An aberrant genotype revealed in recombinant hepatitis B virus strains from Vietnam

Charles Hannoun,1 Heléne Norder2 and Magnus Lindh1

1 Department of Clinical Virology, Göteborg University, Guldhedsgatan 10B, 413 46 Göteborg, Sweden
2 Department of Virology, Swedish Institute for Infectious Disease Control, Stockholm, Sweden

Six genotypes of hepatitis B virus (HBV) have been described. However, relatively few complete genomes originating from East Asia, where most of the world’s HBV carriers live, have been studied. We analysed five complete HBV genomes of Vietnamese origin, which in our previous studies had produced atypical genotyping patterns. All five strains had HBsAg sequences with markers for serotype adw. In phylogenetic tree analysis, two of the genomes clustered with genotype C, and three clustered on a separate branch between genotypes A, B and C, suggesting a new genotype. However, these three strains showed signs of recombination in similarity plot and bootscanning analysis. Phylogenetic tree analysis of two segments separately supported recombination between genotype C and a putative new genotype (or possibly a subgroup of genotype A). The segment between nt 1801 and 2865 was clearly of genotype C origin, while the major part of the genome (nt 2866–1800) was placed on a branch close to genotype A. The findings encourage further study of genotypes and recombination in HBV from this geographical region.

Introduction

Chronic hepatitis B virus (HBV) infection affects more than 350 million people and is a major cause of liver disease worldwide. HBV was originally divided into four genotypes, A–D, on the basis of a more than 8% nucleotide (nt) difference of the whole genome (Okamoto et al., 1988). Phylogenetic tree analyses later confirmed these genotypes, and two further genotypes, E and F, were described (Norder et al., 1992, 1994). These six genotypes have a relatively distinct geographical distribution (Lindh et al., 1997; Norder et al., 1993). Recently, a putative seventh genotype with a unique core insertion was observed in HBV carriers from southern France and Georgia, USA (Stuyver et al., 2000).

In East Asia, where the majority of the world’s HBV carriers live, genotypes B and C prevail. However, considering that few strains from this part of the world have been studied, in particular in relation to the large number of carriers, as yet undiscovered genotypes might exist. In previous studies of HBV genotypes we observed strains with aberrant features in Vietnamese carriers. One group of strains appeared from pre-S and S region restriction fragment length polymorphism (RFLP) to be of genotype A (Lindh et al., 1997), and sequencing of the S region supported this classification. This finding was surprising because, with the exception of the Philippines, genotype A is rarely observed in East Asia. Moreover, these strains had thymine at nt position 1858, while essentially all genotype A strains described so far have cytosine at position 1858. This co-variation has been recognized because in genotype A strains C1858 effectively prevents the emergence of mutations at nt 1896 in the precore region, explaining the low prevalence of these mutations in geographical areas like north-western Europe, where genotype A prevails (Li et al., 1993).

The second group of strains was investigated because it could not be genotyped by pre-S or S region RFLP analysis. Sequencing of the pre-S region indicated that these strains belonged to genotype C, but they also showed similarities with genotype A (Lindh et al., 1998).

By analysing five complete genomes representing the two groups of aberrant strains we investigated whether any of...
them might represent a new genotype. Considering that our previous analyses had showed similarities with genotype A we thought that the study might also provide information about the origin and phylogenetic relatedness of genotype A.

Methods

- **Patients and samples.** Group I was represented by isolates from three HBeAg-positive carriers, producing RFLP patterns typical of genotype A in both the pre-S and S region, but carrying T1858. Two of these patients were relatives (niece/aunt, samples 14118 and 11141); one was unrelated (sample 6871). Group II was represented by isolates from two unrelated HBeAg-positive carriers (samples 3270 and 8290). These strains were identified because they produced atypical but identical pre-S and S region RFLP patterns. All five patients originated from the Hanoi region in northern Vietnam.

- **Sequencing.** PCR amplifying eight overlapping fragments of the HBV genome was done as described previously (Hannoun et al., 1999). Cycle sequencing was done by the chain-termination method using fluorescent dye terminators (ddNTPs) and the same primers as in PCR, analysing each amplicon in both sense and antisense direction. The sequence was read in an ABI Prism 310 automated capillary sequence reader (Applied Biosystems) and then processed using the Sequence Navigator software (Applied Biosystems).

- **Database sequences.** The following complete genomes (represented by their accession numbers) were used in SimPlot and phylogenetic tree analyses: genotype A, Z35717, X72478, S50225, X02763, X70185 and L13994; genotype B, D00329, D50521, D50522 and D23678; genotype C, AB014374, AB014376, D23681, D23683, D23684, D12980 and X75665; genotype D, J02203 and M32138; genotype E, X75664 and X75657; genotype F, X75663. In addition, sequences representing southern African and East Asian genotype A strains were used in trees based on S region comparison.

- **Phylogenetic analysis.** Phylogenetic trees were constructed by maximum likelihood analysis by quartet puzzling (Strimmer & Haeseler, 1996) using TreePuzzle, available at http://www.tree-puzzle.de. The following settings were used: 1000 puzzling steps, no clock-like branch lengths, nt Hasegawa-Kishono-Yano substitution model, transition/transversion parameter estimated from data, and both uniform and gamma distributed (alpha parameter from data set) rate heterogeneities. Trees were also constructed by distance matrix and parsimony analysis using the DNAdist, Neighbor and DNApars software of PHYLIP 3.5c (Felsenstein, 1989; phylogenetic inference package, distributed by the evolution.genetics.washington.edu). Bootstrapping of 1000 replicates was then done by using Seqboot (PHYLIP). Recombination was investigated using SimPlot (Lole et al., 1999; distributed by the author, S. Ray, at http://www.welch.jhu.edu/), and by bootscanning (Salminen et al., 1995).

Results

**Group I**

Two of the sequences, originating from a 21-year-old female and her 6-year-old niece, were almost identical (0·1% difference), while a sequence from an unrelated 43-year-old Vietnamese male differed by 1·5%. Compared with known genotypes, these genomes were most similar to genotype C (6·3–7·9% difference) and genotype A (7·6–8·6% difference).

Phylogenetic tree analyses of the complete genome placed these sequences on a unique branch, between genotype A and C. However, further analysis using SimPlot (Fig. 1) and bootscanning (not shown) indicated recombination, because the part of the genome ranging from nt 1801 to 2865 (33% of the genome) was more similar to genotype C than to the other genotypes, while the rest of the genome (nt 2866–1800, 67% of the genome) appeared to be unique, resembling neither of the known genotypes. In agreement, in phylogenetic trees (Fig. 2) based on nt 1801–2865 these sequences always clustered with genotype C, while in 70% of trees based on nt 2866–1800 they were placed on a unique branch close to genotype A, and in the remaining 30% were just proximal of the node dividing genotypes A and B, or, in a few cases, close to the base of the genotype C branch. Comparison of these trees (i.e. nt 1801–2865 versus nt 2866–1800) using TreePuzzle (option: tree comparison, user defined trees) showed that the topology was significantly different (P < 0·05; Kishino & Hasegawa, 1989). This was not found when the group I sequences were excluded from the analysis, indicating that the difference in topology was indeed due to the divergent placement of the group I strains, thus supporting the presence of recombination. Trees as analysed by DNA distance matrix and neighbour joining (not shown) produced nearly identical topologies to those based on maximum likelihood (TreePuzzle). In trees created by maximum likelihood analysis of amino acid sequences of the polymericase, which represents about 75% of the genome (nt 2307–1620), the group I strains were placed between genotype A and C.

**Group II**

The two genomes, originating from a 14-year-old boy and a 24-year-old female, both of Vietnamese origin but not related, differed by 1·6% from each other. Compared with other genotypes, they were most similar to genotype C (4·5–5·7% difference) and genotype A (8·1–9·1% difference). In the phylogenetic trees they were placed close to the base of the genotype C branch but more peripheral than X75665, a sequence originating from New Caledonia.

Fig. 3 shows the HBsAg amino acid sequences of representatives of the group I and II strains in comparison with database sequences of genotypes A–F. All five genomes carried I110, T126 and K160, which are characteristic for serotype adr and thus differed from most genotype C strains which are of serotype adr. The three sequences representing the putative recombinant showed no unique amino acids, but a unique combination of residues at genotype-related positions, including K24, I110, S114, T126, N131, S143, K160, Y161, A168, V194, N207 and L209.

Table 1 shows the genetic difference between group I and II strains and the sequences of all genotypes, for the segment between nt 1801 and 2865, as well as the segment between nt 2866 and 1800. In the former segment the difference between
Fig. 1. Nucleotide similarity comparison of group I strains with consensus sequences representing each of the genotypes A–F using SimPlot.

Fig. 2. Phylogenetic trees based on comparison of nt 2866–1800 (A) and nt 1801–2865 (B). The trees were created by maximum likelihood analysis using the TreePuzzle software with X75663 as outgroup and with 1000 puzzling steps. The figures at the nodes represent the reliability value, i.e. in percent how often the corresponding cluster was found among 1000 intermediate trees.
Fig. 3. HBsAg amino acids of group I and II strains aligned with database sequences of genotypes A–F.

Table 1. Mean nucleotide differences (%) between the analysed HBV isolates (group I and II) and sequences representing genotypes A–F

<table>
<thead>
<tr>
<th>Genotype*</th>
<th>nt range</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>Group I</th>
<th>Group II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group I (n = 3)</td>
<td>1801–2865</td>
<td>10-6%</td>
<td>10-9%</td>
<td>5-2%</td>
<td>11-9%</td>
<td>12-0%</td>
<td>12-4%</td>
<td>—</td>
<td>6-9%</td>
</tr>
<tr>
<td>Group II (n = 2)</td>
<td>2866–1800</td>
<td>6-6%</td>
<td>8-6%</td>
<td>7-7%</td>
<td>8-8%</td>
<td>8-4%</td>
<td>12-5%</td>
<td>—</td>
<td>7-3%</td>
</tr>
</tbody>
</table>

* Consensus sequences of each genotype.
† Sequence of genotype C most similar to the nt 1801–2865 segment of group I sequences.
‡ Sequence of genotype A most similar to the nt 2866–1800 segment of group I sequences.

group I sequences and genotypes A and C was 10-6% and 5-2%, respectively, while in the latter segment it was 6-6% and 7-7%, respectively. The table also shows the nt differences between genotypes A–F (consensus sequences) and the sequences Z35717 (genotype A) and D12980 (genotype C), which by tree analysis were the sequences most similar to the nt 2866–1800 and 1801–2865 segments (respectively) of the group I genomes. Thus, in relation to the intra-genotype variability, the nt 2866–1800 part is more distant from published genotype A sequences than the nt 1801–2865 part is from genotype C. The S regions of group I strains (included in the nt 2866–1800 segment) were also compared to further non-European, S region sequences (not available as complete genomes), showing that they were distant from genotype A from southern Africa and East Asia (Fig. 4).

Discussion

In the present study five HBV isolates from patients originating from northern Vietnam were analysed. These isolates were recognized because they produced atypical or unexpected genotyping patterns by RFLP analysis. Sequencing of the complete genome showed that two of the isolates belonged to genotype C, but were relatively divergent from
previously reported strains, including an overall nt difference of 4.5–5.7% and an HBsAg region sequence representing serotype adw. The three other isolates were first thought to represent a new genotype. However, a segment of these genomes (nt 1801–2865) was found to be similar to genotype C, while the remaining major part (nt 2866–1800) differed substantially from all genotypes, although it was most similar to genotype A. The findings suggest that these aberrant strains represent recombination between genotype C and a new genotype, or possibly between genotype C and a subgroup of genotype A. This conclusion was supported by SimPlot and bootscanning analysis, and by phylogenetic trees constructed by comparing the two segments of the genome separately. Moreover, topology of the trees from the two parts of the genome was significantly different when compared using TreePuzzle.

To some extent the assessment of recombination was hampered by the lack of available databases of genotype A and C sequences from Vietnam or the surrounding areas. Both the recombinant, genotype C segment (nt 1801–2865) of the group I sequences and the two group II isolates were placed between Japanese and New Caledonian sequences on the genotype C branch. The absence of C1858 and a 6 nt insertion in the core region, both of which are typical for genotype A, agrees with the interpretation that the nt 1801–2865 segment of the group I strains originates from genotype C.

The similarity of the nt 2866–1800 segment with genotype A agrees with our previous finding that S region RFLP classified these strains as genotype A (Lindh et al., 1997). However, in the SimPlot analysis the nt 2866–1800 segment appeared to differ from all genotypes, with genotype A being only slightly less different than the other genotypes, supporting the fact that this segment represents a new genotype rather than genotype A. This agrees with the finding that in phylogenetic analysis the S region of group I strains was clearly distinct from available genotype A sequences of Asian and African origin. Accordingly, a detailed analysis of the HBsAg region of the group I strains revealed similarities with all genotypes, in particular A, B and C. These similarities as well as the tree topology indicate that this putative new genotype (or subgroup of genotype A) may be a link between the European/African genotype A and the Asian genotypes B and C, suggesting that all these genotypes may have a common ancestor, as previously suggested (Norder et al., 1996).

The fact that one of the three genomes in group I was 1.2% different from the other two but also contained the recombinant segment suggests that the recombination may not be recent. Analysis of more sequences from southeast Asia (and in particular from Vietnam) is essential for investigating whether this recombinant is widespread and for identifying complete genomes representing the putative new genotype or subgroup of genotype A. Such studies could possibly also identify the genotype C strains from which the recombinant segment originates. Moreover, further study of recombination should be undertaken, in particular of isolates from this geographical region, where the HBV prevalence is high and co-infection with different genotypes may exist.

Recombination of segments from different genotypes of HBV has so far not been much discussed. However, signs of recombination have been found in integrated HBV DNA (Georgi-Geisberger et al., 1992), and analysis of HBV genomes from databases suggests that recombination may be relatively frequent (Bollyky et al., 1996; Bowyer & Sim, 2000). These studies describe genotype B strains that by recombination probably have acquired a portion of the core gene from genotype C. Of interest, this segment largely overlaps with the recombinant (genotype C) segment described in the present study, possibly indicating that this part is particularly prone to recombination. It is unclear at what stage of replication recombination is most likely to occur. Recombination during reverse transcription of the pregenomic RNA, a process taking place simultaneously with encapsidation, appears unlikely...
because it would require two pregenomic RNA molecules to be encapsidated, and such phenomena have not been observed in studies of HBV particle sedimentation rates. It seems more likely that recombination would take place in the nucleus where co-existing covalently closed circular DNA copies derived from different genotypes could exchange genomic segments. Regardless of mechanism, recombination requires coinfection with different strains (genotypes) and this has not been well documented either. It is possible that coinfection is more frequent than previously thought, but that it is rarely detected because infection with a second strain is suppressed or merely cannot be established due to the very high virus load of the first strain. In such cases, coinfection may become visible only when the first strain disappears. Such shift in genotypes (from A to D) has indeed been observed during HBe seroconversion in European children (Gerner et al., 1998).

Further study of East Asian HBV sequences might clarify the detailed relatedness between these genotypes and is required for establishing the existence of the putative new genotype and to further evaluate recombination events in HBV.

We thank Mika Salminen for help with SimPlot analysis, Peter Horal for critical review of the manuscript and Anki Gusdal for her technical expertise. The project was supported by grants from the Swedish Medical Research Council and the Swedish Medical Association.

References


Received 14 March 2000; Accepted 24 May 2000