Hepatitis B virus (HBV), a member of the family Hepadnaviridae, has a circular, partially double-stranded DNA genome of about 3200 nt (Magnius & Norder, 1995). The genome is highly compact, containing four open reading frames (ORFs) encoding the polymerase, surface antigen (S), nucleocapsid (core) and X proteins. These ORFs show significant overlap, with almost half of the HBV genome encoding more than one protein product. Substitution rates in the overlapping and non-overlapping genomic regions vary (Bollyky & Holmes, 1999; Zhou & Holmes, 2007).

HBV genomes are segregated into eight genotypes based on sequence divergences (Norder et al., 1992; Bowyer & Sim, 2000; Stuyver et al., 2000; Arauz-Ruiz et al., 2002; Akuta & Kumada, 2005). Genotypes are defined as sequences within a phylogenetic clade having less than 4% sequence divergence among the members of the clade and more that 8% sequence divergence with extra-clade sequences (Okamoto et al., 1988; Norder et al., 1994; Bowyer & Sim, 2000). Divergences as high as 13.6 and 15.5% have been noted for genotypes E and F with respect to genotypes A–D (Norder et al., 1994). Phylogenetic analysis of different regions and genes within the HBV genome has shown that, although the branching order of genotypes differs from the whole genome, the sequences belonging to specific genotypes consistently cluster together (Norder et al., 1994; Kato et al., 2002).

Recombination within and between genotypes has created complex patterns of evolution (Bollyky & Holmes, 1999; Bowyer & Sim, 2000). Recombination can alter the cladistic structure of genotypes. Within the core region alone, recombination between genotypes B and C is considered to lead to the formation of the Ba subgenotype (Sugauchi et al., 2002) and the apparent merging of genotypes D and E (Bowyer & Sim, 2000). A recent report on the potential recombinant nature of genotype G adds another example of a genotype, in addition to genotype E, all member strains of which originate from one ancestral intergenotype recombinant strain (Simmonds & Midgley, 2005). These observations emphasize a need for systematic investigation of genealogical relationships among genotypes along the HBV genome. In the present study, we conducted such an investigation of HBV genotypes using different regions of the HBV genome based on information available from a large database of sequences.

A set of 311 full-length human HBV sequences was obtained from GenBank. Non-human primate sequences and sequences with deletions were excluded. When multiple sequences came from a single patient, only one sequence was chosen. HBV strains with intergenotypic recombination have been found by many researchers (Bollyky & Holmes, 1999; Bowyer & Sim, 2000; Fares & Holmes, 2002; Simmonds & Midgley, 2005). SimPlot was used to remove individual sequences with evidence of intergenotypic recombination (Lole et al., 1999). Identification of recombination by SimPlot requires the use of reference sequences representing each HBV genotype. These sequences were selected by summing the square of the genetic distances between a given sequence and all other sequences in the database belonging to the same genotype, and selecting the sequence with the lowest sum-squared distance for each genotype. This approach to selecting sequences ensured that the identified benchmark sequences were representative of an entire genotype and, as such, were devoid of any elements that were uncharacteristic for the selected genotype. The use of actual instead of consensus sequences guaranteed against any potential
artefactual structural features being introduced by generating the consensus. The GenBank accession numbers for the sequences selected were: AY128092 (genotype A), D23679 (genotype B), AF458664 (genotype C), AF121240 (genotype D), AB032431 (genotype E), AB036905 (genotype F), AF160501 (genotype G), AY090457 (genotype H), and AF046996 (woolly monkey HBV). These sequences were tested with SimPlot and RDP (Martin & Rybicki, 2000) to ensure they were free from recombination. No attempts were made to remove genotype- and subgenotype-wide recombinants such as genotype E (Bowyer & Sim, 2000) and subgenotype Ba sequences (Sugauchi et al., 2002). Only single intergenotypic recombinant sequences were removed from further analysis.

Sequences from the resulting dataset were aligned using CLUSTAL_X (Thompson et al., 1997), and all positions with gaps in the aligned sequences and the 36 nt insert in the core gene of genotype G sequences (Stuyver et al., 2000) were removed. No attempt was made to ensure that codon integrity was maintained. Woolly monkey HBV (GenBank accession no. AF046996) was used as an outgroup. The final dataset contained 241 human HBV sequences.

Originally, the HBV sequence dataset was examined using a sliding window of 700 bases with a 50-base step to create phylogenetic trees for each window (data not shown) to determine phylogenetic relationships between different regions of the genome in a systematic manner. However, evaluation of variations in topologies of a large set of trees is difficult. To facilitate analysis of changes of intergenotypic distances among HBV genotypes within different regions across the genome, a graphical method was devised.

This method is based on plots that depict the distances from two different genotypic sequences, called query sequences Q_1 and Q_2, to a third sequence, called the reference sequence R. A sliding window of 700 bases with a 50-base step was used to examine the set of nine benchmark sequences described above. The 700-base window was chosen to minimize multifurcations and intergenotypic mixing of sequences (Nei et al., 1998). A distance matrix was generated for each window using TREEPUZZLE (Strimmer & von Haeseler, 1996). Graphs of the matrix distances were created by plotting the distances for Q_1 and Q_2 with respect to R. The distance values for each window for a specific pair of Q_1 and Q_2 were plotted as d(R,Q_2) versus d(R,Q_1) (Fig. 1). The plot shows the change in this value within each sliding window across the HBV genome. It was assumed that, if rates of mutation are constant between Q_1 and Q_2 with respect to R across the genome, a straight-line plot would be expected. As the rate of mutation across the HBV genome is not constant (Bollyky & Holmes, 1999; Hannoun et al., 2000; Fares & Holmes, 2002), there will be deviations from a straight line. Large deviations between sequences may represent either large deviations in the rate of mutation between the query sequences or recombination events. The distance matrix plot alone cannot make a distinction between these two explanations. Additionally, if R is phylogenetically at an equal distance from both Q_1 and Q_2, the distance matrix plot for all sliding windows along the HBV genome should be spread along the straight line at 45°. Since each sequence for this analysis was specifically selected to represent the entire genotype, it was expected that all plots would follow this pattern.

However, using these distance plots, HBV genotypes were found to be clustered into two distinct groups. The first group comprised genotypes A–E and the second genotypes F and H. Whenever Q_1 and Q_2 were from genotypes A–E, the value of the ratios approximated 1 (Fig. 1b). The same outcome was achieved when genotype F and H sequences formed the query pair (Fig. 1a). These results indicated that the members of each group evolved at a similar rate to the other members within the group. Whenever one sequence in a query pair was genotype A–E and the second member was either genotype F or H, the values of the ratios were...
about 2 or 0.5, depending on how the distance values were plotted (Fig. 1c, d). This outcome indicated that the two groups, one being composed of genotypes A–E and the other of genotypes F and H, were evolving at different rates. Such deviation suggests differences in hierarchical structure among genotypes. When genotype G was used as one of the query sequences, deviation from linearity was observed on all plots in the X/core region (between nt 1331 and 2291) (Fig. 1e, f). This finding suggests that the X/core region of genotype G has experienced a higher rate of evolution compared with the corresponding region of any other query sequence used in the distance matrix plot. This observation could be due to either a higher rate of mutation in genotype G or a recombination event. Without the X/core region included, the genotype G sequence behaved like the genotype A–E sequences.

To examine this deviation from expected behaviour in more detail, the frequency distribution of phylogenetic distances for the 242 sequence dataset (241 human sequences plus the woolly monkey sequence) was determined. To do this, DNADIST with the Kimura two-parameter model was used to generate distance matrices (Felsenstein, 1989). Frequency distributions were generated by segregating these phylogenetic distances into 80 equal-sized distance categories and plotting the frequency of occurrence for all distances per distance bin using the frequency function in Microsoft Excel. If HBV was stratified only as isolates (subgenotypes) and genotypes, a two-peak frequency distribution should become evident (excluding the outgroup peak) when the frequency of phylogenetic distances was plotted for all genotypes. However, a three-peak distribution was observed when complete HBV genome sequences were used (data not shown).

Because of the aberrant behaviour seen with genotype G (Fig. 1e, f), the X/core region (nt 1331–2291) was first excluded from analysis. With the X/core region excluded, the three-peak structure of the distribution was better delineated (Fig. 2a). A phylogenetic tree created using DNADIST and NEIGHBOR (Felsenstein, 1989) with the Kimura two-parameter model revealed two groups of genotypes: group I was composed of genotypes A–E and G, and group II of genotypes F and H (Fig. 2c).

To examine this segregation further, we analysed frequency distribution footprints (Fig. 2b). These footprints showed segregation of sequence distances separated intragenotypically (box 1), intergenotypically (box 2) and as an intergroup (box 3). The left-most bimodal peak consisted of intra-subgenotype and individual isolate sequence distances (with divergence of less than 4%) and intersubgenotype distances (sequence divergence between 4 and 8%). The bimodal pattern of these subgenotypic peaks was due to genotypes A, D and F splitting into subgenotypes (Bowyer & Sim, 2000; Norder et al., 2004).

Similar analysis with the X/core region alone (nt 1331–2291) yielded a phylogenetic tree in which HBV genotypes fell into three groups. The first group comprised genotypes A–E, the second genotypes F and H, and the third genotype G (Fig. 3c). Similar results were obtained using maximum-likelihood phylogenetic reconstructions (Guindon & Gascuel, 2003) on the set of nine benchmark sequences (data not shown). The frequency distribution footprint showed that all genotype G distances, other than those within genotype G, were now within the range of intergroup distances (Fig. 3b, box 3, light shaded bars). The movement of genotype G from group I to group III in the X/core region explains the deviation seen in the distance matrix plots, which included genotype G as one of the query sequences (Fig. 1e, f). The data are consistent with genotype G having resulted from a recombination event between the polymerase gene of a parental genotype G and the non-polymerase region of an as-yet-undiscovered HBV isolate, for which a complete nucleotide
Recombination has played a significant role in shaping some extant subgenotypes and genotypes (Bowyer & Sim, 2000; Sugauchi et al., 2002). Noticeable traces of inter-genotypic recombination can be found within any region of the HBV genome (Simmonds & Midgley, 2005), with a ‘hot spot’ being located at nt 1627–3252, which includes the X/core region (Zhou & Holmes, 2007). However, it is important to note that the majority of these recombinants are represented by a single HBV strain, which suggests that all of these strains have a very recent origin and/or limited fitness benefits. Data presented here on genotype G recombination, as well as published data on recombination between genotypes D and E (Bowyer & Sim, 2000) and genotypes B and C (Sugauchi et al., 2002), indicate that only recombination between strains from distant HBV clades within the genomic region outside the polymerase gene (within the X/core region) may generate progeny strains with significant evolutionary advantages over the parental strains.

In conclusion, present-day HBV research is orientated towards examining HBV genotypes and subgenotypes. The finding of HBV genotypes being organized into three higher-order supragenotype groups leads to the understanding that HBV sequences are organized in a more complex hierarchy. This finding may have significant implications for HBV biology, pathobiology, vaccine-implementation strategies and diagnostic assay development.

Acknowledgements

Disclaimer: This information is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry. It does not represent and should not be construed to represent any agency determination or policy.

References


Nei, M., Kumar, S. & Takahashi, K. (1998). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci U S A 95, 12390–12397.


