# Establishment of reference sequences of hepatitis B virus genotype C subgenotypes

**H.L. Zhu[1], C.T. Wang[1], J.B. Xia[2], X. Li[1] and Z.H. Zhang[1]**

[1]Virology Laboratory, Department of Infectious Diseases,
The First Affiliated Hospital, Anhui Medical University, Hefei, Anhui, China
[2]Virology Laboratory, Department of Clinical Laboratory,
Hubei Province Maternal and Child Health Hospital, Wuhan, Hubei, China

Corresponding author: Z.H. Zhang
E-mail: zzh1974cn@163.com

Genet. Mol. Res. 14 (4): 16521-16534 (2015)
Received July 2, 2015
Accepted September 25, 2015
Published December 9, 2015
DOI http://dx.doi.org/10.4238/2015.December.9.24

**ABSTRACT.** Hepatitis B virus genotype C (HBV/C) has the largest number of subgenotypes (C1-C16) that vary with geography and isolates. HBV/C prevails in Southeast Asia (C1, C5-C16), East Asia (C2), Oceania (C3), and Australia (C4). Suitable reference strains for different subgenotypes could greatly facilitate research into HBV/C, but unfortunately they are scarce. We retrieved 974 HBV/C full-length sequences from the GenBank database and subgenotyped them by phylogenetic analysis. Reference sequences of each subgenotype from different locations were established with the most frequent nucleotide present at each position of the isolates that belonged to the same subgenotype. The reference sequences of subgenotypes C1, C2, C5, and C6 have been constructed and deposited in GenBank (KM999990-KM999993). The homology between the reference sequences and almost all the isolates belonging to the corresponding subgenotype was higher than 96%. Similarly, bootstrap values in phylogenetic trees supported clustering of reference strains with isolates belonging to the same subgenotypes. Moreover, both homology and phylogeny analyses showed that reference sequences had significant heterogeneity with isolates from other genotypes and subgenotypes. Sequence analysis

further revealed that the mutation rate in the basal core promoter (BCP) region was extremely high in HBV/C2, relatively high in HBV/C1, but lower in HBV/C5 and HBV/C6. Mutations in the pre-core (Pre-C) region were common in HBV/C but the mutation rate was lower than in the BCP. HBV/C5 has the oldest ancestral age, followed by C6, which is much more ancient than C1 and C2. This study successfully established references for HBV/C subgenotypes.

**Key words:** Hepatitis B virus; Genotype; Sequence analysis

## INTRODUCTION

Hepatitis B virus (HBV) infection is a major public health problem worldwide; with more than 240 million people chronically infected, it causes approximately 6 million deaths annually (Ott et al., 2012). Owing to a lack of proofreading activity in the HBV reverse sequence, the nucleotide substitution rate of the complete HBV genome has been estimated at 2.2 x $10^{-6}$ substitutions/site/year, which is intermediate between the rates for RNA and other DNA viruses (Okamoto et al., 1987; Paraskevis et al., 2013). As a result, HBV is prone to mutation, may exist as a quasispecies, and gives rise to new genotypes, subgenotypes, and recombinants (Kim et al., 2011).

Currently, HBV is classified into eight genotypes (A to H) based on greater than 8% nucleotide variation of the genome. Genotypes A-F have been further split into subgenotypes with a divergence of over 4% throughout the genome (Okamoto et al., 1988; Norder et al., 1994). Novel genotypes I and J have also been described but remain controversial (Tatematsu et al., 2009; Phung et al., 2010). HBV genotypes and subgenotypes vary with geography as well as virological and clinical properties.

Asia is one of the largest global pools of chronic HBV infections, with a carriage rate of hepatitis B surface antigen (HBsAg) greater than 8% among the general population (Ott et al., 2012). With 16 subgenotypes and related recombinants having been reported, HBV/C is predominant in Asia. However, some recently described subgenotypes have not been fully accepted (Shi et al., 2012). HBV/C1 (Cs) is mainly found in Southeast Asia, whereas C2 (Ce) is predominant in East Asia (Huy et al., 2004). HBV/C3 is confined to Oceania, while C4 (Caus) is exclusively found in Australia and is regarded as the most divergent subgenotype within HBV/C (Davies et al., 2013). Subgenotypes C5 and C7 are found in the Philippines, while C6 and C8-C16 have been isolated from Indonesian sources (Sakamoto et al., 2006; Lusida et al., 2008; Mulyanto et al., 2010; Mulyanto et al., 2011; Juniastuti et al., 2011; Davies et al., 2013). Most researchers focus on subgenotypes C1 and C2 as they are the two major groups within HBV/C (Huy et al., 2004). With small quantities and rare subsequent studies, subgenotypes C3 and C7-C16 sometimes serve as outgroup isolates in phylogenetic analysis.

Reference sequences of HBV that are closely related to strains of the same clade (genotypes and subgenotypes) are important in sequence studies. A number of reference sequences representing HBV genotypes are currently in use. However, these reference sequences were either determined from the first isolated strain or established from a limited number of isolates (Bichko et al., 1985; Naumann et al., 1993; Norder et al., 1994; Stuyver et al., 2000; Owiredu et al., 2001; Arauz-Ruiz et al., 2002; Sugauchi et al., 2002; Zhang et al., 2009). As new HBV strains are constantly isolated in large numbers and reference strains for subgenotypes are still scarce, there is an urgent need to take these diversities into account and eliminate biases to establish reference strains that are more representative.

In this article, a number of updated reference sequences for different HBV/C subgenotypes were established by large-scale phylogenetic analysis of full-length DNA sequences obtained from GenBank. These reference sequences may contribute to studies on the epidemiological and virological features of HBV/C.

## MATERIAL AND METHODS

### Sequences collection and subgenotyping

Full-length HBV sequences (3435) were retrieved from original publications by searching the GenBank nucleotide database using the terms: "HBV, genotype, complete, genome" in November 2013, and sequences that belonged to genotype C were selected for further analysis. Identical entries and sequences with polynucleotides of less than 3100 bases or more than 3400 bases were excluded. Recombinants and HBV/C4 isolates were also excluded. Subgenotypes were determined by phylogenetic comparison of entire genomic sequences. The phylogenetic trees were constructed using the neighbor-joining algorithm with 1000 bootstrap replicates, and all other parameters were set to default in the MEGA 5.2 software. Subgenotypes were determined if sequences phylogenetically clustered with any of the known subgenotypes, as described previously (Hall, 2013).

### Establishment of HBV/C subgenotype reference strains

Multiple alignments of all entire genomic sequences of each HBV/C subgenotype from different regions or countries were performed with ClustalW, which is included in the MEGA 5.2 software package, and the results of alignments were further confirmed by visual inspection. The reference sequences of each HBV/C subgenotype from the different regions or countries were established with the most common nucleotides at each position of all isolates belonging to that subgenotype from the corresponding location. The frequency of base substitution in each position of each HBV/C subgenotype from different regions or countries was analyzed and mutations in hotspot positions such as 1762, 1764, 1814, 1858, and 1896 in the BCP/Pre-C region were recorded.

### Phylogenetic trees and divergence time

The MEGA 5.2 software package was used to construct the phylogenetic trees and calculate relative divergence times. The phylogenetic trees were constructed either among reference sequences or with reference sequences and 20 randomly selected isolates of the same subgenotypes using the neighbor-joining algorithm with 1000 bootstrap replicates and all other parameters were set to default. For a subgenotype with a total number of less than 20 isolates, all isolates were selected. The maximum likelihood algorithm was used for the calculation of divergence time. The estimate of the substitution rate of $2.2 \times 10^{-6}$ per site per year for the full-length viral genome was used for divergence time calculation.

### Nucleotide and amino acid analysis of the reference strains

Homology analysis was performed using the DNAMAN V6 software package (Zhang et

al., 2009). For a subgenotype with a total number of greater than or equal to 20 isolates, homology analysis was conducted with three strains that were most similar to, and another three strains that were most phylogenetically divergent from, the corresponding reference sequence. Otherwise, homology analysis was conducted with the reference sequence and all isolates belonging to that subgenotype. Statistics such as maximum, minimum, average, median, and quartile of homology were obtained using SPSS software (SPSS Company, Chicago, IL) and the relative box-chart was constructed using GraphPad software (GraphPad Company, La Jolla, CA). Homology analyses at the nucleotide and amino acid levels were conducted for complete genomes and partial genomic fragments of newly established and previous reference sequences. Partial genomic fragments such as the pre-S1, pre-S2, S, pre-C, C, P, and X regions were prepared in advance according to the genome of HBV/C subgenotypes.

## RESULTS

### Geographic distribution of HBV/C subgenotypes

Among the 3435 HBV full-length strains, 1082 isolates were HBV/C including 61 recombinants, 15 HBV/C4, 19 HBV/C7-C16, and 13 unclassified. Recombinants and HBV/C4 strains were excluded because they are still controversial owing to the potential roles they play in phylogenetic analysis-based subgenotyping (Shi et al., 2012; Davies et al., 2013). No sequences were grouped into HBV/C3 and fewer than five isolates were classifiable into each subgenotype from C7 to C16, thus there was no need to establish reference sequences for these subgenotypes. Therefore, 974 HBV/C strains were selected for reference sequence establishment. In keeping with previous reports that HBV/C is hyperendemic in Asia, the overwhelming majority of these strains were from Asia, and C2 and C1 were the major subgenotypes. Most C2 strains (86.46%) were from China. Southeast Asia accounted for 48% of C1, 90.01% of C5, and all C6 strains, mirroring the endemic distribution of these four subgenotypes (Table 1). In these regions, HBV/C1 (Southeast Asia) and HBV/C2 (China) were also the main local subgenotypes (Table 2).

**Table 1.** Geographic distribution of hepatitis B virus genotype C (HBV/C) subgenotypes.

| Subgenotype | Location | N | Percent (%) | Total |
|---|---|---|---|---|
| C1 | **Southeast Asia** | **96** | **48.00** | 200 |
| | Hong Kong | 39 | 19.50 | |
| | China | 31 | 15.50 | |
| | India | 11 | 5.50 | |
| | Others | 23 | 11.50 | |
| C2 | **China** | **645** | **86.46** | 746 |
| | Japan | 58 | 7.77 | |
| | South Korea | 10 | 1.34 | |
| | Southeast Asia | 10 | 1.34 | |
| | Taiwan | 10 | 1.34 | |
| | Others | 13 | 1.74 | |
| C5 | **Southeast Asia** | **10** | **90.91** | 11 |
| | Taiwan | 1 | 9.09 | |
| C6 | **Southeast Asia** | **17** | **100.00** | 17 |

Locations where certain subgenotype(s) dominate are highlighted in bold. Southeast Asia includes Thailand, Malaysia, Indonesia, and the Philippines.

**Table 2.** Hepatitis B virus genotype C (HBV/C) subgenotypes in different locations.

| Location | Subgenotype | N | Percent (%) | Total |
|---|---|---|---|---|
| China | C1 | 31 | 4.59 | 676 |
| | **C2** | **645** | **95.41** | |
| Japan | C1 | 6 | 9.38 | 64 |
| | **C2** | **58** | **90.63** | |
| Hong Kong | **C1** | **39** | **88.64** | 44 |
| | C2 | 5 | 11.36 | |
| India | **C1** | **11** | **78.57** | 14 |
| | C2 | 3 | 21.43 | |
| Taiwan | C1 | 3 | 21.43 | 14 |
| | **C2** | **10** | **71.43** | |
| | C5 | 1 | 7.14 | |
| Southeast Asia | **C1** | **96** | **72.18** | 133 |
| | C2 | 10 | 7.52 | |
| | C5 | 10 | 7.52 | |
| | C6 | 17 | 12.78 | |
| South Korea | **C2** | **10** | **100.00** | 10 |

Dominant subgenotypes in each location are indicated in bold. Southeast Asia includes Thailand, Malaysia, Indonesia, and the Philippines.

## Establishment of reference sequences of HBV/C subgenotypes

By aligning 974 different HBV/C full-length strains, reference sequences of subgenotypes C1, C2, C5, and C6 from different locations were constructed and submitted to GenBank with accession Nos. KM999990-KM999993 and KP017266-KP017272. For convenience, reference sequences of HBV/C2 in China (C2 for short) and subgenotypes C1, C5, and C6 in Southeast Asia (C1, C5 and C6 for short, respectively) were chosen to represent the corresponding subgenotypes (KM999990-KM999993).

## Nucleotide and amino acid analysis of reference strains

The established references shared greater than 96% homology with all isolates of the same subgenotype. In the case of HBV/C5, differences among all sequences were within 4% (Figure 1). The phylogenetic tree showed that all reference strains had monophyletic clustering with good bootstrap support for isolates of the same subgenotype (Figure 2). Although both previously published C and our HBV/C2 reference sequences clustered with C2 strains, our reference strain was phylogenetically more similar to them, suggesting that previous C sequences also belonged to subgenotype C2 while our newly established reference sequence defined HBV/C2 more accurately.

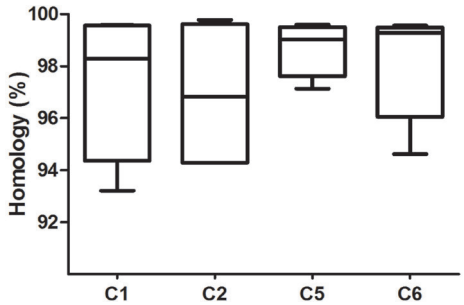



**Figure 1.** Homology percentages between reference strains and isolates of the corresponding subgenotypes.

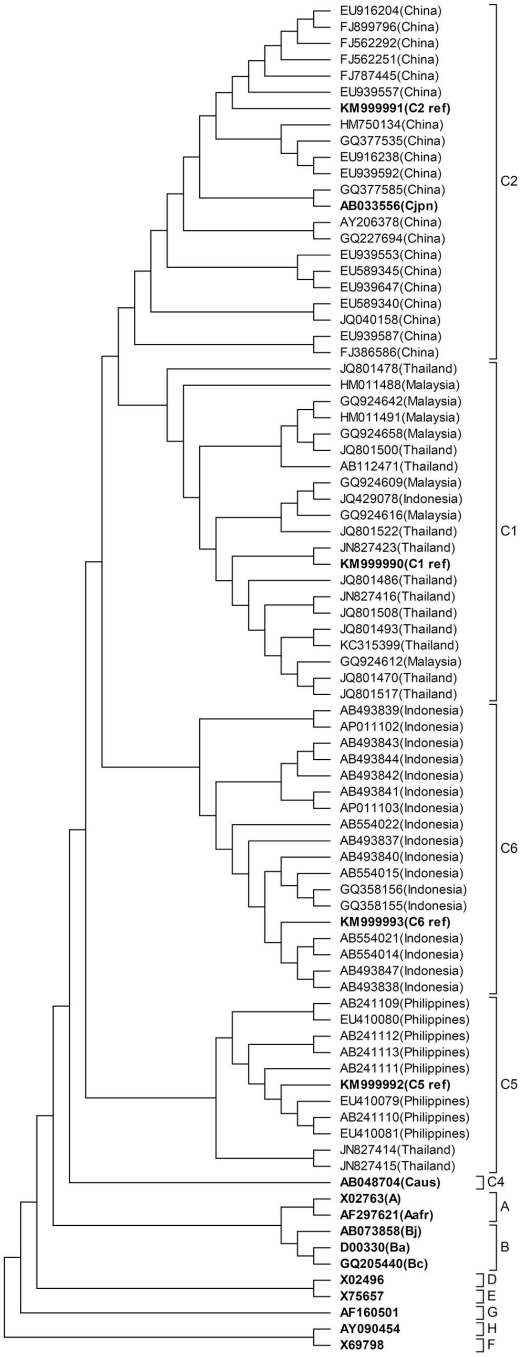



**Figure 2.** Phylogenetic trees of reference sequences and some selected strains constructed by the neighbor-joining method. Accession numbers and the locations are shown in each branch and reference strains are highlighted in bold. Hepatitis B virus (HBV) genotypes and subgenotypes are listed on the right of each respective cluster.

Compared with the previous reference strains of HBV/C, our newly established reference strains showed high homology with the C strain (min 95.52%), followed by C4 (Caus) (min 92.29%), but significant heterogeneity with the other genotypes (Table 3). The homology of S genes among the HBV/C reference strains was slightly higher than among the complete genomes. Moreover, all new reference strains, especially HBV/C2, were more closely related to the previous C strain than C4 at both the nucleotide and amino acid levels. The P gene and polymerase gene showed the least homology in all four open reading frames (ORFs).

**Table 3.** Homology analysis of complete genomes and S genes between reference strains of established hepatitis B virus genotype C (HBV/C) subgenotypes and genotypes A-H.

| Reference strains | Accession Nos. | C1 | | C2 | | C5 | | C6 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Full | S gene | Full | S gene | Full | S gene | Full | S gene |
| A | X02763 | 91.07% | 91.85% | 91.54% | 92.44% | 91.01% | 91.94% | 91.10% | 92.27% |
| Aafr | AF297621 | 91.26% | 92.48% | 91.64% | 92.99% | 91.20% | 92.82% | 91.20% | 92.91% |
| Ba | D00330 | 91.73% | 91.19% | 91.85% | 91.69% | 91.10% | 91.35% | 91.17% | 91.02% |
| Bc | GQ205440 | 91.98% | 91.35% | 92.16% | 91.85% | 91.42% | 91.52% | 91.38% | 91.19% |
| Bj | AB073858 | 90.23% | 90.94% | 90.30% | 91.35% | 89.42% | 91.02% | 89.98% | 90.52% |
| C | AB033556 | **96.83%** | **97.42%** | **98.66%** | **98.84%** | **95.52%** | **96.26%** | **96.14%** | **96.84%** |
| Caus | AB048704 | 93.16% | 93.35% | 94.09% | 94.35% | 92.29% | 92.68% | 93.84% | 94.18% |
| D | X02496 | 90.25% | 90.68% | 90.57% | 91.37% | 90.19% | 91.11% | 89.65% | 90.51% |
| E | X75657 | 89.82% | 89.50% | 90.29% | 89.92% | 90.13% | 90.00% | 89.76% | 89.50% |
| F | X69798 | 86.81% | 85.79% | 87.09% | 86.20% | 86.84% | 86.62% | 87.06% | 86.12% |
| G | AF160501 | 88.04% | 88.92% | 88.20% | 89.17% | 87.95% | 89.50% | 87.76% | 89.08% |
| H | AY090454 | 86.27% | 86.28% | 86.39% | 86.37% | 85.52% | 86.62% | 86.61% | 86.78% |

Full indicates complete genome.

Among the four new reference sequences, subgenotypes C1, C5, and C6 were most closely related to HBV/C2. HBV/C2 showed greatest similarity to C1, while C1, C2, and C6 had the lowest homology with C5, and this trend held true for both the complete genome and the S gene alone (Table 4). Correspondingly, the phylogenetic trees showed that C2 was most phylogenetically similar to the other three subgenotypes, while C5 was most divergent from the other three subgenotypes (Figure 3A).

**Table 4.** Homology analysis of complete genomes and S genes between reference strains of hepatitis B virus genotype C (HBV/C) subgenotypes.

| Reference strains | C1 | | C2 | | C5 | |
|---|---|---|---|---|---|---|
| | Full | S gene | Full | S gene | Full | S gene |
| C2 | 97.39% | 97.76% | - | - | | |
| C5 | 94.84% | 95.43% | 95.99% | 96.84% | - | - |
| C6 | 95.33% | 96.09% | 96.42% | 97.42% | 94.40% | 95.34% |

Full indicates complete genome.

## Phylogenetic trees and distance calculations of HBV/C complete genomes

With a published nucleotide substitution rate of $2.2 \times 10^{-6}$ per site per year, subgenotype C5 diverged from the most recent common ancestor (tMRCA) about 12,000 years ago and C6 evolved about 10,000 years ago, while C1 and C2 appeared about 6000 years ago (Figure 3A). C1-Southeast Asia and C1-India evolved about 1100 years ago, about 100 years earlier than C1-

China and C1-Hong Kong (Figure 3B). C2-China had a similar divergence time to C2-Taiwan, C2-Southeast Asia, and C2-Japan of about 550 years ago, while C2-South Korea appeared about 1500 years ago (Figure 3C).

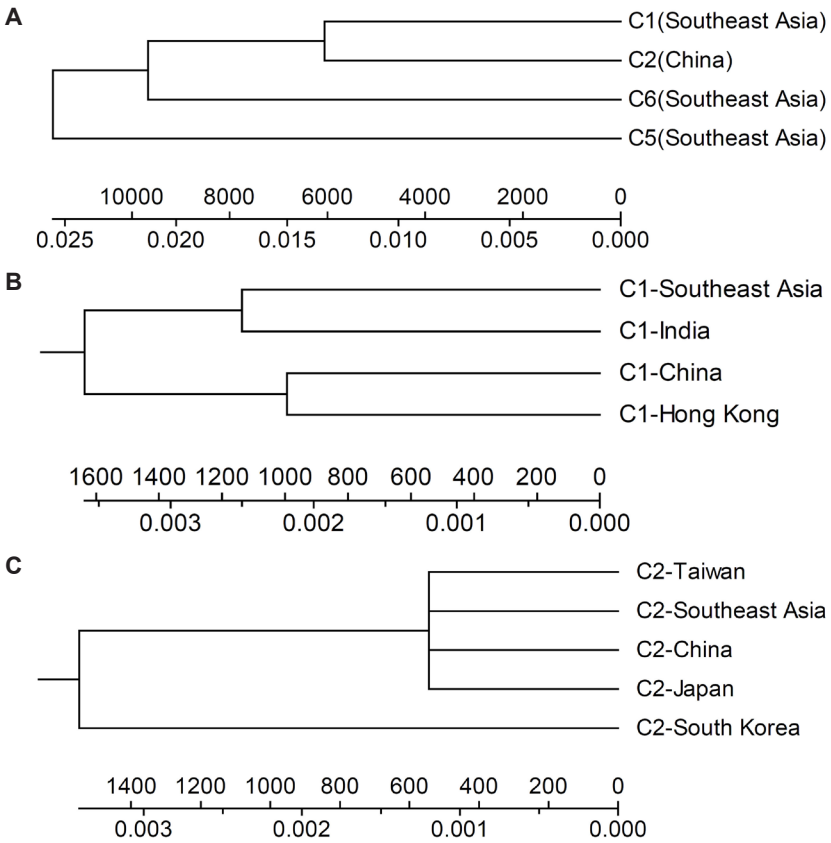



**Figure 3.** Phylogenetic trees of the reference strains. Phylogenetic trees of the reference strains of hepatitis B virus genotype C (HBV/C) subgenotypes (**A**), HBV/C1 in different locations (**B**), and HBV/C2 in different locations (**C**) were constructed using the neighbor-joining method and calibrated with a substitution rate of $2.2 \times 10^{-6}$ per site per year. Subgenotypes with corresponding locations are shown in each branch. The branch lengths correspond to length of time (see the timescale bar at the bottom of the tree).

## Nucleotide substitution of reference sequences of HBV/C subgenotypes

Nucleotide substitutions in the BCP/Pre-C regions are commonly found in HBV/C. The occurrence of A1762T/G1764A in BCP reached 52.08% (C1), 69.77% (C2), 30% (C5), and 11.76% (C6). The substitution rate in the Pre-C region was significantly lower than in the BCP region. No mutation occurred at nucleotide position 1814 in any of the HBV/C strains. The frequency of G1896A reached 21.88% in HBV/C1 and 30.08% in C2. C1858 variation frequency reached 13.54% in C1 and T1858 variation was found in all HBV/C2 strains. No mutation was found in nucleotide positions 1858 and 1896 in both C5 and C6 subgenotypes.

The frequency of base alternation at other nucleotide positions of reference sequences was also analyzed (Figure 4). The reference sequences of C1, C2, C5, C6, and the corresponding analysis of nucleotide substitution were listed in odd-numbered and even-numbered lines, respectively. The nucleotides of reference sequences were probably subgenotype-specific and the sites in accordance with those in C1 reference are represented by black dots. The blue and red dots indicate variable sites with base substitution rates greater than 5% and 20%, respectively. The inter-subgenotype divergences for subgenotypes C1, C2, C5, and C6 were found to be relatively low, but the strains within subgenotypes revealed considerable intra-subgenotype divergences, reflecting the well-conserved area and mutant-rich regions. The area with constant sites sharing identity with the reference sequences was considered well conserved. The region with some red dots was regarded as highly mutated, while that with blue dots was considered prone to mutation and requires further attention. Moreover, the base alternation rate of the S or X gene was found to be lower than that of the C or P genes.

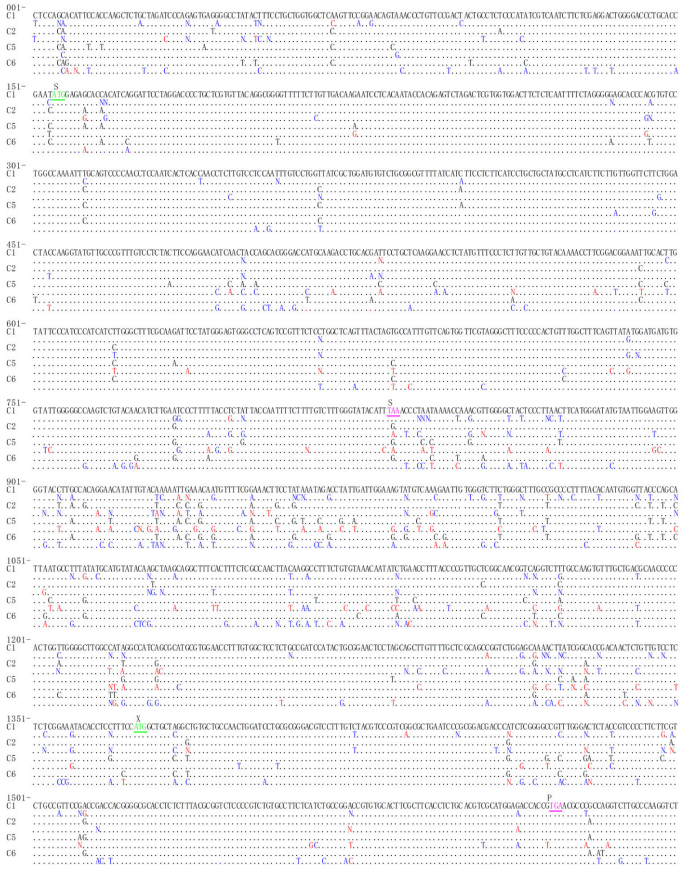
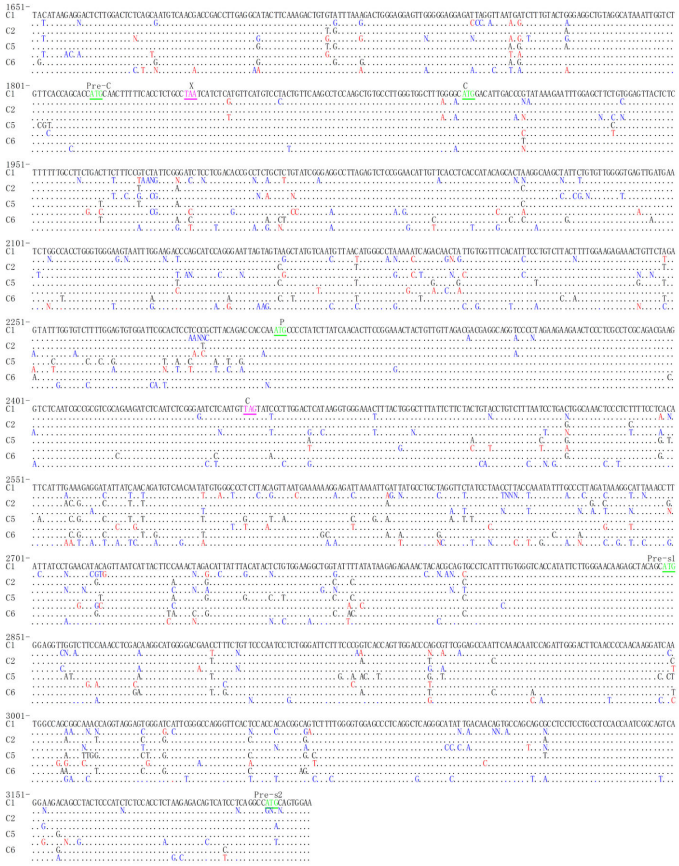


**Figure 4.** Reference strains of hepatitis B virus genotype C (HBV/C) subgenotypes and frequency of nucleotide substitutions. Dots with different colors indicate rate of alternation: position with no alteration, black; position with >5% alternation, blue; position with >20% alternation, red. N in position means multiple base substitution (R = A/G, Y = C/T, M = A/C, K = G/T, S = C/G, W = A/T, H = A/C/T, B = C/G/T, V = A/C/G, D = A/G/T, N = A/C/G/T).

Continued on next page

**Figure 4.** Continued.



## DISCUSSION

Consistent with previous studies, HBV/C was predominant in Asia with subgenotype C2 in China and subgenotypes C1 and C5-C16 in Southeast Asia. Environmental, host, and viral factors contribute to the clinical and virological differences among genotypes and subgenotypes (Kim et al., 2011; Liu and Kao, 2013). Reference sequences that fully represent strains isolated in certain endemic areas are vital in sequence research, especially on mutation. Strategies for constructing HBV reference sequences have been proposed and a number of reference strains have been established with limited numbers of isolates (Zhang et al., 2009; Zhang et al., 2015). However, subgenotypes have always been ignored. The authors of a recent study reported that a chemically synthesized HBV Bc consensus genome was replication-competent *in vitro* and *in vivo*, and genetically similar to isolates of the same subgenotype (Zhang et al., 2015). Similar studies on the reassortment of H1N1 virus and tobacco mosaic virus have demonstrated that viruses can be chemically synthesized and sustain biological functions, although they may not exist in nature (Imai et al., 2012; Cooper, 2014).

In the present study, the reference sequences of HBV subgenotypes C1, C2, C5, and C6 were established on the basis of the sequence data of unprecedented large numbers of isolates from different endemic regions. Accurate classification is important and these isolates were carefully assigned to their respective genotype and subgenotypes (Hall, 2013). The reference sequences of each subgenotype in hyperendemic locations were selected to represent those subgenotypes. Homology and phylogeny analyses together proved that our newly established reference strains were highly representative. Although the previous C reference sequence was found to belong to HBV/C2, our newly established C2 reference sequence, which was phylogenetically more similar to C2 strains, was more representative of the subgenotype. Thus, these reference strains could exhibit characteristic virological, epidemiological, and clinical features of HBV/C subgenotypes.

In each subgenotype, well-conserved areas and mutation-rich regions were determined by aligning a great number of complete genomic strains. It has been suggested that well-conserved areas are related to commonality, whereas mutation regions involve inter-strain variation (Zhang et al., 2015). Therefore, identification of conserved and mutant-rich regions can help dissect the roles of different regions in viral biology, epidemiology, and pathogenesis, among others. Genotype and subgenotype classifications were based on nucleotide variation of the complete genome and S gene (Okamoto et al., 1988; Norder et al., 1994). In this study, the S gene was regarded as relatively conserved by the relative low inter-strain divergences in all ORFs, which could to some extent explain why the S gene, but not other ORFs, was selected for subgenotype and genotype determination. At the same time, identification of well-conserved areas may improve viral detection, primer design, viral gene amplification, and sequencing. Additionally, these established reference sequences in which well-conserved areas can be easily located may facilitate the development and/or improvement of HBV vaccines. As the number of isolates was large in this study, the analyses of alternation in each site and the revealed well-conserved areas as well as mutation-rich regions were of considerable informative value.

HBV/C confers increased risk of clinically severe infection compared with other genotypes. Compared with genotype B, HBV/C has a poorer response to interferon therapy, which may be related to the high propensity of mutations in the BCP/Pre-C regions (Araujo et al., 2011). As mutations in the BCP/Pre-C regions are associated with HBe antigen seroconversion (SC) and viral replication, HBV/C has a higher tendency to develop BCP/Pre-C mutations and a higher viral load than genotype B (Liu and Kao, 2013). It has been reported that HBV/Ba (B2), which features recombination with HBV/C in the Pre-C/C regions, is more likely to cause the development of serious liver disease than HBV/Bj (B1), which is not a recombinant (Liu et al., 2009). A1762T and G1764A mutations have been regarded as risk factors for hepatocellular carcinoma (HCC), whereas G1896A has been associated with decreased incidence of HCC (Yang et al., 2008; Tseng et al., 2015). Nucleotide positions 1858 and 1896 form a pair in the epsilon stem loop structure, which is critical in viral reverse transcription of HBV (Chotiyaputta and Lok, 2009). The C1858 variant is detected frequently in HBV/C1 but not in C2, thus the frequency of G1896A is lower in C1 compared with C2. However, the mutation rate of the Pre-C region is significantly lower than in the BCP region in genotype C (Liu and Kao, 2013). Therefore, the fact that HBV/C2 has a higher BCP variants rate compared with C1 could help explain why both subgenotypes C1 and C2 have been associated with HCC, but only HBV/C2 is regarded as an independent risk factor (Zhang et al., 2008). A prospective cohort of 1006 patients chronically infected with HBV who were followed up for approximately 7.7 years also confirmed that the highest risk for HCC was HBV/C2 infections, followed by HBV/C1 in comparison with genotype B (Chan et al., 2008).

With distinct geographical distributions, the oldest genotype, namely HBV/C, has been associated with anthropological history and has been used to help understand how HBV has spread in Asia and globally (Okamoto et al., 1987). HBV infected humans around 33,600 years ago with an estimated substitution rate of $2.2 \times 10^{-6}$ per site per year (Paraskevis et al., 2013). With the closest neighbor being subgenotype C2, it has been suggested that HBV/C6 most probably evolved from C2 after being introduced into a new genetic environment (Cavinta et al., 2009). Our present study showed that HBV/C5 and HBV/C6 were more ancient than C1 and C2, which is concordant with previous estimated divergence times for these subgenotypes (Paraskevis et al., 2013). Subgenotypes C5 and C6 were isolated in limited numbers from Southeast Asia, which may be their most probable origin. A 16th century AD Korean mummy infected with HBV/C shared greatest similarity with contemporary Korean HBV/C2 sequences (97% similarity), indicating that the mummy-derived HBV sequence may have migrated from Chinese or Japanese communities to Korea and evolved at least 3000 years ago, most probably earlier, with a mutation rate of about $10^{-5}$ per site per year (Kahila Bar-Gal et al., 2012; Locarnini et al., 2013). Similarly, HBV/C strains from eastern India have high similarity to those from Southeast Asia, suggesting an introduction from Asian countries to India (Datta, 2008).

However, several limitations exist in this study. Some isolates might not have been retrieved by the search terms used. The accuracy of collected full-length sequences was not completely guaranteed since they were obtained from the database. Moreover, the number of sequences was large but still limited. For instance, India has intermediate HBV endemicity and HBV/C infection is common, but HBV/C isolates from India were rarely found in the database (Ott et al., 2012). The analysis of the substitutions could not take some factors affecting each individual sequence into account; for example, the patients from whom HBV was isolated may have been at different disease stages and receiving anti-viral treatment or no treatment. Thus, future research on HBV reference strains should try to address these biases.

In conclusion, reference sequences representing subgenotypes C1, C2, C5, and C6 were successfully established by large-scale phylogenetic analysis. This work may greatly facilitate future HBV/C studies on genetic variations, epidemiology, pathogenicity, vaccine development, and therapeutic intervention.

## Conflicts of interest

The authors declare no conflict of interest.

## ACKNOWLEDGMENTS

Research supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China (#20093420120005) and the National Science Foundation of China (#30771907). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## REFERENCES

Araujo NM, Waizbort R and Kay A (2011). Hepatitis B virus infection from an evolutionary point of view: how viral, host, and environmental factors shape genotypes and subgenotypes. *Infect. Genet. Evol.* 11: 1199-1207.

Arauz-Ruiz P, Norder H, Robertson BH and Magnius LO (2002). Genotype H: a new Amerindian genotype of hepatitis B virus

revealed in Central America. *J. Gen. Virol.* 83: 2059-2073.
Bichko V, Pushko P, Dreilina D, Pumpen P, et al. (1985). Subtype ayw variant of hepatitis B virus. DNA primary structure analysis. *FEBS Lett.* 185: 208-212.
Cavinta L, Sun J, May A, Yin J, et al. (2009). A new isolate of hepatitis B virus from the Philippines possibly representing a new subgenotype C6. *J. Med. Virol.* 81: 983-987.
Chan HL, Tse CH, Mo F, Koh J, et al. (2008). High viral load and hepatitis B virus subgenotype ce are associated with increased risk of hepatocellular carcinoma. *J. Clin. Oncol.* 26: 177-182.
Chotiyaputta W and Lok AS (2009). Hepatitis B virus variants. *Nat. Rev. Gastroenterol. Hepatol.* 6: 453-462.
Cooper B (2014). Proof by synthesis of Tobacco mosaic virus. *Genome Biol.* 15: R67.
Datta S (2008). An overview of molecular epidemiology of hepatitis B virus (HBV) in India. *Virol. J.* 5: 156.
Davies J, Littlejohn M, Locarnini SA, Whiting S, et al. (2013). Molecular epidemiology of hepatitis B in the Indigenous people of northern Australia. *J. Gastroenterol. Hepatol.* 28: 1234-1241.
Hall BG (2013). Building phylogenetic trees from molecular data with MEGA. *Mol. Biol. Evol.* 30: 1229-1235.
Huy TT, Ushijima H, Quang VX, Win KM, et al. (2004). Genotype C of hepatitis B virus can be classified into at least two subgroups. *J. Gen. Virol.* 85: 283-292.
Imai M, Watanabe T, Hatta M, Das SC, et al. (2012). Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* 486: 420-428.
Juniastuti, Utsumi T, Nugrahaputra VE, Amin M, et al. (2011). Another novel subgenotype of hepatitis B virus genotype C from papuans of Highland origin. *J. Med. Virol.* 83: 225-234.
Kahila Bar-Gal G, Kim MJ, Klein A, Shin DH, et al. (2012). Tracing hepatitis B virus to the 16th century in a Korean mummy. *Hepatology* 56: 1671-1680.
Kim BK, Revill PA and Ahn SH (2011). HBV genotypes: relevance to natural history, pathogenesis and treatment of chronic hepatitis B. *Antivir. Ther.* 16: 1169-1186.
Liu CJ and Kao JH (2013). Global perspective on the natural history of chronic hepatitis B: role of hepatitis B virus genotypes A to J. *Semin. Liver Dis.* 33: 97-102.
Liu S, Zhang H, Gu C, Yin J, et al. (2009). Associations between hepatitis B virus mutations and the risk of hepatocellular carcinoma: a meta-analysis. *J. Natl. Cancer Inst.* 101: 1066-1082.
Locarnini S, Littlejohn M, Aziz MN and Yuen L (2013). Possible origins and evolution of the hepatitis B virus (HBV). *Semin. Cancer Biol.* 23: 561-575.
Lusida MI, Nugrahaputra VE, Soetjipto, Handajani R, et al. (2008). Novel subgenotypes of hepatitis B virus genotypes C and D in Papua, Indonesia. *J. Clin. Microbiol.* 46: 2160-2166.
Mulyanto, Depamede SN, Surayah K, Tjahyono AA, et al. (2010). Identification and characterization of novel hepatitis B virus subgenotype C10 in Nusa Tenggara, Indonesia. *Arch. Virol.* 155: 705-715.
Mulyanto, Depamede SN, Wahyono A, Jirintai, et al. (2011). Analysis of the full-length genomes of novel hepatitis B virus subgenotypes C11 and C12 in Papua, Indonesia. *J. Med. Virol.* 83: 54-64.
Naumann H, Schaefer S, Yoshida CF, Gaspar AM, et al. (1993). Identification of a new hepatitis B virus (HBV) genotype from Brazil that expresses HBV surface antigen subtype adw4. *J. Gen. Virol.* 74: 1627-1632.
Norder H, Couroucé AM and Magnius LO (1994). Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology* 198: 489-503.
Okamoto H, Imai M, Kametani M, Nakamura T, et al. (1987). Genomic heterogeneity of hepatitis B virus in a 54-year-old woman who contracted the infection through materno-fetal transmission. *Jpn. J. Exp. Med.* 57: 231-236.
Okamoto H, Tsuda F, Sakugawa H, Sastrosoewignjo RI, et al. (1988). Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *J. Gen. Virol.* 69: 2575-2583.
Ott JJ, Stevens GA, Groeger J and Wiersma ST (2012). Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity. *Vaccine* 30: 2212-2219.
Owiredu WK, Kramvis A and Kew MC (2001). Molecular analysis of hepatitis B virus genomes isolated from black African patients with fulminant hepatitis B. *J. Med. Virol.* 65: 485-492.
Paraskevis D, Magiorkinis G, Magiorkinis E, Ho SY, et al. (2013). Dating the origin and dispersal of hepatitis B virus infection in humans and primates. *Hepatology* 57: 908-916.
Phung TB, Alestig E, Nguyen TL, Hannoun C, et al. (2010). Genotype X/C recombinant (putative genotype I) of hepatitis B virus is rare in Hanoi, Vietnam--genotypes B4 and C1 predominate. *J. Med. Virol.* 82: 1327-1333.
Sakamoto T, Tanaka Y, Orito E, Co J, et al. (2006). Novel subtypes (subgenotypes) of hepatitis B virus genotypes B and C among chronic liver disease patients in the Philippines. *J. Gen. Virol.* 87: 1873-1882.
Shi W, Zhu C, Zheng W, Zheng W, et al. (2012). Subgenotyping of genotype C hepatitis B virus: correcting misclassifications and identifying a novel subgenotype. *PLoS One* 7: e47271.

Stuyver L, De Gendt S, Van Geyt C, Zoulim F, et al. (2000). A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. *J. Gen. Virol.* 81: 67-74.

Sugauchi F, Orito E, Ichida T, Kato H, et al. (2002). Hepatitis B virus of genotype B with or without recombination with genotype C over the precore region plus the core gene. *J. Virol.* 76: 5985-5992.

Tatematsu K, Tanaka Y, Kurbanov F, Sugauchi F, et al. (2009). A genetic variant of hepatitis B virus divergent from known human and ape genotypes isolated from a Japanese patient and provisionally assigned to new genotype J. *J. Virol.* 83: 10538-10547.

Tseng TC, Liu CJ, Yang HC, Chen CL, et al. (2015). Higher proportion of viral basal core promoter mutant increases the risk of liver cirrhosis in hepatitis B carriers. *Gut* 64: 292-302.

Yang HI, Yeh SH, Chen PJ, Iloeje UH, et al. (2008). Associations between hepatitis B virus genotype and mutants and the risk of hepatocellular carcinoma. *J. Natl. Cancer Inst.* 100: 1134-1143.

Zhang HW, Yin JH, Li YT, Li CZ, et al. (2008). Risk factors for acute hepatitis B and its progression to chronic hepatitis in Shanghai, China. *Gut* 57: 1713-1720.

Zhang Z, Xia J, Sun B, Dai Y, et al. (2015). *In vitro* and *in vivo* replication of a chemically synthesized consensus genome of hepatitis B virus genotype B. *J. Virol. Methods* 213: 57-64.

Zhang ZH, Zhang L, Lu MJ, Yang DL, et al. (2009). Establishment of reference sequences of hepatitis B virus genotype B and C in China. *Zhonghua Gan Zang Bing Za Zhi.* 17: 891-895.