Virus evolution during chronic hepatitis B virus infection as revealed by ultradeep sequencing data

Leandro R. Jones,1,2 Mariano Sede,1,3 Julieta M. Manrique1,2 and Jorge Quarleri1,3

Correspondence
Jorge Quarleri
quarleri@fmed.uba.ar
Leandro R. Jones
lrj000@gmail.com

1Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917 (C1083ACA) Buenos Aires, Argentina
2Laboratorio de Virología y Genética Molecular, Facultad de Ciencias Naturales sede Trelew, Universidad Nacional de la Patagonia San Juan Bosco, 9 de Julio y Begrano S/N (9100) Trelew, Chubut, Argentina
3Instituto de Investigaciones Biomédicas en Retrovirus y Sida (INBIRS), Facultad de Medicina, Universidad de Buenos Aires, Paraguay 2155-Piso 11 (C1121ABG) Buenos Aires, Argentina

Despite chronic hepatitis B virus (HBV) infection (CHB) being a leading cause of liver cirrhosis and cancer, HBV evolution during CHB is not fully understood. Recent studies have indicated that virus diversity progressively increases along the course of CHB and that some virus mutations correlate with severe liver conditions such as chronic hepatitis, cirrhosis and hepatocellular carcinoma. Using ultradeep sequencing (UDS) data from an intrafamilial case, we detected such mutations at low frequencies among three immunotolerant patients and at high frequencies in an inactive carrier. Furthermore, our analyses indicated that the HBV population from the seroconverter patient underwent many genetic changes in response to virus clearance. Together, these data indicate a potential use of UDS for developing non-invasive biomarkers for monitoring disease changes over time or in response to specific therapies. In addition, our analyses revealed that virus clearance seemed not to require the virus effective population size to decline. A detailed genetic analysis of the viral lineages arising during and after the clearance suggested that mutations at or close to critical elements of the core promoter (enhancer II, epsilon encapsidation signal, TA2, TA3 and direct repeat 1-hormone response element) might be responsible for a sustained replication. This hypothesis requires the decline in virus load to be explained by constant clearance of virus-producing hepatocytes, consistent with the sustained progress towards serious liver conditions experienced by many CHB patients.

INTRODUCTION

Following acute hepatitis B virus (HBV) infection, the risk of progression to chronicity is age dependent, with up to 5 % of adults and almost 95 % of children born to chronically infected mothers becoming chronic carriers (McMahon et al., 1985; Tassopoulos et al., 1987). Chronic HBV infection (CHB) may progress through four phases known as the immune-tolerant, immune-clearance, inactive or non-replicative, and reactivation or immune-escape phases (Kwon & Lok, 2011). These phases do not occur in all individuals and may not be sequential. Those patients exposed to HBV by mother-to-child transmission go through the immune-tolerant phase, which may last for 20–30 years, and is characterized by hepatitis B e antigen (HBeAg) positivity, high HBV DNA levels (> 20 000 IU ml⁻¹), scarce or null changes in alanine aminotransferase (ALT) levels and normal/minimal histological activity. The immune-clearance phase (or HBeAg-positive chronic hepatitis) is characterized by the presence of HBeAg, high serum HBV DNA levels, persistently or intermittently increased ALT levels and active inflammation in the liver. During this phase, spontaneous HBeAg seroconversion occurs at a rate of 10–20 % per year. HBeAg seroconversion is frequently but not always accompanied by a sudden increase in ALT levels. These changes are the result of the immune awakening of the host, culminating in seroconversion from HBeAg to anti-HBe, and entry into the inactive carrier phase. At this time, an extremely low or undetectable level of HBV DNA is found, accompanied by normal ALT levels, and normal or near-normal liver histology. For unclear reasons, a subset of patients experiences virus reactivation – the reactivation or immune-escape phase – which is characterized by

The GenBank accession numbers for the haplotype sequences reported here are KP813865–KP813996.

Five supplementary figures and five tables are available with the online Supplementary Material.
HBV DNA levels >2000 IU ml⁻¹ and fluctuating ALT levels (Kwon & Lok, 2011).

HBV circulates in the blood as a mixture of genetic variants known as quasispecies (Osiowy et al., 2006; Zhou & Holmes, 2007). Many of these variants harbour mutations that are detrimental for virus fitness, while others possess mutations that confer replication advantages, facilitate immune escape or cause resistance to antiviral drugs. During the immune-clearance phase, mutations in the pre-core (pC) and core promoter (CP) regions are frequently selected in HBeAg-negative patients. The CP regulates the transcription of pre-genomic RNA and pC mRNA. It is a single regulatory region that includes two major functional elements: the upper regulatory region and the basic core promoter (BCP). CP variants can be found in all HBV genotypes, although they are most commonly associated with genotype C, whereas the genotype studied here was genotype D1 (Chotiyaputta & Lok, 2009).

Recent studies have shown that viral mutations gradually accumulate along the course of CHB (Cheng et al., 2013; Lim et al., 2007; Sede et al., 2013). In addition, it is well known that viruses isolated from severe liver disease cases usually display characteristic mutations (Kramvis & Kew, 1999). These data strongly suggest that virus diversity assessment might be useful for monitoring disease status and predicting patient outcome. Therefore, the present work set out to study the tempo and mode of CP evolution, a region of the genome deeply involved in regulation of both virus replication and gene expression, in a familiarly transmitted case, in order to evaluate the potential of ultra-deep sequencing (UDS) data for monitoring CHB. To this aim, we performed a strict and thorough filtering of the pyrosequencing data described by Sede et al. (2014), and the processed reads obtained were submitted to in-depth evolutionary analyses. Intrafamilial HBV transmission represents an outstanding scenario for analysing HBV molecular evolution along the course of CHB, because the virus has the opportunity to evolve in multiple but homogeneous host backgrounds. Sequence data were available for two sampling times for the family mother (MA and MB) and single sampling times for her daughter (D) and sons (S1 and S2). The mother experienced a switch from the immune-tolerant to the inactive-carrier phase between sampling times MA and MB, whereas the children were anchored in the immune-tolerant phase. We analysed the intra- and inter-patient phylogenetic patterns and inferred the corresponding virus mutational histories. We also used Bayesian coalescent analyses to incorporate a temporal frame into these analyses.

RESULTS

Virus diversity is heavily enhanced in response to immune clearance

HBV sequences from the mother’s samples (MA during clearance and MB after she entered the inactive-carrier phase) presented 36 and 70 haplotypes, respectively, compared with 24, 21 and 18 haplotypes detected among the sequences from D, S1 and S2 (Fig. 1a). Fig. 1(b) depicts a CorreLogo three-dimensional (3D) snapshot that summarizes virus variability and the mutual information in the whole dataset. Around 85% of the viral sequences from the mother displayed a deletion relative to the rest of the sequences, located between two TA-rich regions at the BCP (TA2 and TA3; Fig. 1c). The deletion encompassed 8 bp (positions 1763–1770) in all the haplotypes except for three low-frequency haplotypes (haplotypes 50, 79 and 107), in which the deletion was extended to a ninth position (position 1762). Other than that, nucleotide substitutions were distributed rather homogeneously along the pC/C region, except where the pX and pC/C genes overlap. Interestingly, 18 pairs of positions that in all cases involved position 1659 at enhancer II or position 1796, which is part of the epsilon encapsidation signal, displayed significant mutual information (Fig. 1b). These correlations included substitutions that might either allow or hamper potential DNA or RNA base pairs (represented by colour codes in Fig. 1b), suggesting the existence of complex interactions between particular nucleotide positions of the CP.

Many viral haplotypes were exclusive to MA and/or MB, i.e. haplotypes 3, 6, 7, 9, 11, 16 and 19 (Fig. 1d, Table S1, available in the online Supplementary Material). In addition, the haplotypes that presented the higher frequencies in the children (haplotypes 1, 2, 4 and 5) were under-represented in the mother. These differences were highly significant after standard statistical contrasts were performed among MA and MB, as well as among MA or MB against D, S1 and S2 (χ² test, P<0.01; Table S1). In addition, the results were confirmed by principal coordinate analysis of the haplotype frequencies distribution, which located MA and MB far from D, S1 and S2 in the plane defined by the first two principal coordinates (Fig. 1e). Phylogenetic analysis revealed a profusion of divergent viral sequences at the sampling times MA and MB from the mother. Some of these sequences were clustered into relatively divergent clades that we named lineages 1–6 (Fig. S1). Bootstrap supports and posterior probabilities were quite low, with the exception of lineages 1, 5 and 6 (Fig. S2, Table S2). However, a detailed cladistic analysis offered additional support to these groupings (see Supplementary Material). Nevertheless, all the phylogenetic tests performed within this study were implemented using bootstrap trees in order to weight the impact of phylogenetic uncertainty.

Immune clearance produces no significant impact on virus effective population size

The seroconverter patient virus load decreased by several orders of magnitude between the sampling times MA and MB, suggesting that the virus effective population size might have been affected by the clearance. However, this
The idea was challenged by the rampant virus diversification observed for this patient (Fig. 1a), which suggested the prevalence of significant replication rates. Thus, we compared the constant population size, exponential growth and Bayesian skyline demographical models by using a Bayes factor analysis. The Bayesian skyline model was slightly favoured, and the corresponding Bayesian skyline plot (BSP) suggested a slight increment of the viral population size in the mother (Table 1, Fig. 2a). However, the Bayes factor analysis did not provide conclusive evidence in favour of the Bayesian skyline model and thus this result must be considered tentative. Notwithstanding this, the constant population size model could not be rejected by the Bayes factor analysis, strongly suggesting that the abrupt viral load drop observed for the mother was not associated with a decrease in the virus effective population size (Table 1). In addition, the Bayesian dated trees showed that the divergent lineages from the mother originated during the period between 54 months prior to and 40 months after the sampling time MA, based on the

---

**Fig. 1.** Virus diversity in familiarly transmitted HBV. (a) Raw number of reads and the corresponding number of haplotypes identified in the children (D, S1 and S2) and the two sampling times from the mother (MA and MB). (b) CorreLogo snapshot of the sequences from the four patients studied. Only the variable positions are shown. The 3D grey arrows point from the 5’ to the 3’ ends of the alignment. The lateral bars [two-dimensional (2D) logos] describe the sequence conservation of each alignment position in terms of bits. The stacks inside the grid (3D logos) represent pairs of alignment columns with high mutual information. The red, orange and yellow stacks indicate complementary base pairs (G-C, A-T/U and G-T/U, respectively). The grey bars indicate SD of the information represented in the 2D logos and the mutual information in the 3D ones. Stacks corresponding to mutual information values of <0.5 bits are not shown. (c) Sequence logo representation of genomic positions 1750–1775. Around 85% of the mother sequences (upper logo) displayed a deletion that affected the positions indicated by thinner symbols. The deletion was absent from the children’s viruses (lower logo). The TA-rich pC/C transcription initiation signals TA1, TA2 and TA3 are indicated by shaded boxes. (d) Haplotype distribution among the study patients and sampling times. Haplotype frequencies are depicted only for haplotypes that presented frequencies >100, for graphical reasons (haplotype numbers are given on the x-axis and the corresponding frequencies on the y-axis). (e) Principal coordinate analysis of the viral populations from the children and the two sampling times of the mother. Colour codes are as in (d). Co1, component 1; Co2, component 2. The upper and lower diagrams correspond to principal coordinates analyses based on Morisita and Raup–Crick indices, respectively.
corresponding 95 % highest posterior density (HPD) intervals (Fig. 2b, Table S3), whereas a much more remote origin was inferred for the deleted strains (mean 247 months, 95 % HPD 136–381; Fig. 2b, Table S3). Thus, the Bayesian analyses supported an active and continued cladogenesis, which requires replication to generate new virus variants.

**Table 1.** Bayesian evidence for constant population (CP), exponential (Exp) and BSP models

Positive values indicate a better model fit of the row’s model compared with the column’s model. Simpler models were preferred in the absence of significant Bayes factors (B \(>4\sim5\); Kass & Raftery, 1995).

| Patient | Model | \(\ln P(\text{model|data})\) | \(\text{SEM}\) | CP | Exp | BSP |
|---------|-------|-----------------|----------|----|-----|-----|
| All*    | CP    | -1660.326       | ±0.303   | –  | –   | –   |
|         | Exp   | -1647.44        | ±0.283   | 5.596 | –    | –   |
|         | BSP   | -1643.24        | ±0.254   | 7.42 | 1.824 | –   |
| Mother  | CP    | -1171.274       | ±0.106   | –  | -1.391 | -2.719 |
|         | Exp   | -1168.072       | ±0.108   | 1.391 | –    | -1.328 |
|         | BSP   | -1165.013       | ±0.107   | 2.719 | 1.328 | –   |

*The analyses were performed with all the sequences (All) or with only the sequences from the mother (Mother).
Evidence of significant amounts of intra-patient evolution

Despite the presence of many haplotypes that were shared among the four patients (e.g. haplotypes 1, 2, 4 and 5; Figs 1d and S1), all the patients also presented virus lineages that were seemingly exclusive, as reflected by the clustering of reads of the same colour in Fig. S1. This indicated that, despite the patients’ proximity, each patient dataset provided independent estimates of CHB evolution. To assess for statistical significance, we evaluated the level of inter-patient phylogenetic structuring. To this aim, we borrowed two metrics used in community ecology, the mean phylogenetic distance (MPD) and the mean nearest taxon distance (MNTD). Here, patients are analogous to communities and virus haplotypes to taxa, as defined by Webb et al. (2002). The whole family corresponded to a regional species pool (Webb et al., 2002). Both weighted and unweighted analyses indicated that the viral populations were structured (Fig. 2c). The significant overdispersion observed for the MB sampling time in the unweighted analysis was a consequence of the emergence of low-frequency variants in association with the clearance process, in agreement with the Bayesian coalescent analyses (Fig. 2b, Table S3).

Immune clearance induces variations in CD4+ T-cell epitopes

The studied HBV genomic region overlaps with a domain encoding parts of the HBx and HBc proteins. These regions have two major CD4+ T-cell antigenic determinants, 126EIRLKVFVLGGCRHK140 and 1MDIDPYKEFGATVELLSFLP20, respectively (Carman et al., 1997; Malmassari et al., 2007). The frequency of the WT HBx epitope and its variants showed an almost exclusive impact on viral sequences from the mother. A high predominance of a deleted HBx epitope was inferred, with the simultaneous presence of several mutated epitopes at positions 127, 130 and 131, at both sampling times from the mother (MA and MB) (Table S4). The HBc epitope appeared mutated only at the second sampling time of the mother (MB), showing three different variants (T5P, S12T and T5P+S12T) of low frequency (0.014, 0.009 and 0.032).

Virus mutations related to advanced liver disease seem to accumulate along the course of CHB

Many of the virus mutations identified here affected genomic positions frequently associated with chronic hepatitis, liver cirrhosis and hepatocellular carcinoma (Table 2) (Yin et al., 2011; Zhang et al., 2013). Remarkably, such mutations were present in both the inactive carrier patient and the immunotolerant ones, supporting the idea of a continuous diversification along the course of CHB, accompanied by the emergence of mutants associated with serious liver conditions. In order to characterize in depth the occurrence of nucleotide substitutions at these positions, we evaluated the proportions of mutated reads at positions related to cirrhosis and hepatocellular carcinoma in HBV genotypes B and C (Table 2). The most recent common ancestor state (interpreted as WT nucleotide) was assigned consensus nucleotides.

<table>
<thead>
<tr>
<th>Position†</th>
<th>MRCA†</th>
<th>Children</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td>1673</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>1719</td>
<td>T</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>1726</td>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>1730</td>
<td>C</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>1731</td>
<td>A</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>1753</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1762</td>
<td>A</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>1774</td>
<td>C</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>1799</td>
<td>C</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1846</td>
<td>C</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>1913</td>
<td>C</td>
<td>A</td>
<td>C</td>
</tr>
</tbody>
</table>

†Genomic position.

‡Most recent common ancestor state (interpreted as WT nucleotide), determined by character optimization in the maximum-likelihood tree. Positions 1762, 1764 and 1766, which presented indels, were assigned consensus nucleotides.
and other positions, we inferred the minimal number of substitutions required by each site given the observed sequences and trees (hereafter referred to as evolutionary paths). In addition, we dissected the substitutions that occurred in the seroconverter patient subpopulation from those that occurred in the immunotolerant patients. Since the evolutionary path of a given site depends on the underlying phylogeny, the entire procedure was multiplied along 100 bootstrap trees in order to weight the effect of phylogenetic uncertainty (see Supplementary Material for further details). Positions 1659, 1762 and 1896 were much more variable than the rest of the positions analysed, although one of these sites (position 1762) was only variable in the mother. As expected, many other sites were more diverse for the mother. Positions 1669, 1673, 1676, 1678, 1701, 1719, 1726, 1727, 1730, 1739, 1752, 1757, 1776, 1783 and 1960 displayed mutations for the viruses from the mother but were invariable among the children’s virus populations. Positions 1753, 1762, 1764, 1771, 1799, 1802, 1803, 1846, 1863, 1913, 1915, 1934 and 1951 were variable for the mother and children but were more variable for the mother’s viruses. However, positions 1685, 1707, 1712, 1761, 1775, 1810, 1849, 1855 and 1942 were more variable for D, S1 and S2 than for MA and MB. Thus, mutations occurred in both the seroconverter patient and the immune-tolerant patients, albeit at a different pace. These results, together with confidence intervals for the mean evolutionary paths (interpreted as plausible ranges of variation), are summarized in Fig. 3 and detailed in Table S5.

**DISCUSSION**

UDS data from four people from the same family (mother and children) were analysed in order to study HBV evolution during CHB. The mother passed through the immune-clearance phase during the follow-up, and thus

---

**Fig. 3.** Position-wise diversity of familiarly transmitted HBV. The y-axis represents the minimum amount of evolutionary change required to explain the nucleotide patterns observed along the studied genomic region (x-axis). The mother and children populations are colour coded as indicated in the figure. The upper panel corresponds to a diversity profile obtained from the maximum-likelihood tree. The lower panel depicts the mean (diamonds) and SD (whiskers) of the same metric obtained from 100 bootstrap trees. The genes, regulatory regions and two CD4+ T-cell recognition sequences spanned by the studied sites are depicted in grey over the diversity profiles: pX, protein X coding region; PC/C, pre-core/core gene coding region; C, core gene coding region; EN II, enhancer II; CURS, core upstream regulatory sequence (the alpha, beta, gamma and delta, CURS motifs are indicated); BCP, basic core promoter; TA1–4, initiation sites of the PC/C (TA1–3) and C (TA4) transcripts, HRE, hormone response element binding site; DR1, direct repeat 1, POLY, unique polyadenylation signal, CD4+, CD4+ T-cell recognition sites.
two sampling points collected over 3 years were available corresponding to the clearance and inactive-carrier stages. A strong viral diversification accompanied the clearance, characterized by the emergence of several exclusive lineages. Virus mutations also accumulated among the children (who were anchored in the immune-tolerant stage), although at a slower pace. Remarkably, many of the affected genomic positions in both the mother and children are frequently mutated in HBV from patients with chronic hepatitis, liver cirrhosis and hepatocellular carcinoma.

The enhanced diversity observed during and after virus clearance (Figs 1d, S1 and S3) is counterintuitive if one considers the concomitant virus load fall, which suggests a diminished viral replication. In addition, previous studies have shown that some CP mutations frequently observed in advanced disease cases can significantly enhance in vitro virus replication (reviewed by Kramvis & Kew, 1999). These data suggest that the diversity augmentation observed in advanced stages of CHB may be driven by higher genome replication rates, which is compatible with the sustained effective population size supported by our Bayesian analyses (Table 1, Fig. 2a). In fact, deletions homologous to those detected for MA and MB (Fig. 1c) have been shown to potentiate virus replication and pre-genomic RNA transcription, despite the presence of low in vivo levels of HBsAg, HBeAg and HBeAg (Kohno et al., 2000; Moriyama, 1997; Buckwold et al., 1996). The deletion observed here affected the TA-rich sites TA2 and TA3, and the which is a binding site for liver-enriched factor (LEF), hepatocyte nuclear factor 4 (HNF4) and testicular orphan receptor 4 (TR4). Thus, the mutation can lead to reduced transcription of pre-core mRNA without repressing transcription of pre-genomic RNA (Kohno et al., 2000; Lin et al., 2003). In addition, three mutations (T1753C/A, 1762T/G and 1764A) were present at MB that can produce a fourfold increase in virus replication rates (Parekh et al., 2003). This scenario requires the decline in virus load to be explained by a constant clearance of virus-producing hepatocytes, consistent with the liver injuries observed in many CHB patients and the presence of CD4+ T-cell epitope mutations observed here (Table S4).

Our evolutionary analyses indicated that the HBV population from the seroconverter patient underwent many genetic changes in response to virus clearance (Figs 1d, e, 2a, b and S1). Furthermore, all the study patients presented virus mutations, many of which have been shown recently to correlate with serious liver conditions (Table 2) (Yin et al., 2011; Zhang et al., 2013). The discovery of such mutations in immune-tolerant patients is novel and possibly has been hampered to date due to the limited sensitivity of traditional genotyping approaches (i.e. molecular cloning followed by Sanger sequencing). Nonetheless, the accumulation of mutations along the course of CHB observed here is consistent with the gradual virus diversification observed previously for genotype B-infected patients (Cheng et al., 2013; Lim et al., 2007). Thus, we propose a model for viral haplotype dynamics in which mutations related to severe liver conditions tend to accumulate along the course of CHB (Fig. 4). These data also suggest that virus diversity and/or the frequency to mutations at particular genomic sites might correlate positively with the risk of developing serious liver pathologies. In this sense, deep sequencing may have the potential to determine the status and predict the outcome of CHB patients. Also, the early administration of antiviral treatment may prevent or delay the emergence of virus mutations, with a potential positive effect on patient prognosis. Further studies are needed to corroborate these hypotheses, including larger numbers of patients, extended periods of time and detailed surveys of clinical parameters.

**METHODS**

**Patients.** The intrafamilial HBV transmission case studied here has been described in detail elsewhere (Sede et al., 2014). HBV sequences were obtained from a mother and her three offspring, a daughter (D) and two sons (S1 and S2). Phylogenetic analyses demonstrated that the viruses belonged to genotype D1 and provided robust evidence for intrafamilial transmission. The mother attended medical consultation in May 2007, displaying elevated ALT levels (233 U l⁻¹) and HBV virus load (> 1.1E7 IU ml⁻¹) levels, and was also seropositive for HBeAg, indicating that the patient was likely to be undergoing the immune-clearance phase. The first sample from the mother (MA) was taken at this time, and a second sample was taken in November 2010 (MB). At the latter time, she presented normal ALT levels, a much lower virus load (~ 8000 IU ml⁻¹) and was anti-HBeAg seropositive, indicating that the patient had probably entered the inactive-carrier phase. Samples from the three children were taken in April (S1) and August (S2 and D) 2011. All displayed HBeAg-positive serology, normal ALT levels (21, 21 and 23 U l⁻¹, respectively) and high virus loads (> 1.1E7 IU ml⁻¹), compatible with the immune-tolerant phase...

![Fig. 4. Model for haplotype dynamics during CHB. Chronically infected patients represent the principal source of new infections (represented by bold arrow in the figure). In this model, acute and early chronic infections are established mainly by viruses that harbour WT sequences. Intra-patient evolution leads to late mutations, those that are common in advanced CHB, which accumulate along the course of the infection. The process is driven by the quasi-species cloud, which provides a constant source of genetic variants. The WT viruses are fitter in acutely infected patients and early stages of chronic infection, whereas the genotypes harbouring late mutations are fitter in advanced CHB stages.](http://jgv.microbiologyresearch.org)
Dataset and haplotype analyses. We reprocessed and reanalysed the pyrosequencing data described by Sede et al. (2014), which encompass genomic positions 1639–1976. The raw sequence output generated by the Roche/454 GS-FLX platform was processed through the native amplicon pipeline. The obtained reads were further processed with the program PRINSEQ, which depurates pyrosequencing data based on the read quality (Schmieder & Edwards, 2011). After this, the data were filtered by the error correction algorithm implemented in the program KEC, which corrects outlier sequences based on k-mer frequencies and identifies haplotypes based on the corresponding results (Skums et al., 2012). Given that KEC returns a corrected set of sequences, we did not use this output directly. Instead, we identified all the original, PRINSEQ-processed reads that matched exactly with the KEC-identified haplotypes. This procedure was implemented to ensure that all the reads submitted to sequence analysis were represented among the original reads, i.e. were confirmed empirically. Prior to this final step, a round of visual inspection was applied to the KEC haplotypes in order to identify and eliminate any possible haplotype containing regions suspicious of being artefacts related to homopolymer length ambiguity (Huse et al., 2007). As a final screening, we searched for potential recombinants among the reads that passed all the filtering steps described. Potentially chimeric sequences were identified with RECCO (Maydt & Lengauer, 2006). This program attempts to explain each sequence in an alignment by introducing substitutions in or proposing recombination events among the rest of the sequences from the dataset. RECCO gives each sequence a score defined in terms of ‘savings’, by counting how many evolutionary steps are saved by assuming recombination. As we were interested in removing any possible trace of recombination from our data, we discarded all the sequences that fulfilled a very low cut-off (two savings). Overall, 18 150 reads were analysed, of which 3576 and 2621 corresponded to MA and MB, respectively, 3733 to D, 3543 to S1 and 4677 to S2. Haplotype frequencies were obtained with the program mothur (Schloss et al., 2009). Principal coordinates analysis was used to compare the haplotype distribution among patients and sampling times, using distance matrices generated with the Raup–Crick and Morisita similarity indices. The Raup–Crick distance is a probabilistic index based on the presence/absence of data, whereas the Morisita distance also takes into consideration the abundance of each haplotype (Woldà, 1981). The analysis was carried out using Past software (http://folk.uio.no/ohammer/past/).

Sequence alignment. Sequence alignments were obtained with MAFFT (Katoh & Standley, 2013). The gap opening and extension parameters were left at their default values. The alignments were performed with iterative refinement and weighted sum-of-pair scores and a consistency score obtained from local alignments (Katoh & Standley, 2014). Sequence alignments were inspected by CorreLogo (Bindewald et al., 2006). This software provides a 3D representation of the properties of a sequence alignment, consisting of a grid delimited by two-dimensional logos equivalent to standard sequence logo representations (Schneider & Stephens, 1990), plus stacks located in each of the grid cells that represent the mutual information of each pair of alignment positions. Thus, the program helps to identify correlations between bases and potential RNA and DNA base pairs.

Phylogenetic and meta-phylogenetic analyses. Maximum-likelihood phylogenetic trees were obtained with the program FastTree (Price et al., 2010). This program explores the space of trees by using up to $4 \times \log_2(N)$ rounds of minimum-evolution nearest-neighbour interchange (NNI), two rounds of subtree–pruning–regrafting moves and up to $2 \times \log_2(N)$ rounds of maximum-likelihood NNIs, where $N$ is the number of unique sequences in the dataset. As described in Results, around 85% of the viral sequences from the mother displayed a deletion that encompassed the same 8 bp segment (positions 1763–1770) in all of the affected sequences, except for three low-frequency haplotypes in which the mutation affected a ninth position (position 1762). Thus, we deduced that the 8 bp deletion evolved once and that it was further extended to a ninth position in the three low-frequency haplotypes, i.e. the deleted reads are monophyletic. However, preliminary phylogenetic analyses did not argue with this assumption, which we attributed to the fact that the phylogenetic methods used treated gaps as missing data, equivalent to dismissing the information provided by gaps, with the consequent risk of generating biased results (Simmons, 2014; Warnow, 2012). Thus, the information provided by the deletion was incorporated into the analysis by imposing a monophyletic constraint on the sequences harbouring the deletion. In addition, 14 potentially recombinant haplotypes (haplotypes 72, 78, 80, 81, 112–114, 116, 119, 120–123 and 126) were removed from the dataset to avoid biases in phylogenetic analyses (Schierup & Hein, 2000). Branch supports were evaluated by the bootstrap and Bayesian posterior probabilities. For poorly supported clades (see Results), we implemented a tree metric with the aim of measuring the proximity of a given set of terminals in a set of bootstrap trees, assuming that a topological proximity in the bootstrap trees would be observed for truly closely related operational taxonomic units (OTUs; please see Supplementary Material).

Standardized effect sizes of phylogenetic structure on MPD and MNTD were obtained by subtracting the distances obtained from randomized communities, i.e. the haplotypes distributed randomly among the patients, from the distances in the patients and dividing by the SD of the metric in the null data, using the routines implemented in the picante package (Kembel et al., 2010). Thus, negative values of these standardized effects corresponded to clumped phylogenetic distributions of virus haplotypes.

Dated phylogenies were obtained with BEAST (Bouckaert et al., 2014). Based on substitution rates and models determined previously for non-overlapping genomic regions, a log-normal molecular clock was set with a rate of $4.8 \times 10^{-5}$ substitutions per site per month (Zhou & Holmes, 2007). Bayes factor analyses were performed with Tracer (http://tree.bio.ed.ac.uk/software/tracer/). We used an equivalent prior on all the dated clades to ensure a minimal, and equivalent, influence of the prior on the outcome of the analyses. We considered that the most likely origin of the virus populations studied was a contagion of the mother in her childhood. Thus, we set this scenario as our a priori belief by a log-normal distribution with $M=6.38$ and $S=1.25$. The huge amount of sequences generated by UDS precluded us from performing coalescent Bayesian analyses with the full dataset. Thus, we performed the analyses with the unique sequences obtained from each patient ($n=155$). Convergence and effective sample size were assessed with Tracer.

Evolutionary paths (Felsenstein, 2004; Fitch, 1971) were inferred by an $r$ script that combined different tools provided in the packages ape, phangorn, adephylo and picante (Jombart et al., 2010; Kembel et al., 2010; Paradis et al., 2004; Schliep, 2011) (see Supplementary Material and Fig. S4). The evolutionary changes that occurred in the mother’s HBV population as compared with those that occurred within the children’s viruses were differentiated by counting separately the substitutions in the tree branches spanned by each of the two groups of sequences (Fig. S5). Ancestral character states were estimated by phangorn, with the exception of positions with gaps that were assigned the corresponding highest frequency nucleotides.
ACKNOWLEDGEMENTS

This work was supported by the Argentinian National Agency of Scientific and Technological Promotion (grant nos PICT 2012-0422 to J. Q. and PICT PRH-14 120 to L. R. J.), the University of Buenos Aires (grant no. 200201101100034) and the Argentine National Scientific and Technical Research Council (grant no. PIP-112201101101089).

REFERENCES


Sede, M., Lopez-Ledesma, M., Frider, B., Pozzati, M., Campos, R. H., Fliechman, D. & Quarleri, J. (2014). Hepatitis B virus depicts a high degree of conservation during the immune-tolerant phase in

http://jgv.microbiologyresearch.org


