Relationships within and between genotypes of hepatitis B virus at points across the genome: footprints of recombination in certain isolates

S. M. Bowyer and J. G. M. Sim

National Institute for Virology and Department of Virology, University of the Witwatersrand, Private Bag X4, Sandringham 2131, Johannesburg, South Africa

Hepatitis B virus (HBV) was partitioned into type, subtype and isolate categories and the average evolutionary distances within and between categories was plotted at each of 54 points along the genome. The graphs showed alternating variable and conserved domains within and between HBV subtypes and revealed that some specimens assigned to different groups are more similar across several contiguous intervals than specimens belonging to the same group. Isolates were screened individually to determine their conformation to type and mosaic structure was identified in 14/65 specimens. Two entire clades (six specimens) of genotype B had a B/C sequence switch in the core gene region, whereas six genotype D specimens showed D/A switching in one or more regions of the genome. Genotype E was not separate from genotype D in the X and C subgenomic regions. The nature and distribution of polymorphic sites in mosaic regions was mapped at both the nucleotide and protein levels and the position of the variant fragments was related to mutational hot spots and linear epitopes of HBV. Mosaic structure was demonstrated statistically in 11 isolates using bootstrap resampling and recombination, rather than random change, appeared to be the mechanism responsible. The sequence between and including the two DR regions was represented in all putative recombinants. The distribution of genetic distances over subgenomic regions showed that substitution rates are not constant among the lineages of HBV in the preS regions. Genotype F is the most diverse group. Only genotypes A, C and F partition consistently into subtypes.

Introduction

Hepatitis B virus (HBV) belongs to the hepadnavirus family of enveloped DNA viruses containing a partially double-stranded genome of 3182–3221 bp depending on genotype. It is the smallest of the DNA viruses which infect man and causes acute hepatitis of varying severity. The virus persists in 2–10% of adult patients and approximately 90% of infected infants leading to chronic liver disease. In highly endemic areas, infection is predominantly acquired during the perinatal/neonatal period or by horizontal transmission in the first few years of life (Beasley & Hwang, 1984; Botha et al., 1984; Vardas et al., 1999). Since this results in a high prevalence of long-term HBV carriers with a low average age at infection (Edmunds et al., 1996), the virus has a long time span in which to evolve within its host.

The genome is read in all three reading frames and viral regulatory elements are all within coding regions which introduces constraints on the ability of the virus to accept mutations and remain viable (Yang et al., 1995). Nevertheless, heterogeneity among the strains of HBV circulating globally is 10-fold greater than that in the majority of DNA viral genomes. This is explained, at least partially, by the fact that hepadnavirus replication takes place via an RNA intermediate and reverse transcriptase is known to have a high error rate (Boyer et al., 1992). A nucleotide exchange rate of between 0·1 and 0·7 per year (Günther et al., 1999) has been estimated for the HBV (Okamoto et al., 1987) and woodchuck hepatitis virus (WHV; Girones & Miller, 1989) genomes, respectively, which is similar to the most slowly evolving gene of retroviruses, the gag gene, and one to two orders of magnitude lower than the
mutation rates previously calculated for the positive-and negative-strand RNA viruses (Girones & Miller, 1989).

Originally, four genotypic groups of HBV (A–D) were defined, based on an inter-genotypic divergence score of 8.5–10.0% between 18 complete genomes, as compared to a score of 1.1–2.7% between isolates within the same genotype (Okamoto et al., 1988). This genotypic classification was extended to six genotypes (A–F) by phylogenetic analysis of 122 surface antigen (HBsAg) genes (Norder et al., 1993). The genotypic groups are geographically arranged (Magnius & Norder, 1995) with genotypes B and C confined to Asia while genotype A predominates in Northern Europe giving way to genotype D as one moves toward the Mediterranean region. Genotype E is mainly found in parts of East, Central and West Africa and genotype F is only found in the New World and the Pacific which is also home to the Cq.

Phylogenetic analyses. Sequences were aligned manually using the DNASIS genetic systems software program (Hitachi software). Multiple HBV DNA phylogenies and bootstrap analyses were performed using programs (DNAPARS, SEQBOOT, DNADIST, FITCH, NEIGHBOR, DNAML, CONSENSE, DRAWTREE and DRAWG RAM) from the PHYLIP phylogeny inference package (version 3.5c, by Joseph Felsenstein). Sequences were identified by their GenBank accession numbers as well as by their position in the DNAPARS consensus tree (see legend to Fig. 1). Subgroups were defined as groups of isolates divergent by 4% or less and clusters were combined to form higher order clades if their component isolates showed a sequence divergence of less than this. Retaining the standard cut-off of 8% divergence to define genotypes, the 65 specimens were classified into a series of 11 subtypes within the six genotypes.

Average genetic distance graphs. The aligned full-genome sequences were split into 53 separate files of 60 nucleotides and one of 42 nucleotides using CLUSTALW (Thompson et al., 1994). These 54 files of interleaved sequence data provided the input to DNADIST which produced 54 distance matrices, one for each 60 nucleotides of the genome. Each matrix was formatted into a single column and imported into the standard spreadsheet QUATTRO PRO (Corel 7; Perfect Office Suite) for ease of subsequent calculation and manipulation. For each of the 11 subgroups at each of the 54 intervals, we then calculated the average pair-wise distance between: (i) isolates from the same subgroup; (ii) isolates from different subgroups within a genotype; (iii) and isolates from subgroups belonging to different genotypes.

Plotting these average distances against the position of the nucleotide interval within the genome generated 11 intra-subgroup, six inter-subgroup and 49 inter-genotype genetic distance graphs.

Isolate screening. Each of the 65 specimens was screened for conformance to type using a simple self-written dBase program based on Siepel’s recombinant identification program, RIP (Siepel & Korber, 1995). Our program compared the sequence of each specimen with each of 11 consensus sequences (one for each subgroup) and recorded the number of matches over each 50 nucleotides. Variant regions within a specimen were identified when the best match within a window switched from the type established from the full-genome consensus tree.

Individual genetic distance graphs. Having identified specimens within the database which contained variant regions, we plotted, in turn on the same axis, the average pair-wise distance of the specimen from subgroups of interest at each of the 54 points along the genome.

Nucleotide/protein maps. The distribution and nature of nucleotide (and amino acid) mutations within an isolate were mapped against the consensus sequence of the parental subtypes. The base (or amino acid) at each variant position was compared to the corresponding consensus base/amino acid from both the original and alternate genotype by listing the three values as a triplet of bases made up of the mutant value with the original and alternate value to its left and right, respectively. If either reference base/amino acid matched the specimen it was replaced by an asterisk. Proteins were mapped in all three reading frames (not shown).
Subgenomic bootstrap trees. The boundaries of the variant regions defined nine mosaic blocks, or fragments, within the specimens and the bootstrapped re-sampling NEIGHBOR-JOINING tree was drawn for each and compared with the full-genome bootstrapped tree.

Histograms. We examined the distribution of the genetic distances between the 65 specimens within conserved and variable domains along the genome. These included the preS2 region, the surface gene (which overlaps the reverse transcriptase/polymerase, RT/Pol, domain of P), the RNase H domain, the X gene, the core gene, the terminal protein (TP) and the preS1 region. The distance matrix for each region was used to calculate the frequency of each successive 0.005 range of genetic distance. This frequency was plotted to show the distribution of genetic distances within each of three categories: intra-subgroup, inter-subgroup and inter-genotype. Database records included information on the source of each distance reading so that a listing of specimens contributing to peaks of interest in the histograms could be generated, sorted and analysed when required.

Results

There have been many genetic classification studies of HBV (Kidd-Ljunggren et al., 1994, 1995; Mizokami et al., 1997; Ohba et al., 1995; Norder et al., 1993) since the first phylogenies appeared (Okamoto et al., 1988; Uy et al., 1992; Norder et al., 1992). To clarify some of the questions raised by these and other studies, we divided the six genotypes into 11 subgroups, in which specimens differed from each other by 4% or less, and used sequence similarity measurements to examine the relationships between 65 complete genomes at 54 points along the genome.

Subgroups of HBV

Using the distance matrix program FITCH, with 100 data sets, the 65 complete genomes grouped into the six conventional genotypes (A–F) all with bootstrap values of 100% (Fig. 1). Genotype E isolates clustered together and away from genotype D. Genotype F was the most diverse group, separate from all other genotypes.

Regardless of the algorithm used, the bootstrapped trees all showed two clusters of genotype A, three of genotype B and five of genotype C. We retained the designations A and A’ to differentiate the subgroups of genotype A (Bowyer et al., 1997). Each of the three subgroups of genotype B clustered with an original prototype of genotype B as recognized by Okamoto et al. (1988) and were designated B, B’ and B” (for the groups containing D00329, D00330 and D00331, respectively, previously designated serotypes adw1, adw2 and adw3). Despite numerous clades of C in the bootstrapped tree, only the subgroups C and Cq− observed by Norder et al. (1994) fitted our subgroup criteria. Genotype D formed a core clade with a very high bootstrap value of 99 plus six outliers which...
Fig. 2. For legend see facing page.
Footprints of recombination in HBV isolates

Full genome sequences were compared within and between subgroups at each of the 54 positions along the genome.

Intra-subgroup. Fig. 2(a) compares the average pair-wise distances between specimens within each of the 11 subgroups in our chosen range of 0–4%. Subgroup A of genotype A is the most conserved subtype showing intra-subgroup distances below 4% over most of the genome (Fig. 2a, i). The effect of treating genotype F (within which specimens varied by 4.7%) as a single subgroup is evident in Fig. 2a, ii. The pattern of this genetic distance graph is more reminiscent of the inter-genotypic graphs (Fig. 2b) confirming that more than one subgroup of this genotype exists and demonstrating the value of the graphs in providing a snapshot of the relationships between isolates from the same or different subgroups.

Inter-subgroup. The inter-subgroup graphs show a series of conserved and variable domains along the genome of HBV (Fig. 2b). Surface gene variation is typically below 4% between specimens in subgroups of the same genotype. The X gene is also well conserved and differences seldom exceed 8%. The preS1 and preS2 regions vary by 8% or more except in the B × B’ (Fig. 2b, ii) and D × Dm (Fig. 2b, vi) graphs. The TP domain of the P gene (intervals 38–48) peaks above 8% except in the D × Dm graph (Fig. 2b, vi). The RNase H domain (intervals 18–27 between the heavy vertical lines in Fig. 2) of the P gene is most variable between subgroup B’ and the other two subgroups of genotype B (Fig. 2b, iii and iv). Differences between subtypes in the core gene (intervals 30–41 between the thin vertical lines in Fig. 2) vary from < 8% between subgroups B’ and B” (Fig. 2b, iv) to > 18% between B and B’ (Fig. 2b, ii) and B and B” (Fig. 2b, iii).

Inter-genotype. After careful examination of the 49 inter-genotypic graphs, we selected eight typical, or anomalous, graphs to illustrate the features of these series (Fig. 2c). The conserved and variable domains are more defined in these graphs. Differences of 12–34% are typical between the genotypes in the preS1, preS2 and TP domains. The heterogeneity of the early RNase H domain persists. The S and X genes, particularly in their early regions, are most conserved. Unconstrained regions of the HBV genome show the most variation. These include the last third of the RT/Pol gene, the first half of the RNase H domain (before it overlaps the X gene) and almost the entire TP domain (the boundaries of the domains of P are as defined by Miller, 1988). The preS1 and preS2 regions which overlap the spacer region of the polymerase gene also fall into this category. Although the second half of the X gene and the first two-thirds of the core...
gene are unconstrained, the former is well conserved and the latter is not uniformly variable between the different types. The anomalous peaks in B × B' and B × B'' (marked with * in Fig. 2b, ii and iii) are dramatically reversed in their corresponding inter-genotype graphs with genotype C [see peaks marked * for B' × C (Fig. 2c, iv), and ** for B'' × C (Fig. 2c, v)]. These two sets of graphs show quite clearly that two entire subgroups of genotype B are more closely related to genotype C over at least 480 nucleotides (intervals 30–37). In contrast, the graph of subgroup B (the third clade of genotype B) versus subgroup C (Fig. 2c, iii) has the typical inter-genotype pattern. Only inter-subgroup differences exist between both subgroups of genotype D and genotype E over the latter part of the X gene and most of the C gene (Fig. 2c, vi and vii).

Mosaic structure within isolates

The screening programme identified mosaic sequence in 14/65 isolates. These isolates were from subgroup D (3/12, D05–D07), subgroup Dm (3/6, Dm16–Dm18), subgroup B' (4/4, B'01–B'04), subgroup B'' (2/2, B''11–B''12) and genotype E (2/2, E01–E02). In all cases, genotype D contained mosaics of genotype A and genotype B contained mosaics of genotype C. Both genotype E specimens displayed only subgroup differences from the subgroups of genotype D between nucleotides 1576 and 2262.

Graphs of genetic distance of variant specimens from the consensus of their parental genotypes located breakpoints (Fig. 3). The first and last variant nucleotide (Fig. 4a) and amino acid (not shown) of each block was precisely identified. Nine mosaic fragments were identified within the 14 specimens (Fig. 4b).

Bootstrap values (with a 75% cut-off) were used to give a statistical measure to the genotype switch within the fragments and to establish the confidence of the groupings (Simmonds et al., 1996). Mosaic structure was demonstrated in seven out of nine fragments (Fig. 5; Table 1a, bootstrap column). Dm16 (Figs 3a, 4b and 5a), Dm17 (Figs 4a, b and 5b) and Dm18 (Figs 4b and 5b, c) have one, or more, extensive and significant D/A segments. A long B/C fragment in the core gene region (Figs 3b, 4b and 5c, d) was common to all specimens in clades B' and B''. Two of the shorter fragments, the D/A fragments in specimen Dm18 (fragment I) and specimens D05–D07 (fragment VII), did not have bootstrap values above the 75% cut-off (Table 1). There is a complete disruption of the conventional phylogeny of HBV in the core subgenomic region (Fig. 5c, d).

In this region, genotype B is represented by subgroup B specimens only, whereas the rest of genotype B clusters with genotype C and genotype E loses its own identity and clusters with genotype D. Genotype A loses its identity as a separate and discrete group between intervals 28 and 30 unless variants (D05–D07 and D16–D18) are excluded from the analysis (not shown).

Effect of parallel evolution

If selection is causing parallel replacements, the evidence for mosaic structure should be stronger when only synonymous change is considered (Smith, 1992). However, the opposite is true within eight out of nine fragments. The difference in divergence between a fragment and its respective parental consensus sequences when total change is considered [mosaic index (MI); Table 1a] is greater than when synonymous change alone is considered [synonymous mosaic index (SMI); Table 1b] with one exception, fragment V in RF1 (see MI – SMI < 1; Table 1b). Thus, mosaic structure does not appear to have evolved through parallel evolution.

Effect of host selective pressure

We examined type/subtype variation at anchor residues of known linear epitopes (Chisari & Ferrari, 1995) relative to the nine fragments and their variant amino acids. Subtype variation at anchor residues was found in 17/31 epitopes examined. Nine of these were represented in the fragments: preS2(44–53), F/S; HBs(185–194), S/A; Pol(816–824), D/V; Core(1–20), S/T; Core(18–27), V/I; Core(28–47), D/E; Core(50–69), N/T; Core(88–96), T/V; and Core(111–125), L/I.

Effect of functional constraints

The ratio of synonymous change to total change is a measure of functional constraint within a gene or the degree to which a gene is conserved. A functionally important gene will be well conserved with a ratio close to 1, since in the main, change is synonymous. Conversely, genes ordered according to their increasing substitution rate are also ordered for decreasing functional importance. The ratios for change from the original genotype (Table 1b, SG1/G1 column) show a difference in the functional constraint between the different reading frames, genes and positions on the genome, whereas the ratios for change from alternate genotype are fairly constant (Table 1b, SG2/G2 column).

---

Fig. 4. The distribution and nature of variation within individual isolates mapped against the consensus sequence of the parental subtypes as described in the text. (a) Mosaic region (nucleotides 822–1775) of specimen Dm17. For clarity, only variant sites are listed. Over these 954 nucleotides, the fragment would require 82 changes [69 A/D (black, right arrowheads) and 13 unique/quasispecies changes (grey, ×)] to conform with the consensus of the original genotype D of Dm16 but only 20 changes [7 D/A (white, left arrowheads) and the 13 unique/quasispecies changes] to conform with the alternate genotype A. (b) By coding each batch of 20 nucleotides in the same way as before, the complete genome, and not just variant positions, of 13/14 putative recombinants and one non-recombinant, A02, is mapped proportionately to show mosaic areas in context. The black areas mark the position and extent of the mosaic regions within each specimen. Roman numerals I–IX show the start of each of the nine mosaic fragments characterized and listed in Table 1.
Fig. 5. For legend see facing page.
The mosaic areas identified in Fig. 4(b) were divided into nine fragments (I–IX) and each fragment was characterized by a bootstrap tree. In (a) and (b), G1 and G2 refer to the original and alternate genotype of the specimen, respectively. G1 (nt) [or G1 (aa)] refers to the number of nucleotides (or amino acids) within the fragment which differ from the original parent; G2 (nt) [or G2 (aa)] refers to the number of nucleotides (or amino acids) within the fragment which differ from the alternate genotype.

(a) The similarity of each fragment within each specimen to both parental subgroups was determined as a percentage [columns '% G1 (nt)' and '% G2 (nt)']. The difference in similarity between the parental subgroups [mosaic index (MI) = % G2 (nt) – % G1 (nt)] is listed. ns, Not significant.

<table>
<thead>
<tr>
<th>Fragment no.</th>
<th>Specimen</th>
<th>G1</th>
<th>G2</th>
<th>Position</th>
<th>Interval</th>
<th>Length (nt)</th>
<th>% G1 (nt)</th>
<th>% G2 (nt)</th>
<th>Bootstrap (%)</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Dm18</td>
<td>D</td>
<td>A</td>
<td>129–402</td>
<td>3–7</td>
<td>274</td>
<td>12</td>
<td>95.6</td>
<td>6</td>
<td>97.8</td>
</tr>
<tr>
<td>II</td>
<td>Dm16</td>
<td>D</td>
<td>A</td>
<td>495–780</td>
<td>9–13</td>
<td>286</td>
<td>19</td>
<td>93.4</td>
<td>4</td>
<td>98.6</td>
</tr>
<tr>
<td>III</td>
<td>Dm18</td>
<td>D</td>
<td>A</td>
<td>737–1219</td>
<td>13–21</td>
<td>483</td>
<td>43</td>
<td>91.1</td>
<td>8</td>
<td>98.3</td>
</tr>
<tr>
<td>IV</td>
<td>Dm17</td>
<td>D</td>
<td>A</td>
<td>822–1775</td>
<td>15–30</td>
<td>954</td>
<td>82</td>
<td>91.4</td>
<td>20</td>
<td>97.9</td>
</tr>
<tr>
<td>V</td>
<td>Dm18</td>
<td>D</td>
<td>A</td>
<td>1320–2082</td>
<td>23–35</td>
<td>757</td>
<td>40</td>
<td>94.7</td>
<td>13</td>
<td>98.3</td>
</tr>
<tr>
<td>VI</td>
<td>E01</td>
<td>A</td>
<td>D</td>
<td>1576–2339</td>
<td>27–39</td>
<td>764</td>
<td>55</td>
<td>92.8</td>
<td>18</td>
<td>97.6</td>
</tr>
<tr>
<td>VII</td>
<td>D05</td>
<td>A</td>
<td>D</td>
<td>1637–1775</td>
<td>28–33</td>
<td>139</td>
<td>10</td>
<td>92.8</td>
<td>3</td>
<td>97.8</td>
</tr>
<tr>
<td>D06</td>
<td>Dc</td>
<td>A</td>
<td>D</td>
<td>1637–1759</td>
<td>28–33</td>
<td>123</td>
<td>9</td>
<td>92.7</td>
<td>2</td>
<td>98.4</td>
</tr>
<tr>
<td>D07</td>
<td>Dc</td>
<td>A</td>
<td>D</td>
<td>1637–1759</td>
<td>28–33</td>
<td>123</td>
<td>9</td>
<td>92.7</td>
<td>2</td>
<td>98.4</td>
</tr>
<tr>
<td>VIII</td>
<td>B‘01</td>
<td>B</td>
<td>C</td>
<td>1742–2200</td>
<td>30–37</td>
<td>459</td>
<td>46</td>
<td>90.0</td>
<td>14</td>
<td>96.9</td>
</tr>
<tr>
<td>B‘02</td>
<td>Bc</td>
<td>B</td>
<td>C</td>
<td>1742–2200</td>
<td>30–37</td>
<td>459</td>
<td>41</td>
<td>91.1</td>
<td>14</td>
<td>96.9</td>
</tr>
<tr>
<td>B‘03</td>
<td>Bc</td>
<td>B</td>
<td>C</td>
<td>1742–2200</td>
<td>30–37</td>
<td>459</td>
<td>41</td>
<td>91.1</td>
<td>14</td>
<td>96.9</td>
</tr>
<tr>
<td>B‘04</td>
<td>Bc</td>
<td>B</td>
<td>C</td>
<td>1742–2200</td>
<td>30–37</td>
<td>459</td>
<td>41</td>
<td>91.1</td>
<td>14</td>
<td>96.9</td>
</tr>
<tr>
<td>B‘11</td>
<td>Bc</td>
<td>B</td>
<td>C</td>
<td>1742–2200</td>
<td>30–37</td>
<td>459</td>
<td>45</td>
<td>90.2</td>
<td>9</td>
<td>98.0</td>
</tr>
<tr>
<td>B‘12</td>
<td>Bc</td>
<td>B</td>
<td>C</td>
<td>1742–2200</td>
<td>30–37</td>
<td>459</td>
<td>45</td>
<td>90.2</td>
<td>9</td>
<td>98.0</td>
</tr>
<tr>
<td>IX</td>
<td>Dm18</td>
<td>D</td>
<td>A</td>
<td>2191–2337</td>
<td>38–39</td>
<td>147</td>
<td>15</td>
<td>89.8</td>
<td>1</td>
<td>99.3</td>
</tr>
</tbody>
</table>

**Histograms**

The eight frequency histograms which plot the distribution of genetic distances at specific regions of the genome (Fig. 6) give an indication of the rate at which the genotypes are changing in relation to one another. The genetic distances were plotted in three separate series (intra-subgroup [black], inter-subgroup [white] and inter-genotype [grey]) and these formed three distinct but overlapping distributions in most of the histograms. The median and range of the isolate and subtype distributions did not vary greatly across the genome with the notable exception of the subgenomic fragment corresponding to the surface gene where the subtype/isolate distinction is minimal (Fig. 6a, ii). On the other hand, the range and median of the type distribution varied greatly across the genome. Multiple distributions and a wide range of genetic distances were present in the preS1 (Fig. 6a, i) and preS2 (Fig. 6a, vii) subgenomic regions. Examination of the specimens contributing to these multiple distributions confirmed that the genotypes are not evolving at a constant rate over this part of the genome. The genotypic median does not exceed 8% in the surface (5–8%; Fig. 6a, i and preS2 (Fig. 6a, vii) subgenomic regions. Examination of the specimens containing mosaic blocks are ‘wrongly’ grouped in variant subgenomic regions so their presence is detected by inter-subgroup distances in the inter-genotype range (or vice
versa) and appear as a right-hand bias of the white inter-subgroup distribution (or left-hand bias of the grey inter-genotype distribution). This effect is most prominent in the core gene subgenomic fragment (Fig. 6a, v). The right-hand bias of the grey inter-genotype distribution, present in many of the histograms, was formed primarily by the pair-wise-distances between specimens from genotypes A–E and specimens from genotype F. This is evident in all the histograms but is most marked in the X region (Fig. 6a, iv) and the TP domain (Fig. 6a, vi) and forms a discrete distribution in the full-genome histogram (Fig. 6b).

**Discussion**

Type, subtype and isolate categories of HBV were determined from the full-genome bootstrapped tree (Fig. 1). We plotted average evolutionary distances within and between categories at each of 54 points along the genome (Fig. 2) and determined the distribution of evolutionary distances within subgenomic domains of HBV (Fig. 6). The genetic distance graphs showed a distinct pattern of alternating variable and conserved domains within and between HBV subtypes. Genotype B does not have the same topology over different subgenomic regions. Two of the subgroups of B (B' and B") clustered with genotype C in the core gene region (Fig. 5c, d). B and B', and D and Dm did not appear to be true subgroup partitions and the lack of preS variation in their intra-subgroup graphs appeared to confirm this (Fig. 2i and vi). Genotype E did not partition as a separate monophyletic group over all subgenomic regions, often clustering with genotype D (Fig. 5c, d). Genotype F is extremely different from the other genotypes and, although some variable sites are shared between genotypes F and B, none of the three inter-genotypic graphs between the clades of genotype B and genotype F, or the trees, showed a recent relationship between these two genotypes. Only genotypes A (subtypes A and A'), C (subtypes C and Cq') and F (not defined) partition consistently into subtypes.

The histograms showing the distribution of pair-wise distances (Fig. 6) over subgenomic regions indicate that substitution rates are not constant among the lineages in the preS regions and between genotype F and all other genotypes,

**Table 1 (cont.)**

(b) The extent of each fragment in each reading frame (RF column) was determined at the amino acid level and again the similarity of each specimen to both parental subgroups was determined. The latter was used to determine the synonymous change (SG) in each fragment of each specimen in each reading frame and the difference in synonymous change between the parental subgroups [synonymous mosaic index (SMI) = % SG2 - % SG1] is shown. A positive value in column 16, MI - SMI, indicates that the mosaic character is less enhanced when synonymous change alone is considered (and vice versa). The ratio of synonymous change versus total change for each specimen in each reading frame from original genotype and alternate genotype are given in columns SG1/G1 and SG2/G2, respectively.
Footprints of recombination in HBV isolates

Fig. 6. The distribution of pair-wise genetic distances (a) over each of seven named subgenomic genes/domains, the intervals covered by each region are shown in parentheses, or (b) over the complete genome. Successive ranges of genetic distance along the X-axis progress in steps of 0.005, whereas the Y-axis represents the frequency of that range within the distance matrix for the region being studied. The contribution of each of the three distributions is separate: intra-subgroup distances (black); inter-subgroup distances (white); and inter-genotype distances (grey).
A–E, over most of the genome. This is contrary to the findings of Yang et al. (1995), but genotypes E and F and subgroups B, B' and Cq were not represented in their study.

Screening isolates individually, 50 nucleotides at a time, against the consensus sequence of the 11 subgroup partitions identified possible mosaic structure in 14 specimens. Mosaic structure was demonstrated statistically in 11 isolates using bootstrap resampling (Table 1). The nature and distribution of polymorphic sites within the fragments was mapped at both the nucleotide (Fig. 4) and protein levels (not shown). The position of the fragments was related to mutational hot spots and linear epitopes of HBV. Fragments were found in all except the preS1 coding region. The sequence between and including the two DR regions (nucleotides 1592–1840) is represented in all recombinant specimens. A common region, fragment VIII, was involved in all subtype B and B'' specimens. This same fragment was found in four additional subgroup B' isolates from GenBank (X98073–X98076) which were excluded from the main study because their genomes had large insertions and/or deletions.

Sequence variation may be due to chance point mutations or recombination of DNA segments (Stephens, 1985). At the same time, functional constraints, host–virus interactions and selective pressures determine the mutations which are lost and those which are retained and this can lead to gene conservation or parallel evolution in independently arising strains. Frequent recombination and/or mutational hot spots can confuse evolutionary relationships, but in the simplest case, mosaic blocks of sequence identical to an alternate type (or subtype) within a specimen of established type is considered unequivocal evidence (Smith, 1992) that recombination has taken place. Mosaic structure caused by random parallel replacements would be more evident when only synonymous change is considered but this was not found within the identified fragments. Functional constraints limit non-synonymous variation and result in variable subtype/type differences in HBV at the protein level (Mizokami et al., 1997). This was evident when the translation products of the fragments were compared with the consensus proteins of their original genotype. However, when compared with the consensus of their alternate genotype, a constant (quasispecies) difference was observed. This would only be expected when like proteins of the same subtype are compared and further supports the mosaic structure within the fragments. Some of the changes to the linear epitopes observed within the fragments, e.g. preS2–(44–53) (F/S), would be expected to alter the binding characteristics of host HLA antigens whereas many, e.g. V/I, T/V, L/I or S/A, would not be expected to cause a major change. Although the ends of fragments V and VI correspond approximately to the end of the major epitopes of the core gene, many epitopes of P [Pol(61–69)] and all preS1 epitopes are not represented in the fragments. Thus, different MHC backgrounds in human populations in different parts of the world are unlikely to be entirely responsible for the heterogeneity we have observed. Bootstrap re-sampling confirmed the mosaic structure in 11 specimens and recombination, rather than random change, appears to be the dominant mechanism for this structure.

None of the mosaics observed in this study breach the known geographical boundaries of the genotypes, as established by molecular epidemiological studies (Magnius & Norder, 1995). This is a necessary condition of our mechanism of choice since recombination implies a relatively high frequency of superinfection. Superinfection has been reported for HBV but this has always been considered rare and unimportant (Heijtink et al., 1982; Tabor et al., 1977). HBV replication involves template switches during both minus- (Wang & Seeger, 1993; Tavis et al., 1994) and plus- (Will et al., 1987) strand synthesis and intra- or inter-molecular template switching is a common mechanism for homologous recombination (Pathak & Wei-Shau, 1997). However, this is thought to be an unlikely mechanism in HBV since the HBV pregenome replicates only after encapsidation (Ganem, 1991) and, unlike the retroviruses which have a dimeric genome, HBV is thought to package a single RNA pregenome. Nevertheless, Raimondo et al. (1988) have reported the presence of replicative intermediates sensitive to DNase I digestion in the liver of a patient and suggested that unencapsidated molecular forms of HBV DNA can accumulate in chronic HBV carriers. Chronic carriage has a complex pathology progressing from replicative to non-replicative disease and often resulting in virus integration and/or hepatocellular carcinoma. This progression is not always linear and non-replicative carriers can re-activate and return to an earlier stage of the disease (Dusheiko et al., 1985). Further studies are needed to confirm that template switching is impossible at all stages of disease progression.

However, replication is not the only stage at which HBV has an opportunity to recombine. Initiation of infection and hepadnavirus replication involve conversion of genomic relaxed circular DNA (RC DNA) into covalently closed circular DNA (cccDNA) within the nucleus of the infected cell in a manner not fully understood, but which is thought to utilize cellular DNA-modifying enzymes (Köck & Schlicht, 1993). It has also been speculated that cellular enzymes could be responsible for changes to the episome, making it a better substrate for integration (Schirmacher et al., 1995). Increasingly, it is being suggested that the processes of mutation and integration are linked in some instances. Identical mutations have been reported in free and integrated WHV (Kew et al., 1993) and HBV (Georgi-Geisberger et al., 1992) from a single patient. A recent study used an in vitro duck hepatitis B virus (DHBV) system to map the plus- and minus-strand cleavage sites of topoisomerase I (top I) which is considered a likely candidate for both conversion of RC DNA to cccDNA and for integration of episomes into the host DNA (Pourquier et al., 1999). This model showed that top I was capable of converting RC DNA to cccDNA in vitro and that this was achieved via non-homologous recombination. An earlier study which
defined illegitimate replication of linear DHBV DNA in primary hepatocyte cultures also found that the 3' end of the minus-strand efficiently participates in intra- and intermolecular non-homologous recombination to produce monomeric cccDNA or oligomeric forms in which monomers are joined near the ends in random orientation (Yang & Summers, 1995). Although these oligomeric forms have not been found in viral particles nor shown to take part in illegitimate replication, they would have a similar mosaic structure to that which we have observed.

Recombination has been documented previously in HBV. A 196 bp region in the preCore/Core was found to enhance recombination in vitro in the presence of extracts from actively dividing cells (Hino et al., 1991). Georgi-Geisberger et al. (1992) found evidence of homologous recombination, very similar to what we have found at the population level, when studying integrated and episomal HBV from a single patient. Bollyky et al. (1996) used bootstrapped maximum-likelihood trees and a randomization test to demonstrate mosaic structure statistically in 2/25 complete genome sequences and concluded that the heterogeneity which they observed was the result of recombination between viruses of different genomic and antigenic types.

Further study is required to produce direct evidence for recombination in HBV and to clarify the role of recombination in the heterogeneity of HBV. The mosaic structure which we and others have observed affects entire clades and alters the phylogeny of HBV over extensive subgenomic fragments. Many enigmas of HBV persist, including geographical differences in the pathology and evolution of HBV and the variety of host responses which infection with HBV takes in different individuals (Foster & Thomas, 1993), and recombination could be the mechanism responsible.

This investigation was supported by a research grant from the Poliomyelitis Research Foundation. I would like to thank Drs Lynn Morris, Karin Kidd-Ljunggren and Alistair Kidd for reading this manuscript and for their valuable advice.

References


Received 23 August 1999; Accepted 18 October 1999