Differences in Cpg Island Distribution Between Subgenotypes of the Hepatitis B Virus Genotype

Lin Chen, Yi Shi, Wanrong Yang, Yafei Zhang, Qinxiu Xie, Yunsong Li, Xu Li, Jun Li, Zhenhua Zhang

Background: Hepatitis B virus (HBV) genotypes show genomic variations, resulting in different CpG islands in each HBV genotypes or subgenotype. This study aimed to establish reference sequences for each HBV subgenotype of A–H genotypes and to analyze the characteristics of the CpG islands.

Material/Methods: There were 3,037 retrieved whole-genome sequences of HBV genotypes A–H from GenBank, 28 subgenotype reference sequences were established for these genotypes. CpG islands of the subgenotype reference sequences were analyzed, and 939 strains were selected from the 3,037 genomic sequences. Differences in CpG islands between subgenotypes were compared using the chi-squared and non-parametric tests.

Results: Of the 28 subgenotype reference sequences established, 11 subgenotype reference sequences lacked CpG island I, and only F4 contained a new CpG island. Of all selected strains, 48.35% (454/939) contained three traditional CpG islands I, II, and III (no new islands); 45.05% (423/939) lacked CpG island I; 38.98% (366/939) contained only CpG islands II and III; and 12.46% (117/939) contained new islands (genotypes A1, D7) (genotype G had no new islands). Strains with or without CpG island I, or new islands between subgenotypes of each HBV genotype were significantly different (P<0.05). Strains containing CpG islands I, II, and III new islands among different subtypes in HBV genotypes A, C, and F were significantly different (P<0.05).

Conclusions: Different HBV genotypes and subgenotypes had characteristic CpG island patterns. Strains with or without CpG island I, or new islands among subgenotypes of each HBV genotype, were significantly different.

MeSH Keywords: CpG Islands • Genotype • Hepatitis B virus • Methylation

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/910049
**Background**

Worldwide, hepatitis B virus (HBV) infection has high rates of morbidity and mortality and represents a serious public health issue with changes in epidemiological features resulting from factors including migration, genetic variation, and the effects of treatment [1–3]. Because of the absence of ‘proofreading’ activity in HBV DNA polymerase, random nucleotide misincorporation into the replicating DNA strand leads to a high mutation rate, which is estimated at between 1.4–3.2×10⁻⁶ substitutions per site per year in the whole HBV genome [4]. Currently, HBV isolates worldwide have been divided into ten well-accepted genotypes (A to J), based on an inter-genotypic difference of greater than 8% in sequences [5–7]. The establishment of representative reference sequences might facilitate studies on HBV infection and its pathogenicity. Certain HBV references sequences had been reported previously. However, these HBV references sequences are not truly representative, since they were either simply based on the first identified isolates or isolates of small sample sizes [8–15]. Considering representativeness and consistency, we previously established reference sequences of HBV genotypes A (A1, A2, A3, A5), B2, and C (C1, C2, C5, C6) on the basis of large numbers HBV sequences, and deposited them in GenBank, with GenBank accession codes being KP234050–KP234053 (A1, A2, A3, A5) and KM999990–KM999993 (C1, C2, C5, C6) [15–17]. The synthesized consensus genome of subgenotype (subtype) B2 is replication-competent upon transfection into hepatoma cells in vitro and expression and replication in mice [18]. Representative reference sequences of other genotypes still need to be established.

Viral gene expression is believed to be partially regulated by DNA methylation, which usually occurs in the promoter region leading to transcriptional silencing and gene repression, via multiple mechanisms in human tissues [19]. DNA methylation frequently occurs in the CpG dinucleotide-rich region known as the CpG island. Currently, six CpG islands, including three newly identified CpG islands (IV, V, and VI), are generally accepted to exist. These CpG islands are located at the transcription start site or are located upstream and downstream of the promoters [20–22]. Host DNA methyltransferase mediates CpG methylation, and methylated HBV CpG islands limit their regulation of viral protein expression [23,24]. Agents promoting methylation of nuclear HBV DNA may constitute another new antiviral treatment modality than nucleotide analog drugs and interferons [25].

This study aimed to establish standard sequences between HBV subtypes and to further clarify CpG-enriched sites in the HBV genome in subtypes of HBV genotypes A–H. The location, length, and distribution of CpG islands in the genome of individual subtypes and their representative strains were investigated. Comparative analysis of CpG islands was performed among the reference sequences and representative strains of HBV genotypes A–H. The data from this study might provide insights into the clinicopathological and virological characteristics of distinct HBV genotypes and subtypes.

**Material and Methods**

Sequence sources and criteria

The GenBank Nucleotide Database was searched up to December 1, 2013, at the National Center for Biotechnology Information, using the keywords, ‘hepatitis B virus,’ ‘genotype,’ and ‘complete genome.’ Complete genome sizes ranging from 3100–3300 bp were included.

**Establishment of hepatitis B virus (HBV) subtype references**

Reference sequences of eight subtypes of HBV genotypes C and A were created in 2015 and 2016, respectively [16,17]. Using a similar method, phylogenetic and molecular phylogenetic analysis was conducted using the Maximum Likelihood statistical method and Tamura-Nei model, with 1000 bootstrap replicates, using the MEGA version 7 software (Kumar, Stecher, and Tamura 2015). All sequences were compared with known HBV references including A1 (KP234050), Aafr (AF297621), Bj (AB073858), C2 (KM999991), C (AB033556), D (X02496), E (X75657), F (X69798), G (AF160501), and H (AY090454), to determine the subtypes they belonged to. Finally, 1,680 strains belonging to 20 subtypes of HBV genotypes B, D–H were identified.

All strains in each subtype were aligned using the AlignX software, a component of Vector NTI Advance11.5 (Thermo Fisher Scientific, Waltham, MA, USA), using the ClustalW algorithm. The nucleotide in each position of the reference sequence of each subtype was determined on the basis of the nucleotide with the highest frequency in the corresponding position. Phylogenetic trees of the whole genome and 5 gene of 28 HBV subtype reference sequences were constructed using the Unweighted Pair Group Method with Arithmetic Mean method with 1000 bootstrap replicates, using MEGA 7 software.

**Computation of CpG islands**

A pool of strains was selected from the 3,037 strains based on which reference sequences were established, and all CpG islands were identified in them. To construct this pool, 50 strains were randomly selected from a certain subtype if that subtype had more than 50 strains, and all were selected if a...
subtypes were selected. The CpG islands were computed using two online methods, the MethPrimer (www.urogene.org/cgi-bin/methprimer/methprimer.cgi) and the CpG Plot (www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot). The criteria used to distinguish a CpG island included a GC content of \( \geq 50\% \), the observed/expected CpG dinucleotide of \( \geq 0.6 \), and a window size of \( \geq 100 \) bp [23,26]. Information regarding the size, number, location, and other features of CpG islands in the reference sequences of HBV subtypes and selected strains were further collected. CpG islands were classified primarily by their positions in the reference sequences. Identities of CpG islands proximal to one another, separated by a boundary, were determined on the basis of positions where major regions of those islands were located.

### Statistical analysis

The chi-squared (\( \chi^2 \)) test was performed using SPSS version 16.0 (IBM, New York, USA) for the composition of the CpG islands among the various subtypes. Non-parametric tests were performed using SPSS version 16.0 (IBM, New York, USA) for

| Genotype | Subgenotype | Strains number of subgenotype | Size range of strains (bp) | GeneBank number of references | Size of references (bp) | Nuclotide size of references (bp) | Amino acid size of references (aa) |
|----------|-------------|-----------------------------|---------------------------|-------------------------------|------------------------|-----------------------------------|-----------------------------------|
| A        | A1          | 155                         | 3117-3253                 | KP234050                      | 3221                   | 681 465 2538 558 226 154 845 185 |
|          | A2          | 231                         | 3115-3228                 | KP234051                      | 3221                   | 681 465 2538 558 226 154 845 185 |
|          | A3          | 22                          | 3119-3221                 | KP234052                      | 3221                   | 681 465 2538 558 226 154 845 185 |
|          | A5          | 25                          | 3215-3226                 | KP234053                      | 3221                   | 681 465 2538 558 226 154 845 185 |
| B        | B1          | 24                          | 3125-3227                 | KP341007                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | B2          | 277                         | 3101-3248                 | KP341008                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | B3          | 23                          | 3128-3218                 | KP341009                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | B4          | 17                          | 3200-3215                 | KP341010                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | B6          | 39                          | 3215-3218                 | KP341011                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | B7          | 13                          | 3215                      | KP341012                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | B9          | 19                          | 3179-3215                 | KP341013                      | 3215                   | 681 465 2538 558 226 154 845 185 |
| C        | C1          | 199                         | 3110-3239                 | KM999990                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | C2          | 699                         | 3101-3254                 | KM999991                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | C5          | 9                           | 3215                      | KM999992                      | 3215                   | 681 465 2538 558 226 154 845 185 |
|          | C6          | 17                          | 3119-3220                 | KM999993                      | 3215                   | 681 465 2538 558 226 154 845 185 |
| D        | D1I         | 131                         | 3110-3215                 | KP322599                      | 3182                   | 681 465 2538 558 226 154 845 185 |
|          | D1M         | 301                         | 3101-3191                 | KP322600                      | 3182                   | 681 465 2538 558 226 154 845 185 |
|          | D2          | 156                         | 3149-3218                 | KP322601                      | 3182                   | 681 465 2538 558 226 154 845 185 |
|          | D3          | 133                         | 3110-3182                 | KP322602                      | 3182                   | 681 465 2538 558 226 154 845 185 |
|          | D5          | 31                          | 3119-3188                 | KP322603                      | 3182                   | 681 465 2538 558 226 154 845 185 |
|          | D7          | 33                          | 3170-3194                 | KP322604                      | 3182                   | 681 465 2538 558 226 154 845 185 |
| E        | E           | 198                         | 3185-3212                 | KX186584                      | 3212                   | 681 465 2538 558 226 154 845 185 |
|          | F1          | 68                          | 3161-3217                 | KX264496                      | 3212                   | 681 465 2538 558 226 154 845 185 |
|          | F2          | 18                          | 3182-3215                 | KX264497                      | 3212                   | 681 465 2538 558 226 154 845 185 |
|          | F3          | 21                          | 3131-3219                 | KX264498                      | 3212                   | 681 465 2538 558 226 154 845 185 |
|          | F4          | 28                          | 3214-3227                 | KX264499                      | 3212                   | 681 465 2538 558 226 154 845 185 |
| G        | G           | 28                          | 3234-3251                 | KX264500                      | 3248                   | 681 465 2538 558 226 154 845 185 |
| H        | H           | 22                          | 3187-3218                 | KX264501                      | 3248                   | 681 465 2538 558 226 154 845 185 |

D1I is D1 India, D1M is D1 Middle East. S, X, P, C is S, X, P, C regions of HBV DNA genome. ‘–’ The same as the number before the first occurrence of the symbol in this column.
Results

Characteristics of the strains of hepatitis B virus (HBV)

In total, 3,037 whole-genome sequences of HBV A–H genotypes met the selection criteria and were included in the analysis. Among them, 1,357 belonged to genotypes A and C; 1,680 belonged to genotypes B, D–H (Table 1). The geographic distribution has been summarized in Supplementary Table 1. We established 28 genotype reference sequences of HBV genotypes A–H and deposited them in GenBank (Table 1). To reduce statistical bias resulting from differences in sample size, each subtype comprised no more than 50 strains. Therefore, 939 HBV strains were further selected for CpG island analysis, with genome size ranging 3117–3225 bp (A), 3128–3227 bp (B), 3104–3220 bp (C), 3101–3212 bp (D), 3185–3215 bp (E), 3129–3227 bp (F), 3234–3251 bp (G), and 3234–3251 bp (H) (Table 2).

Table 2. The CpG islands’ features in strains of HBV subgenotypes A–H.

| Genotype | Sub-genotype | Selected strains No. | CGI No. of each strain | Lack of CGI I No. of strains | Ratio | Number of strains | Strains contain 3 new CGIs | Ratio of new CGIs | New CGIs |
|----------|--------------|----------------------|------------------------|-------------------------------|-------|-------------------|---------------------------|-------------------|---------|
| A        | A1           | 50                   | 2–3                    | 34%                           | 17(0) | 33(0)             | 0                         | 0                 |         |
|          | A2           | 50                   | 2–4                    | 20%                           | 10(1) | 40(2)             | 6% IV                     |                   | IV, V   |
|          | A3           | 22                   | 2–3                    | 82%                           | 18(2) | 4(0)              | 0                         |                   |         |
|          | A5           | 25                   | 2–3                    | 76%                           | 19(1) | 6(0)              | 4% VI                     |                   | IV, V   |
|          | B1           | 24                   | 2–4                    | 4%                            | 1(0)  | 23(3)             | 13% IV, V                 |                   |         |
|          | B2           | 50                   | 2–4                    | 10%                           | 5(2)  | 45(5)             | 14% IV                    |                   |         |
|          | B3           | 23                   | 2–4                    | 22%                           | 5(3)  | 18(7)             | 43% IV, V                 |                   |         |
|          | B4           | 17                   | 2–4                    | 35%                           | 6(1)  | 11(1)             | 12% IV                    |                   |         |
|          | B6           | 39                   | 2–5                    | 3%                            | 1(0)  | 38(14)            | 36% IV                    |                   |         |
|          | B7           | 13                   | 2–4                    | 0%                            | 0     | 13(5)             | 38% IV, VI                |                   |         |
|          | B9           | 19                   | 2–4                    | 16%                           | 3(0)  | 16(5)             | 26% IV                    |                   |         |
| C        | C1           | 50                   | 2–4                    | 92%                           | 46(1) | 4(1)              | 14% IV, V, VI             |                   |         |
|          | C2           | 50                   | 2–4                    | 80%                           | 40(4) | 10(3)             | 14% IV, V, VI             |                   |         |
|          | C3           | 9                    | 2–3                    | 100%                          | 9(1)  | 0                 | 11% IV                    |                   |         |
|          | C6           | 17                   | 2–4                    | 12%                           | 2(1)  | 15(1)             | 12% IV                    |                   |         |
|          | D1           | 50                   | 2–4                    | 22%                           | 11(1) | 39(4)             | 10% IV                    |                   |         |
|          | D2           | 50                   | 2–4                    | 22%                           | 11(0) | 39(2)             | 4% IV                     |                   |         |
|          | D3           | 50                   | 2–3                    | 14%                           | 7(1)  | 43(0)             | 2% IV                     |                   |         |
|          | D5           | 31                   | 3–4                    | 3%                            | 1(1)  | 30(3)             | 13% IV, V                 |                   |         |
|          | D7           | 33                   | 1–3                    | 42%                           | 14(0) | 18(0)             | 0                         |                   |         |
| E        | E1           | 50                   | 2–4                    | 16%                           | 8(3)  | 42(2)             | 10% IV                    |                   |         |
|          | F1           | 50                   | 2–3                    | 100%                          | 50(1) | 0                 | 2% V                      |                   |         |
|          | F2           | 18                   | 2–3                    | 100%                          | 18(1) | 0                 | 6% IV                     |                   |         |
|          | F3           | 21                   | 2–3                    | 100%                          | 21(1) | 0                 | 5% V                      |                   |         |
|          | F4           | 28                   | 2–3                    | 100%                          | 28(22)| 0                 | 79% V                     |                   |         |
| G        | G1           | 28                   | 2–3                    | 93%                           | 26(0) | 0                 | 0                         |                   |         |
| H        | H1           | 22                   | 2–4                    | 100%                          | 22(4) | 0                 | 18% IV, V                 |                   |         |

CGI – CpG island. * Number of strains containing new CpG islands. '–' Absence of new CpG island.
Among the 28 reference sequences, genome lengths were different. However, the lengths of S and X genes were the same, with 681 and 465 bases coding for 226 and 465 amino acids, respectively (Table 1). The phylogenetic relationship between whole genomes and S genes among 28 reference sequences is shown in Figure 1A. The phylogenetic relationship of the S region is very similar to that of the whole genome, but with some minor differences. Their clustered tendency was similar to those reported previously [21].

**Characteristics of the CpG islands**

Of 28 reference sequences, each had 2–3 CpG islands. Among them, 17 (60.71%) contained three traditional CpG islands I, II, and III, while 11 other subtypes (39.29%) lacked a CpG island I (Table 3, Figure 1B). Only subtype F4 contained a new island, CpG island V (located at sites 1933–2036). In most of the reference sequences, the position and size of CpG islands II and III were similar. However, significant differences existed in those of CpG island I (Table 3).

Of the 939 selected strains, each strain contained 1–5 CpG islands. Furthermore, 48.35% (454/939) of strains included only CpG islands I, II, and III, but no new islands and 12.46% (117/939) contained new islands (Table 2). Among selected strains, only one strain contained a single CpG island (CpG island II, D7, FJ90442), 366 (38.98%) contained only CpG islands II and III, 62 (6.60%) contained four CpG islands, and
only one strain contained five CpG islands (CpG islands I–V, B6, DQ463802).

The number of strains containing CpG islands I–VI was 515, 939, 938, 65, 47, and 8, respectively. The distribution of CpG islands I–III is shown in Table 4 and Figure 2. A non-parametric test (rank sum test) was performed for the start sites, end sites, and lengths of CpG islands I, II, and III, respectively. All other sites displayed a significant difference (P<0.05) between the length of CpG island I in each subtype of A and between the length of CpG island II in each subtype of A were not significant. Differences between CpG island I of each subtype of C (start sites, end sites, and lengths), and between the length of CpG island II in each subtype of A were not significant. All other sites displayed a significant difference (P<0.05) (Supplementary Table 2).

CpG islands I and IV had a wide range of start and end sites and overlaps frequently occurred between the start and end sites of these CpG islands among different strains. The standard

Figure 1. Phylogenetic trees and CpG island distribution among hepatitis B virus (HBV) subtype reference sequences. (A) Phylogenetic tree diagram for the whole genome and S gene. (B) CpG island distribution. The horizontal axis denotes HBV genome sequence. The vertical blue strips indicate CpG islands I, II, V, and III, respectively. Vertical red bars below the horizontal axis indicate CpG dinucleotides, and the more intensive the red bars, the higher the enrichment of CpG dinucleotides.
Table 3. The number, size and position of CpG islands in HBV subgenotype references.

| References genotype | No. of CGIs | CGI I | CGI II | CGI III |
|---------------------|------------|-------|--------|---------|
|                     |            | Size (bp) | Position | Size (bp) | Position | Size (bp) | Position |
| A1                  | 3          | 186    | 99–284 | 417      | 1247–1663 | 161      | 2282–2442 |
| A2                  | 3          | 101    | 185–285 | 436      | 1228–1663 | 156      | 2294–2449 |
| A3                  | 2          |        |        | 436      | 1228–1663 | 144      | 2299–2442 |
| A4                  | 2          |        |        | 436      | 1228–1663 | 146      | 2299–2442 |
| B1                  | 3          | 154    | 112–265 | 451      | 1221–1671 | 165      | 2298–2462 |
| B2                  | 3          | 101    | 112–212 | 431      | 1228–1664 | 156      | 2298–2449 |
| B3                  | 3          | 103    | 110–212 | 450      | 1223–1672 | 156      | 2299–2455 |
| B4                  | 3          | 103    | 110–212 | 444      | 1223–1671 | 155      | 2299–2454 |
| B5                  | 3          | 103    | 110–212 | 448      | 1228–1664 | 156      | 2298–2449 |
| C1                  | 2          |        |        | 425      | 1247–1671 | 167      | 2280–2446 |
| C2                  | 2          |        |        | 451      | 1215–1665 | 149      | 2298–2446 |
| C3                  | 2          |        |        | 518      | 1215–1732 | 212      | 2234–2445 |
| C4                  | 3          |        |        | 158      | 1247–1730 | 163      | 2298–2449 |
| D1                  | 3          | 101    | 186–286 | 444      | 1228–1671 | 158      | 2289–2446 |
| D1M                 | 3          | 101    | 186–286 | 441      | 1228–1668 | 171      | 2276–2446 |
| D2                  | 3          | 101    | 186–286 | 441      | 1228–1668 | 171      | 2276–2446 |
| D3                  | 3          | 103    | 184–286 | 434      | 1228–1667 | 167      | 2280–2446 |
| D4                  | 3          | 101    | 186–286 | 437      | 1228–1664 | 183      | 2276–2458 |
| D5                  | 3          | 101    | 186–286 | 437      | 1228–1664 | 183      | 2276–2458 |
| D6                  | 3          | 101    | 186–286 | 437      | 1228–1664 | 183      | 2276–2458 |
| E                   | 3          | 100    | 184–283 | 428      | 1240–1667 | 123      | 2334–2456 |
| F1                  | 2          |        |        | 421      | 1242–1669 | 158      | 2298–2455 |
| F2                  | 2          |        |        | 410      | 1242–1651 | 162      | 2300–2461 |
| F3                  | 2          |        |        | 423      | 1243–1667 | 124      | 2335–2458 |
| F4                  | 3          | 104*   | 1933–2036* | 416   | 1242–1665 | 160      | 2299–2458 |
| G                   | 2          |        |        | 416      | 1216–1628 | 115      | 2257–2454 |
| H                   | 2          |        |        | 620      | 1106–1728 | 118      | 2336–2453 |

CGI – CpG island. * This is CpG island V. The blank represents CpG island absences. The first T of the EcoRI cleavage site is position 1 which by genotypes B/C as the standard sequences.

Table 4. CpG islands distribution characteristics in selected sequences.

| CGI | Start position | End position | Length |
|-----|----------------|--------------|--------|
| I   | 184* (76–187*) | 285 (211–304) | 102 (100–193) |
| II  | 1228 (1109–1248) | 1667 (1624–1725) | 439 (406–560) |
| III | 2298 (2234–2349) | 2453 (2405–2492) | 157 (108–200) |
| IV  | 330 (257–557) | 436 (374–659) | 112 (100–226) |
| V   | 1933 (1921–1945) | 2036 (2024–2054) | 104 (100–124) |
| VI  | 2877 (2800–2891) | 2986 (2942–2991) | 105 (100–166) |

CGI – CpG island. * Median; % the range of percentiles 2.5–97.5.
Figure 2. Frequency distribution of the start and end sites and the size of CpG islands in selected strains of hepatitis B virus (HBV). The left and middle panels show location frequency distribution of the start and end of each CpG island, respectively. The right panel shows the frequency distribution of the size of the CpG island. The horizontal axis refers to the site or size of CpG islands. The vertical axis refers to the frequency. Different colors represent different hepatitis B virus (HBV) genotypes. Annotations on the histogram indicate the major sites of individual subtypes (criteria: number of strains ≥20 and accounts for ≥50% of all strains from the corresponding subtype). The class interval of the start and end sites for CpG island IV are 50 nt and 20 nt for other CpG islands. N is the sum of all events.
**Figure 3.** Characteristics of CpG islands in selected strains from different subtypes of hepatitis B virus (HBV) (A) Percentages of hepatitis B virus (HBV) strains with differences in CpG island I, or new island status between different subtypes. CGI represents the CpG island. Green and blue colors represent strains lacking CpG island I. Orange and brown colors represent strains containing CpG island I. Blue and orange colors represent the presence of new islands. (B) The frequency composition of strains containing CpG islands I, II, III, and new islands, respectively, in each HBV subtype or genotype. The P-value represents the chi-squared analysis of the composition ratios between the subtypes (within the square brackets).
deviation (SD) values of the start site of CpG islands I–VI were determined to be 42.66, 34.49, 27.44, 88.39, 6.01, and 38.87, respectively, on using SPSS version 16.0. Based on the range and SD of the start site, CpG islands I and IV showed the highest dispersion, consistent with the distribution of CpG island I among different reference sequences. The start site of CpG island IV with or without island I displayed no significant difference in a rank sum test (P=0.413). Because CpG island IV was downstream of CpG island I, a similar analysis was performed for the start sites of CpG island IV and end sites of CpG island I of all strains containing CpG island IV displaying a significant difference (P<0.001). Therefore, CpG islands I and IV were considered not to be the same island.

The absence of CpG island I

Among 939 selected strains, 423 (45.05%) lacked CpG island I. Specifically, 100% of strains of subtypes C5, F1–F4 and genotype H, and more than 75% strains of subtypes A3, A5, C1, C2, and genotype G, respectively, but less than 48% strains of other (sub)genotypes lacked CpG island I. B7 was the only subtype where all strains contained CpG island I (Table 2) (Figure 3A). With respect to the genotype, CpG island I was lacking in 43.54% (A), 11.35% (B), 76.98% (C), 25.76% (D), 16% (E), 100% (F), 92.86% (G) and 100% (H). Overall, CpG island I was most frequently absent in genotypes C, F, G, and H. Strains with new islands accounted for 13.24% (56/423) of those without a CpG island I, and 11.82% (61/516) of those with a CpG island I. A chi-squared test was performed, but the differences were not statistically significant (P=0.513).

Characteristics of new CpG islands

Of the 939 selected strains, 117 (12.46%) contained 120 new CpG islands. The percentage of strains containing new CpG islands were 4.08% (6/147), 24.86% (46/185), 11.9% (15/126), 6.06% (16/264), 10% (5/50), 21.37% (25/117), 0%, and 18.18% (4/22) in HBV genotypes A–H, respectively. Subtypes A1, D7, and genotype G displayed no new CpG islands (Table 2, Figure 3). Sixty-five strains of HBV genotypes A–F and H contained CpG island IV. Among genotypes B, C, D, F, and H, 47 strains contained CpG island V, primarily distributed in B6 (12/47) and F4 (22/47). Eight strains from genotypes A–C contained CpG island VI (two in genotypes A5 and B7, six in genotype C). Q924641 (B3, no CpG island I), DQ463802 (B6, having CpG island I) and ABS16395 (H, no CpG island I) contained two new islands: CpG islands IV and V. The distribution, median, and 95% range of the start site, end site, and length of CpG islands IV, V, and VI are shown in Table 4 and Figures 2, 4.

Since the differences in CpG island I and new islands were greater than those in CpG islands II and III among each subtype, the chi-squared analysis was performed for the composition ratios of strains with or without CpG island I or new islands between subtypes of the same HBV genotype in this study (Figure 3A). The strains with CpG islands I, II, and III, and new islands among subtypes of each genotype were also
analyzed via a similar chi-squared test (Figure 3B). Figure 3A shows that the composition of strains with or without CpG island I or new CpG islands in each subtype of the same genotype was significantly different (P<0.05). However, there was no significant difference between some subtypes (P>0.05). The information provided in Figure 3B is not the same as in Figure 3A. It is evident that there were no significant differences in the composition of strains containing CpG islands I, II, and III and new islands among the subtypes of HBV genotypes B and D (P>0.05), while genotypes A, C, and F showed the opposite trend (P<0.05).

Discussion

Methylation is a major form of epigenetic modification of genomic DNA, which serves as an important means for functional regulation of the genome and is believed to be involved in the cellular resistance to viral DNA invasion into the nucleus [27]. CpG island methylation of hepatitis B virus (HBV) DNA has been shown to play important roles in regulating the adaptability of the virus, silencing transcription, and down-regulating viral replication [23,28–30]. DNA methylation is related to the specific location of CpG in mammals [31], and the distribution of CpG islands might affect HBV genome methylation. Also, studies have shown that CpG islands of HBV DNA are divergent among different genotypes, which might partially account for the difference in clinical outcome among different HBV genotypes [21].

The absence of CpG island I is common in genotypes A, C, F, G and H, and new CpG islands are common in genotypes B, C, E, F, H, but rare in A, D, and G [21,22]. However, the findings of the present study showed that there were significant differences between specific subtypes, and that every subtype may possess a different CpG island distribution within the same genotype (Figure 1B). Also, the results of the chi-squared test also showed that there were significant differences in CpG island composition among the various subtypes in the same genotype. For example, the absence rates of CpG island I were very high in subtypes C1, C2, and C5, but very low in subtype C6. The CpG island I absence rate in genotype B ranged from 0 (B7) to 35% (B4), genotype D range from 3% (D5) to 48% (D2). Therefore, the characteristics of CpG islands in HBV subtypes may be more accurate than that of its genotypes. Previously, the synthesized genome of the B2 subtype has been shown to be fully replication-competent by in vitro transfection into hepatoma cells, in addition to expression and replication in mice [18]. Therefore, the large number of features of HBV subtypes reference sequences and CpG islands obtained in this study may provide a more substantial theoretical basis for HBV methylation research.

Previous studies that have compared infections caused by HBV genotypes B and C found that infections with genotype B resulted in earlier hepatitis B e antigen (HBeAg) seroconversion, whereas infections with genotype C had an increased risk of developing cirrhosis and hepatocellular carcinoma (HCC) [32–34]. Other studies have speculated that this may be related to the increased absence rate of CpG island I in HBV genotype C [21].

From the previously published data and the findings of this study, HBV genotypes B and C are mostly distributed in China, and mainly include the B2, C1, and C2 subtypes (Supplementary Table 1). Consistent with previous findings, the present study showed that the rate of absence of CpG island I in HBV subtypes C1 and C2 was greater when compared with that in subtype B2. Therefore, it is possible to infer that the infection with HBV subtypes A3, A5 and genotypes F, G and H will result in similar clinical outcomes when compared with subtype C, showing a higher risk of developing cirrhosis and hepatocellular carcinoma (HCC) (Figure 3A). Two previously published clinical studies have shown an association between HBV genotype F and severe liver disease and HCC, and have shown that the risk of the development of HCC associated with HBV infection was significantly greater for genotype F than for genotypes A–D [35,36]. However, the function of CpG island I remains unclear, and how the absence of CpG island I facilitates the development of cirrhosis and HCC in HBV infection in patients requires further research.

The data obtained in the present study showed that CpG islands I–III among different subtypes in the almost all of HBV genotype showed significant differences (P<0.05), except for the CpG island I of genotype C (Supplementary Table 2). Figure 2 illustrates the same phenomenon in which the start sites of HBV subtypes B2 and B6 strains are focused around the same site, with the end sites being scattered. This phenomenon also appears widely in other CpG islands of other HBV subtypes, and the relationship between these occurrences and the different outcomes of infection with different HBV genotypes or the same genotype remain unclear.

This study had several limitations. All data analyzed were downloaded from GenBank, and the number of strains in some HBV genotypes and subtypes was small. Also, this study was a theoretical analysis of the differences in CpG islands, which are potential methylation sites between the various genotypes, and the possible impact of the findings on clinical outcomes was not specifically studied.

Although CpG islands of HBV strains from a specific subtype were quite different, the major distribution tendency of the start and end sites and the lengths of CpG islands were consistent with those of corresponding subtype reference sequences. This finding indicated that the reference sequences

Chen L et al.: CpG island distribution in HBV © Med Sci Monit, 2018; 24: 6781-6794
used in this study were reliable, representative of the general characteristics of strains from respective HBV subtypes, and suitable to serve as models and tools for evaluation of subtype-specific CpG islands as possible targets for methylation. Importantly, these sequences could be used for the investigation of potential roles of HBV DNA methylation in the clinical course of HBV infection.

**Conclusions**

The present study established 28 subtype reference sequences of HBV genotypes A–H. The composition of strains with or without CpG island I, or new islands between different subtypes within the same HBV genotype, showed significant differences. CpG islands I–III among different subtypes in almost all HBV genotypes showed significant differences, except for the CpG island I in genotype C. The findings of this study might provide a foundation for further studies on the role of HBV DNA methylation in determining subtype-specific HBV viral biology, immune reactions against HBV infection, the clinical course of HBV infection, and drug sensitivity.

**Conflict of interest**

None.

**Supplementary Tables**

**Supplementary Table 1.** The main geographic distribution of each HBV subgenotype strains.

| Region     | A1 | A2 | A3 | A5 | B1 | B2 | B3 | B4 | B6 | B7 | B9 | C1 | C2 | C5 | C6 | D11 | D1M | D2 | D3 | D5 | D7 | E  | F1 | F2 | F3 | F4 | G  | H  |   |
|------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Asia       | 40 | 47 | 24 | 362| 15 | 13 | 19 | 185| 9  | 17 | 98 | 217| 83 | 69 | 31 | 4   | 9  |    |    |    |    |    |    |    |    |    |    |    |    |
| India      | 16 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Japan      | 40 | 21 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| China      | 276| 70 | 605|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Indonesia  | 14 | 13 | 11 | 17 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Vietnam    | 12 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Thailand   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Iran       |    | 122|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Syria      |    | 60 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Europe     | 3  | 159| 1  | 10 | 2  | 6  | 2  | 12 | 84 | 62 | 39 | 70 | 4  | 9  |    |    |    |    |    |    |    |    |    |    |    |    |
| Belgium    | 69 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Poland     | 45 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Turkey     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Russia     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Sweden     |    | 16 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Africa     | 64 | 10 | 21 | 4  | 3  | 21 | 1  | 7  | 33 | 198|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| South      |    | 38 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Africa     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Cameroon   |    | 13 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Tunisia    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 33 |
| Guinea     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 70 |
| Niger      |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    | 66 |

**Supplementary Table 6792.** Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]
This table only lists countries which have 10 strains at least and ≥20% of the total number of that continents. Bold number indicates the total number of this subgenotype strains in this continents. China includes mainland China and Hong Kong, but not Taiwan region.

**Supplementary Table 2.** The P values of non-parametric tests of CpG islands among different subtypes of HBV genotypes.

| Region     | A1 | A2 | A3 | A5 | B1 | B2 | B3 | B4 | B6 | B7 | C9 | C1 | C2 | C5 | C6 | D1 | D1M | D2 | D3 | D5 | D7 | E | F1 | F2 | F3 | F4 | G | H |
|------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| America    | 48 | 15 | 21 | 5  | 36 | 5  | 3  | 10 | 18 | 68 | 18 | 21 | 28 | 14 | 13 |    |    |    |    |    |    |    |    |    |    |
| Haiti      | 36 | 21 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Canada     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Greenland  | 10 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Argentina  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Chile      |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| United States |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

The blank represents CpG island absences.

**References:**

1. Lozano R, Naghavi M, Foreman K et al: Global and regional mortality from 235 causes of death in 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. Lancet, 2012; 380: 2095–128
2. Schwetzler A, Horn J, Mikolajczyk RT et al: Estimations of worldwide prevalence of chronic hepatitis B virus infection: A systematic review of data published between 1965 and 2013. Lancet, 2015; 386: 1546–55
3. European Association for the Study of the Liver, European Association for the Study of the Liver: EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. J Hepatol, 2017; 67: 370–98
4. 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: 231–36
5. Hannoun C, Norder H, Lindh M: An aberrant genotype revealed in recombinant hepatitis B virus strains from Vietnam. J Gen Virol, 2000; 81: 2267–72
6. Kramvis A, Kew M, Francois G: Hepatitis B virus genotypes. Vaccine, 2005; 23: 2409–23
7. Tatematsu K, Tanaka Y, Kurbanov F et al: 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 Virol, 2009; 83: 10538–47
8. Bichko V, Pushko P, Dreilina D et al: Subtype ayw variant of hepatitis B virus: DNA primary structure analysis. FEBS Lett, 1985; 185: 208–12
9. Natural N, Schaefer S, Yoshida CF et al: Identification of a new hepatitis B virus (HBV) genotype from Brazil that expresses HBV surface antigen subtype adw4. J Gen Virol 1993; 74(7b): 1627–32
10. Norder H, Courouce AM, Magnus I: Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology, 1994; 198: 489–503
11. Stuyver L, De Gendt S, Van Geyt C et al: A new subtype of hepatitis B virus: Complete genome and phylogenetic relatedness. J Gen Virol, 2000; 81: 67–74
12. Owiredu WK, Kramvis A, Kew MC: Molecular analysis of hepatitis B virus genomes isolated from black African patients with fulminating hepatitis B. J Med Virol, 2001; 65: 485–92
13. Arauz-Ruiz P, Norder H, Robertson BH, Magnus I: Genotype H: A new Amerindian genotype of hepatitis B virus revealed in Central America. J Gen Virol, 2002; 83: 2059–73
14. Sugauchi F, Orito E, Ichida T et al: Hepatitis B virus of genotype B with or without recombination with genotype C over the precore region plus the core gene. J Virol, 2002; 76: 5985–92
15. Zhang ZH, Zhang L, Lu MJ et al: Establishment of reference sequences for sub-genotypes. Virus Genes, 2016; 52: 325–33
20. Vivekanandan P, Thomas D, Torbenson M: Hepatitis B viral DNA is methylated in liver tissues. J Viral Hepat, 2008; 15: 103–7
21. Zhang Y, Li C, Zhang Y et al: Comparative analysis of CpG islands among HBV genotypes. PLoS One, 2013; 8: e56711
22. Zhong C, Hou Z, Huang J et al: Mutations and CpG islands among hepatitis B virus genotypes in Europe. BMC Bioinformatics, 2015; 16: 38
23. Vivekanandan P, Thomas D, Torbenson M: Methylation regulates hepatitis B viral protein expression. J Infect Dis, 2009; 199: 1286–91
24. Vivekanandan P, Daniel HD, Kannangai R et al: Hepatitis B virus replication induces methylation of both host and viral DNA. J Virol, 2010; 84: 4321–29
25. Hong X, Kim ES, Guo H: Epigenetic regulation of hepatitis B virus covalently closed circular DNA: Implications for epigenetic therapy against chronic hepatitis B. Hepatology, 2017; 66: 2066–77
26. Fazzari MJ, Greally JM: Epigenomics: Beyond CpG islands. Nat Rev Genet, 2004; 5: 446–55
27. Jaenisch R, Bird A: Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet, 2003; 33(Suppl): 245–54
28. Curradi M, Izzo A, Badaracco G, Landsberger N: Molecular mechanisms of gene silencing mediated by DNA methylation. Mol Cell Biol, 2002; 22: 3157–73
29. Guo Y, Li Y, Mu S et al: Evidence that methylation of hepatitis B virus covalently closed circular DNA in liver tissues of patients with chronic hepatitis B modulates HBV replication. J Med Virol, 2009; 81: 1177–83
30. Kim JW, Lee SH, Park YS et al: Replicative activity of hepatitis B virus is negatively associated with methylation of covalently closed circular DNA in advanced hepatitis B virus infection. Intervirology, 2011; 54: 316–25
31. Robertson KD: DNA methylation and human disease. Nat Rev Genet, 2005; 6: 597–610
32. Chu CI, Hussain M, Lok AS: Hepatitis B virus genotype B is associated with earlier HBeAg seroconversion compared with hepatitis B virus genotype C. Gastroenterology, 2002; 122: 1756–62
33. Gunther S: Genetic variation in HBV infection: Genotypes and mutants. J Clin Virol, 2006; 36(Suppl. 1): 53–11
34. Yang HQ, Yeh SH, Chen PJ et al: Associations between hepatitis B virus genotype and mutants and the risk of hepatocellular carcinoma. J Natl Cancer Inst, 2008; 100: 1134–43
35. Sanchez-Tapias JM, Costa J, Mas A et al: Influence of hepatitis B virus genotype on the long-term outcome of chronic hepatitis B in Western patients. Gastroenterology, 2002; 123: 1848–56
36. Livingston SE, Simonetti JP, McMahon BJ et al: Hepatitis B virus genotypes in Alaska Native people with hepatocellular carcinoma: preponderance of genotype F. J Infect Dis, 2007; 195: 5–11