Virologica Sinica (2019) 34:42–49
https://doi.org/10.1007/s12250-018-0070-x

www.virosin.org
www.springer.com/12250

RESEARCH ARTICLE

The Variability of Amino Acid Sequences in Hepatitis B Virus

Jianhao Cao[1] · Shuhong Luo[1] · Yuanyan Xiong[2]

Received: 16 July 2018 / Accepted: 13 November 2018 / Published online: 4 January 2019

Abstract

Hepatitis B virus (HBV) is an important human pathogen belonging to the *Hepadnaviridae* family, *Orthohepadnavirus* genus. Over 240 million people are infected with HBV worldwide. The reverse transcription during its genome replication leads to low fidelity DNA synthesis, which is the source of variability in the viral proteins. To investigate the variability quantitatively, we retrieved amino acid sequences of 5,167 records of all available HBV genotypes (A–J) from the Genbank database. The amino acid sequences encoded by the open reading frames (ORF) S/C/P/X in the HBV genome were extracted and subjected to alignment. We analyzed the variability of the lengths and the sequences of proteins as well as the frequencies of amino acids. It comprehensively characterized the variability and conservation of HBV proteins at the level of amino acids. Especially for the structural proteins, hepatitis B surface antigens (HBsAg), there are potential sites critical for virus assembly and immune recognition. Interestingly, the preS1 domains in HBsAg were variable at some positions of amino acid residues, which provides a potential mechanism of immune-escape for HBV, while the preS2 and S domains were conserved in the lengths of protein sequences. In the S domain, the cysteine residues and the secondary structures of the alpha-helix and beta-sheet were likely critical for the stable folding of all HBsAg components. Also, the preC domain and C-terminal domain of the core protein are highly conserved. However, the polymerases (HBpol) and the HBx were highly variable at the amino acid level. Our research provides a basis for understanding the conserved and important domains of HBV viral proteins, which could be potential targets for anti-virus therapy.

Keywords Hepatitis B virus (HBV) · Amino acid · Sequence characterization · Variability and conservation

Introduction

Hepatitis B virus (HBV) is one of the important human pathogens that causes hepatitis. Over 240 million people are estimated to be infected with HBV. HBV belongs to the *Hepadnaviridae* family, *Orthohepadnavirus* genus. It has three forms of viral particles, infectious 44 nm-diameter Dane particles and non-infectious 22 nm-diameter spherical or tubular subviral particles (SVPs). Its genome is only about 3.2 kb and contains four overlapping opening reading frames (ORFs) which encode seven structural or nonstructural proteins (Fig. 1A). Although HBV is a double-stranded DNA virus, it replicates the genome by an intermediate template of pregenome RNA and thus presents a high mutation rate and low fidelity (Seeger and Mason 2015).

There are 3 forms of hepatitis B surface antigens (HBsAg) in the virus envelope with some domains exposed to lumen or cytosol (Bruss 2004). They are large, middle and small HBsAg, respectively and share the same C-terminal domain but different N-terminal domains (Heermann *et al*. 1984; Seeger and Mason 2000). They form the components of both Dane particles and SVPs. The HBsAg can elicit strong protection reaction of individuals against HBV or induce immune tolerance through persistent expression (Buynak *et al*. 1976; Seeger and Mason 2015). Moreover, it has been used as a therapeutic vaccine for

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12250-018-0070-x) contains supplementary material, which is available to authorized users.

✉ Jianhao Cao
jianhaoc2@gmail.com

✉ Shuhong Luo
sluo815@gmail.com

✉ Yuanyan Xiong
xyyan@mail.sysu.edu.cn

[1] Institute of Antibody Engineering, School of Laboratory Medicine and Biotechnology, Southern Medical University, Guangzhou 510515, China

[2] School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

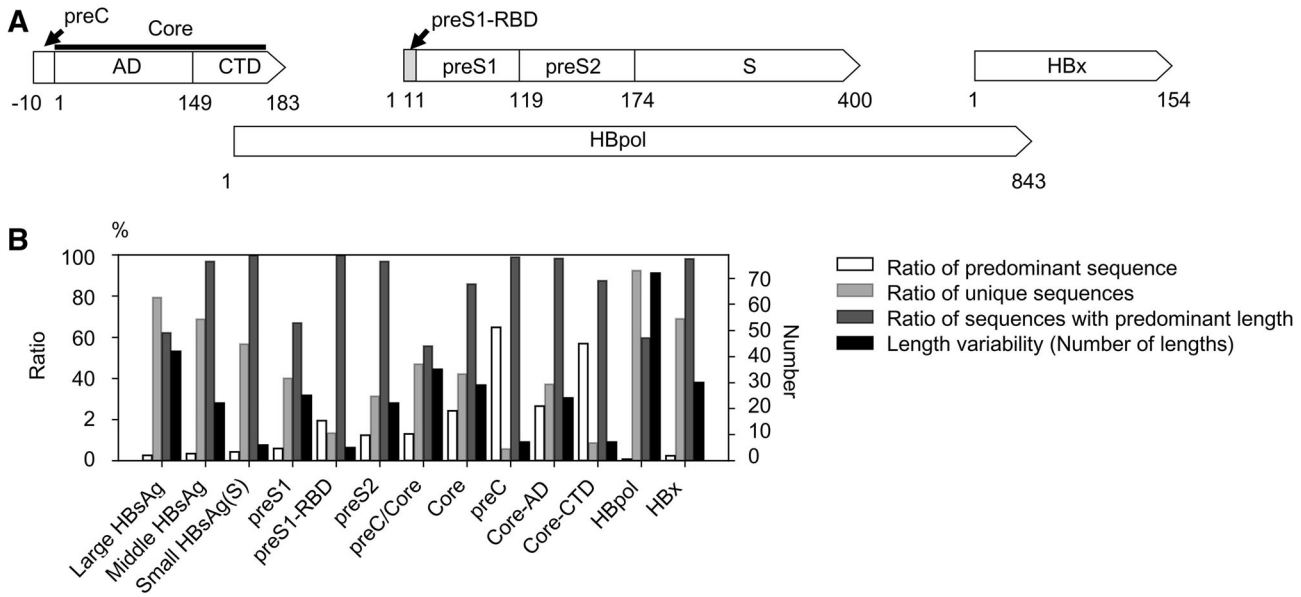

Fig. 1 **A** The illustration of proteins encoded by HBV genome and the position of domains. **B** The statistical properties of different amino acid sequences/domains in all HBV strains. The ratio of sequences containing predominant sequence are smaller than 0.7, even that of HBcAg is not more than 30% (white bar). However, the sequences of the preC domain and the core-CTD are highly conserved (more than 50%). The ratio of sequences containing all unique sequences indicates variability of the sequence/domain (light grey bar). The ratio of sequences with predominant length indicates the degree that sequences or domains confine to the length of amino acid residues (grey bar). The most conserved domains are the domains of preS1-RBD, preS2, S, preC and core-CTD, in which over 90% sequences have the same length. The length variability indicates the variability of sequence/domain in length (black bar). It also suggests the small HBsAg, the preC domain and the core-CTD are the most conserved. The left y-axis indicates the ratio. The right y-axis indicates the length variability.

clinical trial (Dembek *et al*. 2018), but is not efficient enough.

It may provide better information about potential antiviral targets to analyze the variability of HBV proteins at the level of amino acids rather than nucleotides. In this study, to investigate the difference of viral protein sequences/domains, we reported the amino acid sequence variability of HBV strains (genotypes A–J). We also further characterized the features of HBsAg, which may help to understand its function.

Materials and Methods

HBV Sequence Acquisition and Alignment

In November 2017, a total of 5,167 HBV genome records with confirmed genotypes were retrieved from the Genbank Nucleotide Database of National Center for Biotechnology Information (NCBI). The whole dataset was divided into different categories according to areas, China, Southeast-Asia (SE-Asia, including Indonesia, Malaysia, Myanmar, South Korea, Thailand, Vietnam, Japan and Korea), America (including Argentina, Brazil, Canada, Chile, Colombia, Mexico, USA and Venezuela), Europe (including Belgium, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Poland, Russia, Serbia, Spain, Sweden, Turkey and UK), and the other area (including India, Iran, Saudi Arabia and Syria). Then, the amino acid sequences of 4 HBV ORFs (S, C, P and X) were extracted according to their start and end sites and were then subjected to sequence alignment with Clustal Omega (Sievers *et al*. 2011).

Subsequent analysis was focused on those functional sequences or domains annotations. The sequences labelled with "nonfunctional" or "truncated" were discarded. The final items for analysis were 3,208 sequences of ORF C, 3,064 sequences of ORF S, 3,701 sequences of ORF P and 4,265 sequences for ORF X. The domains for further analysis were the preS1 receptor binding domain (preS1-RBD), the preS1 domain, the preS2 domain, the S domain, the preC domain, the core protein assembly domain (core-AD, usually 149AAs) and the core protein C-terminal domain (core-CTD, also known as the arginine-rich domain, ARD). The positions of the preS1-RBD is according to the sequence previously reported (Yan *et al*. 2012). The preS1/S2/S domains are encoded by ORF S, while the preC, core-AD and core-CTD are encoded by ORF C. The ORF C and S encode the structural proteins of

HBV which are important antigens and were selected for further analysis since they may provide basis for the development of new vaccines.

Features of HBV Sequences and Domains

We analyzed the variability of sequences from different aspects. All domains were extracted from aligned sequences. Aligned sequences with only gaps across the domain were discarded in subsequent analysis. Hence, the remaining items are the total number of sequences for calculation of the percentage of different indices mentioned below. Among all the sequences or domains, the most frequent one was defined as the predominant sequence, which represents the conserved sequence. Sequences or domains with any differences in amino acid sequence were defined as unique sequences. Similarly, the most frequent length of sequences or domains was defined as the predominant one. The ratio of sequences with the predominant length suggests the length conservation of a sequence or domain. The length variability is measured by calculation of number of sequence lengths, which means how many kinds of length a sequence or domain could adopt.

Statistics of the Frequency of Amino Acids of Each Site in Sequences

After sequence alignment, we computed the proportion of amino acids residues (including gaps) for each position as below.

$$R_{\text{amino acid}} = \frac{N_{\text{sequences with specified amino acid}}}{N_{\text{total sequences}}}.$$

The proportion was the number of sequences with specified amino acid divided by the total number of available sequences. We defined the predominant amino acid residue as the one with the highest proportion at that site and evaluated the levels of sequence homology to the predominant residue.

Prediction of the Secondary Structure of HBsAg

A representative sequence record (Accession number of its genome record: AM295797) was subjected to the prediction of secondary structure of the large/middle/small HBsAg. The prediction was performed in PSIPRED website (http://bioinf.cs.ucl.ac.uk/psipred/) (Jones 1999).

Analysis of Similarities of Pair-Wise Sequences

The pair-wise sequence's similarity was calculated based on the aligned sequences. The similarity was defined as the ratio of the number of positions with identical amino acid residues to the length of the sequences. The similarity profile was shown as a histogram, and the total ratio of pair-wise similarity in each panel was 1. Pearson correlation coefficient (R) was computed between two profiles using Python SciPy library (https://www.scipy.org). $R > 0.8$ was regarded as correlation, while $R > 0.9$ was regarded as high correlation.

Results

Amino Acid Sequences Analysis

The geographical sources of the 5,167 sequences of documented HBV strains are listed in Table 1. A total of 3,657 items (70.8%) of them are from Asia, 2181 items (42.2%) are China strains, and 921 items (17.8%) are Europe strains.

We compared 4 features of HBV sequences, the ratio of the predominant sequence, the ratio of unique sequences, the ratio of sequences with the predominant length and the length variability (the Number of lengths) (Fig. 1B). These values can characterize both the variability and the conservation of sequences or domains. The preC domain presented the highest ratio of predominant sequence, indicating a highly conserved domain, followed by the core-CTD. The HBpol and large HBsAg had the top two highest ratios of unique sequences.

The lengths of amino acid sequences or domains of HBV were variable. The number of alternative lengths of small HBsAg was the lowest (Fig. 1B and Supplementary Figure S1), while that of the HBpol was the highest. The preC domain and core-CTD were the next lowest after small HBsAg, which revealed their lengths were highly conserved. The preS1/S2 domains and the HBcAg are the most conserved. However, considering the length, the small and middle HBsAg, the preS1-RBD, the preS2 domain, the core-AD and the HBx are the most conserved. It's probable that the length is critical for assembly, so structural proteins, especially the small HBsAg and the core-AD, were conserved except the preS1 region which functions as receptor binding (Yan *et al*. 2012). Moreover, for the HBpol, the sequence length isn't critical. Some domains of the HBpol may provide the flexibility of length while not affecting function.

We analyzed different HBV proteins at the level of amino acid and found that the predominant sequence of HBcAg accounts for more than only around 20% of total sequences (Fig. 1B). However, almost all small HBsAg sequences are of the same length. The preS2 is also highly conserved in length, while the preS1 is more variable. In comparison, the HBpol was variable both in length and in amino acid sequence.

Table 1 The geographical distribution of 5,167 HBV records.

Asia		Europe		North/South America		Africa		Oceania	
China	2181	UK	349	USA	185	South Africa	73	Australia	24
Japan	516	Belgium	151	Argentina	150	Gabon	1	NewZealand	1
India	308	Luxembourg	124	Canada	66				
Malaysia	237	Sweden	97	Brazil	33				
Vietnam	134	Poland	53	Chile	22				
Iran	86	France	43	Mexico	17				
Syria	70	Netherlands	34	Venezuela	13				
Thailand	48	Germany	22	Colombia	4				
Indonesia	40	Italy	17						
Myanmar	15	Turkey	12						
South Korea	10	Spain	11						
Korea	6	Serbia	5						
SaudiArabia	6	Ireland	2						
		Russia	1						
Total	3,657		921		490		74		25

Amino Acid Sequence Homology in HBV Sequences or Domains

We analyzed the conservation of amino acid residues in four HBV ORFs by calculating the proportion of positions that shared the predominant amino acid. As a whole, it's interesting that some sites were highly conserved while some were extremely variable (Fig. 2 and Supplementary Figure S2). For most amino acid residues, sequence homology levels were above 50%. AAs in the ORF S were more variable. Only ~ 20% of preS1 sites and ~ 50% of preS2 sites had sequence homology levels above 95%, while the 95% sequence homology level was reached in more than 60% of sites in other sequences or domains. Nearly 70% of sites in the HBpol reaching this high level of concordance. Even in the ORF C, there were more than 60% sites with predominant AAs ($\geq$ 0.950). The amino acid residues in the preS2 domain and the S domain were more variable than those in the core-AD and the core-CTD. This indicates the surface antigen could be more variable and provides the basis for immune escape. Among all sites, L350 in the large HBsAg and H83 in the HBpol were the most conserved. No mutation was found at the two sites of all HBV strains in our study. Sites with the highest conservation have the most potential as targets for development of new anti-virus strategies.

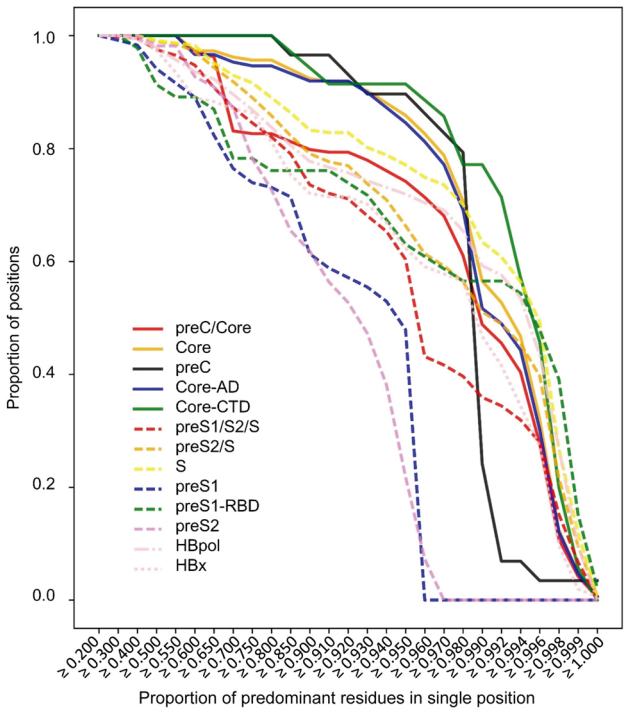

Fig. 2 The similarity of amino acid sequence. The proportion of sites in several genetic regions of HBV that share sequence similarity at various levels of conservation. All amino acids share sequence similarity at 20% or greater (far left). Moving from left to right, sequence similarity for each genetic region is evaluated at increasing levels. Very few amino acid positions share 100% sequence similarity (far right). The proportion of sites with high ratio of predominant residues is lower in the preS1/S2 domain than the S domain. It indicates higher variability of the preS1/S2 domains.

Sequence Variant Analysis

After sequence alignment, we could further obtain the profile of pair-wise similarity in a whole picture. To compare the similarity of sequences or domains in different areas, we extracted subsets of data from different

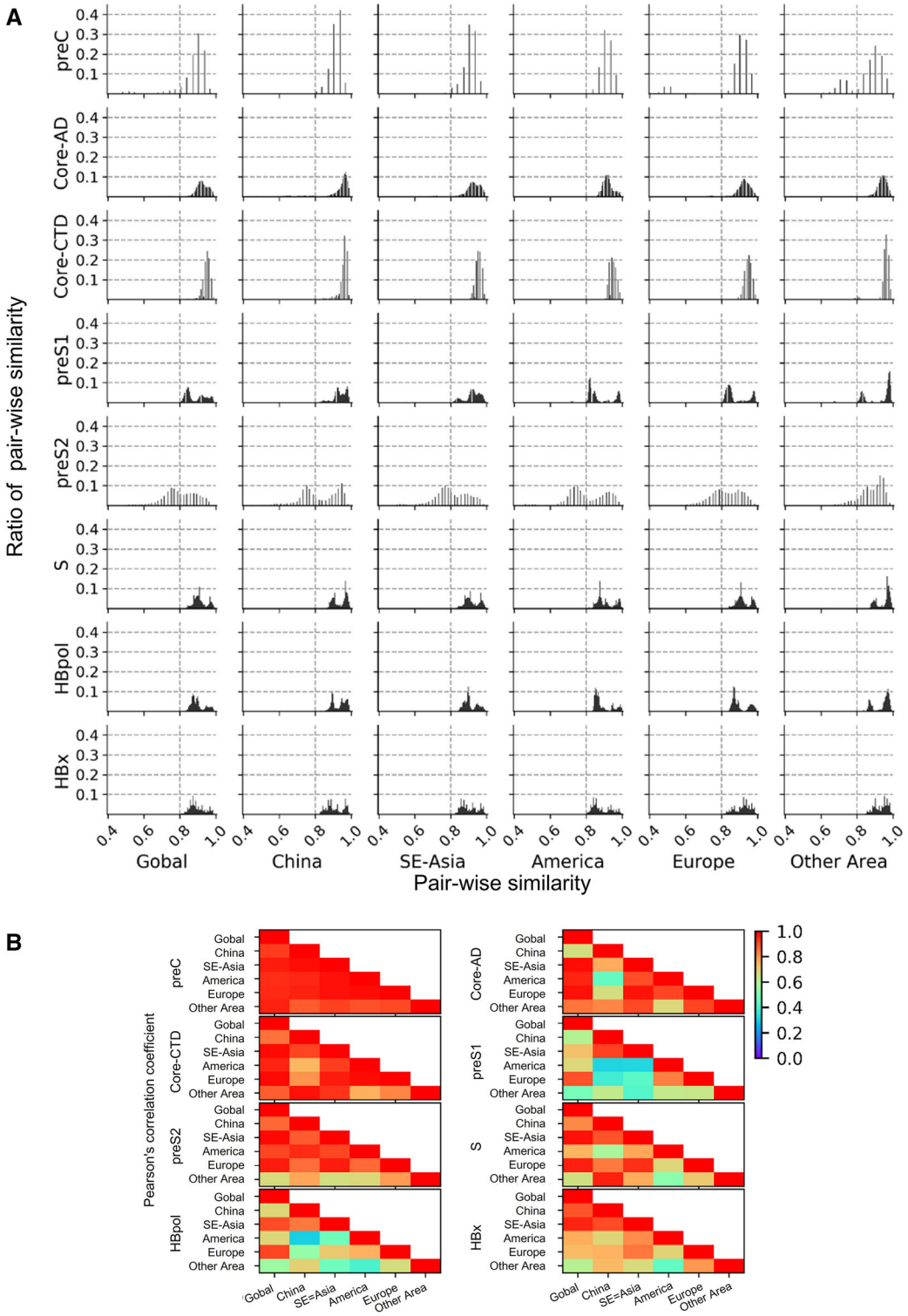

Fig. 3 **A** The histogram profile of pair-wise similarity in different HBV genomic regions, grouped by geographical source of each isolate. The list of countries and the number of sequences from each geographical region are found in Table 1 and section "Materials and Methods". X represents the pair-wise similarity which was defined as the ratio of the number of positions with identical amino acid residues to the length of the sequences. Y represents the ratio of different similarity. The total ratio of pair-wise similarity in each panel was 1. **B** The Pearson cross-correlation coefficient of the profiles of pair-wise similarity in Fig. 3A ($P < 0.001$).

geographic locations, especially China which is a very important area with high prevalence of HBV infection. Surprisely, that different areas exhibited different profiles of pair-wise similarity (Fig. 3A). Almost all pair-wise similarities were more than 80% except that some similarities of preS2 were around 60%. The profiles of both the preC domain and the core-CTD in different areas were highly similar.

When these profiles were compared using the Pearson cross-correlation coefficient, the preS2 domain, the preC domain and the core-CTD show highly coefficient. However, the profiles of the preS1 domain, the S domain, the core-AD, the HBpol and the HBx were different among different areas. When compared among different areas, the similarity profiles are significantly different ($P < 0.001$) for HBpol, preS1, Core-AD (America-China), S (America-China/Other area) and HBx (America-Other area) (Fig. 3B). We also noted that the profile between China and America differ most significantly in Core-AD, preS1, S, HBpol and HBx ($P < 0.001$) (Fig. 3B). The biological difference of the geographical difference in sequence population at amino acid level needs to be further addressed. Maybe it could advance the development of the vaccine or other anti-virus drugs.

Prediction of Secondary Structure of the HBsAg

The HBsAg components include important proteins in immune escape and persistence of immune tolerance. Currently, little is known about the structure of it. From the prediction of its secondary structure, it was surprising that alpha-helices and beta-sheet only exist in the S domain, while the preS1/S2 domains that are only found in the large and middle HBsAg contains only unstructured coils (Fig. 4A). While comparing the different amino acid frequencies of each sites in the HBsAg, we also found that cysteine residues exist in the S domain (Supplementary Figure S3A). Considering the conserved length of S domain, we speculate that the stability of its structure is critical for the assembly of both SVPs and the envelope of Dane particles. There were 40–46 amino acid residues in the N-terminus of preS1 critical for the receptor binding of HBV (Yan *et al*. 2012), so the flexibility in the preS1/S2 domain could provide for allosteric regulation during virus-host recognition.

The Distribution of Predominant Amino Acids in the HBsAg

For the HBsAg sequence, it's interesting that cysteine only exists in S domain (Supplementary Figure S3A). While for HBcAg, another structural protein of virions, only 3 highly conserved cysteine residues exist in the core-CTD (Supplementary Figure S3B). However, cysteine residues exist throughout sequences of the non-structural proteins, the HBpol and the HBx (Supplementary Figure S3C, S3D). A special function of cysteine is to form intra- or inter-molecular disulfide bonds, which can stabilize the structure of proteins. Thus, it's postulated that these disulfide bonds keep the stability of the S domain in the envelope, while the preS1/S2 domain, without any cysteine residues, could provide a highly flexible conformation to facilitate binding to the host receptor.

It has been reported that 40–46 amino acid residues in the N-terminus of the preS1 domain are critical for binding to host receptor (Chouteau *et al*. 2001; Le Duff *et al*. 2009; Yan *et al*. 2012). We found that this domain is highly conserved. However, it's surprising that over 50% of HBV genome records encode the preS1 sequence with an N-terminal extension of the RBD of up to 11 amino acid residues (Fig. 4B). It has never been reported that this sequence is associated with receptor binding. Its amino acid sites were also highly conserved, excepting Lys10.

Discussion

Although HBV carriers in Africa account for almost 25% of those in the world (See Global hepatitis report, WHO 2017), the data about Africa is very few in the Genbank, which is the drawback of our study (Table 1). It also reflects the lack of HBV research in Africa. However, most of HBV sequences come from China. It suggests that China is still a major epidemic area and pays more attention to the research of HBV. The amino acid sequence is a direct determinant of the folding of proteins. Furthermore, during the low-fidelity replication of the HBV genome by reverse transcription, a high degree of sequence variation results. Thus, it's necessary to have a comprehensive investigation on the amino acid sequences of HBV proteins, which explores the variability and conservation of domains that determine their function, as well as differences among different geographical strains.

Our study revealed the variability of HBV ORF S, C, P and X at the amino acid level. It showed the sequence variability both in length and amino acid sequences. The variability of virus proteins doesn't seem to hamper the normal functions. The four ORFs of HBV are partially overlapping. When a mutation happens in an ORF, it also likely happens in another one. It will affect either the function or the transcription of viral proteins. Hence, the conserved regions in the HBV genome are functionally important. However, it also reveals functional flexibility in highly variable regions.

Importantly, the variability of the HBsAg at the level of amino acid provides many potential epitopes for immune

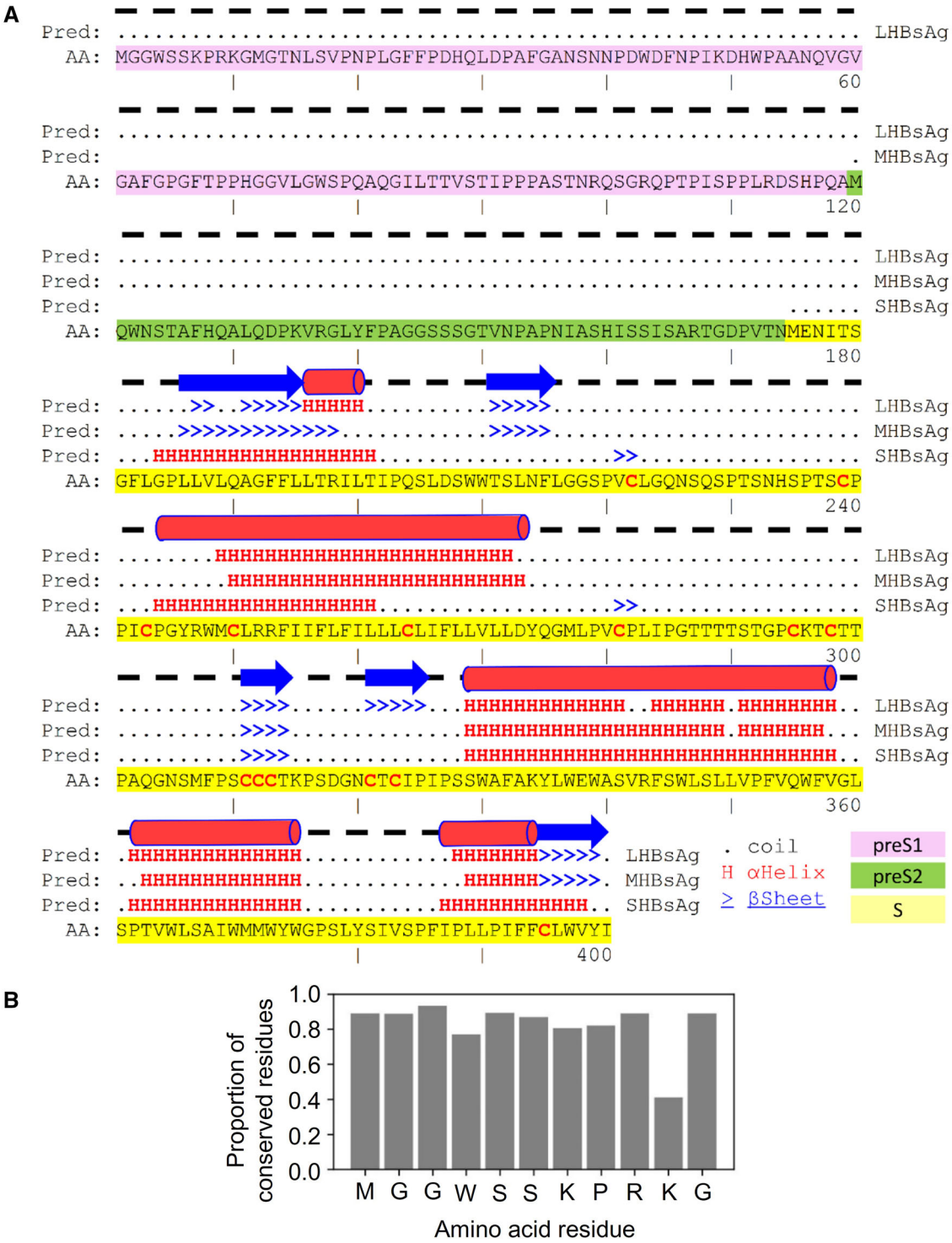

Fig. 4 **A** Secondary structure prediction of the three HBsAg isoforms. Representation is the prediction of 3 forms of HBsAg components. Alpha-helices and beta-sheets only exist in the S domain. The preS1, preS2 and S domains are colored in purple, green and yellow respectively. Cysteine residues are colored in red. Red cyslinders stand for alpha-helices, blue arrows stand for beta-sheet and dash lines stand for unstructured colis in the cartoon for possible unified foding the 3 forms of HBsAg. **B** The 11 amino acid residues upstream of the preS1-RBD. Most residues were highly conserved except Lys10.

recognition. This is a potential mechanism exhausting the immunocytes or antibodies which recognize the HBsAg. It has been reported that the HBsAg in tubular SVPs organizes regularly in crystalline-like pattern (Short *et al.* 2009). As the C-terminal part of the HBsAg, the S domain is transmembrane and folds as the protrusion on the periphery (Bruss 2004; Short *et al.* 2009). Moreover, there were a total of 12 highly conserved cysteine residues in the S domain (Supplementary Figure S3A), far more than those in the core-AD of HBcAg (Supplementary Figure S3B). It indicates that these cysteine residues are probably critical for the stability of the protein structure, which could help the protrusions on SVPs to arrange in a regular way. Even the cysteine residues in HBcAg are not necessary for the disulfide bond (Yu *et al.* 2013). The core-CTD is located in the internal side of the HBV capsid, which helps the virus enclose the genome during assembly (Zlotnick *et al.* 1997). We postulate that those cysteine residues and the lengths of the S domain are very critical for the stability of the HBsAg structure, especially, since the S domain forms the transmembrane region. Furthermore, the N-terminal extension of the preS1-RBD was also a highly conserved sequence. It's probably associated with the function of receptor binding in some unknown way.

In conclusion, we studied the viral proteins of HBV at the level of amino acid. Quantitative investigation revealed the conservation and variability among different sequences and domains. The critical sequences for virus assembly, the small HBsAg and the core-AD, are the most conserved, as well as the preC domain. However, preS1 domain and HBpol show the highest variability by geographical location. It would be helpful to further study the variant epitopes of the HBsAg in immune escape and recognition of HBV and for the development of new vaccines and antiviral drugs.

Acknowledgements The authors would like to thank Prof. Ping Zhu (Institute of biophysics, Chinese Academy of Sciences) and Prof. Jingqiang Zhang (Sun Yat-sen university), who provided help in this research. This work was partially supported by the National Natural Science Foundation of China (Nos. U1611265, 81773271 and 31672536) and the Key Projects of Department of Education of Guangdong Province (No. 2017KZDXM088). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions JC, SL and YX designed the study. JC conducted computational work, JC and YX performed data analysis. JC, SL and YX wrote the manuscript draft.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Animal and Human Rights Statement This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Bruss V (2004) Envelopment of the hepatitis B virus nucleocapsid. Virus Res 106:199–209

Buynak EB, Roehm RR, Tytell AA, Bertland AU, Lampson GP, Hilleman MR (1976) Vaccine against human hepatitis B. JAMA 235:2832–2834

Chouteau P, Le Seyec J, Cannie I, Nassal M, Guguen-Guillouzo C, Gripon P (2001) A short N-proximal region in the large envelope protein harbors a determinant that contributes to the species specificity of human hepatitis B virus. J Virol 75:11565–11572

Dembek C, Protzer U, Roggendorf M (2018) Overcoming immune tolerance in chronic hepatitis B by therapeutic vaccination. Curr Opin Virol 30:58–67

Heermann KH, Goldmann U, Schwartz W, Seyffarth T, Baumgarten H, Gerlich WH (1984) Large surface proteins of hepatitis B virus containing the pre-s sequence. J Virol 52:396–402

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Le Duff Y, Blanchet M, Sureau C (2009) The pre-S1 and antigenic loop infectivity determinants of the hepatitis B virus envelope proteins are functionally independent. J Virol 83:12443–12451

Seeger C, Mason WS (2000) Hepatitis B virus biology. Microbiol Mol Biol Rev 64:51–68

Seeger C, Mason WS (2015) Molecular biology of hepatitis B virus infection. Virology 479:672–686

Short JM, Chen S, Roseman AM, Butler PJG, Crowther RA (2009) Structure of hepatitis B surface antigen from subviral tubes determined by electron cryomicroscopy. J Mol Biol 390:135–141

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mole Syst Biol 7:539

Yan H, Zhong G, Xu G, He W, Jing Z, Gao Z, Huang Y, Qi Y, Peng B, Wang H (2012) Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. eLife 1:e00049

Yu X, Jin L, Jih J, Shih C, Zhou ZH (2013) 3.5 Å cryoEM structure of hepatitis B virus core assembled from full-length core protein. PLoS ONE 8:e69729

Zlotnick A, Cheng N, Stahl SJ, Conway JF, Steven AC, Wingfield PT (1997) Localization of the C terminus of the assembly domain of hepatitis B virus capsid protein: implications for morphogenesis and organization of encapsidated RNA. Proc Natl Acad Sci 94:9556–9561