Molecular epidemiology and genetic diversity of hepatitis B virus genotype E in an isolated Afro-Colombian community

Mónica Viviana Alvarado Mora,1 Camila Malta Romano,2 Michele Soares Gomes-Gouveia,1 Maria Fernanda Gutierrez,3 Flair José Carrilho1 and João Renato Rebello Pinho1

Correspondence
Mónica Viviana Alvarado Mora
monica.viviana@usp.br

1Laboratory of Gastroenterology and Hepatology, São Paulo Institute of Tropical Medicine and Department of Gastroenterology, School of Medicine, University of São Paulo, Brazil
2Laboratory of Virology, São Paulo Institute of Tropical Medicine, Department of Infectious and Parasitic Diseases, School of Medicine, University of São Paulo, Brazil
3Laboratory of Virology, Department of Microbiology, Pontificia Javeriana University Bogotá, Colombia

Hepatitis B virus (HBV) infection is a significant public health concern with 350 million chronic carriers worldwide. Eight HBV genotypes (A–H) have been described so far. Genotype E (HBV/E) is widely distributed in West Africa and has rarely been found in other continents, except for a few cases in individuals with an African background. In this study, we characterized HBV genotypes in Quibdó, Colombia, by partial S/P gene sequencing, and found, for the first time, HBV/E circulating in nine Afro-Colombian patients who had no recent contact with Africa. The presence of HBV/E in this community as a monophyletic group suggests that it was a result of a recent introduction by some Afro-descendent contact or, alternatively, that the virus came with slaves brought to Colombia. By using sequences with sampling dates, we estimated the substitution rate to be about 3.2×10⁻⁴ substitutions per site per year, which resulted in a time to the most recent common ancestor (TMRCA) of 29 years. In parallel, we also estimated the TMRCA for HBV/E by using two previously estimated substitution rates (7.7×10⁻⁴ and 1.5×10⁻⁵ substitutions per site per year). The TMRCA was around 35 years under the higher rate and 1500 years under the slower rate. In sum, this work reports for the first time the presence of an exclusively African HBV genotype circulating in South America. We also discuss the time of the entry of this virus into America based on different substitution rates estimated for HBV.

INTRODUCTION

Hepatitis B virus (HBV) is estimated to cause chronic infection in more than 350 million people worldwide and death in 1 million per year. Nearly 20% of the chronic carriers live in Africa with excessively high prevalence rates reported, especially from the sub-Saharan region (Hübschen et al., 2008). HBV strains have a distinct geographical distribution, and are traditionally classified into eight genotypes, A to H, on the basis of genome diversity (Stuyver et al., 2000). Genotype A is found mainly in North and West Europe, North America and Africa. Genotypes B and C are prevalent in south-east Asia and the Far East. Genotype D has a worldwide distribution and is found predominantly in the Mediterranean region. Genotypes E and F are prevalent in West Africa and in the Amerindian population, respectively (Magnius & Norder, 1995). Recently, genotype G has been reported in the USA and France (Stuyver et al., 2000), and genotype H has been found in Central America (Arauz-Ruiz et al., 2002).

In Africa, viruses belonging to five genotypes, A (HBV/A) to E (HBV/E), have been found. An extensive study of genotypes A, B, C and D in South Africa has been reported (Bowyer et al., 1997). Genotype E was first described in 1992 (Norder et al., 1992). Despite its wide geographical spread and high prevalence, genotype E viruses in Africa reveal a surprisingly low diversity. The mean diversity over the whole genome is 1.75%, compared with 4% diversity for African genotype A (Andernach et al., 2009).

HBV/E is by far the most prevalent genotype in Mali, Burkina Faso, Nigeria, Togo, Benin, Central African Republic and the Democratic Republic of the Congo
Genotype E has not been found outside of Africa, except for a few rare cases mostly in individuals with an African background. In Europe, sporadic genotype E strains were reported from France (Halfon et al., 2006), Italy (Palumbo et al., 2007), Spain (Toro et al., 2006), Belgium (Liu et al., 2001) and The Netherlands (van Steenbergen et al., 2002). Others cases were reported in Argentina (Mathet et al., 2007), Brazil (Sitnik et al., 2007) and the USA (Kato et al., 2004).

The analysis of DNA polymorphisms in the β-globin gene cluster mainly focuses on those linked to the HBB*S gene, given that this mutation is absent among native American populations and was introduced into the American continent basically by the gene flow from Africa during the Atlantic slave trade from the 16th to the 19th century (Lemos Cardoso & Farias Guerreiro, 2006). Data described by Cuellar-Ambrosi et al. (2000) showed that the distribution of haplotypes from the western region of Colombia (which has the largest Afro-descendent population in the country) had typical African haplotypes: Bantu, 55.5%; Benin, 34.8%; Senegal, 4.3%; and Cameroon, 5.4%. These results closely agree with the historical data that indicate that most African slaves brought to Colombia originated from Angola (Bantu population) and São Tomé island in the Bight of Benin (Central West Africa).

Although these studies demonstrate the African origin of this population, they have not addressed the origins of HBV in this population. Also, there are no previous studies reporting the distribution of HBV genotypes in the Quibdó community. The aim of this study is to report the presence of genotype E of HBV in an Afro-descendent community from Quibdó, Colombia, and to determine the possible origin of its presence in this population using phylogenetic and Bayesian analysis.

**RESULTS**

**Genotyping analysis**

Nine samples were positive for the S region (416 bp) by nested PCR, but only five of them were also amplified in the S/POL region (734 bp) (GenBank accession nos GQ487690–GQ487698). To perform the phylogenetic analysis, the missing nucleotides were coded as ‘missing characters’ in nexus block. The longest fragment available from each sample was sequenced and classified into genotype E by maximum-likelihood (ML) reconstruction (Fig. 1). Interestingly, the reconstruction of the nucleotides that change along the ML tree showed that all the Quibdó genotype E sequences have a synapomorphy of 2 nt, strongly suggesting that these sequences originated from a unique lineage introduced into this community in the past (Fig. 2). On the S ORF, these substitutions were non-synonymous and were found at position 133 [‘a’ determinant; Met (ATG) for Leu (TTG)] and at position 161 [Phe (TTC) for Tyr (TAC)]. On the Pol ORF, the first substitution at position 141 was also non-synonymous [Tyr (TAT) for Phe (TTT)], while the second substitution (162 aa) was synonymous [Ile (ATT) for Ile (ATA)]. Nevertheless, the two substitutions were localized out of all the known polymerase domains.

**Bayesian analysis**

**Evolutionary dynamics of HBV/E.** The mean rate of nucleotide substitution for HBV/E using 51 sequences for which sampling dates were known was estimated to be around 3.2 × 10⁻⁴ substitutions per site per year within highest probability densities (HPDs) (2.2 × 10⁻⁴–4.52 × 10⁻⁴). Under this substitution rate, the estimated TMRCA for the origin of this genotype was around 29 years ago when only HBV S/POL sequences were used; this increased to 90 years ago when non-recombinant regions of HBV genomes were utilized (Table 1).

As already expected, a huge variation of TMRCAs was found for HBV/E by using the two different previously estimated substitution rates of 7.7 × 10⁻⁴ and 1.5 × 10⁻⁵ substitutions per site per year (Table 1). Under the higher rate, the TMRCA of genotype E estimated from the complete genomes was 30 years, but this value went up to 1536 years with the slower rate. The variation was also large when the TMRCA was evaluated using the non-recombinant region (the first 1799 nt of the HBV genome), ranging from 37.4 to 1907 years. Finally, the two datasets comprising the most conserved region of the virus (S/POL) revealed a TMRCA ranging between 24.4 and 1224 years or 9.7 and 469.5 years, when only genotype E sequences were utilized. The estimates in Table 1 correspond to the values obtained from the best-fit rate and molecular clock chosen by Bayes Factor comparison. The strict molecular clock best-fitted the dataset of the non-recombining region, and the relaxed uncorrelated lognormal was the best molecular clock for all the other HBV datasets. Therefore, there was little variation among estimates under distinct clock models (data available upon request).

**DISCUSSION**

**HBV/E – what is the origin?**

We showed, for the first time, the presence of HBV genotype E circulating inside a South American popu-
Fig. 1. ML phylogenetic tree of the S/Pol region of the HBV genome (n=188). The collapsed clades correspond to the non-E genotypes (A, B, C, D, F, G and H). The values of posterior probability (>0.9) are shown for key nodes. The nine HBV/E Quibdó sequences are indicated by a bracket. Cen Afr Rep, Central African Republic; SEN, Senegal; DRC, the Democratic Republic of the Congo; IVORY, the Ivory Coast.
lation. There have been many studies to characterize the HBV genotype prevalence in Afro-descendent populations in South America; in the Afro-Venezuelan population, genotypes A and F have already been found. These results might reflect the introduction of genotype A during the slavery period and the absence of the African genotype E supports the hypothesis of a recent origin for this genotype, which is after the time of slavery (Quintero et al., 2002). Also, in Afro-Brazilian communities, all HBV isolates belonged to genotype A, subtype A1 (Motta-Castro et al., 2005, 2008), which suggests that HBV was introduced from Africa to different South American countries during the slave trade.

Fig. 2. Multiple alignment of amino acid sequences of HBsAg of the nine Colombian sequences compared with others obtained from GenBank. The major hydrophilic region (100–160), including the ‘a’ determinant (124–147), is boxed (dotted line). The two synapomorphies are located at positions 133 and 161 of the HBV S gene (boxed, solid line).
Moreover, as far as we know, the Afro-Colombian Congo, the Ivory Coast and Cameroon, as seen in Fig. 1, from Nigeria, Togo, Benin, the Democratic Republic of the Colombian viruses constitute a closer cluster to sequences introduced in the times of slavery, we found that the genotype in Colombia. First, under our estimate of support the two hypotheses about the origin of this evolutionary rate of HBV, it is difficult to estimate the substitution rate of HBV/E. Furthermore, we also estimated TMRCA using the two previously published substitution rates. However, due to the large difference in these rates, our results on the evolutionary dynamics can support the two hypotheses about the origin of this genotype in Colombia. First, under our estimate of 3.2 × 10^−4 substitutions per site per year, the TMRCA for genotype E sequences was around 30 years ago. The similar substitution rate estimated by Zhou & Holmes (2007) also suggested a recent entry of HBV/E into this community. In fact, this hypothesis is more consistent with the accepted time estimated for the origin of genotype E in general. However, by considering a recent origin for HBV/E and consequently, a more recent introduction in America (less than 30 years ago), it is quite unexpected that the viruses found in Colombia are different from any other lineage found in Africa so far. Critically, this last hypothesis is also supported by the estimates of TMRCA using the slower rate and, by the monophyletic cluster (with high posterior probability) of nine sequences isolated from Quibdó (Fig. 2), which shows that a single virus entered the community, establishing a founder effect, and has remained there until now.

Variability in the substitution rate

The substitution rate estimated (3.2 × 10^−4 substitutions per site per year) in this work for genotype E only was obtained using the same methodology used by Zhou & Holmes (2007) and the results were very similar to their findings and also to those of Zaaijer et al. (2008), even though they used a different approach. The first study estimating the substitution rate of HBV reported a rate between 1.4 and 3.2 × 10^−5 substitutions per site per year (Okamoto et al., 1987), obtained by dividing the amount of divergence accumulated in sequences obtained from a single patient by the total length of the sequence, assuming the common ancestor for the viruses at the patient’s birth. By using a similar approach, other studies based on pairwise sequence comparison sampled at different time points estimated a median substitution rate of 5.1 × 10^−4 (Zaaijer et al., 2008) and 7.9 × 10^−5 (Osiowy et al., 2006) substitutions per site per year. Similar evolutionary rates have also been estimated for other Hepadnaviruses such as duck hepatitis B virus: 0.8 and 4.5 × 10^−5 substitutions per site per generation (Pult et al., 2001).

In this work, we used a coalescent-based approach to estimate the substitution rate and the TMRCA of HBV/E. This method was also used previously to estimate the overall substitution rate for HBVs using different datasets (Zhou & Holmes, 2007). There are several advantages to using Bayesian approaches to calculate the rate of nucleotide substitution of pathogens. The first reason is that the estimates are based on thousands of phylogenetic trees instead of a unique distance value based on pair-wise
distances. Also, the method allows the use of matrices of substitution models that correct for distinct probabilities of substitution through the sequences (Huelsenbeck et al., 2001). However, determining the rate of sequence change for HBV is difficult due to its complex organization. More than half of the HBV genome consists of overlapping frame-shifted ORFs. Moreover, overlapping genes and secondary structures of the HBV genome involved in the regulation of replication impose a constraint on the number and nature of substitutions occurring in the genome (Zaaijer et al., 2008). Consequently, the best nucleotide substitution rate would be estimated from the third codon positions, with no overlapping regions. In fact, this has already been done and the values were around 9.63 × 10^{-4} substitutions per site per year. However, due to the large variation in the HPDs obtained from this dataset (from 0.4 to 19.2 × 10^{-4}), this specific rate may not be suitable (Zhou & Holmes, 2007). Another important factor influencing the evolution of HBV is the phase (chronic or acute) of natural infection. Zaaijer et al. (2008) noticed that the substitution rate in the HBV S gene was inversely related to the level of viraemia, with no substitutions occurring in the majority of highly viraemic HBV carriers during decades of cumulative follow-up.

In summary, this work reports, for the first time, the presence of HBV/E circulating in a small community in Colombia. Moreover, this virus appears to have experienced a unique entry into this population, since we found two particular nucleotide changes (synapomorphies) only in Quibdó sequences. We also attempt to estimate the evolutionary rate of HBV/E. The current available data and the substitution rate estimated in this work suggest a recent origin for HBV, but the great variability and the worldwide distribution of HBV genotypes argue against this theory. As an accurate substitution rate for this virus is crucial to understand its evolutionary origins, we argue that further analysis of HBV evolutionary patterns should be done, including relevant clinical and serological data.

METHODS

Genotyping analysis

Study population. Nine positive samples for the surface marker of HBV (HBsAg) were obtained from sera stored at -20 °C using the HBsAg and anti-hepatitis B core antigen serological test. Samples came from different hospitals from this city and confirmed these samples were positive in the previous step, using the primers described by Gomes-Gouvea et al. (2009).

HBV DNA extraction. HBV DNA extraction was carried out from 100 μl serum for each sample using the acid guanidinium thiocyanate (GT)/phenol/chloroform method described by Chomczynski & Sacchi (1987). Briefly, 300 μl GT solution was added to each sample. Ice-cold chloroform (50 μl) was added, followed by homogenization and centrifugation. The supernatant was transferred to a conical tube and precipitated with 300 μl cold ethanol. After discarding the ethanol, samples were dried at 94 °C for 1 min, resuspended in 50 μl ultrapure MilliQ water and stored at -20 °C.

HBV DNA amplification: S and S/Pol regions. To avoid false-positive results, strict procedures for nucleic acid amplification diagnostic techniques were followed (Kwok & Higuchi, 1989). Samples were first amplified with the primers described by Sitnik et al. (2004) in order to get a 416 bp fragment partially covering the HBsAg coding region (S). A fragment of 734 bp partially comprising HBsAg and Pol coding regions (S/Pol) was then amplified from the samples that were positive in the previous step, using the primers described by Gomes-Gouvea et al. (2009).

HBV sequencing. Amplified DNA was purified using ChargeSwitch PCR clean-up kit. Sequencing was performed in an ABI Prism 377 Automatic Sequencer (Applied Biosystems) based on the protocol described by Sanger et al. (1977), using dideoxy nucleoside triphosphates containing fluorescent markers (Big Dye terminator v3.1 cycle sequencing ready reaction kit; Applied Biosystems). The quality of each electropherogram was evaluated using the Phred-Phrap software (Ewing et al., 1998; Ewing & Green, 1998) and consensus sequences were obtained by alignment of both sequenced strands (sense and antisense) using CAP3 software, available at http://asparagin.cenargen.embrapa.br/phph/.

Evolutionary analysis

Phylogenetic analysis. Initially, the nine sequences obtained in this work were genotyped by phylogenetic reconstructions using reference sequences from each HBV genotype obtained from GenBank (n=139) (data available upon request). After genotyping, another dataset was constructed comprising a range of the most variable sequences of genotype E in order to obtain the best classification of the Colombian sequences within this group (n=188). These sequences comprising partial S/Pol coding regions were aligned using CLUSTAL_X software (Thompson et al., 1997) and manually edited in the Se-Al software (available at http://tree.bio.ed.ac.uk/software/seal/). ML phylogenetic trees were inferred using the GARLi program (Zwickl, 2006), which employs an extensive branch-swapping protocol and optimizes the substitution model iteratively during the search. The evolutionary model of DNA substitution and initial parameters used in GARLi were estimated by MODELTEST v.3.7 (Posada & Crandall 1998).

We also mapped the nucleotide changes along the ML tree using the MacClade v.4.7 software (Maddison & Maddison, 2003).

Bayesian analysis

Data preparation. Initially, a large dataset comprising 256 complete genomes of all HBV genotypes was constructed with sequences obtained from GenBank (M. V. Alvarado Mora and others, unpublished data). Since there is evidence of recombination among HBV sequences, we undertook a detailed search for recombination using the RDP3 program (Martin, 2009) to exclude potential recombinants from the study. From this, additional datasets were obtained comprising different genome regions: (i) 232 sequences of the HBV complete genome without recombining sequences; (ii) 256 sequences with 1799 nt (without the recombining region of pre-core/C); (iii) 256 sequences with 834 nt of the S/Pol region from all genotypes; and (iv) 156 HBV/E sequences with 834 nt of the S/Pol region (datasets available from authors upon request). In order to estimate the substitution rate for genotype E, an additional dataset comprising 51 HBV/E sequences containing information about the sampling date was constructed.

Estimating evolutionary dynamics. The TMRCA in years was estimated by using the Bayesian Markov Chain Monte Carlo
ACKNOWLEDGEMENTS

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; 2007/53457-7 and 2008/50461-60), São Paulo, SP, Brazil, and by Pontificia Universidad Javeriana, Bogotá, Colombia. We thank the Laboratorio de Salud Publica, Quibdó, Colombia, for kindly providing the samples for this study.

REFERENCES


