

Bayesian Estimates of the Evolutionary Rate and Age of Hepatitis B Virus

Yang Zhou · Edward C. Holmes

Received: 28 March 2007 / Accepted: 24 May 2007
© Springer Science+Business Media, LLC 2007

Abstract Accurately estimating the evolutionary rate and age of hepatitis B virus (HBV) has proven to be one of the most difficult problems in studies of viral evolution. To help resolve these issues we employed a recently developed Bayesian coalescent approach to globally sampled human and avian hepadnavirus genome sequences, accounting for lineage-specific rate variation, the presence of overlapping reading frames, and the potential impact of recombination. Our analysis revealed an unexpectedly high rate of evolutionary change—up to 10^{-4} nucleotide substitutions (subs) per site per year and always more than $\sim 10^{-6}$ subs/site/year. These rates suggested a time to the most recent common ancestor (tMRCA) of the sampled isolates of consistently less than ~ 1500 years ago for human HBV and less than 6000 years ago for the avian hepadnaviruses. Notably, the evolutionary rate of nonoverlapping regions of the viral genome was ~ 2 -fold greater than that of overlapping genome regions, reflecting the complex patterns of selective constraint inherent in the former. We also reveal that most recombination events in both human and avian HBV tend to fall in a specific region of the viral genome, which contains all four viral open reading frames and which may therefore represent a “hot

spot” for recombination. However, while recombination affects estimates of both evolutionary rate and tMRCA, in no case was this sufficient to challenge the hypothesis that the dominant mode of HBV evolution is by recent cross-species transmission. We conclude that HBV exhibits rapid evolutionary dynamics, typical of other viruses dependent on reverse transcriptase-mediated replication.

Keywords Hepatitis B virus · Relaxed molecular clock · Phylogeny recombination · Most recent common ancestor

Introduction

Hepatitis B virus (HBV) is a common and serious viral pathogen of humans that belongs to the family *Hepadnaviridae*. It is estimated that there are approximately 300 million HBV carriers globally, with Asia particularly badly affected (WHO 2005). Chronic carriers may eventually succumb to serious disease manifestations such as cirrhosis and hepatocellular carcinoma. Despite its importance, the evolutionary origins of HBV and the timescale of its spread, remain uncertain. For example, it is still unclear whether HBV co-diverged with the various hominoid primate species with which it is associated—chimpanzees, gorillas, humans, orang-utans, and gibbons—over many millions of years, or that its spread reflects more recent patterns of contact among human populations following cross-species transmission (Simmonds 2001). This evolutionary complexity is evidently due, in part, to the fact that HBV is a unique enveloped double-stranded (ds) DNA virus with a circular genome of ~ 3200 bp that employs the error-prone polymerase reverse transcriptase (RT) for part of its replication process and that its four open reading frames (ORFs)—encoding the polymerase (P), surface

Electronic supplementary material The online version of this article (doi: 10.1007/s00239-007-0054-1) contains supplementary material, which is available to authorized users.

Y. Zhou · E. C. Holmes
Center for Infectious Disease Dynamics, Department of Biology,
The Pennsylvania State University, Mueller Laboratory,
University Park, PA 16802, USA

E. C. Holmes (✉)
Fogarty International Center, National Institutes of Health,
Bethesda, MD 20892, USA
e-mail: ech15@psu.edu

antigen (S), and nucleocapsid (C) and X proteins—exhibit substantial overlap with each other.

A central aspect in resolving the origin of HBV is obtaining an accurate rate of nucleotide substitution from which divergence times can be estimated. Although variable, previous estimates of evolutionary rates in HBV are intermediate between those of RNA and DNA viruses, ranging from 1.5 to 7.9×10^{-5} nucleotide substitutions (subs) per site per year (substitutions/site/year) (Fares and Holmes 2002; Hannoun et al. 2000; Okamoto et al. 1987; Orito et al. 1989; Osioy et al. 2006). On the basis of these rates, the upper bound for the time to the most recent common ancestor (tMRCA) of sampled isolates of human HBV is in the region of 3000–6000 years (Fares and Holmes 2001; Orito et al. 1989). Such a timescale implies that modern human dispersal is responsible for the phylogeographic pattern of HBV, with those viruses observed in other hominoid primates generated by cross-species transmission rather than long-term host-virus codivergence (Starkman et al. 2003). Related hepadnaviruses are also found in some avian species (geese, ducks, herons) and assigned to the genus *Avihepadnavirus*. However, as yet, no work has been done to estimate evolutionary parameters in these avian viruses, even though they make an excellent counterpart to the study of human HBV.

Most studies of rates of nucleotide substitution in HBV undertaken to date have used serially sampled sequences from one or several individuals who have carried the virus for extended time periods (Okamoto et al. 1987; Orito et al. 1989; Osioy et al. 2006) or from mother-child pairs (Fares and Holmes 2002; Hannoun et al. 2001). Consequently, these results are only indicative of the short-term, and often intrahost, evolutionary rate in chronic HBV carriers, which does not necessarily reflect the process of nucleotide substitution in the long-term. Indeed, extensive intrahost genetic diversity has been observed on occasion (Osioy et al. 2006). Similarly, no previous study has convincingly accounted for the occurrence of lineage-specific variation in rates of nucleotide substitution. Such a process may be particularly important in HBV as it has been demonstrated that viruses that do not express the e antigen (HBeAg⁻) evolve more rapidly than those that are HBeAg⁺ (Hannoun et al. 2000).

Several other characteristics of HBV also contribute to the complexity of inferring the timescale of its evolutionary history. Perhaps of most importance is that the hepadnaviruses are characterized by a complex pattern of overlapping reading frames. In particular, the P ORF occupies 75% of the genome, with the S ORF entirely embedded within it, and nucleocapsid and X ORFs overlapping in its distal regions. Such overlap obviously acts to constrain viral evolution, as synonymous substitutions in one frame are often nonsynonymous in another

(and vice versa), leading to complex patterns of purifying selection (Bollyky and Holmes 1999; Fares and Holmes 2002; Mizokami et al. 1997). It is also likely that the relatively high rate of recombination in HBV has adversely affected estimates of substitution rate and hence divergence times. Since its discovery, recombination has been widely identified across the HBV genome and in the eight circulating human genotypes (Bollyky et al. 1996; Bowyer and Sim 2000; Fares and Holmes 2002; Kurbanov et al. 2005; Luo et al. 2004; Olinger et al. 2006; Simmonds and Midgley 2005; Suwannakarn et al. 2005; Wang et al. 2005; Yang et al. 2006). Indeed, some recombinant HBV isolates have reached high prevalence; for example, the dominant HBV type in Tibet is a C/D recombinant, and the genotype B isolates sampled from China similarly have a recombinant origin (Cui et al. 2002; Luo et al. 2004).

In an attempt to infer the dynamics and timescale of evolution in the human and avian hepadnaviruses with as much accuracy as possible, we employed a recently developed Bayesian Markov chain Monte Carlo (MCMC) method that can account for lineage-specific rate variation through the use of a “relaxed” molecular clock. This method analyzes the distribution of mutational differences among viruses sampled at different times (serial sampling) and provides a natural measure of statistical uncertainty as all estimates are based on a very large number of sampled trees (Drummond et al. 2002, 2005). We employ these methods to human and avian HBV for the first time.

Materials and Methods

Sequence Data

Complete genomes of human HBV where the date (year) of sampling was available were downloaded from GenBank and aligned manually (a full list is provided in Supplementary Table 1). Where dating information was unavailable from GenBank we either referred to the primary literature or contacted the relevant authors. Known recombinants were excluded from the data, as were those viruses sampled from patients subject to anti-viral therapy or suffering immunosuppression (i.e., AIDS patients or organ transplant patients), as these are likely to have “artificially” altered evolutionary rates. This resulted in a data set of 145 human HBV sequences. A number of different data subsets were then constructed for evolutionary analysis (see below). Using the same criteria we collected the complete genomes of 35 avian hepadnaviruses for which the year of sampling was known (full details given in Supplementary Table 2).

Identification of HBV Recombinants

We used the RDP2 program (Martin et al. 2005a) to identify recombinant HBV sequences. This employs six different algorithms to identify recombinants: RDP, Si-Scan, Bootscan, Chimeric, MaxChi, and GENECONV. As there is no consensus on the optimal method for detecting recombination, we deemed that sequences identified as recombinant by three or more methods within RDP2 should be excluded from any further analysis (a list of recombinant sequences is provided in Supplementary Table 1). This approach therefore represents a balance between being overly conservative and including false-positive results. Following the exclusion of recombinants, our data set comprised 76 complete genome sequences of human HBV, representing genotypes A, C, D, E, F, and H, which may be considered our “global” human HBV data set (all viruses assigned to genotypes B were found to be recombinant and no dated G genotype sequences were available for study). For these 76 complete genome sequences we also constructed alignments of the (concatenated) overlapping and nonoverlapping regions of the viral genome. Similarly, we also extracted and concatenated the third codon positions of the nonoverlapping reading frame regions, as these are expected to comprise the least constrained set of nucleotide sites in the viral genome. Finally, because sufficient sequences were available, we also extracted all those sequences assigned to genotypes C ($n = 26$) and D ($n = 22$). All sequence alignments are available from the authors on request.

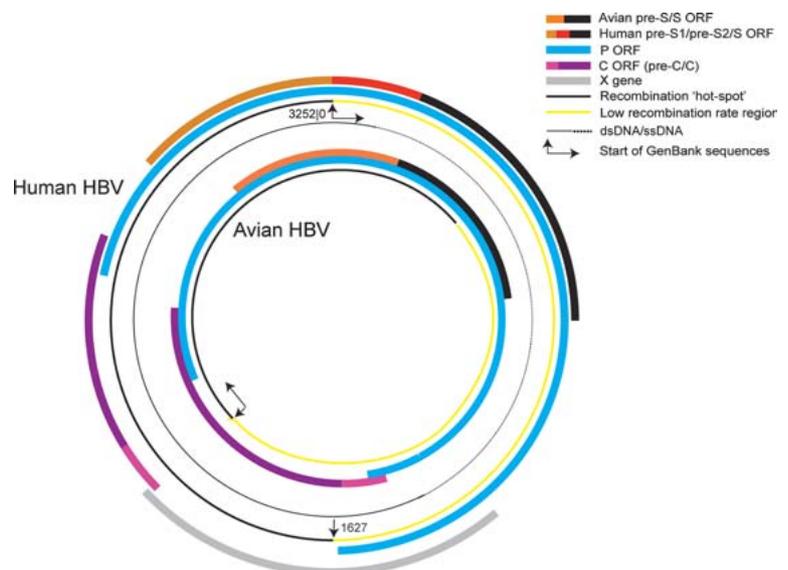
During this recombination screen we also observed that most recombination events occurred in one region of the viral genome. Therefore, to improve the power of our phylogenetic analysis, we divided the HBV genome into

two halves, reflecting their differing frequencies of recombination (Fig. 1): (i) nucleotides 1–1626, covering the pre-S2 region, the S ORF, 1623 bp of the P ORF, and the first 226 nucleotides of the X ORF, and which had a lower rate of recombination, (ii) nucleotides 1627–3252, covering the whole pre-C/C gene, the pre-S1 region, the first 915 bp of the P ORF, and the last 228 bp of the X ORF, and which contained the majority of the recombination events. By removing recombinant region ii from our initial data set of 145 human HBV genomes, and excluding identical sequences, we were able to produce a data set of 108 sequences comprising region i described above and characterized by a relatively low recombination rate (Supplementary Table 1).

Analysis of Evolutionary Dynamics

Overall rates of evolutionary change (nucleotide substitutions per site per year) and the tMRCA for each data set were estimated using the Bayesian MCMC approach available in the BEAST package (<http://www.evolve.zoo.ox.ac.uk/Beast/>) (Drummond and Rambaut 2003). In all cases, we employed the most general GTR+I+ Γ_4 model of nucleotide substitution, as this consistently provides the best description of viral evolution. Data sets were also analyzed under demographic models of (a) constant population size, (b) exponential population growth, (c) expansion population growth, and (d) logistic population growth. For all models, we used both strict and relaxed (uncorrelated exponential) molecular clocks. In each case, MCMC chains were run for sufficient time to achieve convergence (assessed using the TRACER program: <http://www.evolve.zoo.ox.ac.uk/software.html?id=tracer>), and models were compared using Akaike’s information crite-

Fig. 1 Genome structure of human and avian HBV. The outer circle represents the human genome, while the inner circle depicts the avian hepadnavirus genome. Individual ORFs and the recombination “hot spot” and region of low recombination rate are also indicated (although, because of the variable distribution of recombination break-points, the demarcation of these regions is only approximate)



rion (AIC), with uncertainty in parameter estimates reflected in the 95% highest probability density (HPD).

Analysis of Avian Hepatitis B Virus

For the 35 avian hepadnavirus genome sequences collected herein, we performed the same process of recombinant identification as with the human HBV data. Strikingly, the majority of recombination events in the avihepadnavirus occurred in the same (homologous) region as human HBV (Fig. 1) (although these regions are numbered differently in the GenBank accessions). Hence, we were again able to construct a “low recombination rate” data set reflecting approximately 50% of the viral genome for the avian viruses in the same manner as for the human HBV sequences. These data were then analyzed using the same Bayesian coalescent methods as in the analysis of the human hepadnavirus data sets.

Phylogenetic Analysis

Maximum likelihood (ML) phylogenetic trees were inferred both for the complete genome data sets of 76 human and 35 avian HBV sequences and for the low recombination rate region of 108 human HBV and 35 avian HBV sequences. In all cases phylogenetic trees were estimated using the ML method implemented in the PAUP* package (Swofford 2003), assuming the most general GTR+I+ Γ_4 model of nucleotide substitution. A neighbor-joining bootstrap resampling analysis (1000 replications) was performed to assess the support for specific nodes, again utilizing the ML substitution model. Finally, average pairwise distances among data sets were calculated under the Tamura-Nei model with a Γ -distribution (shape parameters taken from the ML phylogenetic analysis) using the MEGA package (Kumar et al. 2004).

Results and Discussion

Recombination in Human and Avian HBV

The six methods contained within the RDP2 package identified different subsets of HBV sequences as recombinants (Supplementary Table 1). On average, across all six methods, 41% of 145 human HBV sequences were suspected as recombinant, suggesting that this process occurs frequently in the hepadnaviruses. The highest number of recombinants—92—was identified by the RDP method, while Chimera gave the lowest number (27). Notably, nine sequences were identified as recombinant in all methods, all of which belong to genotype B. A further seven genotype B sequences were identified as recombinant by five of

the six methods, four of which were sampled from a Vietnamese family following mother-children transmission (Hannoun et al. 2000). Overall, we found only 21 sequences with no detectable recombination and 29 identified as recombinant by only one method.

In the case of the 35 avian hepatitis B viruses, no genome sequence was entirely clear of recombination. Two sequences were identified as recombinant by all six methods, with GENECONV identifying the highest number of recombinant sequences ($n = 30$). Although recombination is clearly frequent within avian HBV, the small number of HBV sequences available necessitated that all were retained for further analysis.

Perhaps the most notable result of our recombination analysis was its bias toward a specific genomic region (Table 1). In human HBV, approximately 80% of the recombination break points were located in a region (approximately 50%) of the viral genome that covers the pre-S1 region, the pre-C/C gene, the second half of the X gene, and the first part of the polymerase gene (Fig. 1). Hence, our results suggest that this region comprises a “hot spot” for recombination. A similar recombination hot spot was identified in the avian hepadnaviruses and covers orthologous gene regions (pre-S gene and partial S gene, pre-C/C gene, and the first part of the polymerase gene) as human HBV, although avian viruses do not contain the X gene (Fig. 1). We therefore designate the remaining ~50% of the HBV genome as a “low recombination rate” region, although (i) sporadic recombination events may still occur in this region, and (ii) the varied distribution of recombination break-points across the HBV genome means that the demarcation of these regions is only approximate.

Table 1 Analysis of recombination frequency in human and avian HBV^a using RDP2

Method	Human HBV		Avian HBV	
	Positions 1–1626	Positions ^b 1627–3252	Positions ^b 1–1529	Positions 1530–3067
RDP	12 ^c	210	54	4
GeneConv	16	93	99	10
MaxChi	17	61	39	6
BootScan	1	84	70	0
SiScan	47	169	54	19
Chimera	34	113	40	16
Average	21.2	121.7	59.3	9.2

^a Small putative recombinant regions (<100 bp) were excluded

^b Despite their different GenBank numbering, these recombination “hot spot” regions are approximately homologous (see Fig. 1 for the numbering scheme)

^c Number of recombination breakpoints detected

There are a number of reasons why recombination could occur more frequently in our putative recombination hot spot. First, it could be subject to weaker selective constraints than other regions, such that recombinants are more likely to survive the selective sieve once they are generated. However, this scenario seems unlikely given that the recombination hot spot has a high concentration of overlapping reading frames, which will increase selective constraints, and the average pairwise distance (π) in this region is not substantially higher than the genome average (0.121 versus 0.105 and 0.215 versus 0.183 for the human and avian viruses, respectively), as expected with relaxed selective constraints. Alternatively, it may be that recombination rates are higher in the hot spot due to increased positive selection, such that recombination accelerates the generation of advantageous genotypes, although adaptive evolution is only rarely documented in HBV (Fares and Holmes 2002). Finally, it is possible that recombination rates reflect patterns of genome modularity, such that most recombination events occur at gene boundaries (Martin et al. 2005b; Simmonds and Midgley 2005). Although the use of overlapping reading frames in HBV undoubtedly complicates this phenomenon, it is notable that the notional boundary of the recombination hot spot in human HBV is within one nucleotide of the boundary of the P ORF and that the low recombination rate region roughly coincides with the single-stranded (ss) DNA region of the HBV genome (Fig. 1).

Rates of Evolutionary Change in Human HBV

After recombinant sequences were excluded, we were left with a data set of 76 human HBV genome sequences, from which we could estimate rates of nucleotide substitution and the tMRCA using a Bayesian coalescent approach. Given the inherent difficulties in inferring evolutionary rates in HBV we focus not only on the mean rates evolutionary change, but also on the lower 95% HPD values, as these represent the lowest (i.e., most conservative) possible rates of viral evolution.

In all cases except the concatenated nonoverlapping regions of the viral genome, a model of exponential population growth had the highest posterior probability (detailed results presented in Supplementary Table 3). For the nonoverlapping regions of the HBV genome, a logistic growth model was best-fit to the data. Further, in all cases the strict molecular clock was rejected a valid description of the evolutionary dynamics of HBV in favor of a relaxed (uncorrelated) molecular clock. Under the best-fit models the mean evolutionary rate of the 76 complete genomes (global human HBV data) was 7.72×10^{-4} subs/site/year, with a lower 95% HPD value of 8.36×10^{-5} to subs/site/year. These rates are approximately one order of magnitude

higher than those estimated previously and fall closer to the substitution rates observed in RNA than DNA viruses (Hanada et al. 2004; Jenkins et al. 2002). Notably, broadly similar (and high) mean substitution rates were observed in different data subsets: 2.60×10^{-4} subs/site/year for the low recombination rate region of the HBV genome (lower 95% HPD = 3.22×10^{-5} subs/site/year) (Fig. 2) and 3.73×10^{-4} subs/site/year (lower 95% HPD = 4.68×10^{-5} subs/site/year) and 4.30×10^{-4} subs/site/year (lower 95% HPD = 1.16×10^{-4} subs/site/year) for genotypes C and D, respectively (Table 2).

We next considered evolutionary rates in the overlapping and nonoverlapping regions of the HBV genome. As expected, the average pairwise distance of the concatenated nonoverlapping regions ($\pi = 0.14$) is approximately 1.5 times greater than the overlapping regions ($\pi = 0.09$), indicating that the latter is subject to stronger selective constraints. Likewise, the substitution rates in the nonoverlapping region are approximately twice those of the overlapping region (Table 2) but still close to those estimated for the remainder of the viral genome. The π value for the concatenated third codon positions from the nonoverlapping regions was greater still (0.48), and produced the highest rate of evolutionary change observed in our analysis of human HBV, at 9.63×10^{-4} , although with very wide HPD values (lower 95% HPD = 4.13×10^{-5} subs/site/year). Finally, it is noteworthy that our estimates of mean substitution rates under the strict molecular clock (although a less favored description of the data) range from 2.29×10^{-5} to 1.57×10^{-4} subs/site/year, so that they are both lower and closer to those HBV rates estimated previously.

Our observation that HBV evolves relatively rapidly is not unexpected given its reliance on reverse transcriptase-mediated replication, in which intrinsic mutation rates of ~ 0.2 per genome per replication have been reported (Drake et al. 1998). Thus, although arranging genomes into overlapping reading frames is likely to reduce the long-term substitution (fixation) rate in HBV, it is unlikely to be sufficient to reduce this rate by the orders of magnitude required for this to have codiverged with hominoid species (Bollyky and Holmes 1999). Indeed, the evolutionary rate in the overlapping region of the viral genome is only about half that in the nonoverlapping region.

Evolutionary Dynamics of Avian Hepadnaviruses

Our study also considered, for the first time, the evolutionary dynamics of avian hepadnaviruses, focusing on the rates of evolutionary change and the tMRCA of the available samples. The evolutionary rates in avian HBV estimated here are similar to those observed in human HBV (Table 2). Specifically, our mean estimate for the substi-

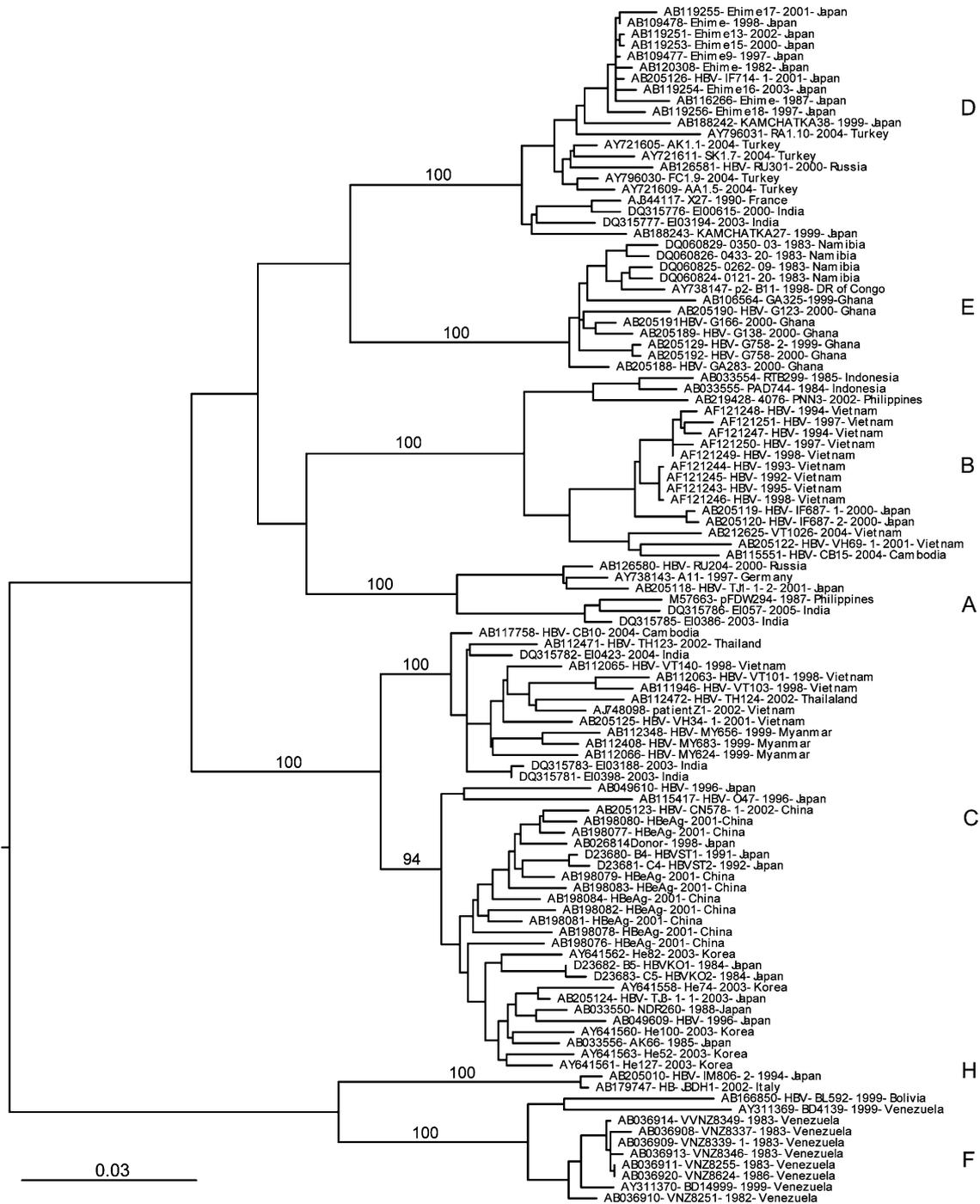


Fig. 2 Phylogenetic tree of the low recombination rate region (~50%) of the human HBV genome ($n = 108$). The tree is midpoint rooted for clarity only and all horizontal branch lengths are drawn to a scale of nucleotide substitutions per site. Bootstrap support values (>90%) are shown for key nodes and the genotypes of human HBV are indicated. The year of sampling is given in the isolate name

tution rate in the complete genome of avian HBV is 7.32×10^{-4} subs/site/year (lower 95% HPD = 1.56×10^{-5} subs/site/year) with an estimated tMRCA at 811 years (upper 95% HPD = 2335 years). Substitution rates were again lower in the low recombination rate region of the viral genome at 4.85×10^{-4} subs/site/year (lower 95%

HPD = 9.20×10^{-6} subs/site/year), which resulted in a deeper mean tMRCA of 1351 years (upper 95% HPD = 5389 years) for the viruses sampled here. Further evidence for a recent evolutionary history in avian hepadnaviruses is that those isolates sampled from duck species (the major hosts of avian HBV) are intermixed with isolates sampled

Table 2 Substitution rates and tMRCA for human and avian HBV sequences

Data set	Substitution rate, $\times 10^{-4}$ subs/site/year (95% HPD)	Age, yr (95% HPD)
Human global ($n = 76$)	7.72 (0.84–8.61)	229 (64–580)
Overlapping regions ($n = 76$)	2.97 (0.55–5.54)	282 (69–697)
Nonoverlapping regions ($n = 76$)	5.72 (2.01–9.12)	237 (85–461)
3rd codon positions, nonoverlapping regions ($n = 76$)	9.63 (0.41–19.20)	477 (57–1509)
Genotype C ($n = 26$)	3.73 (0.47–7.15)	96 (23–271)
Genotype D ($n = 22$)	4.30 (1.16–7.26)	66 (26–140)
Low recombination rate region ($n = 108$)	2.60 (0.32–5.19)	394 (88–1127)
Avian global ($n = 35$)	7.32 (0.16–17.04)	811 (65–2335)
Avian low recombination rate region ($n = 108$)	4.85 (0.09–12.60)	1350 (69–5389)

from other species, including geese and storks, indicative of frequent cross-species transmission (Fig. 3). However, it is also clear that a far larger sample of viruses is required to fully understand the evolutionary dynamics of avian hepadnaviruses.

The tMRCA of Human HBV

We next inferred the tMRCA of human HBV using the same Bayesian coalescent approach (Table 2). Under the strict (constant) molecular clock, estimates for the tMRCA of the sampled isolates of human HBV were highly variable, although most fell in the range 2000–4000 years. These dates are similar to those previously estimated for HBV, all of which similarly utilized a strict molecular clock (Fares and Holmes, 2002; Hannoun et al. 2000; Okamoto et al. 1987; Orito et al. 1989; Osioy et al. 2006). However, far more recent tMRCA estimates were obtained using our best-fit relaxed molecular clock: 229 years (upper 95% HPD = 580 years) for our 76 whole-genome sequence data set, 394 years (upper 95% HPD = 1127 years) for the low recombination rate region, 282 years (upper 95% HPD = 697 years) for the concatenated overlapping regions, 237 years for the concatenated nonoverlapping regions (upper 95% HPD = 461 years), 477 years (upper 95% HPD = 1509 years) for the third codon positions from the concatenated nonoverlapping regions, 96 years (upper 95% HPD = 271 years) for genotype C, and, finally, 65 years (upper 95% HPD = 140 years) for genotype D. Thus, our estimate of the tMRCA for the low recombination rate region is almost twice that of the whole genome (as was also the case in the analysis of avian HBV), indicating that recombination, especially at the intergenotypic level, strongly affects estimates of both the rate of nucleotide substitution and the age of genetic diversity. Hence, it is clearly important to purge recombinant sequences from any evolutionary analysis of HBV gene sequence data. As a corollary, we suggest that

the results of the low recombination rate region of the HBV genome are more reliable.

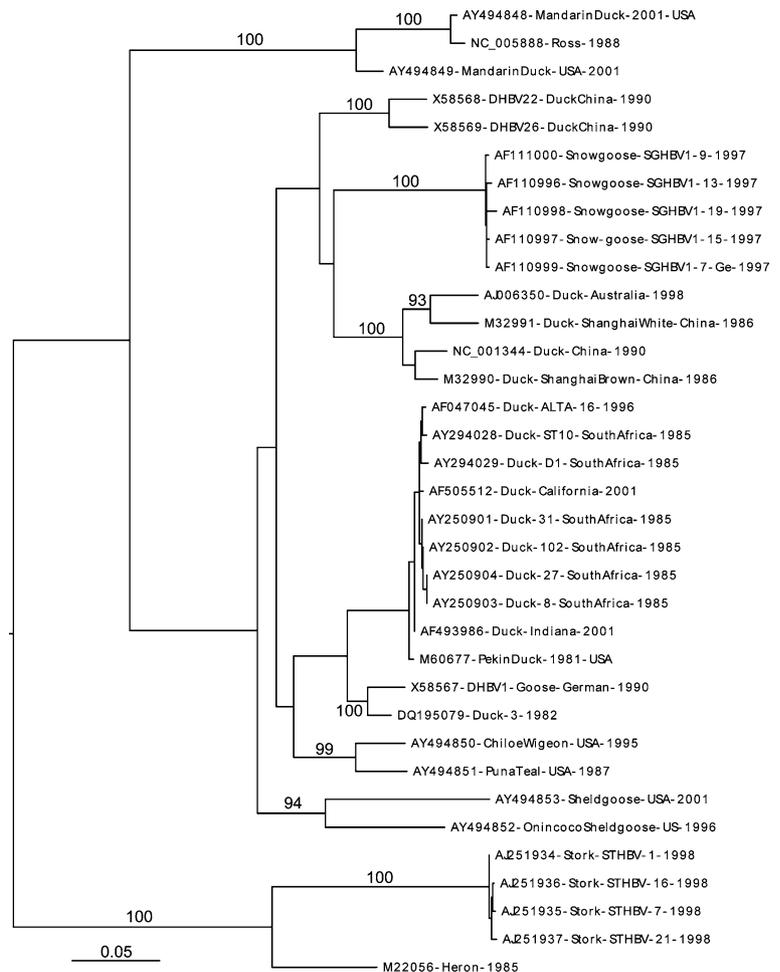
Although there is considerable variation in our estimates of the tMRCA of human HBV, the oldest possible estimate from our analysis (under a nonfavored model) was 7228 years (third codon positions, nonoverlapping regions under a model of constant population size, and a relaxed molecular clock; upper 95% HPD = 23,610 years). Although this represents a deeper time estimate than most studies of HBV evolution, it is still clearly incompatible with the hypothesis that the timescale of HBV evolution follows patterns of hominoid speciation and the subsequent dispersal of anatomically modern humans. Further, the mix of human and nonhuman primate isolates on phylogenetic trees of HBV provides strong evidence against the codivergence theory (Starkman et al. 2003). As such, the results presented here strongly suggest that the evolutionary history of HBV is one in which cross-species transmission has been the norm, with the current distribution of relatively rapidly evolving genotypes following patterns of modern human dispersal. Finally, a similarly shallow evolutionary history was observed in the case of avian HBV, although the implications of this recent history are less clear and the small sample size means that the full effects of recombination have yet to be accounted for.

Acknowledgments We thank Dr. K. Abe, Dr. R. Chakravarty, Dr. Y. Tanaka, Dr. H. Okamoto, Dr. F. Pujol, Dr. Y. Muraki, Dr. K. Fujiwara, Dr. Y. Wen, Dr. P. Karayiannis, Dr. J. Newbold, Dr. W. Hans, Dr. J. Giambone, and Dr. M. Kew for kindly providing additional information concerning the sequences they generated.

References

- 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* 42:97–192

Fig. 3 Phylogenetic tree of the avian hepadnaviruses ($n = 35$). The tree is midpoint rooted for clarity only and all horizontal branch lengths are drawn to a scale of nucleotide substitutions per site. Bootstrap support values (>90%) are shown for key nodes. The year of sampling is given in the isolate name



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

Cui C, Shi J, Hui L, Xi H, Zhuoma, Quni, Tsedan, Hu G (2002) The dominant hepatitis B virus genotype identified in Tibet is a C/D hybrid. *J Gen Virol* 83:2773–2777

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148:1667–1686

Drummond AJ, Rambaut A (2003) BEAST v1.0. Available at: <http://www.evolve.zoo.ox.ac.uk/beam/>

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22:1185–1192

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 88:699–710

Fares MA, Holmes EC (2002) A revised evolutionary history of hepatitis B virus (HBV). *J Mol Evol* 54:807–814

Hanada K, Suzuki Y, Gojobori T (2004) A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol* 21:1074–1080

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

Jenkins GM, Rambaut A, Pybus OG, Holmes EC (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol* 54:156–165

Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 5:150–163

Kurbanov F, Tanaka Y, Fujiwara K, Sugauchi F, Mbanya D, Zekeng L, Ndemi N, Ngansop C, Kaptue L, Miura T, Ido E, Hayami M, Ichimura H, Mizokami M (2005) A new subtype (subgenotype) Ac (A3) of hepatitis B virus and recombination between genotypes A and E in Cameroon. *J Gen Virol* 86:2047–2056

Luo K, Liu Z, He H, Peng J, Liang W, Dai W, Hou J (2004) The putative recombination of hepatitis B virus genotype B with pre-C region of genotype C. *Virus Genes* 29:31–41

Martin DP, Williamson C, Posada D (2005a) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21:260–262

Martin DP, van der Walt E, Posada D, Rybicki EP (2005b) The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet* 51:475–479

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

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. *Japan J Exp Med* 57:231–236
- Olinger CM, Venard V, Njayou M, Oyefolu AO, Maiga I, Kemp AJ, Omilabu SA, le Faou A, Muller CP (2006) Phylogenetic analysis of the precore/core gene of hepatitis B virus genotypes E and A in West Africa: new subtypes, mixed infections and recombinations. *J Gen Virol* 87:1163–1173
- 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
- Osiowy C, Giles E, Tanaka Y, Mizokami Y, Minuk GY (2006) Molecular evolution of hepatitis B virus over 25 years. *J Virol* 80:10307–10314
- Simmonds P (2001) Reconstructing the origins of human hepatitis viruses. *Phil Trans R Soc Lond B* 356:1013–1026
- Simmonds P, Midgley S (2005) Recombination in the genesis and evolution of hepatitis B virus genotypes. *J Virol* 79:15467–15476
- Suwannakarn K, Tangkijvanich P, Theamboonlers A, Abe K, Poovorawan Y (2005) A novel recombinant of Hepatitis B virus genotypes G and C isolated from a Thai patient with hepatocellular carcinoma. *J Gen Virol* 86:3027–3330
- Swofford D (2003) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4 ed. Sinauer Associates, Sunderland, MA
- Starkman S, MacDonald DM, Lewis JCM, Holmes EC, Simmonds P (2003) Geographic and species association of hepatitis B virus genotypes in non-human primates. *Virology* 314:381–393
- Wang Z, Liu Z, Zeng G, Wen S, Qi Y, Ma S, Naoumov NV, Hou J (2005) A new intertype recombinant between genotypes C and D of hepatitis B virus identified in China. *J Gen Virol* 86:985–990
- World Health Organization (2000) Hepatitis B: World Health Organization fact sheet 204. Available at: <http://www.who.int/mediacentre/factsheets/fs204/en/>
- Yang J, Xing K, Deng R, Wang J, Wang X (2006) Identification of Hepatitis B virus putative intergenotype recombinants by using fragment typing. *J Gen Virol* 87:2203–2215