Available online at www.sciencedirect.com

ScienceDirect

Progress in Natural Science 18 (2008) 13–20

Progress in Natural Science

Review

Overview of HBV whole genome data in public repositories and the Chinese HBV reference sequences

Guanghua Wu [a,b], Huiguo Ding [c], Changqing Zeng [b,*]

[a] *Graduate School of Chinese Academy of Sciences, Beijing 100049, China*
[b] *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China*
[c] *Beijing Youan Hospital, Capital Medical University, Beijing 100069, China*

Received 5 July 2007; received in revised form 16 July 2007; accepted 22 August 2007

Abstract

The number of Hepatitis B virus (HBV) whole genomic sequences in public nucleotide databases (GenBank, EMBL, and DDBJ) had reached 866 by January 1, 2007. Coming from 46 countries and regions, these sequences were categorized as eight genotypes (A–H). With the statistical and phylogenetic analysis on all available complete genomic data of HBV, we here present an overview of HBV sequences in public databases. From all registered 229 HBV genomes in Chinese regions as well as 59 sequencing data from our research group, we report the establishment of reference sequences of HBV strains prevailing in China. These analyses provide clues for the effects of HBV genotypes in host clinical progressions, geographic distribution of the infection, and the viral evolutionary history. Moreover, the viral sequence reference would be helpful in the identification of various HBV mutations. Based on the analysis of various public databases, we suggest that the Chinese HBV database with the clinical information should be constructed.

Keywords: Public databases; HBV genomic sequences; Genotype; Subtype; Chinese reference sequences

1. Introduction

Hepatitis B virus (HBV) infection is a major clinical problem. According to World Health Organization, about 30% of world population, approximately two billions, are estimated to carry detectable HBV antigens. Nearly 350 millions are chronically infected. At least one million deaths annually are caused by hepatic failure, hepatocirrhosis, and liver cancer as a result of HBV infection. In life cycle of HBV, there is a reverse transcription process which leads to the four orders higher in its mutation rate than other DNA viruses (up to 10^{-5}) due to the lack of proof reading activity by viral reverse transcriptase [1,2]. Such a high mutation rate of HBV makes the coexistence of various sequences within one infected individual, called as "quasispecies". Generally speaking, under the pressure of host immune system or clinical antiviral treatment, the dominant HBV strain in quasispecies may be depressed but the virus populations are not completely eliminated and remain a low level of replication. Thus, the strains that are resistant to medication and are escaped from the immune surveillance may become a new dominant type, resulting in drug resistance or virus rebound.

The first HBV DNA was sequenced by Galibert in 1979, which initiated HBV genomic research [3]. In 1988, researchers in Japan cloned three HBV strains of *adw* subtype from sera of chronic asymptomatic HBV carriers. The sequence divergence of 3.9–5.6% was shown in these three strains, whereas 8.3–9.3% of difference was seen between the strains from Japan and the United States. Based on the divergence of 18 HBVs they classified these viruses into A, B, C, D genotypes [4], taking a sequences difference of

[*] Corresponding author. Tel.: +86 10 80481146; fax: +86 10 80498676.
E-mail address: czeng@genomics.org.cn (C. Zeng).

1002-0071/$ - see front matter © 2007 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2007.08.002

8% as a cutoff. With the increase of HBV sequences identified worldwide, further phylogeny analysis found four more genotypes, resulting in the present eight HBV genotypes and several subtypes.

Up to January 1, 2007, there had been 866 HBV genomic sequences deposited in public nucleotide sequence databases (GenBank, EMBL, and DDBJ) [5], 387 out of which are with genotype identifications and 436 were indicated by countries and regions. These data include sequences of wild types, natural mutants, dominant drug resistant mutants, dominant immune resistant mutants, as well as all types of clone sequences in quasispecies, providing important information to the research on HBV polymorphisms, genotype features, virus epidemiology, and evolution [6]. China is a country with a large number of hepatitis patients. In public databases, sequences from China count approximately one quarter of the total. These data are valuable resource to study HBV spreading in China and to provide the knowledge of HBV genomics to the clinical research.

Detection of HBV mutants is very important in research of hepatitis B pathogenesis, prevention, and treatment. Reference sequences are fundamental for mutation identification. Due to the variation of HBV sequences obtained from different localities, reference sequences from a certain area are very necessary as the representative to that locality.

In this study, 866 sequences retrieved from public databases were analyzed with statistical and bioinformatic softwares including CLUSTALW, PHYPLIP, MEGA3.1, Perl script, etc., to obtain an overview of these HBV genomes. Moreover, 229 from Chinese regions and 59 whole genomes sequenced by our laboratory were genotyped from which Chinese reference sequences of genotypes B and C were further established.

2. HBV genomic sequences in public databases

Until January 1, 2007, 866 HBV genomic sequences were deposited in GenBank, EMBL, and DDBJ. These sequences were from 46 countries and regions in Asia, Africa, North America, Latin America, Europe, and Middle East. As shown in Fig. 1, the number grew much faster after the completion of the draft sequence of the Human Genome in 2001 [7].

HBV infection and epidemic in Asia appear more severe. China has the largest infected population (8%). Similar in public databases, sequence number from Mainland China ranks the highest, and then in order are China Hong Kong, Japan, Africa, Europe, and North America. The number of reported HBV sequences is closely related with infected size (Fig. 2 and Table 1) [8].

About 50% of the HBV genomic sequences in public databases have geographic and genotype annotations, from which information including viral epidemiology, demographic distribution of genotypes could be obtained easily. However, little information about host clinical status is provided. If sufficient clinical information from the patients were included in the databases, the integration of clinical and genomic research would lead to the discovery of more significant HBV mutations through high throughput bioinformatics analysis. Therefore, better clinical interpretations and solutions could be reached [9]. For instance, researchers in England are making effort on establishing the International Public Health Repository for Hepatitis B-HepSeQ [10]. Such a new database would include detailed clinical information as well as molecular biology information of HBV. HepSeQ is a pilot project demonstrating the direction integrating results of HBV genomic and clinical studies.

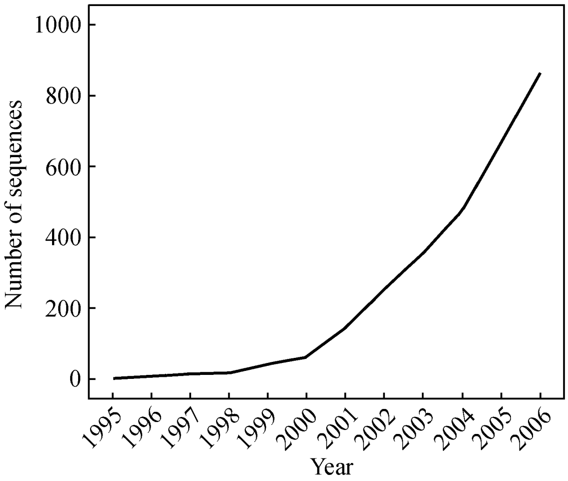


Fig. 1. The increase of HBV genomic sequences in public databases.

3. HBV genotypes associated with host clinical status and viral geographic distribution

Based on the divergence of genomic sequences, HBV are classified into eight genotypes (A–H). These genotypes were found to correlate with demographic features and clinical symptoms [11]. Therefore, it is crucial to study HBV genotype epidemiology and the relationship between genotype and clinical outcomes.

We used two methods to estimate HBV genotype distribution in public databases. For the sequences with identified genotypes, a simple Perl program was applied to count and then to calculate the percentages of each type. In another approach, BLAST [12] was used to align eight genotype reference sequences from NCBI (http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi) with all entries from public databases to group different HBV genotypes followed by the percentage calculation. Both methods yielded similar results. Genotype C counts the most in the databases (about 1/3), and then Genotype B, A, D in a descending order. These four genotypes represent about 80–90% of HBV sequences in databases and the rest are E, F, G, and H. In addition, there are also a few CD and GC types.

Various studies have indicated that HBV genotypes have different geographic and demographic distributions.

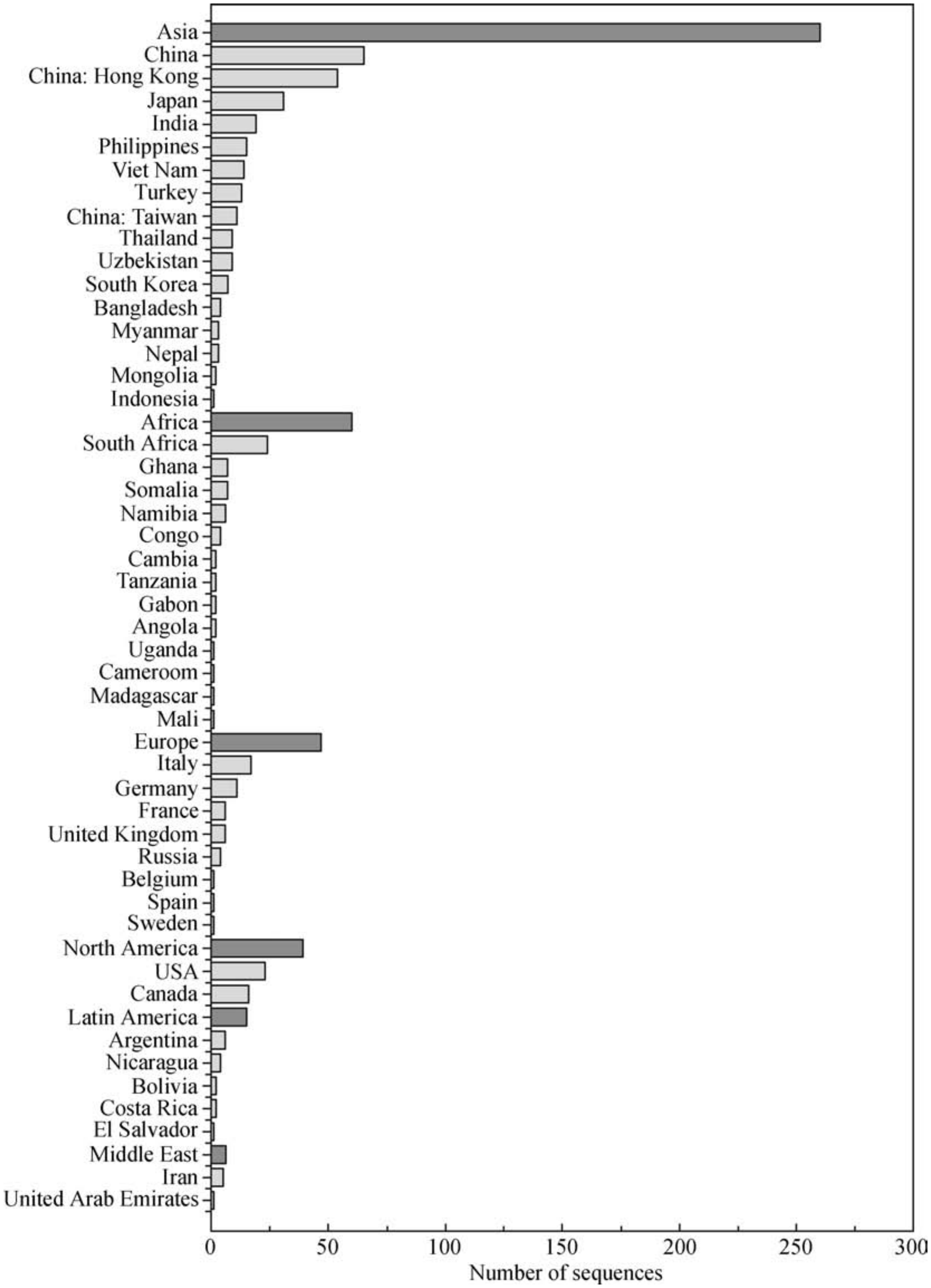


Fig. 2. Geographic distribution of HBV sequences (names of countries and regions are from GenBank annotations).

Table 1
World prevalence of HBV

Area	HBsAg (% of population positive for infection)
Northern, Western, and Central Europe, North America, Australia	0.2–0.5
Eastern Europe, Mediterranean, Russia and the Russian Federation, Southwest Asia, Central and South America	2–7
Parts of China, Southeast Asia, tropical Africa	8–20

Genotype A is widely distributed and found dominant in West Europe, North America and Central Africa, and genotypes B and C are dominant in East Asia, Southeast Asia, China, and Japan. Genotype D is actually the most widely distributed type although with less infected populations in patients of types A, B, and C. This genotype is found dominant in Mediterranean area, Middle East, and India. Genotype E is the major type in Africa and F mainly distributes in American Indians and Central America. Type G is seen in West Europe and North America and H is distributed in USA, Mexico, and Central America such as Nicaragua [13] (Fig. 3).

The most prevalent genotypes in China are B and C. Type C is dominant in North and B is mainly in South China. D is found in certain remote areas including Xinjiang, Tibet and littoral regions. Only a few cases of A and F types are found and no E, G, and H have ever been detected [14].

The clinical outcome of HBV infection varies as a result of different host immune status, infection pathways, as well as the virus genotypes. In Asia where genotypes B and C are dominant, it is proved that genotype C is closely related with more advanced liver diseases and poorer prognosis in comparison with genotype B [15]. In India where genotypes A and D are prevalent, genotype D is more commonly detected in severe liver diseases and young liver cancer patients [16]. In Europe, chronic hepatitis B is closely related with genotype A and acute hepatitis is closely related with D [17]. On the other hand, different reactions upon antiviral treatments were also seen among various genotypes. Research on the interaction between HBV genotypes and interferon medication indicates that seroconversion is more commonly found in genotype B in comparison with type C, and similarly more in genotype A while comparing with D. Although more and more evidence supports above correlations between genotypes and epidemiology and clinical status, it is still unclear whether the HBV genotype distribution pattern is the consequence of susceptibility difference among populations [17].

The circular form of HBV genome results multiple start points in sequencing data from different research groups. The EcoRI recognition site is the most commonly used start point although no unanimous one is appointed. The circular genome with different start points in the public databases hinders high throughput sequence analysis. Here we suggest an unanimous start point be selected or annotated when uploading sequences to the public databases to facilitate multiple sequence alignment and homologous analysis.

4. HBV sequences from China

Currently there are totally 229 HBV genomic sequences from China in public databases. We did multiple sequence alignment on these 229 genomes with 23 references from NCBI [18,19]. Kimura's two parameter model was used to calculate genetic distances [20]. A neighbor-joining [21] tree was constructed and 2000 bootstrap tests were then carried out to verify the phylogenic analysis. The final result was displayed by MEGA software [22]. As shown in Fig. 4, B (75 sequences) and C (134 sequences) are the two dominant genotypes in China, representing 32.8% and 58.8% of the total, respectively. Furthermore, three genotype A and 17 C/D hybrids were reported in China (Fig. 4). The genotyping results based on phylogenetic analysis was then verified with NCBI genotyping tool [23]. In the three sequences of genotype A, AY707087

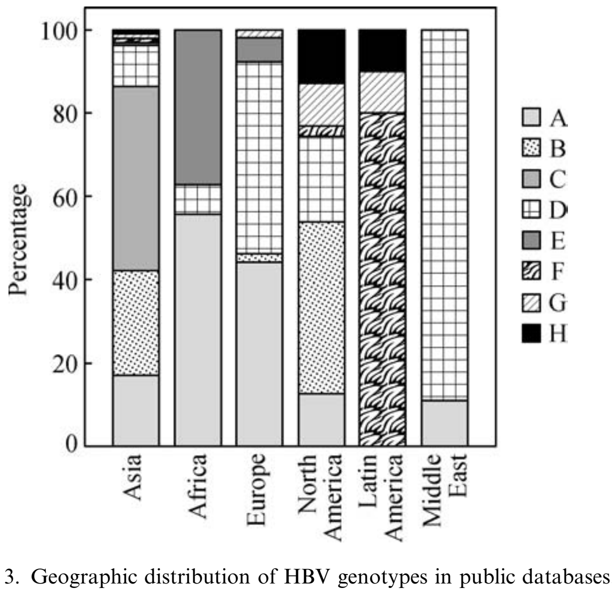


Fig. 3. Geographic distribution of HBV genotypes in public databases.

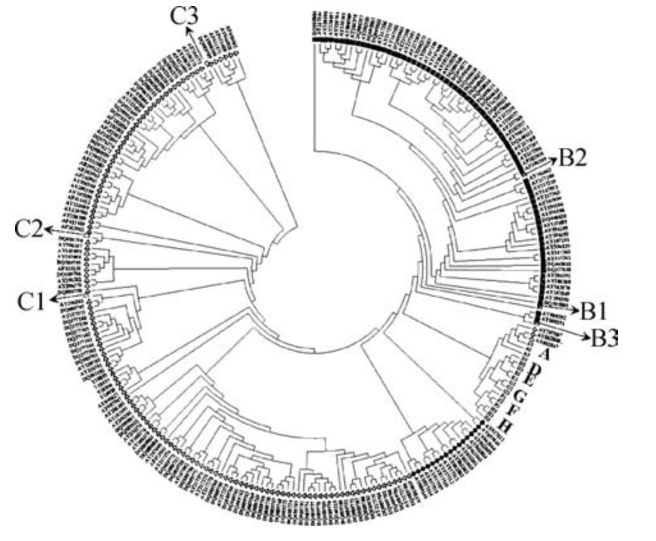
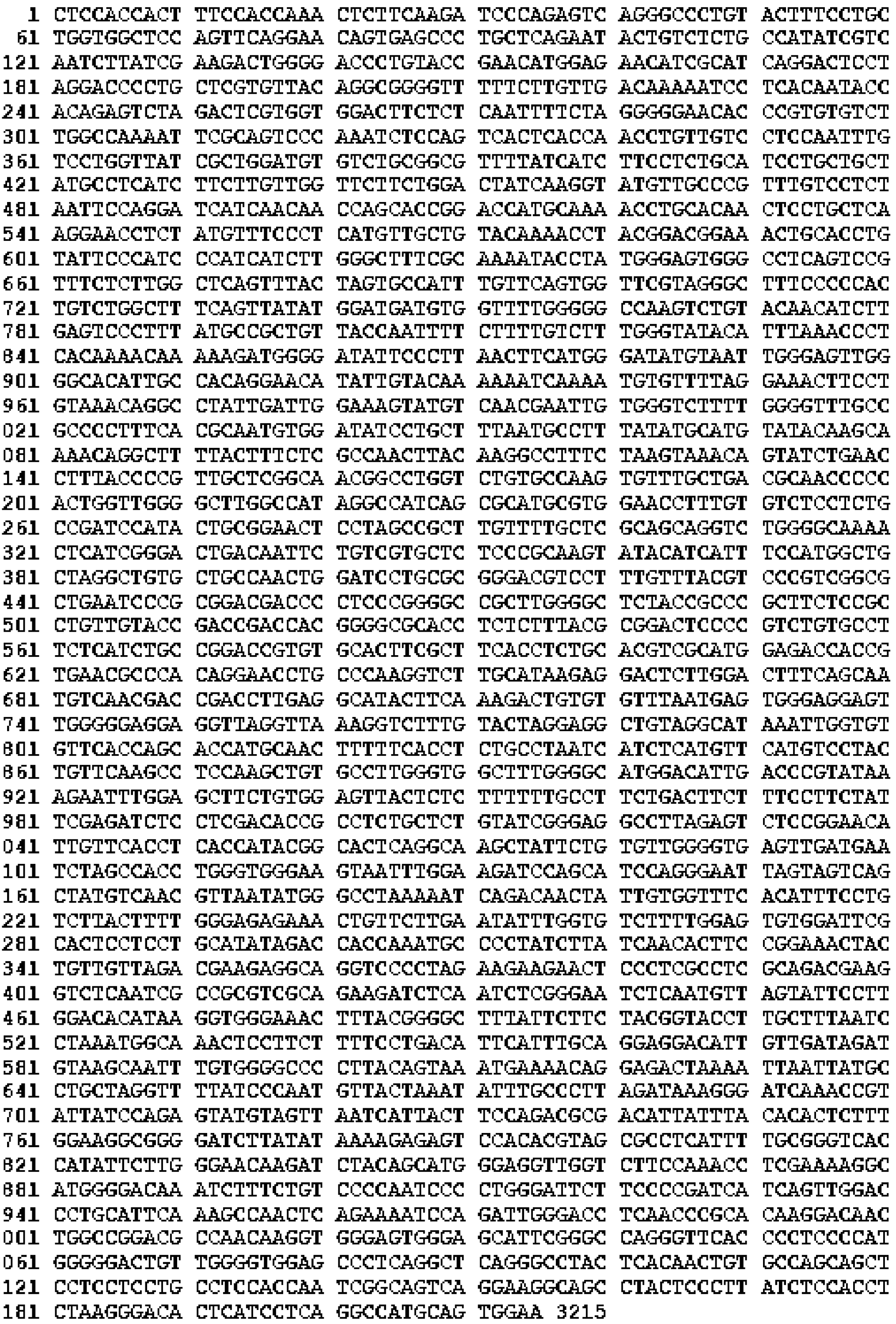


Fig. 4. Phylogeny tree of 229 HBV genomes from China and 23 reference sequences of NCBI. The alignment resulted 75 B, 134 C, 3 A, and 17 C/D genotypes, respectively. NCBI reference sequences include A1 (X02763), A2 (X51970), A3 (AF090842); B1 (D00329), B2 (AF100309), B3 (AB033554); C1 (X04615), C2 (M12906), C3 (AB014381); D (X65259, M32138, and X85254); E (X75657 and AB032431); F (X69798, AB036910, and AF223965); G (AF160501, AB064310, and AF405706); H (AY090454, AY090457, and AY090460). △, Genotype C; ■, genotype B; ◇, genotype A; ◆, C/D types.

was from Fujian, and AY862868 and AY862867 from Qinghai Province. The 17 C/D hybrids were from Tibet, Qinghai, and some other parts of northwest China [24–26].

The epidemiology surveys on HBV infection in China from 1992 to 1995 reported that an average infection rate of 57.6% and carrying rate of 9.75%, converting as a population of 690 million infected, 120 million of carriers, and 20 million of chronic hepatitis patients. From the phylogenic analysis one can draw the picture of genotype and subtype distribution, as well as the possible evolution history of hybrid types [27]. However, if detailed host information such as geo-



```
   1 CTCCACCACT TTCCACCAAA CTCTTCAAGA TCCCAGAGTC AGGGCCCTGT ACTTTCCTGC
  61 TGGTGGCTCC AGTTCAGGAA CAGTGAGCCC TGCTCAGAAT ACTGTCTCTG CCATATCGTC
 121 AATCTTATCG AAGACTGGGG ACCCTGTACC GAACATGGAG AACATCGCAT CAGGACTCCT
 181 AGGACCCCTG CTCGTGTTAC AGGCGGGGTT TTTCTTGTTG ACAAAAATCC TCACAATACC
 241 ACAGAGTCTA GACTCGTGGT GGACTTCTCT CAATTTTCTA GGGGGAACAC CCGTGTGTCT
 301 TGGCCAAAAT TCGCAGTCCC AAATCTCCAG TCACTCACCA ACCTGTTGTC CTCCAATTTG
 361 TCCTGGTTAT CGCTGGATGT GTCTGCGGCG TTTTATCATC TTCCTCTGCA TCCTGCTGCT
 421 ATGCCTCATC TTCTTGTTGG TTCTTCTGGA CTATCAAGGT ATGTTGCCCG TTTGTCCTCT
 481 AATTCCAGGA TCATCAACAA CCAGCACCGG ACCATGCAAA ACCTGCACAA CTCCTGCTCA
 541 AGGAACCTCT ATGTTTCCCT CATGTTGCTG TACAAAACCT ACGGACGGAA ACTGCACCTG
 601 TATTCCCATC CCATCATCTT GGGCTTTCGC AAAATACCTA TGGGAGTGGG CCTCAGTCCG
 661 TTTCTCTTGG CTCAGTTTAC TAGTGCCATT TGTTCAGTGG TTCGTAGGGC TTTCCCCCAC
 721 TGTCTGGCTT TCAGTTATAT GGATGATGTG GTTTTGGGGG CCAAGTCTGT ACAACATCTT
 781 GAGTCCCTTT ATGCCGCTGT TACCAATTTT CTTTTGTCTT TGGGTATACA TTTAAACCCT
 841 CACAAAACAA AAAGATGGGG ATATTCCCTT AACTTCATGG GATATGTAAT TGGGAGTTGG
 901 GGCACATTGC CACAGGAACA TATTGTACAA AAAATCAAAA TGTGTTTTAG GAAACTTCCT
 961 GTAAACAGGC CTATTGATTG GAAAGTATGT CAACGAATTG TGGGTCTTTT GGGGTTTGCC
1021 GCCCCTTTCA CGCAATGTGG ATATCCTGCT TTAATGCCTT TATATGCATG TATACAAGCA
1081 AAACAGGCTT TTACTTTCTC GCCAACTTAC AAGGCCTTTC TAAGTAAACA GTATCTGAAC
1141 CTTTACCCCG TTGCTCGGCA ACGGCCTGGT CTGTGCCAAG TGTTTGCTGA CGCAACCCCC
1201 ACTGGTTGGG GCTTGGCCAT AGGCCATCAG CGCATGCGTG GAACCTTTGT GTCTCCTCTG
1261 CCGATCCATA CTGCGGAACT CCTAGCCGCT TGTTTTGCTC GCAGCAGGTC TGGGGCAAAA
1321 CTCATCGGGA CTGACAATTC TGTCGTGCTC TCCCGCAAGT ATACATCATT TCCATGGCTG
1381 CTAGGCTGTG CTGCCAACTG GATCCTGCGC GGGACGTCCT TTGTTTACGT CCCGTCGGCG
1441 CTGAATCCCG CGGACGACCC CTCCCGGGGC CGCTTGGGGC TCTACCGCCC GCTTCTCCGC
1501 CTGTTGTACC GACCGACCAC GGGGCGCACC TCTCTTTACG CGGACTCCCC GTCTGTGCCT
1561 TCTCATCTGC CGGACCGTGT GCACTTCGCT TCACCTCTGC ACGTCGCATG GAGACCACCG
1621 TGAACGCCCA CAGGAACCTG CCCAAGGTCT TGCATAAGAG GACTCTTGGA CTTTCAGCAA
1681 TGTCAACGAC CGACCTTGAG GCATACTTCA AAGACTGTGT GTTTAATGAG TGGGAGGAGT
1741 TGGGGGAGGA GGTTAGGTTA AAGGTCTTTG TACTAGGAGG CTGTAGGCAT AAATTGGTGT
1801 GTTCACCAGC ACCATGCAAC TTTTTCACCT CTGCCTAATC ATCTCATGTT CATGTCCTAC
1861 TGTTCAAGCC TCCAAGCTGT GCCTTGGGTG GCTTTGGGGC ATGGACATTG ACCCGTATAA
1921 AGAATTTGGA GCTTCTGTGG AGTTACTCTC TTTTTTGCCT TCTGACTTCT TTCCTTCTAT
1981 TCGAGATCTC CTCGACACCG CCTCTGCTCT GTATCGGGAG GCCTTAGAGT CTCCGGAACA
2041 TTGTTCACCT CACCATACGG CACTCAGGCA AGCTATTCTG TGTTGGGGTG AGTTGATGAA
2101 TCTAGCCACC TGGGTGGGAA GTAATTTGGA AGATCCAGCA TCCAGGGAAT TAGTAGTCAG
2161 CTATGTCAAC GTTAATATGG GCCTAAAAAT CAGACAACTA TTGTGGTTTC ACATTTCCTG
2221 TCTTACTTTT GGGAGAGAAA CTGTTCTTGA ATATTTGGTG TCTTTTGGAG TGTGGATTCG
2281 CACTCCTCCT GCATATAGAC CACCAAATGC CCCTATCTTA TCAACACTTC CGGAAACTAC
2341 TGTTGTTAGA CGAAGAGGCA GGTCCCCTAG AAGAAGAACT CCCTCGCCTC GCAGACGAAG
2401 GTCTCAATCG CCGCGTCGCA GAAGATCTCA ATCTCGGGAA TCTCAATGTT AGTATTCCTT
2461 GGACACATAA GGTGGGAAAC TTTACGGGGC TTTATTCTTC TACGGTACCT TGCTTTAATC
2521 CTAAATGGCA AACTCCTTCT TTTCCTGACA TTCATTTGCA GGAGGACATT GTTGATAGAT
2581 GTAAGCAATT TGTGGGGCCC CTTACAGTAA ATGAAAACAG GAGACTAAAA TTAATTATGC
2641 CTGCTAGGTT TTATCCCAAT GTTACTAAAT ATTTGCCCTT AGATAAAGGG ATCAAACCGT
2701 ATTATCCAGA GTATGTAGTT AATCATTACT TCCAGACGCG ACATTATTTA CACACTCTTT
2761 GGAAGGCGGG GATCTTATAT AAAAGAGAGT CCACACGTAG CGCCTCATTT TGCGGGTCAC
2821 CATATTCTTG GGAACAAGAT CTACAGCATG GGAGGTTGGT CTTCCAAACC TCGAAAAGGC
2881 ATGGGGACAA ATCTTTCTGT CCCCAATCCC CTGGGATTCT TCCCCGATCA TCAGTTGGAC
2941 CCTGCATTCA AAGCCAACTC AGAAAATCCA GATTGGGACC TCAACCCGCA CAAGGACAAC
3001 TGGCCGGACG CCAACAAGGT GGGAGTGGGA GCATTCGGGC CAGGGTTCAC CCCTCCCCAT
3061 GGGGGACTGT TGGGGTGGAG CCCTCAGGCT CAGGGCCTAC TCACAACTGT GCCAGCAGCT
3121 CCTCCTCCTG CCTCCACCAA TCGGCAGTCA GGAAGGCAGC CTACTCCCTT ATCTCCACCT
3181 CTAAGGGACA CTCATCCTCA GGCCATGCAG TGGAA 3215
```

Fig. 5. Chinese HBV reference sequence CHNHBV07-B.

graphic location, ethnic background, and particularly the clinical reports could also be included, phylogenic analysis with these data will greatly help us understand the infection pathways, evolution history of HBV in China, as well as assist us in disease prevention, control, and treatment.

5. Construction of Chinese HBV reference sequences

Detection of HBV mutants is very important in research of pathogenesis, prevention, and treatment. Reference sequences are fundamental for the mutation identification. Because of the variation among HBV sequences obtained

```
   1 CTCCACAACA TTCCACCAAG CTCTGCTAGA CCCCAGAGTG AGGGGCCTAT ACTTTCCTGC
  61 TGGTGGCTCC AGTTCCGGAA CAGTAAACCC TGTTCCGACT ACTGCCTCAC CCATATCGTC
 121 AATCTTCTCG AGGACTGGGG ACCCTGCACC GAACATGGAG AACACAACAT CAGGATTCCT
 181 AGGACCCCTG CTCGTGTTAC AGGCGGGGTT TTTCTTGTTG ACAAGAATCC TCACAATACC
 241 ACAGAGTCTA GACTCGTGGT GGACTTCTCT CAATTTTCTA GGGGGAGCAC CCACGTGTCC
 301 TGGCCAAAAT TCGCAGTCCC CAACCTCCAA TCACTCACCA ACCTCTTGTC CTCCAATTTG
 361 TCCTGGCTAT CGCTGGATGT GTCTGCGGCG TTTTATCATA TTCCTCTTCA TCCTGCTGCT
 421 ATGCCTCATC TTCTTGTTGG TTCTTCTGGA CTACCAAGGT ATGTTGCCCG TTTGTCCTCT
 481 ACTTCCAGGA ACATCAACTA CCAGCACGGG ACCATGCAAG ACCTGCACGA TTCCTGCTCA
 541 AGGAACCTCT ATGTTTCCCT CTTGTTGCTG TACAAAACCT TCGGACGGAA ACTGCACTTG
 601 TATTCCCATC CCATCATCCT GGGCTTTCGC AAGATTCCTA TGGGAGTGGG CCTCAGTCCG
 661 TTTCTCCTGG CTCAGTTTAC TAGTGCCATT TGTTCAGTGG TTCGTAGGGC TTTCCCCCAC
 721 TGTTTGGCTT TCAGTTATAT GGATGATGTG GTATTGGGGG CCAAGTCTGT ACAACATCTT
 781 GAGTCCCTTT TTACCTCTAT TACCAATTTT CTTTTGTCTT TGGGTATACA TTTGAACCCT
 841 AATAAAACCA AACGTTGGGG CTACTCCCTT AACTTCATGG GATATGTAAT TGGAAGTTGG
 901 GGTACTTTAC CGCAGGAACA TATTGTACTA AAACTCAAGC AATGTTTTCG AAAACTGCCT
 961 GTAAATAGAC CTATTGATTG GAAAGTATGT CAAAGAATTG TGGGTCTTTT GGGCTTTGCT
1021 GCCCCTTTTA CACAATGTGG CTATCCTGCC TTAATGCCTT TATATGCATG TATACAATCT
1081 AAGCAGGCTT TCACTTTCTC GCCAACTTAC AAGGCCTTTC TGTGTAAACA ATATCTGAAC
1141 CTTTACCCCG TTGCCCGGCA ACGGTCAGGT CTCTGCCAAG TGTTTGCTGA CGCAACCCCC
1201 ACTGGATGGG GCTTGGCCAT AGGCCATCGG CGCATGCGTG GAACCTTTGT GGCTCCTCTG
1261 CCGATCCATA CTGCGGAACT CCTAGCAGCT TGTTTTGCTC GCAGCCGGTC TGGAGCGAAA
1321 CTTATCGGAA CCGACAACTC TGTTGTCCTC TCTCGGAAAT ACACCTCCTT TCCATGGCTG
1381 CTAGGGTGTG CTGCCAACTG GATCCTGCGC GGGACGTCCT TTGTCTACGT CCCGTCGGCG
1441 CTGAATCCCG CGGACGACCC GTCTCGGGGC CGTTTGGGAC TCTACCGTCC CCTTCTTCAT
1501 CTGCCGTTCC GGCCGACCAC GGGGCGCACC TCTCTTTACG CGGTCTCCCC GTCTGTGCCT
1561 TCTCATCTGC CGGACCGTGT GCACTTCGCT TCACCTCTGC ACGTCGCATG GAGACCACCG
1621 TGAACGCCCA CCAGGTCTTG CCCAAGGTCT TACATAAGAG GACTCTTGGA CTCTCAGCAA
1681 TGTCAACGAC CGACCTTGAG GCATACTTCA AAGACTGTTT GTTTAAAGAC TGGGAGGAGT
1741 TGGGGGAGGA GATTAGGTTA ATGATCTTTG TACTAGGAGG CTGTAGGCAT AAATTGGTCT
1801 GTTCACCAGC ACCATGCAAC TTTTTCACCT CTGCCTAATC ATCTCATGTT CATGTCCTAC
1861 TGTTCAAGCC TCCAAGCTGT GCCTTGGGTG GCTTTGGGGC ATGGACATTG ACCCGTATAA
1921 AGAATTTGGA GCTTCTGTGG AGTTACTCTC TTTTTTGCCT TCTGACTTCT TTCCTTCTAT
1981 TCGAGATCTC CTCGACACCG CCTCTGCTCT GTATCGGGAG GCCTTAGAGT CTCCGGAACA
2041 TTGTTCACCT CACCATACAG CACTCAGGCA AGCTATTCTG TGTTGGGGTG AGTTGATGAA
2101 TCTGGCCACC TGGGTGGGAA GTAATTTGGA AGACCCAGCA TCCAGGGAAT TAGTAGTCAG
2161 CTATGTCAAT GTTAATATGG GCCTAAAAAT CAGACAACTA TTGTGGTTTC ACATTTCCTG
2221 TCTTACTTTT GGAAGAGAAA CTGTTCTTGA GTATTTGGTG TCTTTTGGAG TGTGGATTCG
2281 CACTCCTCCC GCTTACAGAC CACCAAATGC CCCTATCTTA TCAACACTTC CGGAAACTAC
2341 TGTTGTTAGA CGACGAGGCA GGTCCCCTAG AAGAAGAACT CCCTCGCCTC GCAGACGAAG
2401 GTCTCAATCG CCGCGTCGCA GAAGATCTCA ATCTCGGGAA TCTCAATGTT AGTATCCCTT
2461 GGACTCATAA GGTGGGAAAC TTTACTGGGC TTTATTCTTC TACTGTACCT GTCTTTAATC
2521 CTGAGTGGCA AACTCCCTCC TTTCCTCACA TTCATTTACA GGAGGACATT ATTAATAGAT
2581 GTCAACAATA TGTGGGCCCT CTTACAGTTA ATGAAAAAAG GAGATTAAAA TTAATTATGC
2641 CTGCTAGGTT CTATCCTAAC CTTACCAAAT ATTTGCCCTT GGACAAAGGC ATTAAACCGT
2701 ATTATCCTGA ACATGCAGTT AATCATTACT TCAAAACTAG GCATTATTTA CATACTCTGT
2761 GGAAGGCTGG CATTCTATAT AAGAGAGAAA CTACACGCAG CGCCTCATTT TGTGGGTCAC
2821 CATATTCTTG GGAACAAGAG CTACAGCATG GGAGGTTGGT CTTCCAAACC TCGACAAGGC
2881 ATGGGGACGA ATCTTTCTGT TCCCAATCCT CTGGGATTCT TTCCCGATCA CCAGTTGGAC
2941 CCTGCGTTCG GAGCCAACTC AAACAATCCA GATTGGGACT TCAACCCCAA CAAGGATCAC
3001 TGGCCAGAGG CAAATCAGGT AGGAGCGGGA GCATTCGGGC CAGGGTTCAC CCCACCACAC
3061 GGCGGTCTTT TGGGGTGGAG CCCTCAGGCT CAGGGCATAT TGACAACAGT GCCAGCAGCA
3121 CCTCCTCCTG CCTCCACCAA TCGGCAGTCA GGAAGACAGC CTACTCCCAT CTCTCCACCT
3181 CTAAGAGACA GTCATCCTCA GGCCATGCAG TGGAA 3215
```

Fig. 6. Chinese HBV reference sequence CHNHBV07-C.

Table 2
Comparison between Chinese HBV reference sequences and NCBI HBV reference sequences

Reference sequences	Source	Number of nt differences compared with CHNHBV07-B	Percentage of difference	Subtype
D00329	Japan	117	3.6	Bj
AF100309	China	27	0.8	B2
AB033554	Indonesia	127	3.9	B3
		CHNHBV07-C		
X04615	Japan	33	1.0	C1
M12906	Japan	51	1.6	C1
AB014381	Japan	41	1.3	C1

from different locations, only the reference sequences from a certain region are representative and reliable in related studies [28].

There are also subtypes found within genotypes. Four subtypes were found in genotype B and B1 is dominant in Japan, B2 dominant in China and Vietnam, B3 dominant in Indonesia, and B4 prevalent in Vietnam. B2 is also known as Bj and B2–B4 are also called Ba. In genotype C, C1 is dominant in Japan, South Korea, and China; C2 in China, Southeast Asia, and Bengal; C3 in Oceania; and C4 in native residents of Australia [29]. In NCBI reference sequences of genotype B and C (X04615, M12906, AB014381 are genotype C and D00329, AF100309, AB033554 are genotype B), AF100309 is from China; D00329, X04615, M12906, and AB014381 are from Japan; AB033554 is from Indonesia. D00329 is type Bj and AB033554 is type B3.

Based on 51 HBV sequences of genotype C and 8 sequences of genotype B identified in our laboratory, as well as 113 sequences of genotype C and 70 genotype B in public databases, we established Chinese HBV reference sequences of genotype B and C, named as CHNHBV07-B and CHNHBV07-C (Figs. 5 and 6).

Comparing CHNHBV07-B and CHNHBV07-C with other NCBI references from other Asia countries, we found higher diversity in genotype B with 3.6–3.9% of sequence difference. As our constructed reference sequences came from the alignment of all currently available public data as well as from our recent analysis of whole genome surveys, we believe these two reference sequences are uniquely representing the local types in China (Table 2).

6. Prospect

As shown in this review, HBV genomic sequences in public databases provide rich resources for both genomic and clinical research. Considering the high mutation rate of HBV, more data with high sequencing quality and particularly more detailed annotations (such as genotype/subtype, serotype, and host information) will be very valuable. With this additional information, the correlation of viral mutation patterns with clinical progress, the evolutionary history and the molecular epidemiology of HBV could be further elucidated. Stratification of genomic sequences based on comprehensive clinical information is crucial in HBV research. Feeding back the genomic implications to clinical research to verify the result from genomic studies is also vital for HBV clinical research. A good interaction between genomic and clinical studies is certainly a promising approach of tackling problems of both basic research and medical treatment.

In the public databases of GenBank, EMBL, and DDBJ, the number of HBV sequences from China ranks the most (about one quarter of the total). This amount of data is a valuable resource for Chinese HBV genomic study. As the country with the largest infected population, it is more important for China to integrate clinical research with genomic study. A good form of interaction of these two scopes of HBV research would be a database including sequencing data and clinical reports as we suggested here. To explore the dynamics and evolution of the host–virus interaction, molecular biological information of HBV such as DNA and protein sequence, and mutation map should be described with clinical information including pathogenesis, treatment, and drug resistant history. Such combinatorial research will greatly promote all our actions to eliminate the virus from our people.

Acknowledgements

This work was supported by CAS National Knowledge Innovation Program (KIP) (KSCX2-SW-207), Beijing Municipal Education Commission Funds Program (KM20070025024), Beijing Integrated Traditional and Western Medicine Key Disciplines.

References

[1] Okamoto H, Imai M, Kametani M, et al. Genomic heterogeneity of hepatitis B virus in a 54-year-old woman who contracted the infection through materno-fetal transmission. Jpn J Exp Med 1987;57(4):231–6.

[2] Orito E, Mizokami M, Ina Y, et al. Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc Natl Acad Sci USA 1989;86(18):7059–62.

[3] Galibert F, Mandart E, Fitoussi F, et al. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. Nature 1979;281(5733):646–50.

[4] Okamoto H, Tsuda F, Sakugawa H, et al. Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. J Gen Virol 1988;69(10):2575–83.

[5] Jenuth JP. The NCBI. Publicly available tools and resources on the Web. Methods Mol Biol 2000;132:301–12.

[6] Simmonds P, Midgley S. Recombination in the genesis and evolution of hepatitis B virus genotypes. J Virol 2005;79(24):15467–76.

[7] Chan EY. Advances in sequencing technology. Mutat Res 2005;573(1–2):13–40.

[8] Department of Communicable Diseases Surveillance and Response, WHO. Hepatitis B; 2002.

[9] Zhang Y, Bo XC, Yang J, et al. HBVPathDB: a database of HBV infection-related molecular interaction network. World J Gastroenterol 2005;11(11):1690–2.

[10] Gnaneshan S, Ijaz S, Moran J, et al. HepSEQ: International public health repository for hepatitis B. Nucleic Acids Res 2007;35(Database issue):D367–70.

[11] Robertson BH, Margolis HS. Primate hepatitis B viruses – genetic diversity, geography and evolution. Rev Med Virol 2002;12(3):133–41.

[12] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.

[13] Starkman SE, MacDonald DM, Lewis JC, et al. Geographic and species association of hepatitis B virus genotypes in non-human primates. Virology 2003;314(1):381–93.

[14] Liu X, Tang H, He F. New advance in genotype of hepatitis B virus. World J Gastroenterol 2006;14(22):6, [in Chinese].

[15] Chan HL, Tse CH, Ng EY, et al. Phylogenetic, virological, and clinical characteristics of genotype C hepatitis B virus with TCC at codon 15 of the precore region. J Clin Microbiol 2006;44(3):681–7.

[16] Chan HL, Tsui SK, Tse CH, et al. Epidemiological and virological characteristics of 2 subgroups of hepatitis B virus genotype C. J Infect Dis 2005;191(12):2022–32.

[17] Guettouche T, Hnatyszyn HJ. Chronic hepatitis B and viral genotype: the clinical significance of determining HBV genotypes. Antivir Ther 2005;10(5):593–604.

[18] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85(8):2444–8.

[19] Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990;183:63–98.

[20] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16(2):111–20.

[21] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4):406–25.

[22] Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 2004;5(2):150–63.

[23] Rozanov M, Plikat U, Chappey C, et al. A web-based genotyping resource for viral sequences. Nucleic Acids Res 2004;32(Web Server issue):W654–9.

[24] Cui C, Shi J, Hui L, et al. The dominant hepatitis B virus genotype identified in Tibet is a C/D hybrid. J Gen Virol 2002;83(11):2773–7.

[25] Schaefer S. Hepatitis B virus taxonomy and hepatitis B virus genotypes. World J Gastroenterol 2007;13(1):14–21.

[26] Wang Z, Liu Z, Zeng G, et al. A new intertype recombinant between genotypes C and D of hepatitis B virus identified in China. J Gen Virol 2005;86(4):985–90.

[27] Guo YB, Hou JL, Dai W. Establishment of the consensus sequence of hepatitis B virus prevailing in the mainland of China. Chin J Microbiol Immunol 1999;19(3):197–200, [in Chinese].

[28] Xu HM, Ren H, Qing YL, et al. Establishment of consensus sequence of PreS/S of hepatitis B virus with genotype B/serotype adw2 or genotype C/serotype adrq+ prevailing in Chongqing of China. Chin J Epidemiol 2003;24(10):913–6.

[29] Norder H, Courouce AM, Coursaget P, et al. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. Intervirology 2004;47(6):289–309.