J Mol Evol (2002) 54:807–814
DOI: 10.1007/s00239-001-0084-z

JOURNAL OF MOLECULAR EVOLUTION



# A Revised Evolutionary History of Hepatitis B Virus (HBV)

**Mario Ali Fares,[1] Edward C. Holmes[2]**

[1] Institut "Cavanilles" de Biodiversitat i Biologia Evolutiva and Department de Genética, Universitat de Valéncia, Valencia, Spain
[2] Department of Zoology, University of Oxford, South Parks Road, Oxford OXI 3PS, UK

Received: 23 March 2001 / Accepted: 27 November 2001

**Abstract.** Previous studies of the evolutionary history of hepatitis B virus (HBV) have been compromised by intergenotype recombination and complex patterns of nucleotide substitution, perhaps caused by differential selection pressures. We examined the phylogenetic distribution of recombination events among human HBV genotypes and found that genotypes A plus D, and genotypes B plus C, had distinct patterns of recombination suggesting differing epidemiological relationships among them. By analyzing the nonoverlapping regions of the viral genome we found strong bootstrap support for some intergenotypic groupings, with evidence of a division between human genotypes A–E from the viruses sampled from apes and human genotype F. However, the earliest events in the divergence of HBV remain uncertain. These uncertainties could not be explained by differential selection pressures, as the ratio of nonsynonymous-to-synonymous substitutions ($d_N/d_S$) did not vary extensively among lineages and there is no strong evidence for positive selection across the whole tree. Finally, we provide a new estimate of the mean substitution rate in HBV, $4.2 \times 10^{-5}$, which suggests that divergence of HBV in humans and apes has occurred only in the last 6000 years.

**Key words:** Hepatitis B virus — Phylogeny — Genotypes — Recombination — Substitution rates — Cross-species transmission

*Correspondence to:* Dr. Edward C. Holmes; *email:* Edward.Holmes@zoo.ox.ac.uk

## Introduction

Hepatitis B virus (HBV) is a chronic infectious pathogen belonging to the family Hepadnaviridae. As well as humans, hepadnaviruses have been isolated from birds, rodents, and a variety of nonhuman primates; chimpanzee, gorilla, orang-utan, gibbon, and woolly monkey (McDonald et al. 2000; Lanford et al. 1998; Norder et al. 1996; Vaudin et al. 1988; Verschoor et al. 2001; Warren et al. 1999). HBV has a circular genome of approximately 3200 bp, comprising double-stranded DNA and single-stranded RNA, and contains four open reading frames encoding the polymerase (P), surface antigen (S), nucleocapsid (C), and X proteins. Significantly, these reading frames show substantial overlap, with approximately half the viral genome encoding more than one protein product. At a genetic level, human HBV strains sampled worldwide have been classified into seven genotypes, denoted A to G (Norder et al. 1992; Stuyver et al. 2000) and displaying a 10–15% sequence divergence. While genotypes A and D have global distributions, genotype E is confined to West Africa, genotypes B and C to East and South East Asia, and genotype F to Central and South America, including among indigenous populations (Arauz-Ruiz et al. 1997; Norder et al. 1994). Only a single genotype G sequence is available to date.

Despite the availability of multiple whole-genome sequences for comparison, there is as yet no consensus on the evolutionary relationships or divergence times of the various primate HBVs and, hence, on its epidemiological history in humans (Grethe et al. 2000; Bollyky and Holmes 1999; Hu et al. 2000; Takahashi et al. 2000). There are a number of causes for this lack of phylogenetic resolution. First, different genomic regions show

very different patterns of evolution, including variable substitution rates, which might lead to conflicting phylogenetic signal (Bollyky and Holmes 1999). In particular, the evolution of overlapping genomic regions is qualitatively different from that of nonoverlapping regions, as synonymous changes in one frame may produce amino acid replacements in another (Mizokami et al. 1997). Consequently, despite utilizing more nucleotides, an analysis combining both overlapping and nonoverlapping genomic regions might compromise phylogenetic accuracy. Although analytical methods have now been developed to consider sequences with multiple reading frames (Pederson and Jensen 2001) and offer much hope for the future, at present these are able to only estimate pairwise distances. It is also evident that substitution rates vary extensively, with viruses that do not express the e antigen (HBeAg–) evolving more rapidly than those that are HBeAg+ (Hannoun et al. 2000a), a difference correlated with the strength of the host immune response. Finally, intergenotype recombination is increasingly recognized as an important factor in HBV evolution (Bollyky et al. 1996; Bowyer and Sim 2000; Hannoun et al. 2000b). Such a process, if not accounted for, could have a major impact on estimates of both phylogenetic relationship and divergence time (Schierup and Hein 2000).

In the current study we present a revised analysis of the evolutionary history of hepatitis B virus by removing all recombinant HBV sequences, including some that are newly identified, by inferring phylogenies using only nonoverlapping regions of the viral genome, and by estimating long-term substitution rates from mutation rates in HBeAg+ individuals. This analysis allows us to determine the evolutionary history of HBV with more precision.

## Materials and Methods

### *Sequence Data*

All complete genome sequences of primate hepadnaviruses, representing humans, chimpanzees, orang-utans, gibbons, and a single isolate from the woolly monkey were collected from GenBank and aligned manually. Only partial genome sequences are available from the gorilla so these were excluded from our study. Also excluded was the chimpanzee isolate Ch195, which clusters closely with human genotype E and therefore most likely represents a recent transmission from humans (see Discussion).

### *Analysis of HBV Recombination*

We initially screened for recombination by reconstructing neighbor-joining (NJ) trees [method available in PAUP* (Swofford 2000)] for windows of 500 bp, slid along whole-genome alignments in 250-bp increments. This approach was used to detect sequences with conflicting phylogenetic positions. After this preliminary search, we constructed diversity plots to examine the pattern of pairwise sequence divergence between the putative recombinants and their closest parental sequences as identified in the phylogenetic analysis. The percentage pairwise divergence between sequences was estimated using the program DIVERT (Gao et al. 1998), in which a window of 200 bp was slid along the alignments in 10-bp increments.

The optimal breakpoints in the recombinants identified above were then determined using a maximum likelihood (ML) method [program LARD (Holmes et al. 1999)]. In this approach, the sequence alignments for the recombinants and their two closest parental sequences were split in two at every possible point along the sequence and their branch lengths optimized in a phylogenetic tree. This enabled us to identify the breakpoint with the highest likelihood. A likelihood ratio test (LRT) was then used to compare this likelihood to that obtained under the hypothesis of no recombination. The significance of the LR value was compared to a null distribution of ratios generated using 500 Monte Carlo simulated sequences, subjected to the same breakpoint analysis as the real data but allowing no recombination [generated using program Seq-Gen (Rambaut and Grassly 1997)].

### *Phylogenetic Analysis*

After recombinant sequences were removed, a more formal phylogenetic analysis was conducted. Because of the large number of sequences available, particularly from genotype C, very similar sequences were removed from the analysis. This resulted in a final data set of 80 complete genome sequences, including human HBV genotypes A (16 sequences), B (7 sequences), C (22 sequences), D (15 sequences), E (2 sequences), and F (3 sequences), gibbon HBV (8 sequences), orang-utan HBV (2 sequences), and chimpanzee HBV (5 sequences). The single genotype G sequence was excluded at this point because its conflicting phylogenetic positions suggests that it is a recombinant virus (Stuyver et al. 2000). A full list of sequences used is available at http://evolve.zoo.ox.ac.uk/.

Phylogenetic trees were reconstructed using the concatenated nonoverlapping regions of the HBV genome. Overlapping regions were excluded because they are subject to complex evolutionary processes which might increase phylogenetic noise (see Introduction). Indeed, we observed highly asymmetric patterns of nucleotide substitution in these regions (data not shown). This resulted in a final alignment of 1753 bp.

Phylogenetic trees were estimated using the ML method implemented in the PAUP* package (Swofford 2000). The GTR+I+$\Gamma$ model of DNA substitution was used to infer evolutionary trees. Not only is this model the most general of those available, but it had a much higher likelihood than a GTR model with codon-specific substitution rates (data not shown). The relevant parameter settings were as follows: relative substitution rates of A$\leftrightarrow$C = 2.081, A$\leftrightarrow$G = 4.239, A$\leftrightarrow$T = 1.178, C$\leftrightarrow$G = 1.008, C$\leftrightarrow$T = 4.115, G$\leftrightarrow$T = 1, a proportion of invariable sites (I) of 0.234, a $\Gamma$ shape parameter ($\alpha$) of among-site rate variation of 0.605 with four rate categories, and a base composition of A = 0.255, C = 0.246, G = 0.210, T = 0.288. A starting tree was constructed using NJ, followed by successive rounds of TBR branch-swapping, identifying the ML substitution parameters at each stage, until the tree of highest likelihood was found. One thousand bootstrap NJ trees were also reconstructed using the ML substitution model.

In an attempt to determine the branching order of HBV, a second phylogenetic analysis was run incorporating the woolly monkey HBV (WMHBV) sequence as an outgroup. To ensure positional homology between WMHBV and the hominoid viruses, some highly divergent regions which were difficult to align were excluded, leaving a region of 1593 bp of nonoverlapping sequence for phylogenetic analysis. To simplify the analysis further, the phylogeny was reconstructed using a sample of 26 sequences representing the full genetic diversity in the hominoid viruses (a maximum of three viruses from each group). Once

again, the GTR+I+Γ model of DNA substitution was employed (parameter values available from the authors on request).

### *Estimating Substitution Rates and Dating the Origin of HBV Genotypes*

To estimate the rate of nucleotide substitution in HBV we analyzed the data of Hannoun et al. (2000a), which depicts mutation patterns among 13 chronic HBV carriers from three Vietnamese and Turkish families that were most likely vertically infected. Since none of the carriers were subjected to interferon or antiviral treatments, we believe that these data are a good representation of the normal mutation frequency of the virus. These data can be further divided into those patients that are HBeAg+ ($n = 7$) and those that are HBeAg– ($n = 6$). We assumed that transmission among individuals occurred only through HBeAg+ donors, as HBeAg+ status is commonly used as a marker of infectiousness, and that the mutations observed within HBeAg+ patients are neutral and hence correspond to long-term substitution rates. Conversely, the change to HBeAg– status most likely occurs within each host independently, is associated with an increased substitution rate caused by an exasperated immune response (Bozkaya et al. 1996, 1997; Hannoun et al. 2000a), and so may not contribute to long-term patterns of substitution in HBV. Within the seven HBeAg+ patients, the mean rate of nucleotide substitution per site per year ($k$) was estimated by simply counting the number of substitutions from the donor to the patient in the nonoverlapping part of the genome and then averaging this rate across all patients.

To test for the constancy of substitution rates among human genotypes and different species groups of HBV, we employed a ML version of the relative rate test [(Muse and Weir 1992); program Hy-Phy (Muse and Pond 2000)], incorporating the GTR+Γ substitution model. In this analysis, the sequence from the woolly monkey was used as the outgroup reference sequence and different combinations of representative sequences from each genotype or species group constituted the two ingroups. Because of shared phylogenetic history, these tests cannot be considered independent but should give a general indication of the extent of rate variation in hominoid HBV.

### *Analysis of Selection Pressures*

The selection pressures acting on HBV were investigated using a ML approach in which the numbers of nonsynonymous ($d_N$) and synonymous ($d_S$) substitutions are determined using different models of codon substitution which allow the $d_N/d_S$ ratio, denoted ω, to vary among sites and lineages (Yang et al. 2000). Seven models of codon substitution were used: (i) M0 calculates a single ω for all sites; (ii) M1 divides codon sites into a conserved category ($p_0$), with $\omega_0$ fixed at 0, and a second ($p_1$), with $\omega_1$ set to 1, representing strictly neutral sites; (iii) M2 adds a third category of sites ($p_2$), with $\omega_2$, estimated from the data, able to take on values >1; (iv) M3 estimates ω from the data for three site classes, each of which may be >1 and therefore provides a powerful test of positive selection; (v) M7 uses a discrete β distribution with 10 categories to describe ω among sites, all constrained to be <1; (vi) M8 differs from M7 only in that it estimates ω for an extra class of sites ($p_{10}$) at which ω can be >1; and (vii) the free ratio (FR) model allows ω to vary among branches of the tree and so can be used to detect lineage-specific changes in selection pressure. Because of the very large number of sequences involved, analysis using the FR model was run only on the sample of 26 sequences including WMHBV. Models that are nested may be compared using a LRT. In this way it is possible to compare directly models which allow positive selection (M2, M3, M8) with those that not (M0, M1, M7). Bayesian methods can also be used to calculate the probability that a particular codon site falls into the positively selected class. All analyses were performed using the CODEML program of the PAML package (Yang 1997).

**Table 1.** Maximum likelihood analysis of recombination events in hepatitis B virus[a]

| Genotype | Sequence | Breakpoint | Genes included | LR | $p$ |
|---|---|---|---|---|---|
| B | AF100308 | 1973 | C | 24.249 | <0.001 |
| B | AF100308 | 2231 | C | 32.721 | <0.001 |
| B | HPBAWITY | 2000 | C | 29.67 | <0.001 |
| B | HPBAWITY | 2248 | C | 29.44 | <0.001 |
| B | HPBADW2 | 1973 | C | 21.86 | <0.001 |
| B | HPBADW2 | 2203 | C | 24.28 | <0.001 |
| B | HPBADW3 | 2033 | C | 18.01 | <0.001 |
| B | HPBADW3 | 2271 | C | 23.80 | <0.001 |
| B | AF100309 | 1761 | X | 22.230 | <0.001 |
| B | AF100309 | 2300 | X–C | 38.016 | <0.001 |
| D | HBVAYWE | 800 | P–S | 32.202 | <0.001 |
| D | HBVAYWE | 2000 | P–S–C | 63.216 | <0.001 |
| D | HBVDNA | 780 | P–S | 22.250 | <0.001 |
| D | HBVDNA | 2558 | P–S–C–X | 82.145 | <0.001 |
| D | HBVAWCI | 136 | P | 17.53 | <0.001 |
| D | HBVAWCI | 498 | P–S | 7.62 | 0.006 |
| D | HBVAWCI | 1635 | X | 21.25 | <0.001 |
| D | HBVAWCI | 2563 | P | 46.88 | <0.001 |

[a] The optimal breakpoints for each recombinant are presented with the likelihood ratio (LR) comparing the hypotheses of recombination and no recombination.

## Results

### *Recombination Among Human HBV Genotypes*

Two groups of recombinant sequences were detected in our analysis. The first contains sequences of genotypes B and C, while the second comprises sequences of genotypes A and D (diversity plots are available from the authors on request). The ML analysis identified various breakpoints in the genotype B recombinants, although all were located in the region 1761–2300, which generally corresponds to the nucleocapsid protein (Table 1). Most notable was that all breakpoints were found in nonoverlapping regions of the genome and the recombinant fragments were usually small in size. Furthermore, the genotype B recombinants were closely related on the phylogenetic tree, suggesting that they are the descendants of a single recombination event. In contrast, fewer recombinants were found among genotypes A and D (although sequence HBVAWCI was found to have two distinct recombinant regions), but the observed breakpoints were present in both the overlapping and the nonoverlapping regions of the genome and the recombinant fragments were larger than the genotype B–C recombinants. Thus, it appears that the pattern of recombination differs according to the genotypes involved.

### *Phylogenetic Relationships Among Primate HBVs*

The unrooted ML phylogenetic tree (with 1000 NJ bootstrap replications) for the nonoverlapping region of the

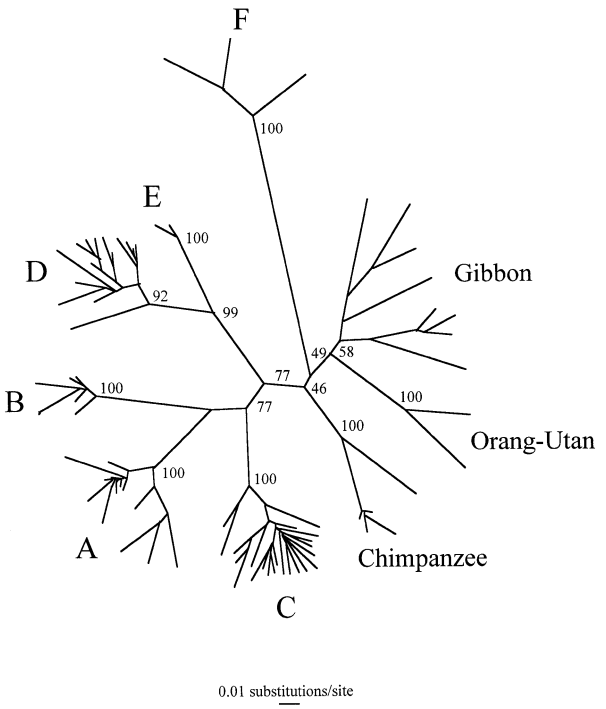


**Fig. 1.** Unrooted maximum likelihood phylogenetic tree of hepatitis B virus (HBV) using 80 nucleotide sequences representing the concatenated nonoverlapping regions of the viral genome. Neighbor-joining bootstrap values (1000 replications) for major groupings are indicated on the relevant nodes. Letters A to F refer to the six genotypes of human HBV. All branch lengths are drawn to scale.

HBV is represented in Fig. 1. The monophyly of viruses from each human genotype or species of nonhuman primate is evident, although those from the gibbon are not supported by high bootstrap values. The ML tree also provides evidence for a variety of intergenotype associations. Most notable was a grouping of all human genotypes with the exception of human genotype F, which instead clusters with the viruses sampled from chimpanzees, orang-utans, and gibbons (77% bootstrap support). Furthermore, genotypes A, B, and C cluster with relatively strong (77%) bootstrap support, and genotypes D and E are found to be closely related. Because genotypes B and C are confined to the Far East and some Pacific regions, whereas genotype A is distributed mainly in western Europe and less so in Asia, Africa, and North America, this phylogeny suggests that genotype A may have also begun its divergence in Asia. Unfortunately the phylogenetic relationships within the ape-genotype F clade are less certain, with the key branches having short lengths and no clear picture coming from the bootstrap analysis. This uncertainty was confirmed in an SH test of likelihood scores (Shimodaira and Hasegawa 1999). Here, a model topology was constructed in which the chimpanzee viruses were more closely related to those from gibbon and orang-utan, rather than those from genotype F. This tree also had the effect of making all human genotypes monophyletic. Although this new tree had a lower likelihood than the ML tree (difference = 2.351), it was not significantly worse ($p = 0.274$). Finally, as human genotype F has previously been shown to have an uncertain phylogenetic position (Bollyky and Holmes 1999), a ML tree was estimated with this genotype excluded (tree not shown). While the branching order was unchanged, the gibbon and orang-utan viruses now grouped together with very strong bootstrap support (94%), while the division between the human and the ape viruses was slightly weaker (68% bootstrap support).

In an attempt to determine the direction of HBV evolution we undertook a second phylogenetic analysis incorporating a subset of the viruses from each genotype/species group using the single woolly monkey sequence as an outgroup (Fig. 2). Unfortunately, this analysis provided no strong support for a particular branching order in the early history of HBV, with human genotype B the first to diverge in the ML tree (although this was not seen in the bootstrap analysis). However, there was continued support for the ape + genotype F grouping (80% bootstrap support) and for a gibbon–orang-utan clade (83% bootstrap support).

### *Rates and Dates of HBV Evolution*

Our first task was to estimate the rate of nucleotide substitution in the nonoverlapping regions of the HBV genome. This was achieved using mutation data from seven HBeAg+ patients compared to their respective donors (Hannoun et al. 2000a). Under this method, the mean rate of nucleotide substitution was estimated to be $4.2 \times 10^{-5}$ nucleotide substitutions/site/year, although with a large standard error, $2.96 \times 10^{-5}$ nucleotide substitutions/site/year, reflecting the fact that the HBV sequences from some patients showed no substitutions over the period of observation.

To determine whether substitution rates are equivalent between the various hominoid HBVs, we performed a series of ML relative rate tests using the woolly monkey as the outgroup reference sequence (Table 2). Four comparisons were conducted to cover as much as the HBV phylogeny as possible, with a single sequence taken as representative of each human genotype or species group (with the exception of genotype E). In no case was the molecular clock rejected, although shared phylogenetic history means that these tests are not independent.

As substitution rates appear to be equivalent among viruses, it is possible to provide a tentative time scale for HBV evolution. However, because of the large variance in our estimated substitution rate we have chosen simply to give an upper bound on divergence times. The maximum distance between any two hominoid HBV sequences under the ML substitution model was ~20%. This would correspond to a divergence time of approximately 2500 years, with a corresponding error of approximately 3500 years. Consequently, the maximum date we estimate for the origin of HBV in hominoid

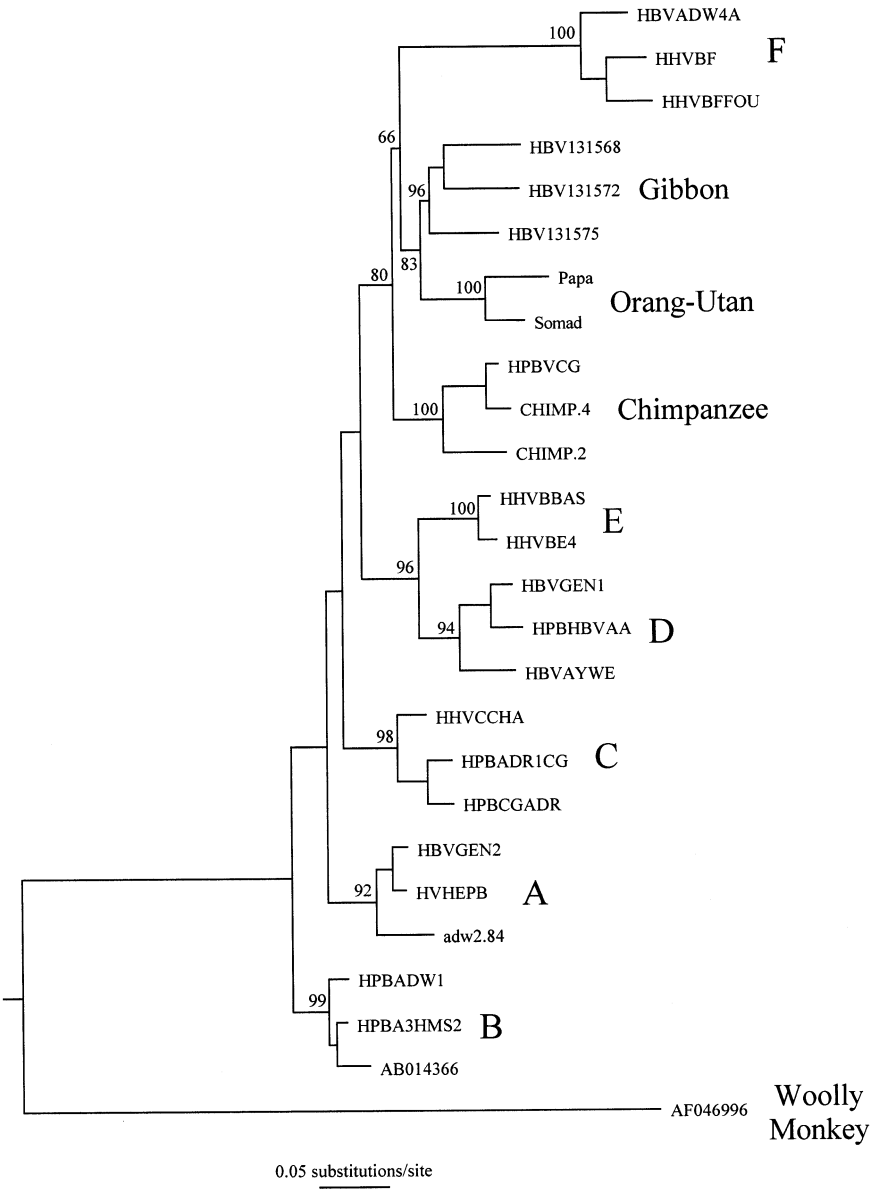


**Fig. 2.** Maximum likelihood phylogenetic tree using the nonoverlapping regions from a representative sample of 26 viruses using the single woolly monkey HBV sequence as an outgroup. Neighbor-joining bootstrap values (1000 replications) for major groupings are indicated on the relevant nodes. Letters A to F refer to the six genotypes of human HBV. Horizontal branch lengths are drawn to scale.

**Table 2.** Maximum likelihood relative rate tests among different phylogenetic groups of hepatitis B virus[a]

| Comparison | $\chi^2$ | $p$ |
|---|---|---|
| Gibbon–A | 1.664 | 0.197 |
| Orang-utan–D | 0.827 | 0.363 |
| Chimpanzee–C | 1.444 | 0.229 |
| B–F | 1.719 | 0.190 |

[a] $\chi^2$ is computed as twice the difference in log likelihood between the clock and the nonclock models for the comparison in question. In every case the woolly monkey sequence was used as the outgroup.

primates is ~6000 years ago. If we assume that the substitution rate in the woolly monkey is the same as that in the hominoid HBVs, then the ~75% distance between these viruses would place their divergence at approximately 9000 years ago, with an upper bound of ~22,000 years. In contrast, for the genetic distances we observe to be compatible with virus–host cospeciation over a period of 40 million years (the hominoid–woolly monkey divergence) requires substitution rates of approximately $10^{-8}$ substitutions/site/year. Hence, the diversification of HBV appears to be a recent event, with frequent cross-species transmission.

### *Natural Selection on HBV Genomes*

Finally, we asked whether changes in selection pressure might have influenced the patterns of sequence evolution observed. In particular, we examined whether selection pressures changed when the virus crossed species boundaries. The results of our ML analyses of selection pressures on all 80 sequences revealed only weak evidence for positive selection in the nonoverlapping regions of the HBV genome (Table 3). Specifically, although the

**Table 3.** Maximum likelihood analysis of selection pressures acting on hepatitis B virus[a]

| Model | Parameter estimates | LRT | $\chi^2$ | $p$ |
|---|---|---|---|---|
| M0 | $\omega = 0.211$ | | | |
| M1 | $p_0 = 0.867$, $p_1 = 0.133$<br>$\omega_0 = 0$, $\omega_1 = 1$ | | | |
| M2 | $p_0 = 0.504$, $p_1 = 0.046$, $p_2 = 0.450$<br>$\omega_0 = 0$, $\omega_1 = 1$, $\omega_2 = 0.072$ | M0 vs M2<br>M1 vs M2 | 2162.033<br>706.021 | <0.00000<br><0.00000 |
| M3 | $p_0 = 0.937$ $p_1 = 0.052$, $p_2 = 0.011$<br>$\omega_0 = 0.031$, $\omega_1 = 0.704$, $\omega_2 = 1.782$ | M0 vs M3<br>M1 vs M3<br>M2 vs M3 | 2221.189<br>765.177<br>59.156 | <0.00000<br><0.00000<br><0.00000 |
| M7 | $p = 0.046$, $q = 0.399$ | | | |
| M8 | $p = 0.321$, $q = 6.637$<br>$p_{10} = 0.034$, $\omega_{10} = 1.267$ | M7 vs M8 | 219.682 | <0.00000 |

[a] LRTs compare twice the difference in log likelihood between competing models to the value obtained under a $\chi^2$ distribution. Degrees of freedom are given by the difference in the numbers of parameters between the models. $p$ and $q$ are parameters of the B distribution.

M3 model, which allows positive selection, is significantly favored over neutral models, only ~1% of sites are subject to relatively weak positive selection ($\omega_2 = 1.782$). Similarly, while M8 model is favored over the M7 model, the highest ω value observed (1.267) does not provide conclusive evidence for adaptive evolution. We also investigated whether selection pressures varied among our sample of 26 sequences. Under the FR model, mean ω values showed relatively little variation among human genotypes/species groups (range 0.078 to 0.207). In particular, there was no evidence of a change in selection pressures on the lineage leading to divergent genotype F ($\omega = 0.211$) or on the lineage leading to the woolly monkey sequence ($\omega = 0.227$). Consequently, these analyses indicate that all primate hepadanviruses are subject to roughly equivalent selection pressures and that there is little evidence for adaptive evolution in these viruses.

## Discussion

### *Recombination in HBV*

The first evidence for homologous recombination in HBV was provided by Georgi-Geisberger et al. (1992). Bollyky et al. (1996) later demonstrated that two complete genome sequences had a mosaic structure, suggestive of recombination, while more recently recombination events involving genotypes B and D (Bowyer and Sim 2000) and genotypes A and C (Hannoun et al. 2000b) have been documented, as well as a potentially recombinant genotype G (Stuyver et al. 2000). In this paper, we demonstrated further recombination events between genotypes A and D and genotypes B and C.

Not only does our study extend the number of known HBV recombinants, but the pattern of recombination might shed some light on the spread of the human genotypes. In particular, genotypes A and D have recombined in both overlapping and nonoverlapping genome regions and the size of the recombinant region varied greatly. This suggests that several distinct recombination events have occurred between the widely geographically distributed genotypes A and D, most likely over a relatively long period of time during which many mixed infections could occur. In contrast, genotypes B and C have seemingly recombined in relatively small segments of the nonoverlapping regions alone. Moreover, the genotype B recombinants form a single monophyletic group, which implies a single main recombination event. This could mean that genotypes B and C have only had relatively recent epidemiological contact, clearly in Asia, during which there has been insufficient time for recombination among longer and more diverse genome regions.

### *The Origins of Primate HBV*

The major goal of our study was to reconstruct the problematic evolutionary history of HBV. To do this we performed phylogenetic analyses of the nonoverlapping regions of the HBV genome in an attempt to minimize phylogenetic error. This analysis was successful in that we were able to provide more phylogenetic resolution than in any previous study. The most striking result of this analysis was the consistent support for a major phylogenetic division between all human HBV genotypes, with the exclusion of genotype F, and the viruses sampled from gibbons, orang-utans, and chimpanzees. Furthermore, despite the isolation of hepadnaviruses from those primate species most closely related to humans, there was no convincing evidence that these viruses have cospeciated with their hosts over millions of years. In particular, the sequences from gibbons and orang-utans group together, which supports cross-species transmission as the geographical ranges of these species overlap in Southeast Asia, and all human viruses do not form a monophyletic group. Recent cross-species transmission is further supported by the observation that some chimpanzee strains cluster within the human genotypes (Hu et al. 2000; Takahasi et al. 2000) and that the single

gorilla sequence available groups closely with those from chimpanzees (Hu et al. 2000). Furthermore, the transfer of other viral infections from humans to chimpanzees in nature has been documented (Ferber 2000).

Our analysis of substitution rates and divergence times provides further evidence that the evolution history of HBV in primates is a relatively recent one. Our observed evolutionary rate—$4.2 \times 10^{-5}$ nucleotide substitutions/site/year—is similar to that obtained for synonymous substitutions per site in a single HBV-infected individual (Orito et al. 1989). Applying this rate to the divergence events depicted in the phylogenetic analysis suggests that the origin of HBV in humans and other hominoid primates has occurred within the last 6000 years. Even with the difficulties intrinsic in analyses of this sort, including multiple substitution in the comparisons involving the woolly monkey and the possibility of subtle selective constraints imposed by RNA secondary structure (Simmonds and Smith 1999), it is difficult to imagine that the error in the substitution rate estimated here could be of the four orders of magnitude needed to reconcile viral divergence times with host speciation dates. Indeed, our estimated substitution rate is likely to represent a lower bound because we have assumed that only HBeAg+ individuals, who accumulate fewer substitutions, are able to transmit the virus and there is some evidence that individuals who have cleared the e antigen may also pass on HBV (Heptonstall et al. 1997). Further, that all the viruses analyzed appear to be under similar selection pressures argues that the colonization of new host species was not accompanied by major changes in the rate of nonsynonymous substitution.

Such a recent evolutionary time scale raises a number of important issues regarding the epidemiological history of HBV. That the virus is found in a wide range of species occupying disparate geographical regions and that the hominoid viruses have diverged within the last 6000 years both point to humans as the most important vector of viral transmission. Indeed, HBV is noteworthy in that the nonhuman primate lineages are no more divergent than those from humans, suggesting that species groups were infected at approximately the same time. This sits in contrast to the pattern in primate lentiviruses in which all human immunodeficiency viruses (HIVs) fall within particular clades of simian immunodeficiency viruses (SIVs), as expected if the latter are the reservoir population. The apparent division of HBV into two major clades, one of which contains the majority of human viruses, with the other containing all those sampled from other ape species as well as human genotype F is also intriguing, although difficult to explain at present.

The most puzzling aspects of the phylogenetic tree of HBV are the divergent positions of the two viruses from the New World; genotype F with respect to the other human genotypes and the woolly monkey sequence with respect to all other primate hepadnaviruses. Due to the lack of contact between the New World and the Old World from the time of the earliest human migrations to the Americas to post-Columbian colonization, the divergence time of New World and Old World viruses must be either older than ~15,000 years ago or more recent than ~500 years ago (Bollyky and Holmes 1999). Given these time constraints, either genotype F originated in the Old World and did not enter the Americas until relatively recently, in which case the WMHBV–hominoid virus divergence may date back to the initial migration of modern humans into the Americas, or genotype F is of American origin and was the first human genotype to diverge from its New World monkey ancestors. The migration of genotype F to the Old World then occurred in post-Colombian times and seeded all subsequent HBV epidemics, including those in other primate species. The survey of HBV infection in a wider range of New World primates is clearly needed to resolve these issues.

*Acknowledgments.* This work was supported by Grant PM97-0060-C02-02 awarded to Dr. Andres Moya by the DGESIS-MEC (Spain). M.A.F. acknowledges a fellowship from the Conselleria de Cultura, Educacion y Ciencia, Generalitat Valenciana, while E.C.H. thanks The Royal Society for financial support. We are also grateful to Dr. David Robertson and two anonymous referees for valuable comments.

## References

Arauz-Ruiz P, Norder H, Visoná KA, Magnius LO (1997) Molecular epidemiology of hepatitis B virus in Central America reflected in the genetic variability of the small S gene. J Infect Dis 176:851–858

Bollyky PL, Holmes EC (1999) Reconstructing the complex evolutionary history of hepatitis B virus. J Mol Evol 49:130–141

Bollyky PL, Rambaut A, Harvey PH, Holmes EC (1996) Recombination between sequences of hepatitis B virus from different genotypes. J Mol Evol 42:97–102

Bowyer SM, Sim JGM (2000) Relationships within and between genotypes of hepatitis B virus at points across the genome: footprints of recombination in certain isolates. J Gen Virol 81:379–392

Bozkaya H, Ayola B, Lok ASF (1996) High rate of mutations in the hepatitis B core gene during the immune clearance phase of chronic hepatitis B virus infection. Hepatology 24:3237–3243

Bozkaya H, Akarca US, Ayola B, Lok ASF (1997) High rate of conservation in the hepatitis B virus core gene during the immune tolerant phase in perinatally acquired chronic hepatitis B virus infection. J Hepatol 26:508–516

Ferber D (2000) Human diseases threaten great apes. Science 289: 1277–1278

Gao F, Robertson DL, Carruthers CD, Morrison SG, Jian B, Chen Y, Barré-Sinoussi F, Girard M, Srinivasan A, Abimiku AG, Shaw GM, Sharp PM, Hahn BH (1998) A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J Virol 72:5680–5698

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

Grethe S, Heckel J, Rietschel W, Hufert F (2000) Molecular epidemiology of hepatitis B virus variants in nonhuman primates. J Virol 74:5377–5381

Hannoun C, Horal P, Lindh M (2000a) Long-term mutation rates in the hepatitis B virus genome. J Gen Virol 81:75–83

Hannoun C, Norder H, Lindh M (2000b) An aberrant genotype revealed in recombinant hepatitis B virus strains from Vietnam. J Gen Virol 81:2267–2272

Heptonstall J, Barnes J, Burton E, Chattopodyhay B, McMillan L, Sullivan K, Tarling R, Viniker D, Boxall E, Cartmill I, Chatterjea M, Neill R, Collins M, Gill N, Ngui SL, Parker C, Ryan M, Teo CG, Coyle P, Craske J, Paver K, Gilson R, Hawkins A, Tedder R, Watts P, Zuckerman M, Morris D, Nazareth B (1997) Transmission of hepatitis B to patients from 4 infected surgeons without hepatitis B e antigen. N Engl J Med 336:178–184

Holmes EC, Worobey M, Rambaut A (1999) Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol 16:405–409

Hu X, Margolis HS, Purcell RH, Ebert J, Robertson BH (2000) Identification of hepatitis B virus indigenous to chimpanzees. Proc Natl Acad Sci USA 97:1661–1664

Lanford RE, Chavez D, Brasky KM, Burns RB III, Rico-Hesse R (1998) Isolation of a hepadnavirus from the woolly monkey, a New World primate. Proc Natl Acad Sci USA 95:5757–5761

McDonald DM, Holmes EC, Lewis JCM, Simmonds P (2000) Detection of hepatitis B virus infection in wild-born chimpanzees (*Pan troglodytes verus*): Phylogenetic relationships within human and other primate genotypes. J Virol 74:4253–4257

Mizokami M, Orito E, Ohba K, Ikeo K, Lau JYN, Gojobori T (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol 44:83–90

Muse SV, Pond SG (2000) Hy-Phy. http://peppercat.statgen.ncsu.edu/~hyphy

Muse SV, Weir BS (1992) Testing for equality of evolutionary rates. Genetics 132:269–276

Norder H, Hammas B, Lofdahl S, Couroucé AM, Magnius LO (1992) Comparison of the amino acid sequences of 9 different serotypes of hepatitis B surface antigen and genomic classification of the corresponding hepatitis B virus strains. J Gen Virol 73:1201–1208

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

Norder H, Ebert JW, Fields HA, Mushahwar IK, Magnius LO (1996) Complete sequencing of a gibbon hepatitis B virus genome reveals a unique genotype distantly related to the chimpanzee hepatitis B virus. Virology 218:214–223

Orito E, Mizokami M, Ina Y, Moriyama EN, Kameshima N, Yamamoto M, Gojobori T (1989) Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc Natl Acad Sci USA 86:7059–7062

Pedersen A-M K, Jensen JL (2001) A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol 18: 763–776

Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS 13:235–238

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–891

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

Simmonds P, Smith DB (1999) Structural constraints on RNA virus evolution. J Virol 73:5787–5794

Stuyver L, Gendt SD, Geyt CV, Zoulim F, Fried M, Schinazi RF, Rossau R (2000) A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. J Gen Virol 81:67–74

Swofford DL (2000) PAUP*. Phylogenetic analysis using parsimony (* and other methods), version 4. Sinauer Associates, Sunderland, MA

Takahashi K, Brotman B, Usuda S, Mishiro S, Prince AM (2000) Full-genome sequence analyses of hepatitis B virus (HBV) strains recovered from chimpanzees infected in the wild: implications for an origin of HBV. Virology 267:58–64

Vaudin M, Wolstenholme AJ, Tsiquaye KM, Zuckerman AJ, Harrison TJ (1988) The complete nucleotide sequence of the genome of a hepatitis B virus isolated from a naturally infected chimpanzee. J Gen Virol 69:1383–1389

Verschoor EJ, Warren KS, Langenhuijzen S, Heriyanto, Swan RA, Heeney JL. (2001) Analysis of two genomic variants of orang-utan hepadnavirus and their relationship to other primate hepatitis B-like viruses. J Gen Virol 82:893–897

Warren KS, Niphuis H, Heriyanto, Verschoor EJ, Swan RA, Heeney JL (1998) Seroprevalence of specific viral infections in confiscated orangutans (*Pongo pygmaeus*). J Med Primatol 27:33–37

Warren KS, Heeney JL, Swan RA, Heriyanto, Verschoor EJ (1999) A new group of hepadnaviruses naturally infecting orangutans (*Pongo pygmaeus*). J Virol 73:7860–7865

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z, Nielsen R, Goldman N, Pedersen AMK 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449