/10.1016/j.jiph.2015.05.004

Subramanian A, Rup Sarkar R (2015) Data in support of large scale comparative codon usage analysis in *Leishmania* and *Trypanosomatids*. Data Brief 4:269–272. https://doi.org/10.1016/j.dib.2015.06.003

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729. https://doi.org/10.1093/molbev/mst197

Taylor JM (2013) Virus entry mediated by hepatitis B virus envelope proteins. World J Gastroenterol 19:6730–6734. https://doi.org/10.3748/wjg.v19.i40.6730

Tian Q, Jia J (2016) Hepatitis B virus genotypes: epidemiological and clinical relevance in Asia. Hepatol Int 10:854–860. https://doi.org/10.1007/s12072-016-9745-2

Torresi J (2002) The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. J Clin Virol 25:97–106. https://doi.org/10.1016/s1386-6532(02)00049-5

Tyagi A, Kumar BTN, Singh NK (2017) Genome dynamics and evolution of codon usage patterns in shrimp viruses. Arch Virol 162:3137–3142. https://doi.org/10.1007/s00705-017-3445-7

van Hemert F, Berkhout B (2016) Nucleotide composition of the Zika virus RNA genome and its codon usage. Virol J 13:95. https://doi.org/10.1186/s12985-016-0551-1

van Hemert F, van der Kuyl AC, Berkhout B (2016) Impact of the biased nucleotide composition of viral RNA genomes on RNA structure and codon usage. J Gen Virol 97:2608–2619. https://doi.org/10.1099/jgv.0.000579

Wai CT, Chu CJ, Hussain M, Lok AS (2002) HBV genotype B is associated with better response to interferon therapy in HBeAg(+) chronic hepatitis than genotype C. Hepatology 36:1425–1430. https://doi.org/10.1053/jhep.2002.37139

Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, Feng W, Zhang G, Song X, Sun X (2018) Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS One 13:e0194372. https://doi.org/10.1371/journal.pone.0194372

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA (2003) TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics 19:671–672. https://doi.org/10.1093/bioinformatics/btg043

Yang HI, Yeh SH, Chen PJ, Iloeje UH, Jen CL, Su J, Wang LY, Lu SN, You SL, Chen DS et al (2008) Associations between hepatitis B virus genotype and mutants and the risk of hepatocellular carcinoma. J Natl Cancer Inst 100:1134–1143. https://doi.org/10.1093/jnci/djn243

Zhou JH, Gao ZL, Zhang J, Ding YZ, Stipkovits L, Szathmary S, Pejsak Z, Liu YS (2013) The analysis of codon bias of foot-and-mouth disease virus and the adaptation of this virus to the hosts. Infect Genet Evol 14:105–110. https://doi.org/10.1016/j.meegid.2012.09.020

Zhou JH, Ding YZ, He Y, Chu YF, Zhao P, Ma LY, Wang XJ, Li XR, LiuYS (2014) The effect of multiple evolutionary selections on synonymous codon usage of genes in the Mycoplasma bovis genome. PLoS One 9:e108949. https://doi.org/10.1371/journal.pone.0108949

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.