Journal of General Virology (2000), **81**, 379–392. Printed in Great Britain

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 10^4-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

Author for correspondence: Sheila Bowyer.
Fax +27 11 882 0596. e-mail sheila@niv.ac.za

0001-6668

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^- subgroup of genotype C (Norder *et al.*, 1994). Two subgroups of genotype A, subgroups A and A′, were found in approximately equal amounts in an urban population from South Africa together with 10% of genotype D (Bowyer *et al.*, 1997).

The initial purpose of this study was to examine the relationships between full genomes of HBV to determine whether further subgroups of the major genotypes exist. In particular, we were interested in the relationship between genotypes D and E, since the existence of genotype E as a unique monophyletic group has been questioned (Kidd-Ljunggren *et al.*, 1995). This phylogenetic analysis of the X gene also reported that, in this region, genotypes B and F branched together. Also, incongruence between trees reconstructed from different parts of the genome of HBV has been documented and we (Bowyer *et al.*, 1997, 1998) and others (Georgi-Geisberger *et al.*, 1992; Bollyky *et al.*, 1996; Mizokami *et al.*, 1997) have discussed the possibility that this was caused by recombination. To address these questions, we first clarified the partitions, and sub-partitions, of HBV and used this information to derive a screening assay to identify mosaics within HBV isolates. Anomalous fragments were then characterized with particular reference to their position in conserved and variant regions of the HBV genome.

Methods

■ **Patients.** Sixty-five HBV full-genome nucleotide sequences, representing each of the genotypes A–F, were obtained from GenBank (see legend to Fig. 1 for accession numbers). The specimens came from 59 patients with acute, chronic or fulminant disease outcomes, three infected chimpanzees and one cell line (Dm14). Two of the chimpanzee specimens, A07 and E03, were sequenced from full-length clones and one, C18, was sequenced from nested PCR fragments. Two patients are represented by two specimens taken at different stages of infection (C05/C06 and C16/C17) and all four of these were sequenced from full-length clones. Only two direct submissions to GenBank, A08 and D12, do not specify the source of the sequence data and a further seven direct submissions (D05–D07, B′01, B05, B06, C04) only specify that the sequence represents a complete genome but do not specify the methodology employed. In 25 (A01, A03–A06, A09, B′02, B07–B09, B′10, B″11, B″12, C03, C07–C12, D01, D08, D11, Dm13, F1) of the remaining 48 cases, the source of the DNA for sequencing is specified as a full-length clone; in 12 (A02, A′01, A′02, B′03, B′04, D03, D04, D10, Cq^-01, Cq^-02, F02, F03), the sequence was pieced together from overlapping PCR fragments; and in 8 (D02, D09, Dm15–Dm18, E01, E02) the information was gleaned from overlapping clones. Three sequences (C13–C15) were generated using nested PCR.

■ **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 DRAWGRAM) 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).

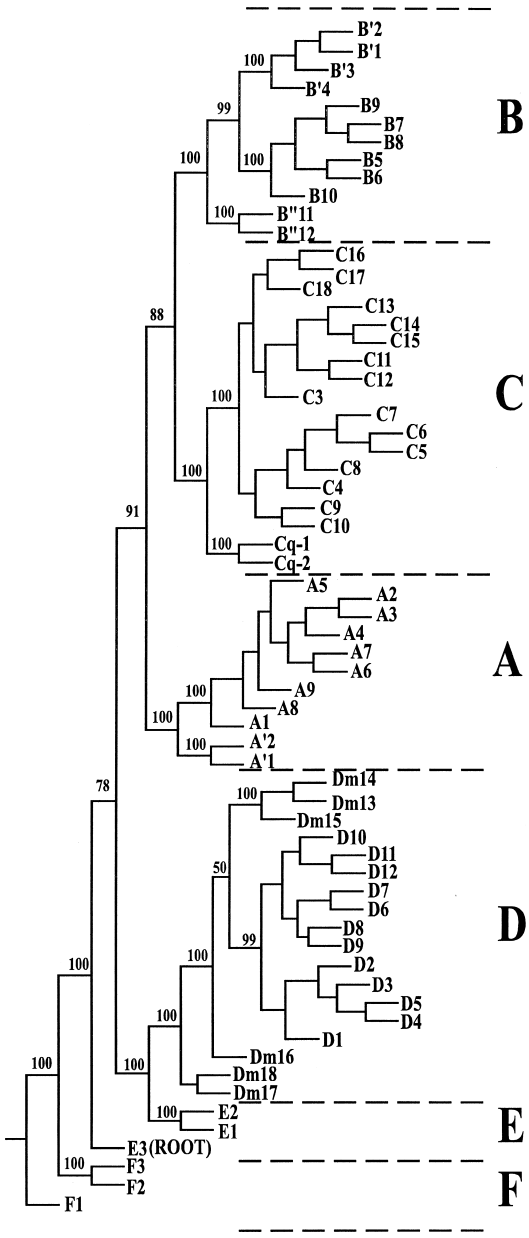

Fig. 1. Full-genome FITCH bootstrapped consensus tree (100 sets) showing the conventional segregation of HBV specimens into the six genotypic groups (A–F). Bootstrap values for the most robust groupings are shown. GenBank accession numbers with corresponding codes used in the text are as follows: A01, X51970; A02, E00010; A03, X02763; A04, X70185; A05, Z72478; A06, Z35717; A07, J02201; A08, L13994; A09, S50225; A′01, M74498 (full sequence in Kremsdorf *et al.*, 1996); A′02, M57663; D01, Z35716; D02, X72702; D03, X97848; D04, X97849; D05, X80925; D06, X80926; D07, X80924; D08, M32138; D09, X59795; D10, L27106; D11, X02496; D12, Y07587; Dm13, J02203; Dm14, U95551; Dm15, X65257; Dm16, X65258; Dm17, X65259; Dm18, X68292; E01, X75664; E02, X75657; F01, X69798; F02, X75663; F03, X75658; E03 (also the ROOT), D00220; B′01, X98077; B′02, D00330; B′03, X97851; B′04, X97850; B05, D50522; B06, D50521; B07, D23677; B08, D23678; B09, D23679; B10, D00329; B″11, D00331; B″12, M54923; Cq⁻01, X75656; Cq⁻02, X75665; C03, D00630; C04, D23684; C05, D23681; C06, D23680; C07, X01587; C08, M38454; C09, V00867; C10, X04615; C11, M12906; C12, D12980; C13, D16667; C14, D16666; C15, D50489; C16, D23682; C17, D23683; C18, L08805.

■ **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

Although all trees are essentially un-rooted, D00220 was chosen as the outgroup and/or seed when the program requested it.

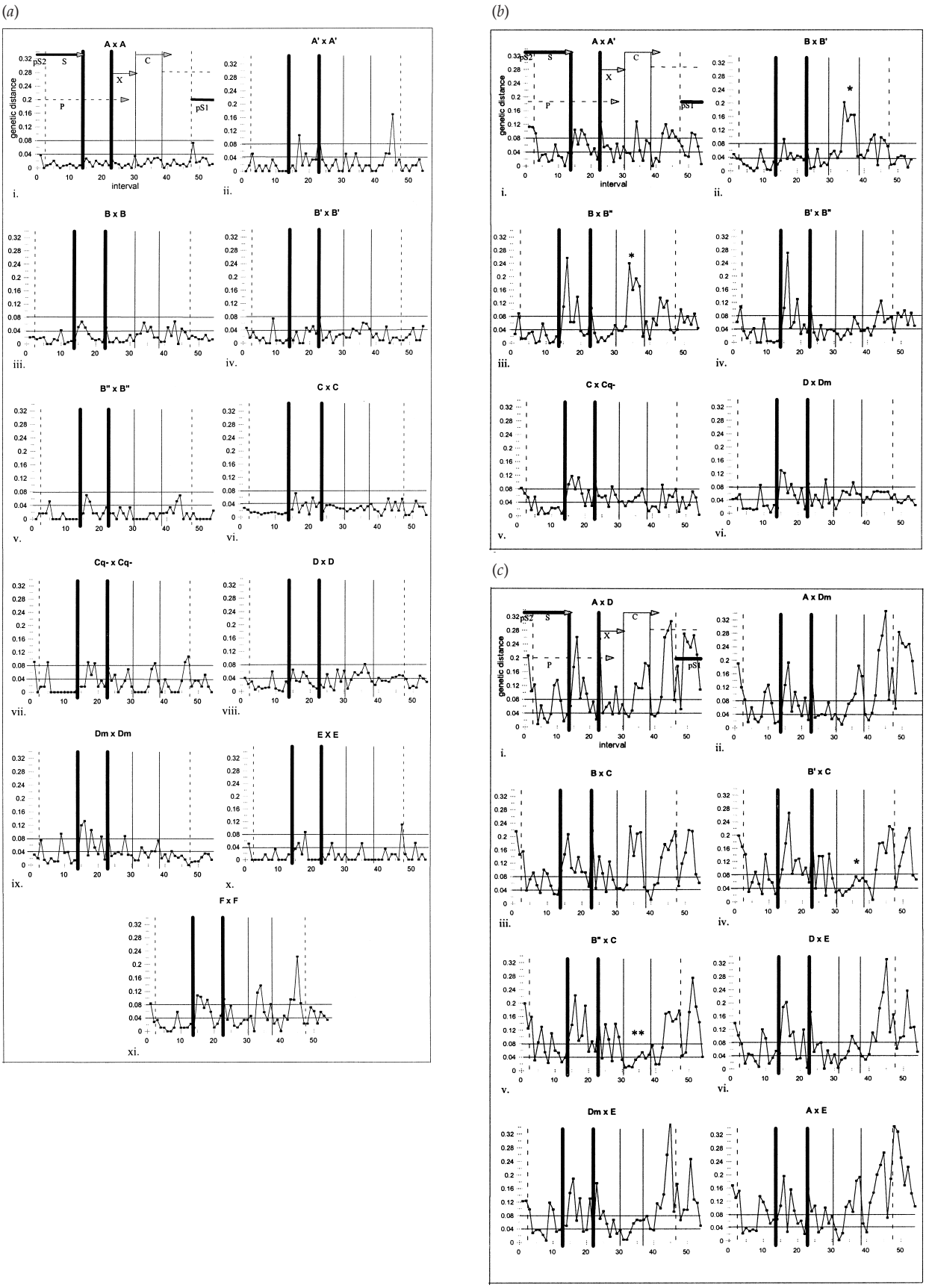

Fig. 2. For legend see facing page.

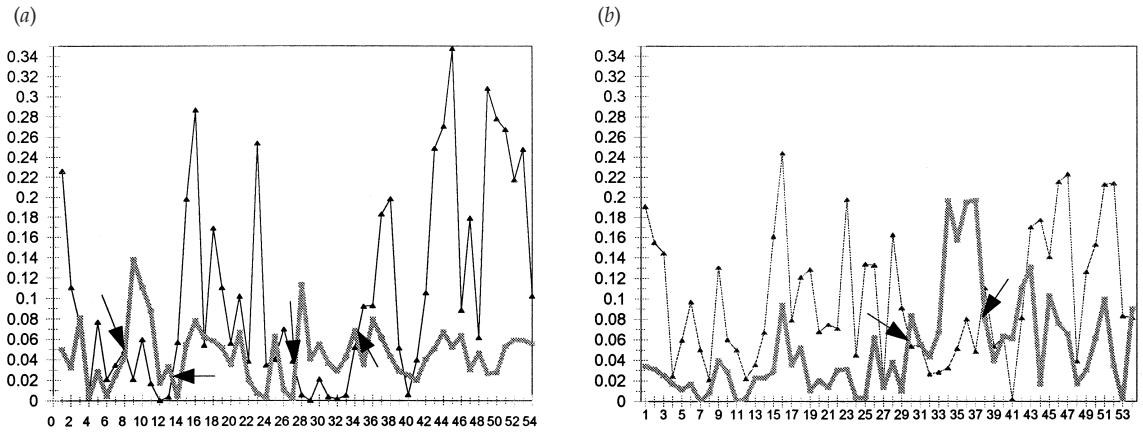

Fig. 3. Graphs of pair-wise distances of putative recombinants from parental subgroups to illustrate breakpoint determination. Average pair-wise distances over each of the 54 intervals between (*a*) isolate Dm16 and (i) subgroup A (black line with triangular markers) and (ii) subgroup D (thick grey line with square markers) and (*b*) isolate B′01 and (i) subgroup C (black line with triangular markers) and (ii) subgroup B (thick grey line with square markers). The graphs in (*a*) intersect within intervals 8, 13, 27 and 34 indicating two separate fragments in specimen Dm16. The graphs in (*b*) intersect within intervals 30 and 37 (see black arrows).

were designated Dm (for mutant group). We treated the three isolates of genotype F as a single clade despite the fact that the average difference between them (calculated from the full-genome matrix) was 4·7%, which is outside the subgroup range. Thus, we defined a total of 11 subgroups (A, A′, B, B′, B″, C, Cq^{-}, D, Dm, E and F) of the 65 isolates within the six conventional HBV genotypes (A–F).

Average genetic distance graphs

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. 2*a*, i). The effect of treating genotype F (within which specimens varied by 4·7%) as a single subgroup is evident in Fig. 2(*a*, xi). The pattern of this genetic distance graph is more reminiscent of the inter-genotypic graphs (Fig. 2*b*) 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. 2*b*). 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. 2*b*, ii) and D × Dm (Fig. 2*b*, vi) graphs. The TP domain of the P gene (intervals 38–48) peaks above 8% except in the D × Dm graph (Fig. 2*b*, 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. 2*b*, 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. 2*b*, iv) to $> 18\%$ between B and B′ (Fig. 2*b*, ii) and B and B″ (Fig. 2*b*, iii).

Inter-genotype. After careful examination of the 49 inter-genotypic graphs, we selected eight typical, or anomalous, graphs to illustrate the features of this series (Fig. 2*c*). 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

Fig. 2. Average genetic distance plotted against interval number. Vertical lines demarcate the beginning and end of genes preS, S, P, X and C of HBV, in the three reading frames (see *a*, i) and horizontal lines highlight the 4 and 8% group category cut-offs. Zero is the *Eco*RI site and each interval consists of 60 nucleotides. (*a*) Average genetic distance within subgroups. (*b*) Average genetic distance between subgroups of the same genotype. (*c*) Average genetic distance between selected subgroups of different genotypes. The asterisks in (*b*) draw attention to the anomalous subregions within the B′ × B and B″ × B inter-subgroup graphs and in (*c*) the corresponding regions within the B′ × C and B″ × C inter-genotype graphs.

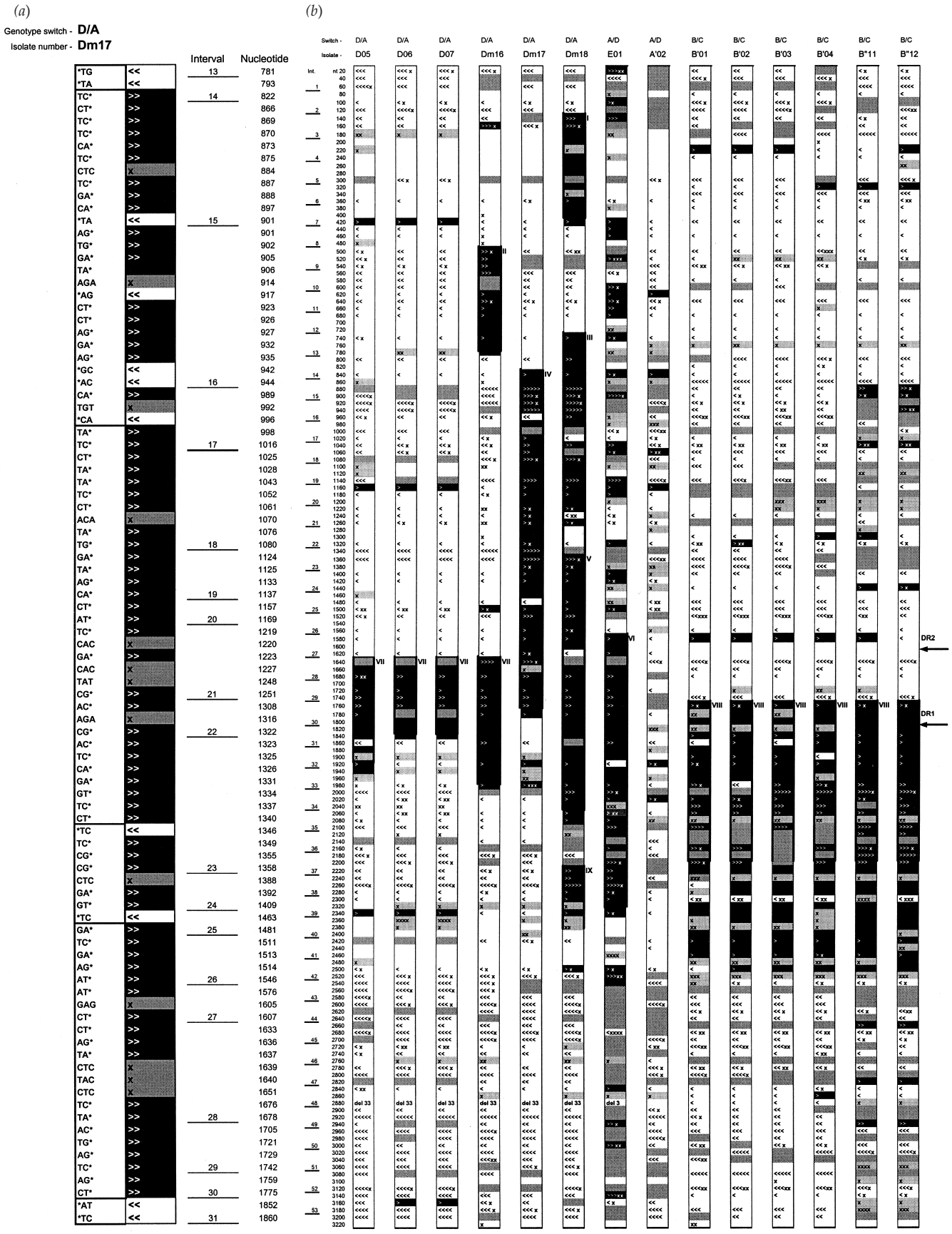

Fig. 4. For legend see facing page.

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. 2*b*, ii and iii) are dramatically reversed in their corresponding inter-genotype graphs with genotype C [see peaks marked * for B′ × C (Fig. 2*c*, iv), and ** for B″ × C (Fig. 2*c*, 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. 2*c*, 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. 2*c*, 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. 4*a*) and amino acid (not shown) of each block was precisely identified. Nine mosaic fragments were identified within the 14 specimens (Fig. 4*b*).

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 1*a*, bootstrap column). Dm16 (Figs 3*a*, 4*b* and 5*a*), Dm17 (Figs 4*a*, *b* and 5*b*) and Dm18 (Figs 4*b* and 5*b*, *c*) have one, or more, extensive and significant D/A segments. A long B/C fragment in the core gene region (Figs 3*b*, 4*b* and 5*c*, *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. 5*c*, *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 1*a*] is greater than when synonymous change alone is considered [synonymous mosaic index (SMI); Table 1*b*] with one exception, fragment V in RF1 (see MI − SMI < 1; Table 1*b*). 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 1*b*, 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 1*b*, 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, A′02, 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.

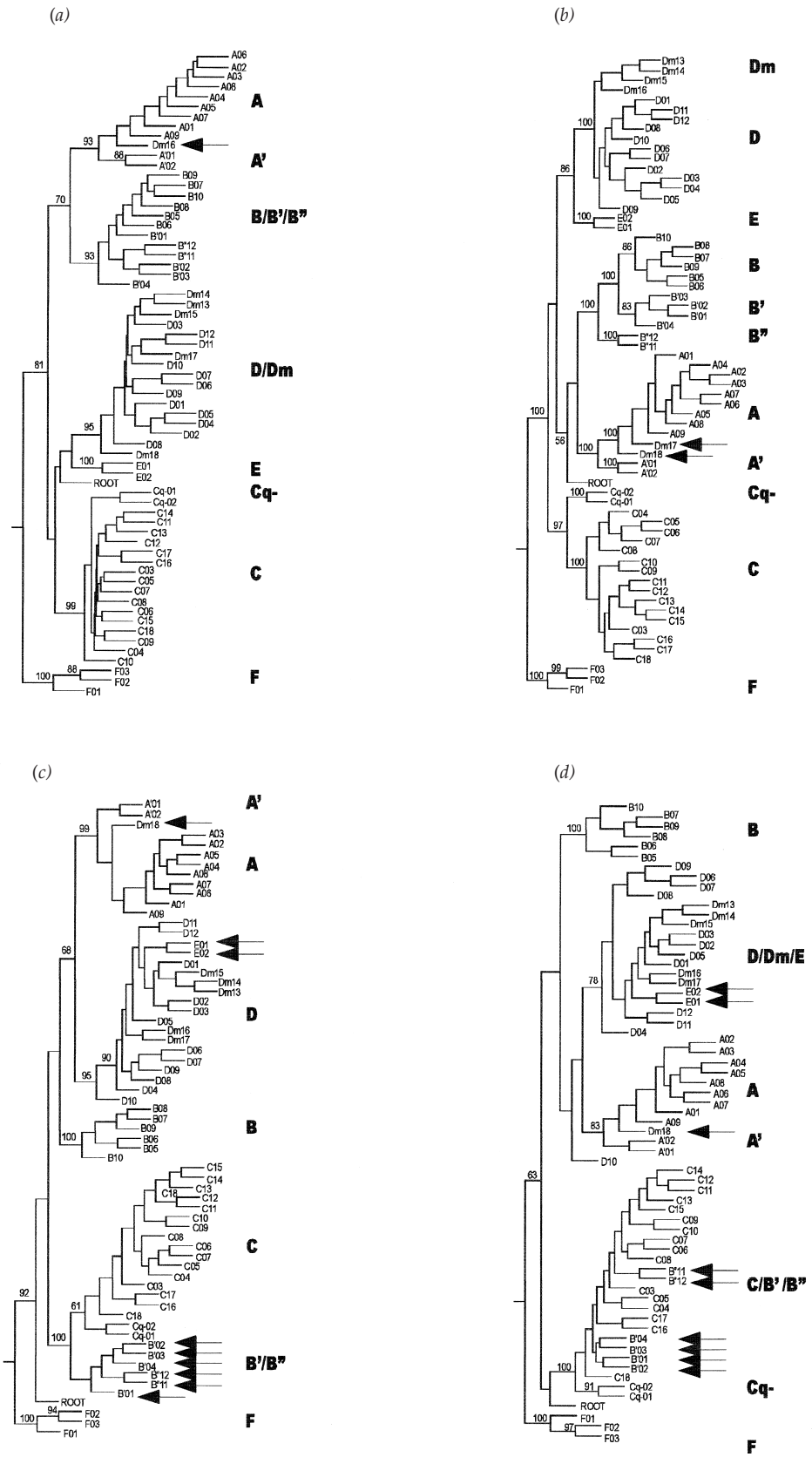

Fig. 5. For legend see facing page.

Table 1. Similarity of fragments derived from mosaic areas to parental subgroups

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)', respectively]. The difference in similarity between the parental subgroups [mosaic index (MI) = % G2 (nt) − % G1 (nt)] is listed. NS, Not significant.

Fragment no.	Specimen	G1	G2	Position		Interval		Length	G1 (nt)	% G1 (nt)	G2 (nt)	% G2 (nt)	Bootstrap (%)	MI
				Start	Stop	Start	Stop							
I	Dm18	D	A	129	402	3	7	274	12	95·6	6	97·8	NS	2·19
II	Dm16	D	A	495	780	9	13	286	19	93·4	4	98·6	93	5·24
III	Dm18	D	A	737	1219	13	21	483	43	91·1	8	98·3	100	7·25
IV	Dm17	D	A	822	1775	15	30	954	82	91·4	20	97·9	100	6·50
V	Dm18	D	A	1326	2082	23	35	757	40	94·7	13	98·3	92	3·57
VI	E01	A	D	1576	2339	27	39	764	55	92·8	18	97·6	95	4·84
VII	D05	D	A	1637	1775	28	33	139	10	92·8	3	97·8	NS	5·04
	D06	D	A	1637	1759			123	9	92·7	2	98·4	NS	5·69
	D07	D	A	1637	1759			123	9	92·7	2	98·4	NS	5·69
VIII	B′01	B	C	1742	2200	30	37	459	46	90·0	14	96·9	100	6·97
	B′02	B	C	1742	2200			459	41	91·1	14	96·9	100	5·88
	B′03	B	C	1742	2200			459	36	92·2	14	96·9	100	4·79
	B′04	B	C	1742	2200			459	47	89·8	16	96·5	100	6·75
	B″11	B	C	1742	2200			459	45	90·2	9	98·0	100	7·84
	B″12	B	C	1742	2200			459	45	90·2	6	98·7	100	8·50
IX	Dm18	A	D	2191	2337	38	39	147	15	89·8	1	99·3	78	9·52

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. 6*a*, 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. 6*a*, i) and preS2 (Fig. 6*a*, 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. 6*a*, ii) and X (7·9%; Fig. 6*a*, iv) gene regions but is 8·9–14·9% over the other subgenomic fragments.

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

Fig. 5. Bootstrap re-sampling trees of 4/9 mosaic fragments. (*a*) Fragment II, intervals 9–13, present in Dm16 only. Dm16 clusters with genotype A with a bootstrap value of 93%. Note that subgroups of genotypes B and D are not present over this subgenomic fragment whereas the subgroups of genotype A (A and A′) and genotype C (C and Cq⁻) are. Genotype E clusters as a separate group with a bootstrap value of 100%.(*b*) Fragment IV, intervals 15–30, is present in Dm17 only but includes parts of fragments III and V of Dm18. Both Dm17 and Dm18 cluster with genotype A with a bootstrap value of 100%. Note the subgroups of genotypes A, B and C are present over this subgenomic fragment whereas genotype D shows Dm13–Dm16 as outliers of the main cluster. Genotype E clusters as a separate group with a bootstrap value of 100%. (*c*) Fragment VI, intervals 27–39, present in E01 and E02. Genotype E clusters with genotype D with a bootstrap value of 95%. The subgroups of genotypes A and C are visible over this subgenomic fragment but two of the three subgroups of genotype B cluster with genotype C. (*d*) Fragment VIII, intervals 30–37, present in B′01–B′04 and B″11–B″12 and including part of fragment V of Dm18. All the genotype B specimens cluster with genotype C with a bootstrap value of 100% and the subfragment of Dm18 clusters with genotype A with a bootstrap value of 83%. Note the subgroups of genotypes A and C are present over this subgenomic fragment and that genotype E clusters with genotype D with a bootstrap value of 78%.

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.

Fragment no.	Specimen	G1	G2	Position Start	Position Stop	RF	Gene	Length	G1 (aa)	G2 (aa)	SG1	SG2	% SG1	% SG2	SMI	MI − SMI	SG1/G1	SG2/G2
I	Dm18	D	A	43	134	1	S	92	6	2	6	4	97·8	98·54	0·73	1·46	0·50	0·33
				43	134	3	P	92	3	2	9	4	96.7	98·54	1·82	0·36	0·75	0·33
II	Dm16	D	A	165	260	1	S	96	9	2	10	2	96·5	99·30	2·80	2·45	0·53	0·11
				165	260	3	P	96	7	3	12	1	95·8	99·65	3·85	1·40	0·63	0·05
III	Dm18	D	A	246	278	1	S	33	3	0	40	8	91·7	98·34	6·63	0·62	0·93	0·19
				246	407	3	P	162	13	0	30	8	93·8	98·34	4·55	2·69	0·70	0·19
IV	Dm17	D	A	459	592	2	X	134	7	1	75	19	92·1	98·01	5·87	0·63	0·91	0·23
				274	540	3	P	267	21	2	61	18	93·6	98·11	4·51	1·99	0·74	0·22
V	Dm18	D	A	606	694	1	C	89	0	1	40	12	94·7	98·41	3·70	−0·13	1·00	0·30
				459	612	2	X	154	7	1	33	12	95·6	98·41	2·77	0·79	0·83	0·30
VI	E01	A	D	606	780	1	C	175	7	3	48	15	93·7	98·04	4·32	0·52	0·87	0·27
				525	612	2	X	88	3	1	52	17	93·2	97·77	4·58	0·26	0·95	0·31
				525	540	3	P	16	0	0	55	18	92·8	97·64	4·84	0·00	1·00	0·33
VII	D05	D	A	547	592	2	X	46	6	1	4	2	97·1	98·56	1·44	3·60	0·40	0·20
	D06	D	A	547	592	2	X	46	5	1	4	1	96·7	99·19	2·44	3·25	0·44	0·11
	D07	D	A	547	592	2	X	46	5	1	4	1	96·7	99·19	2·44	3·25	0·44	0·11
VIII	B′01	B	C	606	734	1	C	129	9	2	37	12	91·9	97·39	5·45	1·53	0·80	0·26
	B′02	B	C	606	734	1	C	129	10	1	31	13	93·2	97·17	3·92	1·96	0·76	0·32
	B′03	B	C	606	734	1	C	129	10	2	26	12	94·3	97·39	3·05	1·74	0·72	0·33
	B′04	B	C	606	734	1	C	129	10	2	37	14	91·9	96·95	5·01	1·74	0·79	0·30
	B″11	B	C	606	734	1	C	129	11	0	34	9	92·6	98·04	5·45	2·40	0·76	0·20
	B″12	B	C	606	734	1	C	129	11	0	34	6	92·6	98·69	6·10	2·40	0·76	0·13
IX	Dm18	A	D	730	779	1	C	50	5	0	10	1	93·2	99·32	6·12	3·40	0·67	0·07

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. 6*a*, 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. 6*a*, iv) and the TP domain (Fig. 6*a*, vi) and forms a discrete distribution in the full-genome histogram (Fig. 6*b*).

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. 5*c*, *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. 2*b*, ii and vi). Genotype E did not partition as a separate monophyletic group over all subgenomic regions, often clustering with genotype D (Fig. 5*c*, *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,

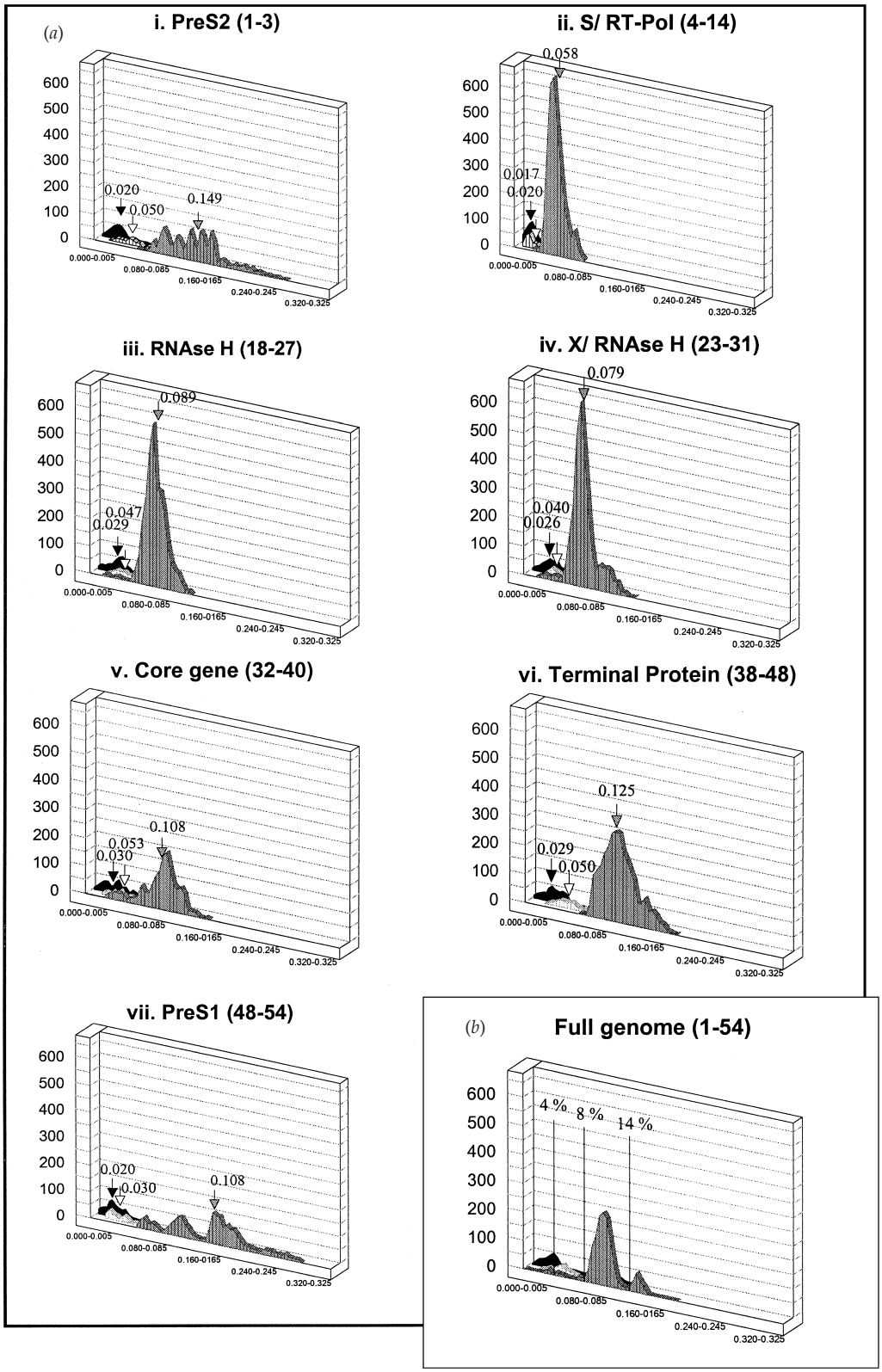

Fig. 6. The distribution of pair-wise genetic distances (*a*) over each of seven named subgenomic genes/domains, the intervals covered by each region are shown in parentheses, or (*b*) over the complete genome. Successive ranges of genetic distance along the X-axis progress in steps of 0·005, whereas 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

Beasley, R. P. & Hwang, L.-Y. (1984). Epidemiology of hepatocellular carcinoma. In *Viral Hepatitis and Liver Disease*, pp. 209–224. Edited by G. N. Vyas, J. L. Dienstag & J. H. Hoofnagle. Orlando, FL: Grune & Stratton.

Bollyky, P. L., Rambaut, A., Harvey, P. H. & Holmes, E. C. (1996). Recombination between sequences of hepatitis B virus from different genotypes. *Journal of Molecular Evolution* **42**, 97–102.

Botha, J. F., Dusheiko, G. M., Ritchie, M. J. J., Mouton, H. W. K. & Kew, M. C. (1984). Hepatitis B virus carrier state in black children in Ovamboland: role of perinatal and horizontal infection. *Lancet* **1**, 1210–1212.

Bowyer, S. M. (1998). Epidemiology and molecular characterization of HBV strains from the northern provinces of Namibia. In *National Institute for Virology, Annual Report* 1997, pp. 38–39. Johannesburg: National Institute for Virology.

Bowyer, S. M., van Staden, L., Kew, M. C. & Sim, J. G. M. (1997). A unique segment of the hepatitis B virus group A genotype identified in isolates from South Africa. *Journal of General Virology* **78**, 1719–1729.

Boyer, J. C., Bebenek, K. & Kunkel, T. A. (1992). Unequal HIV-1 reverse transcriptase error rates with RNA and DNA templates. *Proceedings of the National Academy of Sciences, USA* **89**, 6919–6923.

Chisari, F. V. & Ferrari, C. (1995). Hepatitis B virus immunopathogenesis. *Annual Reviews in Immunology* **13**, 29–60.

Dusheiko, G. M., Bowyer, S. M., Paterson, A., Song, E., Dibisceglie, A. & Kew, M. C. (1985). Clinical and serological events accompanying changes in hepatitis B viral replication: case reports. *Liver* **5**, 77–83.

Edmunds, W. J., Medley, G. F., Nokes, D. J., O'Callaghan, C. J., Whittle, H. C. & Hall, A. J. (1996). Epidemiological patterns of hepatitis B virus (HBV) in highly endemic areas. *Epidemiology and Infection* **117**, 313–325.

Foster, G. R. & Thomas, H. C. (1993). Recent advances in the molecular biology of hepatitis B virus: mutant virus and the host response. *Gut* **34**, 1–3.

Ganem, D. (1991). Assembly of hepadnaviral virions and subviral particles. In *Hepadnaviruses, Current Topics in Microbiology and Immunology*, vol. 168, pp. 61–83. Edited by W. S. Mason & C. Seeger. New York: Springer–Verlag.

Georgi-Geisberger, P., Berns, H., Loncarevic, I. F., Yu, Z.-Y., Tang, Z.-Y., Zentgraf, H. & Schröder, C. H. (1992). Mutations on free and integrated hepatitis B virus DNA in a hepatocellular carcinoma: footprints of homologous recombination. *Oncology* **49**, 386–395.

Girones, R. & Miller, R. L. (1989). Mutation rate of the hepadnavirus genome. *Virology* **170**, 595–597.

Günther, S., Fischer, L., Puli, I., Sterneck, M. & Will, H. (1999). Naturally occurring variants of hepatitis B. In *Advances in Viral Research*, pp. 25–137. Edited by K. Maramorosch, F. A. Murphy & A. J. Shatkin. San Diego: Academic Press.

Heijtink, R. A., van Hattum, J., Schalm, S. W. & Masurel, N. (1982). Co-occurrence of HBsAg and anti-HBs: two consecutive infections or a sign of advanced chronic liver disease? *Journal of Medical Virology* **10**, 83–90.

Hino, O., Tabata, S. & Hotta, Y. (1991). Evidence for increased *in vitro* recombination with insertion of human hepatitis B virus DNA. *Proceedings of the National Academy of Sciences, USA* **88**, 9248–9252.

Kew, M. C., Miller, R. H., Chen, H.-S., Tennant, B. C. & Purcell, R. H. (1993). Mutant woodchuck hepatitis virus genomes from virions resemble rearranged hepadnaviral integrants in hepatocellular carcinoma. *Proceedings of the National Academy of Sciences, USA* **90**, 10211–10215.

Kidd-Ljunggren, K., Couroucé, A.-M., Öberg, M. & Kidd, A. H. (1994). Genetic conservation within subtypes in the hepatitis B virus pre-S2 region. *Journal of General Virology* **75**, 1485–1490.

Kidd-Ljunggren, K., Öberg, M. & Kidd, A. H. (1995). The hepatitis B virus X gene: analysis of functional domain variation and gene phylogeny using multiple sequences. *Journal of General Virology* **76**, 2119–2130.

Köck, J. & Schlicht, H.-J. (1993). Analysis of the earliest steps of hepadnavirus replication: genome repair after infectious entry into hepatocytes does not depend on viral polymerase activity. *Journal of Virology* **67**, 4867–4874.

Kremsdorf, D., Garreau, F., Capel, F., Petit, M.-A. & Bréchot, C. (1996). *In vivo* selection of a hepatitis B virus mutant with abnormal viral protein expression. *Journal of General Virology* **77**, 929–939.

Magnius, L. O. & Norder, H. (1995). Subtypes, genotypes and molecular epidemiology of the hepatitis B virus reflected by sequence variability of the S-gene. *Intervirology* **38**, 24–34.

Miller, R. H. (1988). Close evolutionary relatedness of the hepatitis B virus and murine leukemia virus polymerase gene sequences. *Virology* **164**, 147–155.

Mizokami, M., Orito, E., Ohba, K.-I., Ikeo, K., Lau, J. Y. N. & Gojobori, T. (1997). Constrained evolution with respect to gene overlap of hepatitis B virus. *Journal of Molecular Evolution* **44** (Suppl. 1), 83–90.

Norder, H., Hammas, B., Lofdahl, S., Couroucé, A.-M. & Magnius, L. O. (1992). Comparison of the amino acid sequences of nine different serotypes of hepatitis B surface antigen and genomic classification of the corresponding hepatitis B virus strains. *Journal of General Virology* **73**, 1201–1208.

Norder, H., Hammas, B., Lee, S.-D., Bile, K., Couroucé, A.-M., Mushahwar, I. K. & Magnius, L. O. (1993). Genetic relatedness of hepatitis B viral strains of diverse geographical origin and natural variations in the primary structure of the surface antigen. *Journal of General Virology* **74**, 1341–1348.

Norder, H., Couroucé, A.-M. & Magnius, L. O. (1994). Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology* **198**, 489–503.

Ohba, K.-I., Mizokami, M., Ohno, T., Suzuki, K., Orito, E., Lau, J. Y. N., Ina, Y., Ikeo, K. & Gojobori, T. (1995). Relationships between serotypes and genotypes of hepatitis B virus: genetic classification of HBV by use of surface genes. *Virus Research* **39**, 25–34.

Okamoto, H., Imai, M., Kametani, M., Nakamura, T. & Mayumi, M. (1987). Genomic heterogeneity of hepatitis B virus in a 54-year-old woman who contracted the infection through materno-fetal transmission. *Japanese Journal of Experimental Medicine* **57**, 231–236.

Okamoto, H., Tsuda, F., Sakugawa, H., Sastrosoewignjo, R. I., Imai, M., Miyakawa, Y. & Mayumi, M. (1988). Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *Journal of General Virology* **69**, 2575–2583.

Pathak, V. K. & Wei-Shau, H. (1997). 'Might as well jump!' Template switching by retroviral reverse transcriptase, defective genome formation, and recombination. *Seminars in Virology* **8**, 141–150.

Pourquier, P., Jensen, A. D., Gong, S. S., Pommier, Y. & Rogler, C. E. (1999). Human DNA topoisomerase I-mediated cleavage and recombination of duck hepatitis B virus DNA *in vitro*. *Nucleic Acids Research* **27**, 1919–1925.

Raimondo, G., Burk, R. D., Lieberman, H. M., Muschel, J., Hadziyannis, S. J., Will, H., Kew, M. C., Dusheiko, G. M. & Shafritz, D. A. (1988). Interrupted replication of hepatitis B virus in liver tissue of HBsAg carriers with hepatocellular carcinoma. *Virology* **166**, 103–112.

Schirmacher, P., Wang, W., Stahnke, G., Will, H. & Rogler, C. E. (1995). Sequences and structures at hepadnaviral integration: recombination sites implicate topoisomerase I in hepadnaviral DNA rearrangements and integration. *Journal of Hepatology* **22**, 21–23.

Siepel, A. C. & Korber, B. T. (1995). Scanning the database for recombinant HIV-1 genomes. In *Human Retroviruses and AIDS*, pp. III-35–III-60. Edited by G. Myers, B. H. Mellors, J. W. Hahn, L. E. Henderson, B. Korber, K.-T. Jeang, F. E. McCutchen & G. N. Pavlakis. New Mexico: Los Alamos National Laboratory.

Simmonds, P., Mellor, J., Sakuldamrongpanich, T., Nuchaprayoon, C., Tanprasert, S., Holmes, E. C. & Smith, D. B. (1996). Evolutionary analysis of variants of hepatitis C virus found in South-East Asia: comparison with classifications based upon sequence similarity. *Journal of General Virology* **77**, 3013–3024.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126–129.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular & Biological Evolution* **2**, 539–556.

Tabor, E., Gerety, R. J., Smallwood, L. A. & Barker, L. F. (1977). Coincident hepatitis B surface antigen and antibodies of different subtypes in human serum. *Journal of Immunology* **118**, 369–370.

Tavis, J. E., Perri, S. & Ganem, D. (1994). Hepadnavirus reverse transcription initiates within the stem-loop of RNA packaging signal and employs a novel strand transfer. *Journal of Virology* **68**, 3536–3543.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Uy, A., Wunderlich, G., Olsen, D. B., Heermann, K.-H., Gerlich, W. H. & Thomssen, R. (1992). Genomic variability in the preS1 region and determination of routes of transmission of hepatitis B virus. *Journal of General Virology* **73**, 3005–3009.

Vardas, E., Mathai, M., Blaauw, D., McAnerney, J., Coppin, A. & Sim, J. (1999). Preimmunization epidemiology of hepatitis B virus infection in South African children. *Journal of Medical Virology* **58**, 111–115.

Wang, G.-H. & Seeger, C. (1993). Novel mechanism for reverse transcription in hepatitis B viruses. *Journal of Virology* **67**, 6507–6512.

Will, H., Reiser, W., Weimer, T., Pfaff, E., Büscher, M., Sprengel, R., Cattaneo, R. & Schaller, R. (1987). Replication strategy of human hepatitis B virus. *Journal of Virology* **61**, 904–911.

Yang, Z. & Summer, J. (1995). Illegitimate replication of linear hepadnavirus DNA through nonhomologous recombination. *Journal of Virology* **69**, 4029–4036.

Yang, Z., Lauder, I. J. & Lin, H. J. (1995). Molecular evolution of the hepatitis B virus genome. *Journal of Molecular Evolution* **41**, 587–596.

Received 23 August 1999; Accepted 18 October 1999