Journal of General Virology

# Viral evolution during chronic hepatitis B as revealed by ultra-deep sequencing data

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | VIR-D-15-00323R1 |
| **Full Title:** | Viral evolution during chronic hepatitis B as revealed by ultra-deep sequencing data |
| **Short Title:** | Viral evolution during chronic hepatitis B |
| **Article Type:** | Standard |
| **Section/Category:** | Animal - Small DNA Viruses |
| **Corresponding Author:** | Leandro Roberto Jones<br>Universidad Nacional de la Patagonia San Juan Boaco<br>Trelew, Chubut ARGENTINA |
| **First Author:** | Leandro R Jones |
| **Order of Authors:** | Leandro R Jones |
| | Mariano Sede |
| | Julieta M Manrique |
| | Jorge Quarleri |
| **Abstract:** | Despite chronic hepatitis B virus (HBV) infection (CHB) is a leading cause of liver cirrhosis and cancer, HBV evolution during CHB isn't fully understood. Recent studies indicate that viral diversity progressively increases along the course of CHB and that some viral mutations correlate with severe liver conditions such as chronic hepatitis, cirrhosis and hepatocellular carcinoma. Using ultra deep 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 great genetic changes in response to viral clearance. Altogether, these data point out a potential use of UDS for developing noninvasive biomarkers for monitoring disease changes over time or in response to specific therapies. In addition, our analyses revealed that viral clearance seems not to require viral effective population size to decline. A detailed genetic analysis of the viral lineages arisen 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 DR1-HRE) might be responsible for a sustained replication. This hypothesis requires viral load declination to be explained by a constant clearance of virus-producing hepatocyte, consistent with the sustained progress towards serious liver conditions experienced by many CHB patients. |

Response to Reviewer

Click here to download Response to Reviewer: response.to.reviewers.docx

# Viral evolution during chronic hepatitis B as revealed by ultra-deep sequencing data

Leandro R. Jones[a,b,1], Mariano Sede[a,c], Julieta M. Manrique[a,b]& Jorge Quarleri[a,c,1]

[a] Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917 (C1083ACA) Buenos Aires, Argentina.
[b] Laboratorio 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.
[c] Instituto de Investigaciones Biomédicas en Retrovirus y Sida (INBIRS), Facultad de Medicina, Universidad de Buenos Aires, Paraguay 2155-Piso 11 (C1121ABG) Buenos Aires, Argentina.

[1] Corresponding author. Dr. Jorge Quarleri, phone: +54 1145083689, fax: 54 1145083705, e-mail: quarleri@fmed.uba.ar. Dr. Leandro R. Jones, phone: +54 2804421782, fax: +54 2804421782, e-mail: lrj000@gmail.com.

Total number of words in the main text of the paper: 4920.
Total number of words in the summary: 221.
Number of figures: 4.
Number of tables: 2.

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

**Abstract**

Despite chronic hepatitis B virus (HBV) infection (CHB) is a leading cause of liver cirrhosis and cancer, HBV evolution during CHB isn't fully understood. Recent studies indicate that viral diversity progressively increases along the course of CHB and that some viral mutations correlate with severe liver conditions such as chronic hepatitis, cirrhosis and hepatocellular carcinoma. Using ultra deep 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 great genetic changes in response to viral clearance. Altogether, these data point out a potential use of UDS for developing noninvasive biomarkers for monitoring disease changes over time or in response to specific therapies. In addition, our analyses revealed that viral clearance seems not to require viral effective population size to decline. A detailed genetic analysis of the viral lineages arisen 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 DR1-HRE) might be responsible for a sustained replication. This hypothesis requires viral load declination to be explained by a constant clearance of virus-producing hepatocyte, 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 hepatitis B (CHB) may progress through four phases known as the immune tolerant, the immune clearance, the inactive or non-replicative, and the 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 (> 20000 IU/mL), 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 HBV-DNA is found, accompanied by normal ALT levels and normal or near normal liver histology. For unclear reasons, a subset of patients experience a viral reactivation, namely reactivation or immune escape phase, that is characterized by HBV-DNA levels > 2000 IU/mL and fluctuating ALT levels (Kwon & Lok, 2011).

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

Recent studies have shown that viral mutations gradually accumulate along 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 viral 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 viral 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 in Sede and colleagues (Sede *et al.*, 2014), and the processed reads obtained were submitted to in-deep evolutionary analyses. Intrafamilial HBV transmission represents an outstanding scenario for analyzing 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 analyzed the intra and inter-patient phylogenetic patterns and inferred the corresponding viral mutational histories. We also used Bayesian coalescent analyses to incorporate a temporal frame into these analyses.

## Results

**Viral diversity was heavily enhanced in response to the immune clearance**. The HBV sequences from the mother samples, MA during clearance and MB after she entered the inactive carrier phase, presented 36 and 70 haplotypes, respectively, compared with the 24, 21, and 18 haplotypes detected among the sequences from D, S1, and S2 (Fig. 1A). Figure 1B depicts a *CorreLogo* 3D snapshot that summarizes viral 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 sequences, located between two TA-rich regions at BCP (TA2 and TA3; Fig. 1C). The deletion encompasses 8 bp (positions 1763 to 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 1896, 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 color codes in Figure 1B), suggesting the existence of complex interactions between particular nucleotide positions of the core promoter.

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). In addition, the haplotypes that presented the higher frequencies in the children (haplotypes 1, 2, 4, and 5) were underrepresented 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 (*chi* square test, $p< 0.01$; Table S1). In addition, the results were confirmed by principal coordinate analysis of the haplotypes' frequencies distribution, which located MA and MB far from D, S1 and S2 in the plane defined by the two first 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 to 6 (Fig. S2). Bootstrap supports and posterior probabilities were quite low, with the exception of lineages 1, 5, and 6 (Fig. S3; Table S2). However, a detailed cladistic analysis offered additional support to these groupings (please 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.

**The immune clearance produced no significant impact on viral effective population size ($N_e$).** The seroconverter patient viral load decreased by several orders of magnitude between the sampling times MA and MB, suggesting that the viral effective population size might have been affected by the clearance. However, this idea was challenged by the rampant viral diversification observed for this patient (Figs. 1A and 2B), which suggests the prevalence of significant replication rates. Thus, we compared the Constant Population Size (CPS), Exponential Growth (EG), and Bayesian Skyline (BS)

demographical models by using a Bayes Factor (BF) analysis. The BS model was slightly favored, 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 BF analysis did not provide conclusive evidence in favor of the BS model and thus this result must be considered as tentative. Notwithstanding that, the CPS model could not be rejected by the BF analysis, strongly suggesting that the abrupt viral load drop observed for the mother was not associated with a decrease of the viral effective population size (Table 1). In addition, the Bayesian dated trees showed that the divergent lineages from the mother have originated during the period between 54 months prior to and 40 months after sampling time MA, based on the corresponding 95% 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). This is to say that the Bayesian analyses supported an active and continued cladogenesis, which requires replication for generating new viral variants.

**There were 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 viral lineages which were seemingly exclusive, as reflected by the clustering of reads of the same color in Figure S1. This indicates that, despite patients' proximity, each patient data provided independent estimates of CHB evolution. To asses 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 with communities and viral haplotypes with taxa, as defined by Webb and colleagues (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 is 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).

**The immune clearance induced 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, 126-EIRLKVFVLGGCRHK-140, and 1-MDIDPYKEFGATVELLSFLP-20, respectively (Carman *et al.*, 1997; Malmassari *et al.*, 2007). The frequency of the wild-type 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 in 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).

**Viral mutations related to advanced liver disease seem to accumulate along the course of CHB.** Many of the viral 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 and other positions, we inferred the minimal number of substitutions required by each site given the observed sequences and trees (hereafter evolutionary paths, EP). In addition, we dissected the substitutions that occurred in the seroconverter patient subpopulation from those that occurred in the immunotolerant patients. Since the EP 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 (please see Supplementary Material for further details). Positions 1659, 1762, and 1896 were much more variable than the rest of the positions analyzed, although one of these sites (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 viral populations. Positions 1753, 1762, 1764, 1771, 1799, 1802, 1803, 1846, 1863, 1913, 1915, 1934 and 1951 were variable for mother and children but were more variable for the mother viruses. However, positions 1685, 1707, 1712, 1761, 1766, 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 ones, albeit at a different pace. These results, together with confidence intervals for mean EPs (interpreted as plausible ranges of variation), are summarized in Figure 3 and detailed in Table S5.

## Discussion

Ultra-deep sequencing (UDS) data from 4 people from the same family (mother and children) were analyzed in order to study HBV evolution during CHB. The mother passed through the immune clearance phase during the follow-up, and thus two sampling points collected along 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. Viral 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 mother and children are frequently mutated in HBV from patients with chronic hepatitis, liver cirrhosis and HCC.

The enhanced diversity observed during and after viral clearance (Figs. 1D, S1 and S3B) is counterintuitive if one considers the concomitant viral 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* viral replication [reviewed in (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 pregenomic RNA transcription, despite the presence of low *in-vivo* levels of HBsAg, HBcAg, and HBeAg (Kohno *et al.*, 2000; Moriyama, 1997). The deletion observed here affected the direct repeat 1-hormone response element (DR1-HRE), TA-rich sites TA2 and TA3, and the direct repeat 1-hormone response element (DR1-HRE), 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 precore mRNA without repressing transcription of pregenomic RNA (Kohno *et al.*, 2000; Lin *et al.*, 2003). In addition, three mutations (T1753C/A, 1762T/G, 1764A) were present at MB that can produce a 4-fold increase in viral replication rates (Parekh *et al.*, 2003). This scenario requires viral load declination to be explained by a constant clearance of viral-producing hepatocytes, consistent with 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 great genetic changes in response to viral clearance (Figs. 1D-E, 2A-B and S1). Furthermore, all the studied patients presented viral mutations, many of which have been shown recently to correlate with serious liver conditions (Table 1) (Yin *et al.*, 2011; Zhang *et al.*, 2013). The discovery of such mutations in immunotolerant patients is novel and it possibly has been hampered so far due to 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 viral diversification observed previously for genotype B infected patients (Cheng *et al.*, 2013; Lim *et al.*, 2007). Thus, we propose a model for viral haplotypes'

dynamics in which mutations related to severe liver conditions tent to accumulate along the course of CHB (Fig. 4). These data also suggest that viral diversity and/or the frequency of mutations at particular genomic sites might positively correlate with the risk of developing serious liver pathologies. In this sense, deep sequencing may have the potential of determining the status and predicting the outcome of CHB patients. Also, the early administration of antiviral treatment may prevent or delay the emergence of viral mutations, with a potential positive effect on patients' prognosis. Further studies are needed to corroborate these hypotheses, including larger numbers of patients, extended periods of time and detailed surveys of clinical parameters.

## Materials and 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 girl (D) and two boys (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 (233 U/L) and HBV viral load (>1.1E7 IU/mL) levels, and was also seropositive for HBeAg, indicating that the patient was likely to be traversing 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 viral load (~8000 IU/mL), 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 of them displayed HBeAg positive serology, normal ALT levels (21, 21, and 23 U/L, respectively), and high viral loads (>1.1E7 IU/mL), compatible with the immune tolerant phase (Kwon & Lok, 2011). Liver biopsies were performed on the family members in 2012. The mother had a Metavir score of A3F4, whereas D, S1, and S2 had scores of A2F2, A2F2, and A2F3, respectively. Informed consents were obtained from the adult and from both parents of the three infants, and the study protocol was approved by the Hospital 'Cosme Argerich' Ethics Committee, Buenos Aires, Argentina.

**Dataset and haplotype analyses.** We reprocessed and reanalyzed the pyrosequencing data described in Sede and colleagues (Sede *et al.*, 2014), which encompasses genomic positions 1639–1976. The raw sequence output generated by the Roche/454 GS-FLX platform was processed through the native amplicon pipeline. Then, the obtained reads were further processed with the program *Prinseq*, which depurates pyrosequencing data based on the reads' quality (Schmieder & Edwards, 2011). After that, 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; that is, were empirically confirmed. 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 artifacts 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 quimeric 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 very low cut-off (2 savings). 18150 reads were analyzed, 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 (PCA) was used to compare the haplotypes' 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 presence/absence data whereas the Morisita one also takes into consideration the abundance of each haplotype (Wolda, 1981). The analysis was carried out using the *PAST* software (available at 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 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 2D logos equivalent to standard sequence logo representations (Schneider & Stephens, 1990), plus stacks located into 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 $4x log_2(N)$ rounds of minimum-evolution NNI, 2 rounds of SPR moves, and up to $2x log(N)$ rounds of maximum-likelihood NNIs, where $N$ is the number of unique sequences in the dataset. As mentioned in the Results section, around 85% of the viral sequences from the mother displayed a deletion that encompassed the same 8-bp segment (positions 1763–1770) in all the affected sequences but 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, which is to say that 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 treat 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 monophyly constraint on the sequences harboring 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 (please see the Results section), 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 shall be observed for truly closely related OTUs (please see Supplementary Text).

Standardized effect sizes of phylogenetic structure on MPD and MNTD were obtained by subtracting the distances obtained from randomized communities, that is the haplotypes distributed randomly among the patients, from the distances in the patients and dividing by the standard deviation 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 correspond to clumped phylogenetic distributions of viral haplotypes.

Dated phylogenies were obtained with *BEAST2* (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 4.8E-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 viral 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 UDPS precluded 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 ESS 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) (please see also the Supplementary text 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.

**Acknowledgments**

This work was supported by the Argentinean National Agency of Scientific and Technological Promotion (grant number PICT 2012-0422 to JQ and PICT PRH-14 120 to LRJ); the University of Buenos Aires (grant number 20020110100034); and the Argentinean National Scientific and Technical Research Council (grant number PIP-11220110101089).

## References


**Bindewald, E., Schneider, T. D. & Shapiro, B. A. (2006).** CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic acids research* **34**, W405-411.

**Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. (2014).** BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10**, e1003537.

**Buckwold, V. E., Xu, Z., Chen, M., Yen, T. S. B. & Ou, J.-H. (1996).** Effects of a Naturally Occurring Mutation in the Hepatitis B Virus Basal Core Promoter on Precore Gene Expression and Viral Replication. *Journal of Virology* **70**, 5845–5851.

**Carman, W. F., Boner, W., Fattovich, G., Colman, K., Dornan, E. S., Thursz, M. & Hadziyannis, S. (1997).** Hepatitis B virus core protein mutations are concentrated in B cell epitopes in progressive disease and in T helper cell epitopes during clinical remission. *J Infect Dis* **175**, 1093-1100.

**Cheng, Y., Guindon, S., Rodrigo, A., Wee, L. Y., Inoue, M., Thompson, A. J., Locarnini, S. & Lim, S. G. (2013).** Cumulative viral evolutionary changes in chronic hepatitis B virus infection precedes hepatitis B e antigen seroconversion. *Gut* **62**, 1347-1355.

**Chotiyaputta, W. & Lok, A. S. (2009).** Hepatitis B virus variants. *Nat Rev Gastroenterol Hepatol* **6**, 453-462.

**Felsenstein, J. (2004).** *Inferring Phylogenies*. Sunderlands, Massachusetts: Sinauer Associates, Inc.

**Fitch, W. M. (1971).** Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool* **20**, 406-416.

**Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. (2007).** Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**, R143.

**Jombart, T., Balloux, F. & Dray, S. (2010).** adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* **26**, 1907-1909.

**Kass, R. E. & Raftery, A. E. (1995).** Bayes Factors. *Jour Am Stat Assoc* **90**:773-795.

**Katoh, K. & Standley, D. M. (2013).** MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780.

**Katoh, K. & Standley, D. M. (2014).** MAFFT: iterative refinement and additional methods. *Methods Mol Biol* **1079**, 131-146.

**Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P. & Webb, C. O. (2010).** Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463-1464.

**Kohno, K., Nishizono, A., Terao, H., Hiraga, M. & Mifune, K. (2000).** Reduced transcription and progeny virus production of hepatitis B virus containing an 8-bp deletion in basic core promoter. *J Med Virol* **61**, 15-22.

**Kramvis, A. & Kew, M. C. (1999).** The core promoter of hepatitis B virus. *J Viral Hepat* **6**, 415-427.

**Kwon, H. & Lok, A. S. (2011).** Hepatitis B therapy. *Nat Rev Gastroenterol Hepatol* **8**, 275-284.

**Laskus, T., Rakela, J., Tong, M. J., Nowicki, M. J., Mosley, J. W. & Persing, D. H. (1994).** Naturally occurring hepatitis B virus mutants with deletions in the core promoter region. *Journal of hepatology* **20**, 837-841.

**Lim, S. G., Cheng, Y., Guindon, S., Seet, B. L., Lee, L. Y., Hu, P., Wasser, S., Peter, F. J., Tan, T., Goode, M. & Rodrigo, A. G. (2007).** Viral quasi-species evolution during hepatitis Be antigen seroconversion. *Gastroenterology* **133**, 951-958.

**Lin, W. J., Li, J., Lee, Y. F., Yeh, S. D., Altuwaijri, S., Ou, J. H. & Chang, C. (2003).** Suppression of hepatitis B virus core promoter by the nuclear orphan receptor TR4. *J Biol Chem* **278**, 9353-9360.

**Malmassari, S. L., Deng, Q., Fontaine, H., Houitte, D., Rimlinger, F., Thiers, V., Maillere, B., Pol, S. & Michel, M. L. (2007).** Impact of hepatitis B virus basic core promoter mutations on T cell response to an immunodominant HBx-derived epitope. *Hepatology* **45**, 1199-1209.

**Maydt, J. & Lengauer, T. (2006).** Recco: recombination analysis using cost optimization. *Bioinformatics* **22**, 1064-1071.

**McMahon, B. J., Alward, W. L., Hall, D. B., Heyward, W. L., Bender, T. R., Francis, D. P. & Maynard, J. E. (1985).** Acute hepatitis B virus infection: relation of age to the clinical expression of disease and subsequent development of the carrier state. *J Infect Dis* **151**, 599-603.

**Moriyama, K. (1997).** Reduced antigen production by hepatitis B virus harbouring nucleotide deletions in the overlapping X gene and precore-core promoter. *J Gen Virol* **78 ( Pt 6)**, 1479-1486.

**Okamoto, H., Tsuda, F., Akahane, Y., Sugai, Y., Yoshiba, M., Moriyama, K., Tanaka, T., Miyakawa, Y. & Mayumi, M. (1994).** Hepatitis B virus with mutations in the core promoter for an e antigen-negative phenotype in carriers with antibody to e antigen. *J Virol* **68**, 8102-8110.

**Osiowy, C., Giles, E., Tanaka, Y., Mizokami, M. & Minuk, G. Y. (2006).** Molecular evolution of hepatitis B virus over 25 years. *J Virol* **80**, 10307-10314.

**Paradis, E., Claude, J. & Strimmer, K. (2004).** APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290.

**Parekh, S., Zoulim, F., Ahn, S. H., Tsai, A., Li, J., Kawai, S., Khan, N., Trepo, C., Wands, J. & Tong, S. (2003).** Genome replication, virion secretion, and e antigen expression of naturally occurring hepatitis B virus core promoter mutants. *J Virol* **77**, 6601-6612.

**Price, M. N., Dehal, P. S. & Arkin, A. P. (2010).** FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490.

**Schierup, M. H. & Hein, J. (2000).** Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-891.

**Schliep, K. P. (2011).** phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593.

**Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. & Weber, C. F. (2009).** Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541.

**Schmieder, R. & Edwards, R. (2011).** Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864.

**Schneider, T. D. & Stephens, R. M. (1990).** Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**, 6097-6100.

**Sede, M., Laufer, N., Ojeda, D., Gun, A., Cahn, P. & Quarleri, J. (2013).** Analysis of sequences of hepatitis C virus NS5A genotype 1 in HIV-coinfected patients with a null response to nitazoxanide or peg-interferon plus ribavirin. *Arch Virol* **158**, 1907-1915.

**Sede, M., Lopez-Ledesma, M., Frider, B., Pozzati, M., Campos, R. H., Flichman, D. & Quarleri, J. (2014).** Hepatitis B virus depicts a high degree of conservation during the immune-tolerant phase in familiarly transmitted chronic hepatitis B infection: deep-sequencing and phylogenetic analysis. *J Viral Hepat* **21**, 650-661.

**Simmons, M. P. (2014).** A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. *Mol Phylogenet Evol* **80**, 267-280.

**Skums, P., Dimitrova, Z., Campo, D. S., Vaughan, G., Rossi, L., Forbi, J. C., Yokosawa, J., Zelikovsky, A. & Khudyakov, Y. (2012).** Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* **13 Suppl 10**, S6.

**Tassopoulos, N. C., Papaevangelou, G. J., Sjogren, M. H., Roumeliotou-Karayannis, A., Gerin, J. L. & Purcell, R. H. (1987).** Natural history of acute hepatitis B surface antigen-positive hepatitis in Greek adults. *Gastroenterology* **92**, 1844-1850.

**Warnow, T. (2012).** Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Curr* **4**, RRN1308.

**Webb, C. O., Ackerly, D. D., McPeek, M. A. & Donogue, M. J. (2002).** Phylogenies and Community Ecology. *Annu Rev Ecol Syst* **33**, 475-505.

**Wolda, H. (1981).** Similarity Indices, Sample Size and Diversity. *Oecologia* **50**, 296-302.

**Yin, J., Xie, J., Liu, S., Zhang, H., Han, L., Lu, W., Shen, Q., Xu, G., Dong, H., Shen, J., Zhang, J., Han, J., Wang, L., Liu, Y., Wang, F., Zhao, J., Zhang, Q., Ni, W., Wang, H. & Cao, G. (2011).** Association between the various mutations in viral core promoter region to different stages of hepatitis B, ranging of asymptomatic carrier state to hepatocellular carcinoma. *Am J Gastroenterol* **106**, 81-92.

**Zhang, Q., Yin, J., Zhang, Y., Deng, Y., Ji, X., Du, Y., Pu, R., Han, Y., Zhao, J., Han, X., Zhang, H. & Cao, G. (2013).** HLA-DP polymorphisms affect the outcomes of chronic hepatitis B virus infections, possibly through interacting with viral mutations. *J Virol* **87**, 12176-12186.

**Zhou, Y. & Holmes, E. C. (2007).** Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *J Mol Evol* **65**, 197-205.

**Table 1.** Bayesian evidence for Constant (CP), Exponential (Exp) and Bayesian Skyline Plot (BSP) models[1].

| | | ln P(model \| data) | S.E. | CP | Exp | BSP |
|---|---|---|---|---|---|---|
| All[2] | CP | -1660.326 | +/- 0.303 | - | -5.596 | -7.42 |
| | Exp | -1647.44 | +/- 0.283 | 5.596 | - | -1.824 |
| | 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 | - |

[1] Positive values indicate better model fit of the row's model compared to the column's model. Simpler models were preferred in absence of significant Bayes Factors (B>4-5; Kass & Raftery, 1995).

[2] The analyses were performed with all the sequences (All) and with only the sequences from the mother (Mother).

**Table 2.** Substitutions observed at positions related to cirrhosis and HCC in HBV genotypes B and C. Proportions of mutated reads are given in parenthesis.

| Pos.[a] | MRCA[b] | Children | Mother | Lineage 1 | Lineage 2 | Lineage 3 | Lineage 4 | Lineage 5 | Lineage 6 |
|---|---|---|---|---|---|---|---|---|---|
| 1673 | C | C | C→T (0.2%) | C | C | C | C | C | C→T (100%) |
| 1719 | T | T | T→G (0.3%) | T | T | T | T | T | T→G (100%) |
| 1726 | A | A | A→C (0.3%) | A | A | A | A | A | A→C (100%) |
| 1727 | A | A | T→T (0.3%) | A | A | A | A | A | A→T (100%) |
| 1730 | C | C | C→G (0.3%) | C | C | C | C | C | C→G (100%) |
| 1753 | T | T | T→C (0.2%); T→G (0.03%) | T→C (100%) | T | T | T | T→G (100%) | T |
| 1762 | A | A | A→T (17%); A→G (0.5%); A→- (0.2%) | A→T (100%) | A→T (18.7%) | A→T (20%) | A→T (11.1%) | A | A |
| 1764 | G | G | G→A (1.8%); G→- (85%) | G→A (100%) | G→- (100%) | G→- (100%) | G→- (100%) | G | G |
| 1766 | C | C→A (5%) | C→-(85%) | C | C→-(100%) | C→-(100%) | C→-(100%) | C | C |
| 1773 | T | T→C (2%) | T→C (3%) | T | T | T | T | T→C (100%) | T→C (100%) |
| 1799 | C | C | C→G (0.8 %) | C | C | C | C→G (100%) | C→G (100%) | C |
| 1846 | A | A | A→T (0.8%) | A | A→T (31.2 %) | A | A→T (44.4%) | A | A |
| 1913 | C | C | C→A (0.6%) | C | C | C→A (58.6%) | C | C | C |

[a] Genomic position.

[b] Most recent common ancestor state (interpreted as wild type nucleotide), determined by character optimization in the maximum likelihood tree. Positions 1762, 1764 and 1766, which presented indels, were assigned consensus nucleotides.

**Figure Legends.**

**Figure 1.Viral diversity in familiarly transmitted HBV. (A)** Raw number of reads and the corresponding number of haplotypes identified in the children (D, S1, S2) and the two sampling times from the mother (MA, MB). **(B)** CorreLogo snapshot of the sequences from the four patients studied. Only the variable positions are shown. The 3D gray arrows point from the 5' to the 3' ends of the alignment. The lateral bars (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 gray bars give standard deviations of the information represented in the 2D logos and the mutual information in the 3D ones. Stacks corresponding to mutual information values < 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 viruses (lower logo). The TA-rich pC/C transcription initiation signals TA1,TA 2, and TA3 are indicated by shaded boxes. **(D)** Haplotype distribution among the studied patients and sampling times. Haplotype frequencies are depicted only for haplotypes that presented frequencies greater than 100, for graphical reasons (haplotype numbers are given in the x axis and the corresponding frequencies in the y one). **(E)** Principal coordinate analysis of the viral populations from the children and the two sampling times of the mother. Color codes are as in panel D. *Co1*, component 1; *Co2*, component 2. The upper and lower diagrams correspond to PCAs based on Morisita and Raup–Crick indices, respectively.

**Figure 2. Meta-phylogenetic evidence for familiarly transmitted HBV.** (A) Bayesian Skyline Plot reconstruction of the demographic history of the mother's HBV population. The *x* axis units are months before the last sampling time. The line corresponds to the median estimate and the shaded areas represent 95% HPD limits. The BSP suggests a slight increase of the viral population size, probably caused by the deregulation of BCP genetic elements due to a deletion observed in ~85% of the sequences from the seroconverter patient and point mutations observed in this patient's haplotypes. **(B)** Posterior samples of tMRCAs (months) of the viral populations from the mother (MO; MA plus MB), children (CH; D plus S1 and S2), and the deleted (DE) and divergent (LI) strains from the mother. **(C)** Standardized effects of phylogenetic structure on Mean Nearest Taxon Distances (MNTD) and Mean Phylogenetic Distances (MPD). Data were obtained for 100 bootstrap trees. Significant dispersion is indicated by circles ($p < 0.01$).

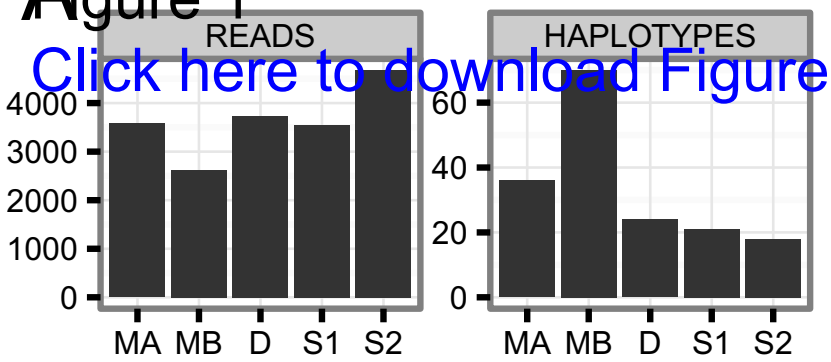
**Figure 3.** Position-wise diversity of familiarly transmitted HBV. The y-axis represents the minimum amounts of evolutionary change required to explain the nucleotide patterns observed along the studied genomic region (x-axis). The mother and children populations are color discriminated as indicated in

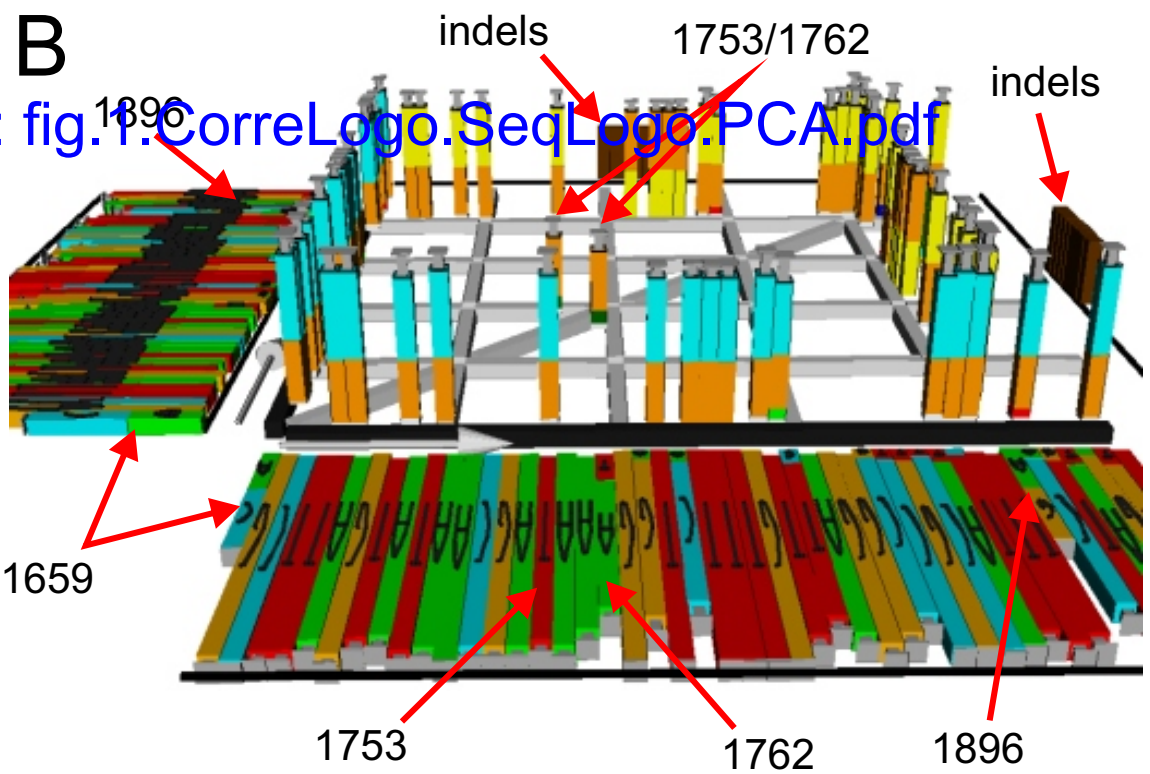
the figure. The upper panel corresponds to a diversity profile obtained from the maximum likelihood tree. The lower panel depicts the mean (diamonds) and standard deviations (whiskers) of the same metric obtained from 100 bootstrap trees. The genes, regulatory regions and two CD4+ T-cell recognition sequences spanned by the studied sites are depicted in gray over the diversity profiles: pX protein X coding region, PC/C precore-core gene coding region, C core gene coding region, EN II enhancer II, CURS core upstream regulatory sequence, BCP basic core promoter, TA1-4 initiation sites of the PC/C (TA1-3) and C (TA 4) transcripts, HRE Hormone Response Element binding site, DR1 direct repeat 1, POLY unique polyadenylation signal, CD4+ CD4+ T-cell recognition sites.

**Figure 4. Model for haplotype dynamics during chronic hepatitis B**. Chronically infected patients represent the principal source of new infections (represented by a ticker arrow in the figure). In this model, acute and early chronic infections are established mainly by viruses that harbor wild-type 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 quasispecies cloud, which provides a constant source of genetic variants. The *wild-type* viruses are fitter in acutely infected patients and early stages of chronic infections, whereas the genotypes harboring *late* mutations are fitter in the advanced CHB stages.

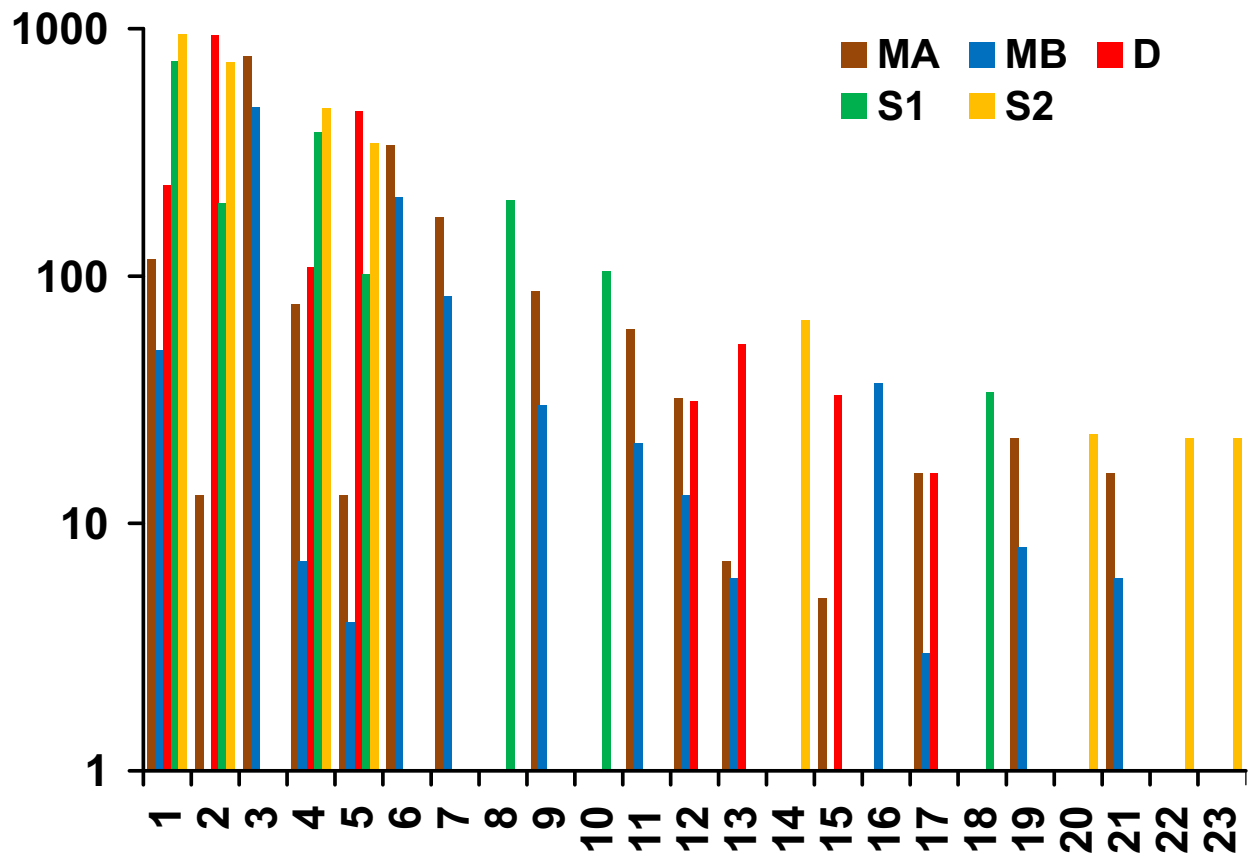
Figure 1

Click here to download Figure: fig.1.CorreLogo.SeqLogo.PCA.pdf







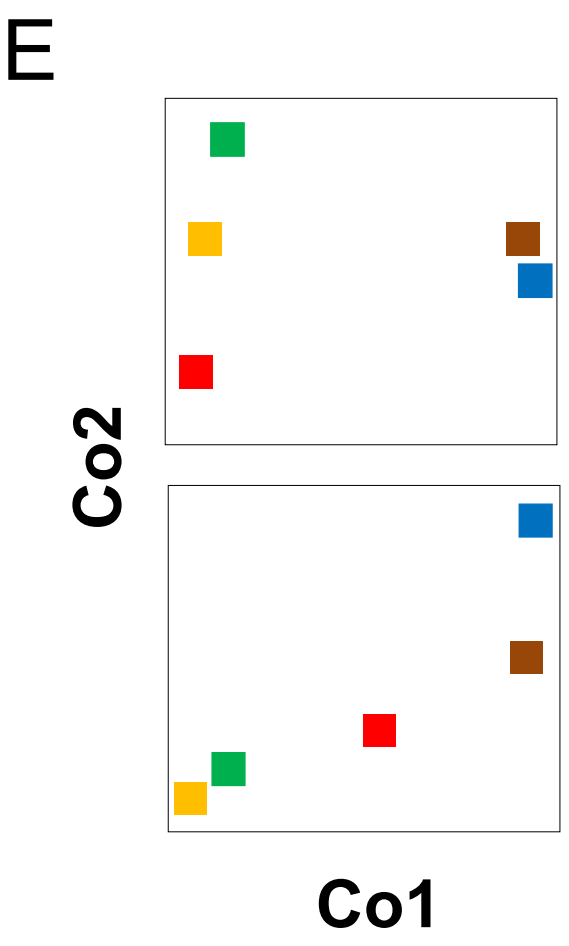



Figure 2

Click here to download Figure: fig.2.R1.Bayes.MNTD-MPD.pdf

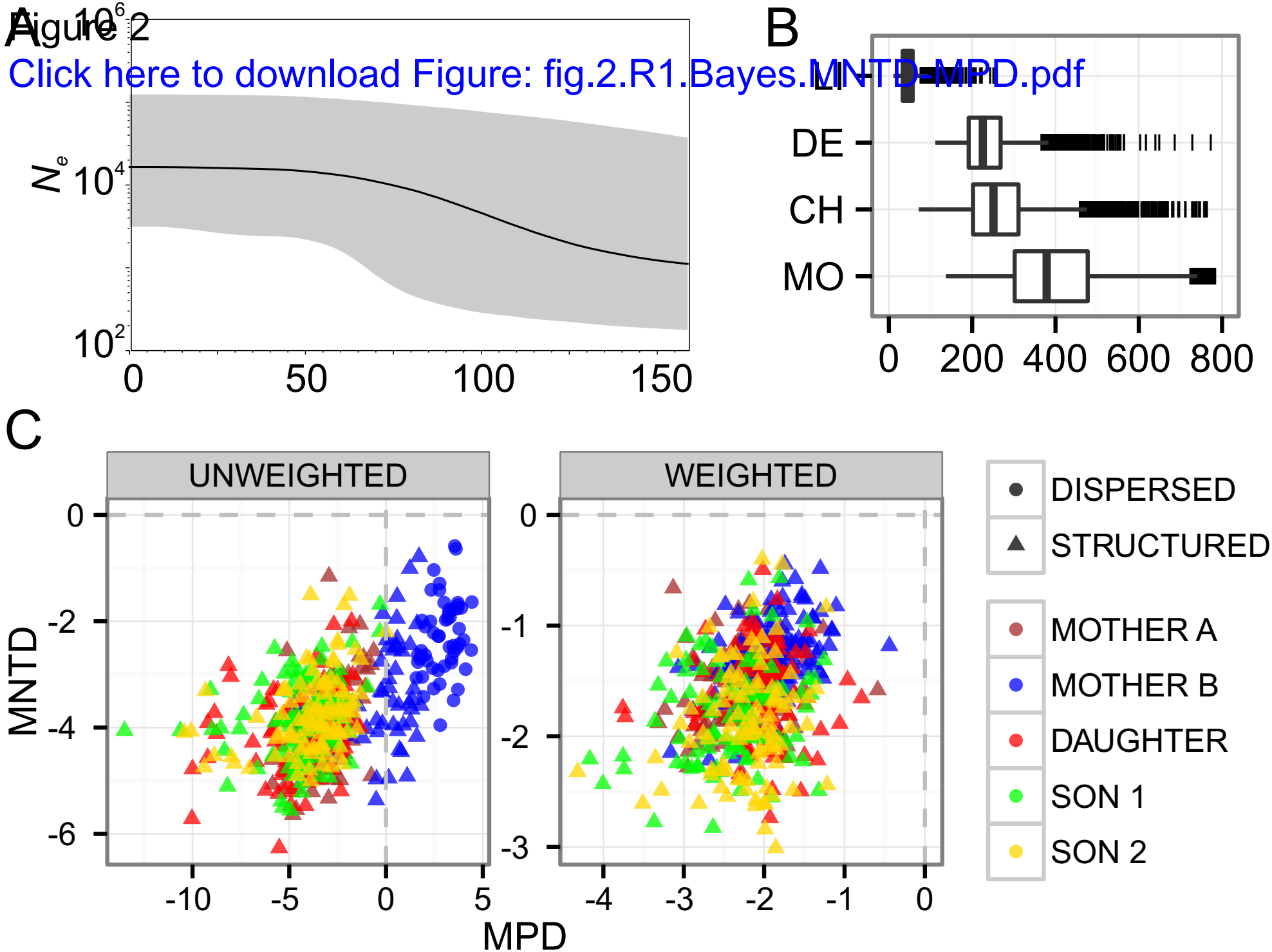


Figure 3

Click here to download Figure: fig.3.div.profile.R1.pdf

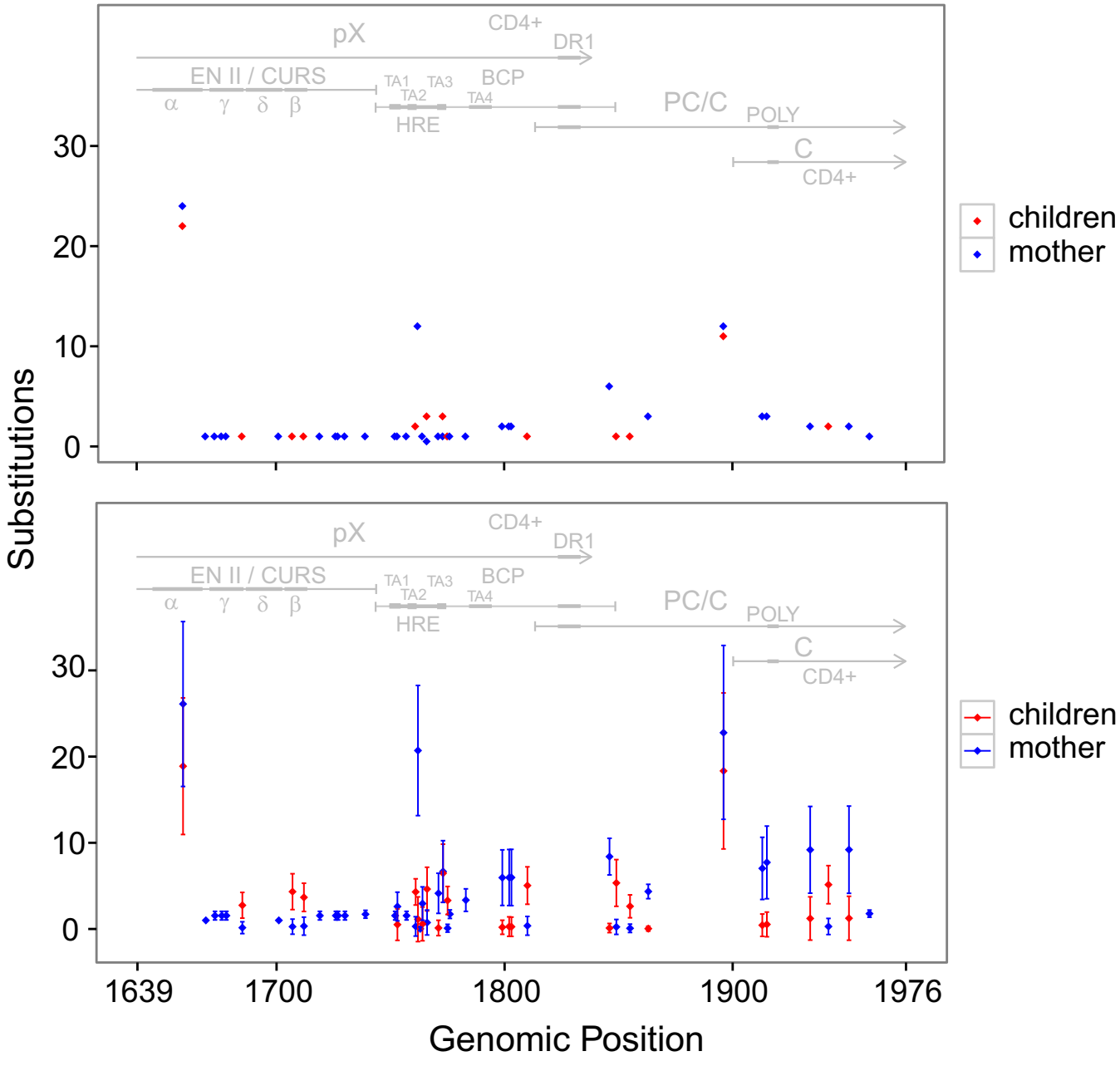
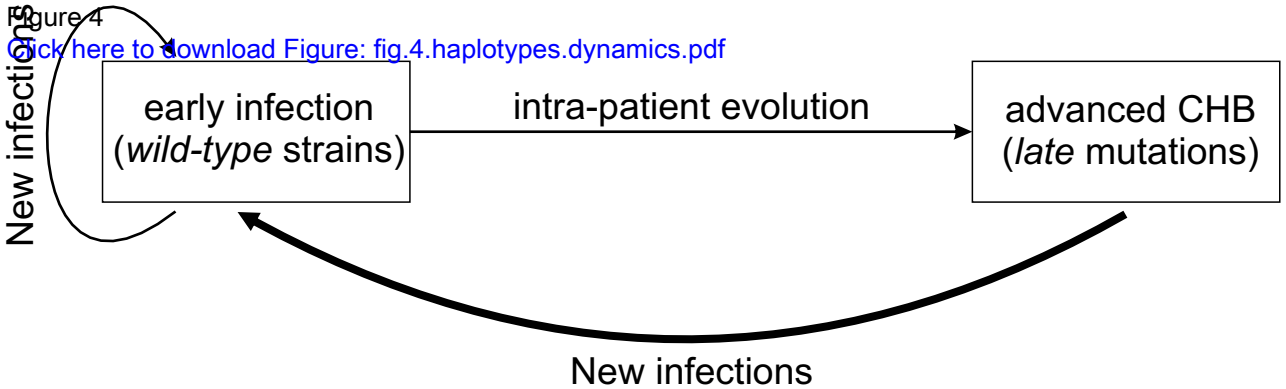


Figure 4

Click here to download Figure: fig.4.haplotypes.dynamics.pdf



Viral evolution during chronic hepatitis B as revealed by ultra deep sequencing data

Leandro R. Jones, Mariano Sede, Julieta M. Manrique & Jorge Quarleri

**Supplementary Material.**

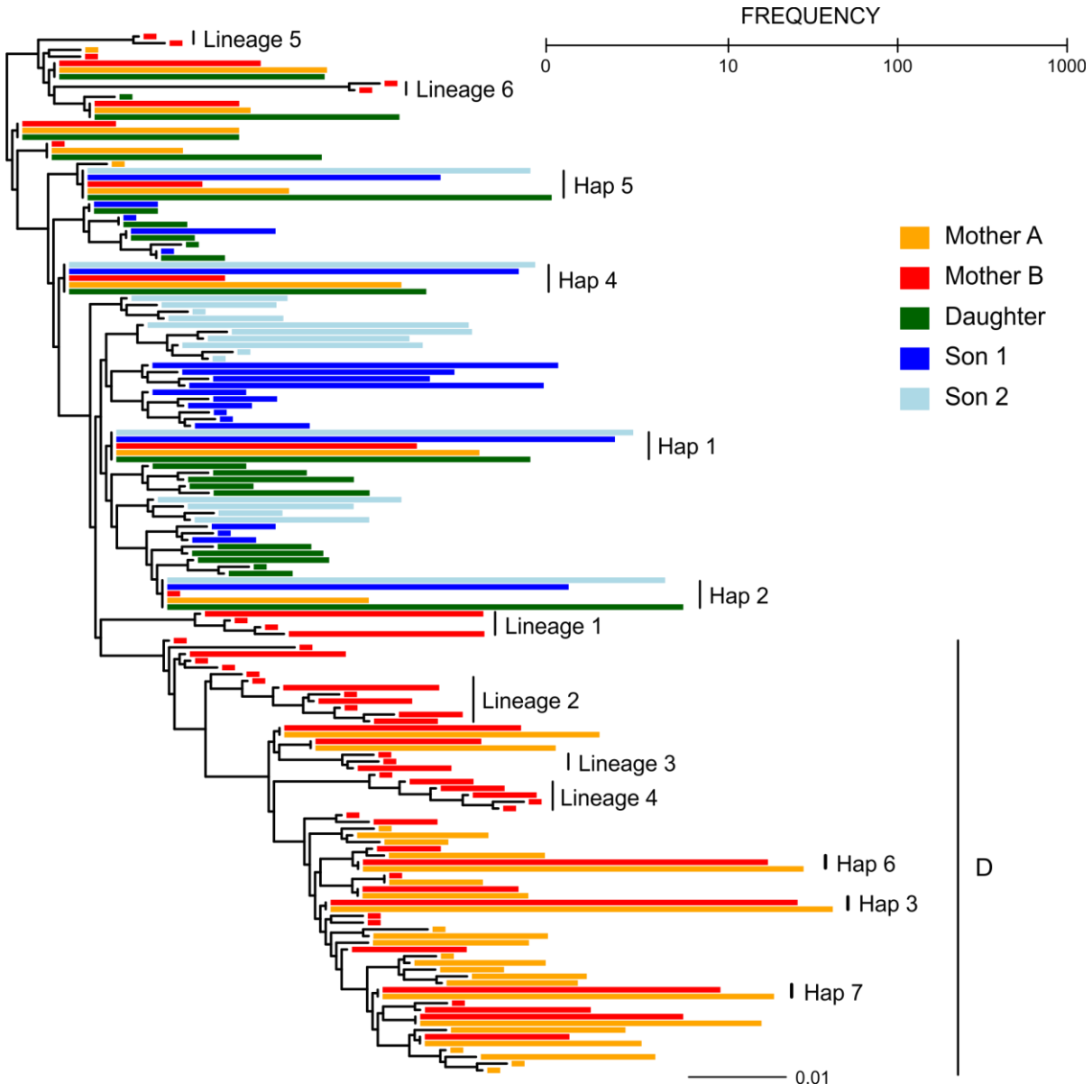



**Figure S1. Maximum Likelihood evidence for familiarly transmitted HBV.** Sequence origins (times A and B of the mother, daughter, and sons 1 and 2) are depicted by colors. The frequency of each individual sequence is depicted by the length of the colored bars, as indicated by the scale located in the upper part of the figure. The vertical lines show the locations of haplotypes 1 to 7 (Hap 1–7) and the divergent (Lineages 1–6) and deleted (D) lineages from the seroconverter patient. The tree scale-bar units are substitutions per aligned position..

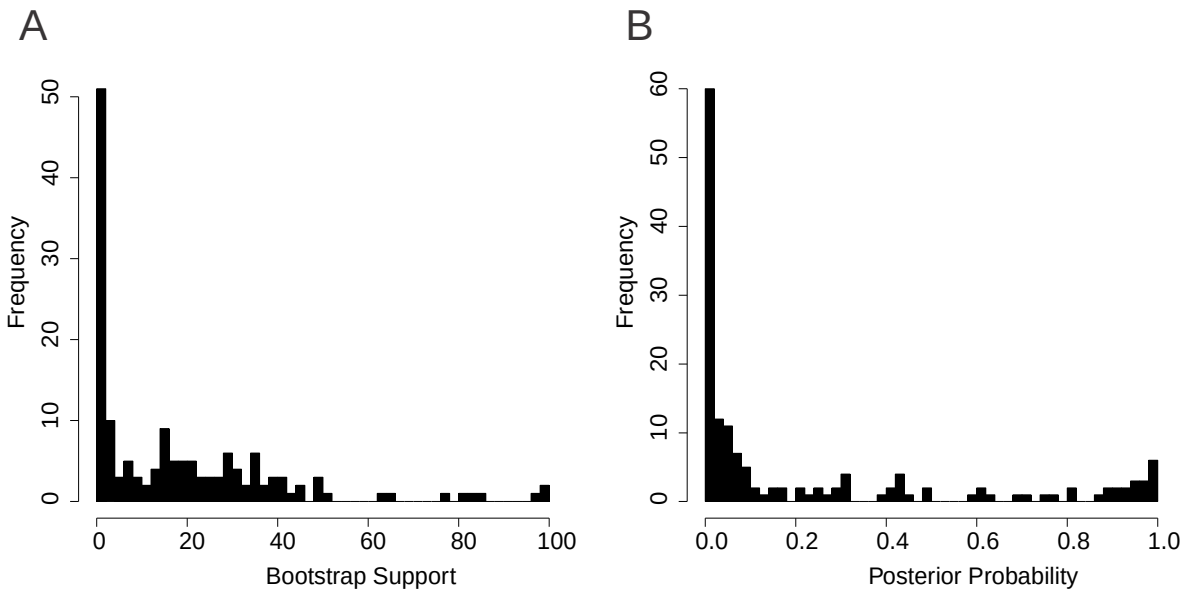



**Figure S2.** Histograms of bootstrap supports (A) and posterior probabilities (B) corresponding to Maximum Likelihood and Bayesian evidence for the viral sequences from the four patients.

**Effect of phylogenetic uncertainty on groupings and diversity estimates.** We implemented a set of simple cladistic metrics that summarizes the proximity of a group of sequences in a set of bootstrap trees in order to evaluate the impact of phylogenetic uncertainty on the clustering of lineages 2, 3, and 4. We define the Intra-Clade Cladistic Distance (ICCD) on a given tree *T*, for all the pairs of sequences $S_{ij}$ belonging to a group of sequences of interest, as:

$$ICCD = \sum_{S_{ij}} F(T).$$

The function $F(T)$ is given by:

$$F(T) = \sum_{b=1}^{B} f(b),$$

where *B* is the number of branches in *T* and $f(b) = 1$ if branch *b* is in the path connecting terminals *i* and *j* or $f(b) = 0$ otherwise. Thus, the smaller the ICCD is, the closer the relatedness of the sequences of interest. In addition, we define three further metrics, the Complement Distances (CoD), Permuted Intra-Clade Cladistic Distances (PtICCD) and Permuted Complement Distances (PtCoD). The CoD measures the relatedness among all the terminals excepting the sequences of interest in the observed tree, thus providing an idea of how small the ICCDs of all the possible groupings from the tree could be. It is defined by the equation:

$$CoD = \sum_{S_{kl}} G(T),$$

and is calculated along all the pairs of sequences $S_{kl}$ that do not belong to the clade of interest. The function $G(T)$ is defined as:

$$G(T) = \sum_{b=1}^{B} g(b),$$

where $g(b) = 1$ if *b* is in the path connecting sequences *k* and *l* and $g(b) = 0$ otherwise. The PtICCD is equivalent to the ICCD but is obtained from a tree in which the terminals are randomly permuted, thus providing an idea of the likelihood of obtaining the observed metric by random chance. The

PtCoD is the CoD obtained from the permuted tree; thus in general it is expected to be very similar to the CoD.

The ICCD indices revealed that the sequences from lineages 2 to 4 were very close to each other among the 100 bootstrap trees analyzed (Fig. S3A). In agreement with that, the average relatedness along the trees and the lineages' relatedness in permuted trees were smaller than the observed lineages' relatedness, as reflected by significantly higher CoD, PtICCD, and PtCoD values ($p$< 1.0E-16, Kolmogorov–Smirnov test).

The significance of the enhanced diversification observed for MA and MB in comparison to D, S1, and S2 was further evaluated by comparing the intra-patient phylogenetic distances observed along 100 bootstrap trees. These analyses confirmed a significantly increased viral diversification for the mother (Fig. S3B ), strengthening the results obtained from the non-phylogenetic haplotypes analysis (Figs. 1D and 1E).

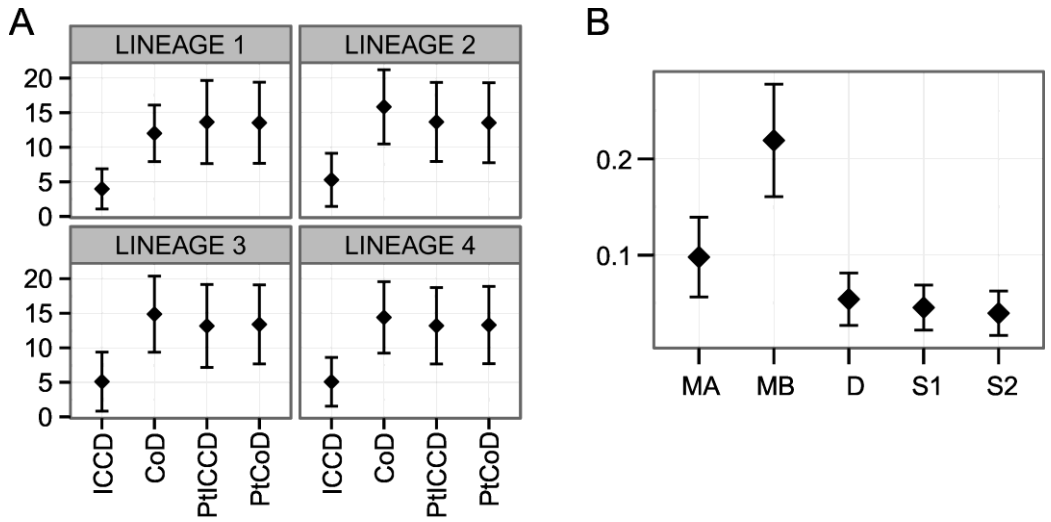



**Figure S3. (A)** Intra-Clade Cladistic Distances (ICCD) among the sequences from lineages 2–4 and the corresponding Complement Distances (CoD), Permuted Intra-Clade Cladistic Distances (PtICCD), and Permuted Complement Distances (PtCoD). Metrics for Lineage 1, which was supported by bootstrap and Bayesian analyses (Table S2), are also included for comparison. All the metrics were calculated over 100 bootstrap trees. The data are summarized by mean values (diamonds) and standard deviations (whiskers). **(B)** Viral phylogenetic diversity in the five family members: *D*, daughter; *MA*, time A of the mother; *MB*, time *B* of the mother; *S1*, son 1; *S2*, son 2. The analysis was multiplied along 100 bootstrap trees. Mean phylogenetic distances are given (diamonds), together with the corresponding standard deviations (whiskers).

**Estimation of Evolutionary Paths.** A nucleotide pattern can require different amounts of evolutionary change depending on the underlying evolutionary history (1, 2). For example, genomic position 1639 displayed 38% of A and 62% of C among all the viral haplotypes identified, whereas position 1762 showed 77% of A, 17% of T, 2% of G, and 3% of haplotypes have the position deleted. Though this indicates a higher diversity for position 1762, the incorporation of the maximum likelihood evidence shows that position 1639 experienced around four times more evolutionary changes than position 1762 (Fig. S5).

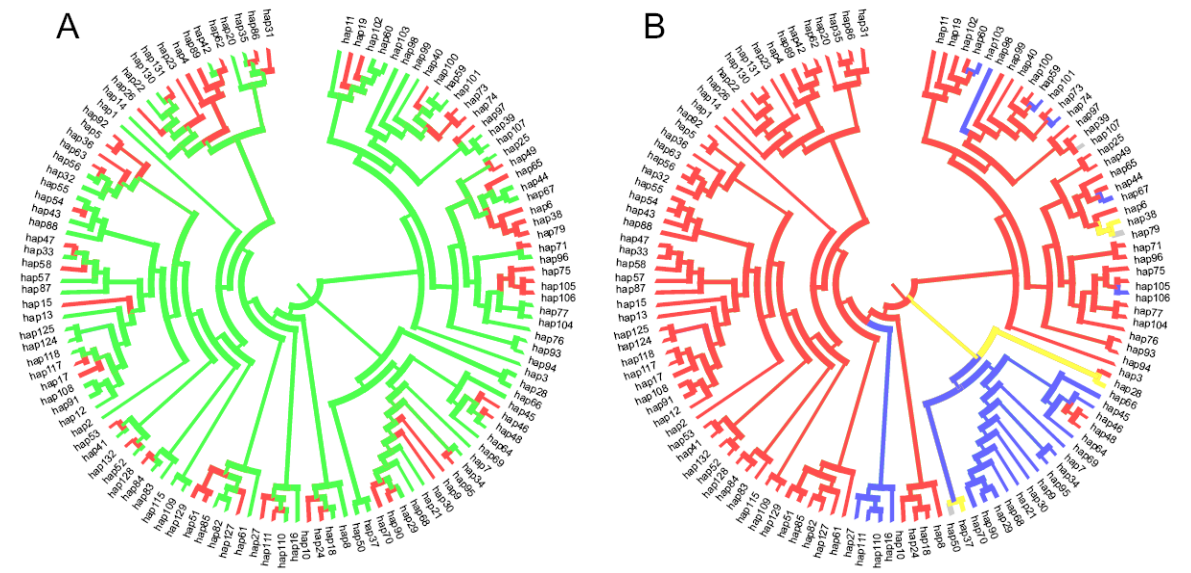



**Figure S4.** Course of evolution of the HBV genomic positions 1639 (A) and 1762 (B). The characters' histories were mapped using the Mesquite software (http://mesquiteproject.org) onto the maximum likelihood tree of all the haplotypes identified among the four family members. Character states are indicated by colors: *green* C, *red* A, *blue* T, *yellow* G, *grey* -. Despite position 1762 displays a higher nucleotide diversity, it requires around four times less evolutionary steps, represented in the trees as color changes along the branches. Equivalent analyses were applied to each alignment position, and multiplied along 100 bootstrap trees as detailed along the main body of the paper.

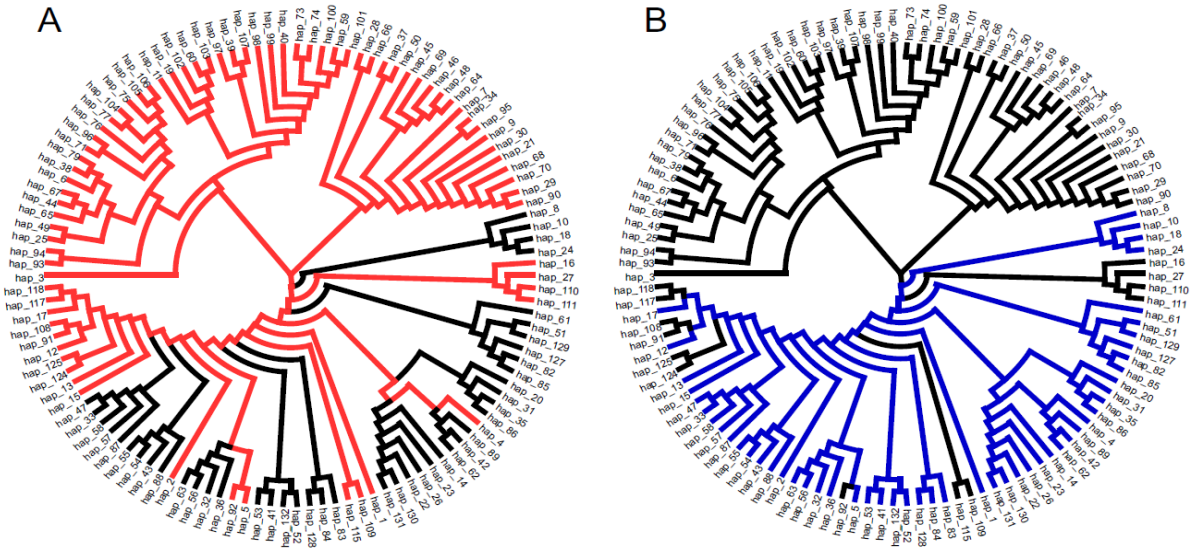



**Figure S5.** Tree branches common to the mother (A) and children (B) haplotypes in the Maximum Likelihood tree. Branch colors represent the parts of the tree used for calculating the site-wise diversity attributed to the mother (red) and the children (blue).

**Supplementary tables.**

**Table S1.** Haplotype frequencies among the two sampling times of the mother (MA and MB), the daughter (D) and the two sons (S1 and S2).

| Haplotype | MA | MB | D | S1 | S2 | Chi.test |
|---|---|---|---|---|---|---|
| 1 | 117 | 50 | 234 | 739 | 947 | ~0 |
| 2 | 13 | 1 | 936 | 197 | 731 | ~0 |
| 3 | 776 | 481 | 0 | 0 | 0 | ~0 |
| 4 | 77 | 7 | 108 | 380 | 475 | 8.8E-173 |
| 5 | 13 | 4 | 462 | 102 | 346 | 2.2E-199 |
| 6 | 336 | 207 | 0 | 0 | 0 | 1.4E-191 |
| 7 | 172 | 83 | 0 | 0 | 0 | 2.7E-98 |
| 8 | 0 | 0 | 0 | 203 | 0 | 1.9E-174 |
| 9 | 87 | 30 | 0 | 0 | 0 | 8.0E-52 |
| 10 | 0 | 0 | 0 | 104 | 0 | 9.7E-89 |
| 11 | 61 | 21 | 0 | 0 | 0 | 4.3E-36 |
| 12 | 32 | 13 | 31 | 0 | 0 | 1.8E-13 |
| 13 | 7 | 6 | 53 | 0 | 0 | 4.1E-32 |
| 14 | 0 | 0 | 0 | 0 | 66 | 6.2E-56 |
| 15 | 5 | 1 | 33 | 0 | 0 | 1.4E-21 |
| 16 | 0 | 37 | 0 | 0 | 0 | 5.4E-31 |
| 17 | 16 | 3 | 16 | 0 | 0 | 5.6E-08 |
| 18 | 0 | 0 | 0 | 34 | 0 | 2.0E-28 |
| 19 | 22 | 8 | 0 | 0 | 0 | 1.5E-12 |
| 20 | 0 | 0 | 0 | 0 | 23 | 4.9E-19 |
| 21 | 16 | 6 | 0 | 0 | 0 | 5.3E-09 |
| 22 | 0 | 0 | 0 | 0 | 22 | 3.5E-18 |
| 23 | 0 | 0 | 0 | 0 | 22 | 3.5E-18 |
| 24 | 0 | 0 | 0 | 16 | 0 | 4.1E-13 |
| 25 | 8 | 7 | 0 | 0 | 0 | 1.4E-4 |
| 26 | 0 | 0 | 0 | 0 | 13 | 1.3E-10 |
| 27 | 0 | 12 | 0 | 0 | 0 | 9.4E-10 |
| 28 | 9 | 0 | 0 | 0 | 0 | 2.8E-07 |
| 29 | 9 | 0 | 0 | 0 | 0 | 2.8E-07 |
| 30 | 9 | 0 | 0 | 0 | 0 | 2.8E-07 |
| 31 | 0 | 0 | 0 | 0 | 9 | 2.8E-07 |
| 32 | 0 | 0 | 8 | 0 | 0 | 1.9E-06 |
| 33 | 0 | 0 | 2 | 6 | 0 | 1.9E-3 |
| 34 | 0 | 8 | 0 | 0 | 0 | 1.9E-06 |
| 35 | 0 | 0 | 0 | 0 | 8 | 1.9E-06 |
| 36 | 0 | 0 | 7 | 0 | 0 | 1.2E-05 |
| 37 | 7 | 0 | 0 | 0 | 0 | 1.2E-05 |
| 38 | 7 | 0 | 0 | 0 | 0 | 1.2E-05 |
| 39 | 0 | 7 | 0 | 0 | 0 | 1.2E-05 |
| 40 | 0 | 7 | 0 | 0 | 0 | 1.2E-05 |
| 41 | 0 | 0 | 0 | 0 | 7 | 1.2E-05 |
| 42 | 0 | 0 | 5 | 0 | 0 | 4.9E-4 |
| 43 | 0 | 0 | 5 | 0 | 0 | 4.9E-4 |
| 44 | 5 | 0 | 0 | 0 | 0 | 4.9E-4 |
| 45 | 5 | 0 | 0 | 0 | 0 | 4.9E-4 |
| 46 | 5 | 0 | 0 | 0 | 0 | 4.9E-4 |
| 47 | 0 | 0 | 2 | 2 | 0 | 0.19 |
| 48 | 4 | 0 | 0 | 0 | 0 | 0.003 |
| 49 | 3 | 1 | 0 | 0 | 0 | 0.074 |
| 50 | 0 | 4 | 0 | 0 | 0 | 0.003 |
| 51 | 0 | 0 | 0 | 4 | 0 | 0.003 |

| Haplotype | MA | MB | D | S1 | S2 | Chi.test |
|---|---|---|---|---|---|---|
| 52 | 0 | 0 | 0 | 0 | 4 | 0.003 |
| 53 | 0 | 0 | 0 | 0 | 4 | 0.003 |
| 54 | 0 | 0 | 3 | 0 | 0 | 0.017 |
| 55 | 0 | 0 | 3 | 0 | 0 | 0.017 |
| 56 | 0 | 0 | 3 | 0 | 0 | 0.017 |
| 57 | 0 | 0 | 2 | 1 | 0 | 0.252 |
| 58 | 0 | 0 | 2 | 1 | 0 | 0.251 |
| 59 | 0 | 3 | 0 | 0 | 0 | 0.017 |
| 60 | 0 | 3 | 0 | 0 | 0 | 0.017 |
| 61 | 0 | 0 | 0 | 3 | 0 | 0.017 |
| 62 | 0 | 0 | 2 | 0 | 0 | 0.091 |
| 63 | 0 | 0 | 2 | 0 | 0 | 0.091 |
| 64 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 65 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 66 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 67 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 68 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 69 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 70 | 2 | 0 | 0 | 0 | 0 | 0.091 |
| 71 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 73 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 74 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 75 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 76 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 77 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 79 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 81 | 0 | 2 | 0 | 0 | 0 | 0.091 |
| 82 | 0 | 0 | 0 | 2 | 0 | 0.091 |
| 83 | 0 | 0 | 0 | 2 | 0 | 0.091 |
| 84 | 0 | 0 | 0 | 2 | 0 | 0.091 |
| 85 | 0 | 0 | 0 | 2 | 0 | 0.091 |
| 86 | 0 | 0 | 0 | 0 | 2 | 0.091 |
| 87 | 0 | 0 | 1 | 0 | 0 | 0.406 |
| 88 | 0 | 0 | 1 | 0 | 0 | 0.406 |
| 89 | 0 | 0 | 1 | 0 | 0 | 0.406 |
| 90 | 1 | 0 | 0 | 0 | 0 | 0.406 |
| 91 | 1 | 0 | 0 | 0 | 0 | 0.406 |
| 92 | 1 | 0 | 0 | 0 | 0 | 0.406 |
| 93 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 94 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 95 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 96 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 97 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 98 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 99 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 100 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 101 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 102 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 103 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 104 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 105 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 106 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 107 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 108 | 0 | 1 | 0 | 0 | 0 | 0.406 |

| Haplotype | MA | MB | D | S1 | S2 | Chi.test |
|---|---|---|---|---|---|---|
| 109 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 110 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 111 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 115 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 117 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 118 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 124 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 125 | 0 | 1 | 0 | 0 | 0 | 0.406 |
| 127 | 0 | 0 | 0 | 1 | 0 | 0.406 |
| 128 | 0 | 0 | 0 | 1 | 0 | 0.406 |
| 129 | 0 | 0 | 0 | 1 | 0 | 0.406 |
| 130 | 0 | 0 | 0 | 0 | 1 | 0.406 |
| 131 | 0 | 0 | 0 | 0 | 1 | 0.406 |
| 132 | 0 | 0 | 0 | 0 | 1 | 0.406 |

**Table S2.** Summary of Bootstrap supports and Bayesian posterior probabilities. Mean and median values along the whole tree are given, together with the bootstrap supports and posterior probabilities of the lineages 1 to 6 (L1-6)[1].

| | Mean | Median | L1 | L2 | L3 | L4 | L5 | L6 |
|---|---|---|---|---|---|---|---|---|
| Bootstrap | 18.7 | 13.0 | 82 | 17 | 38 | 53 | 82 | 100 |
| P. Probability | 0.23 | 0.05 | 1.00 | 0.78 | 0.94 | 1.00 | 1.00 | 1.00 |

[1] A clade was considered as supported if bootstrap and Bayesian posterior probability values were above 80 and 0.9, respectively.

.

**Table S3.** Bayesian estimates of the tMRCA (months) of the strains from the mother (M) daughter (D), sons 1 and 2 (S1 and S2), Lineages 1 to 6 (L1-L6) and the deleted strains from the mother (Del).

| Statistic[1] | M | D | S1 | S2 | Del | L1 | L2 | L3 | L4 | L5 | L6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 445 | 316 | 258 | 257 | 247 | 52 | 43 | 44 | 61 | 33 | 39 |
| median | 414 | 301 | 245 | 244 | 235 | 46 | 40 | 39 | 56 | 28 | 33 |
| HPD | [191, 758] | [148, 504] | [108, 419] | [109, 418] | [136, 381] | [16, 102] | [17, 75] | [12, 87] | [23, 112] | [9, 68] | [9, 84] |
| ESS | 665 | 618 | 653 | 667 | 439 | 1362 | 730 | 489 | 1026 | 3451 | 3564 |

[1] *HPD*, 95% HPD interval; *ESS*, effective sample size.

**Table S4.** Variations in HBx derived CD4+ T cell epitopes.

| | W-T | Del[a] | I127V/T | K130Q | I127V/T+K130M+V121I | Del+mut[b] |
|---|---|---|---|---|---|---|
| MA | 0.155 | 0.835 | 0.000 | 0.000 | 0.000 | 0.010 |
| MB | 0.089 | 0.845 | 0.008 | 0.000 | 0.051 | 0.007 |
| D | 0.987 | 0.000 | 0.000 | 0.013 | 0.000 | 0.000 |
| S1 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| S2 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

[a] Four amino acids deletion (positions 130-133).

[b] Mutation at amino acids other than 127, 130 and 131 (C137Y; G135R).

**Table S5.** Evolutionary path lengths of the nucleotide patterns identified here.

| | Optimal tree[a] | | Bootstrap trees[b] | | CI[c] | |
|---|---|---|---|---|---|---|
| Position | Mother | Children | Mother | Children | Mother | Children |
| 1659 | 24 | 22 | 26.10 | 18.88 | 24.31 - 27.76 | 16.63 - 20.51 |
| 1669 | 1 | 0 | 1.00 | 0.00 | - | - |
| 1673 | 1 | 0 | 1.55 | 0.00 | 1.45 - 1.65 | - |
| 1676 | 1 | 0 | 1.55 | 0.00 | 1.43 - 1.644 | - |
| 1678 | 1 | 0 | 1.55 | 0.00 | 1.42 - 1.62 | - |
| 1685 | 0 | 1 | 0.15 | 2.75 | 0.049 - 0.30 | 2.52 - 3.19 |
| 1701 | 1 | 0 | 1.00 | 0.00 | - | - |
| 1707 | 0 | 1 | 0.27 | 4.33 | 0.10 - 0.42 | 3.78 - 4.65 |
| 1712 | 0 | 1 | 0.33 | 3.67 | 0.13 - 0.59 | 3.25 - 3.98 |
| 1719 | 1 | 0 | 1.55 | 0.00 | 1.44 - 1.65 | - |
| 1726 | 1 | 0 | 1.55 | 0.00 | 1.45 - 1.63 | - |
| 1727 | 1 | 0 | 1.55 | 0.00 | 1.43 - 1.62 | - |
| 1730 | 1 | 0 | 1.55 | 0.00 | 1.45 - 1.63 | - |
| 1739 | 1 | 0 | 1.71 | 0.00 | 1.56 - 1.79 | - |
| 1752 | 1 | 0 | 1.55 | 0.00 | 1.41 - 1.63 | - |
| 1753 | 1 | 0 | 2.60 | 0.51 | 2.34 - 2.96 | 0.22 - 1.00 |
| 1757 | 1 | 0 | 1.55 | 0.00 | 1.46 - 1.64 | - |
| 1761 | 0 | 2 | 0.30 | 4.30 | 0.09 - 0.55 | 3.83 - 4.56 |
| 1762 | 12 | 0 | 20.70 | 1.13 | 19.21 - 22.00 | 0.73 - 1.75 |
| 1764 | 1 | 0 | 2.95 | 0.59 | 2.56 - 3.31 | 0.30 - 1.11 |
| 1766 | 1 | 3 | 0.75 | 4.63 | 0.50 - 1.07 | 4.04 - 5.19 |
| 1771 | 1 | 0 | 4.14 | 0.12 | 3.50 - 4.59 | 0.00 - 0.36 |
| 1773 | 1 | 3 | 6.66 | 6.46 | 5.86 - 7.39 | 5.84 - 7.13 |
| 1775 | 0 | 2 | 0.09 | 3.31 | 0.02 - 0.22 | 2.92 - 3.59 |
| 1776 | 1 | 0 | 1.72 | 0.00 | 1.58 - 1.83 | - |
| 1783 | 1 | 0 | 3.35 | 0.00 | 3.08 - 3.73 | - |
| 1799 | 2 | 0 | 5.95 | 0.20 | 5.25 - 6.37 | 0.10 - 0.43 |
| 1802 | 2 | 0 | 5.97 | 0.28 | 5.44 - 6.56 | 0.06 - 0.50 |
| 1803 | 2 | 0 | 5.97 | 0.26 | 5.33 - 6.68 | 0.10 - 0.53 |
| 1810 | 0 | 1 | 0.37 | 5.04 | 0.22 - 0.61 | 4.69 - 5.44 |
| 1846 | 6 | 0 | 8.39 | 0.11 | 8.04 - 8.78 | 0.04 - 0.23 |
| 1849 | 0 | 1 | 0.24 | 5.34 | 0.12 - 0.40 | 4.91 - 5.90 |
| 1855 | 0 | 1 | 0.08 | 2.63 | 0.02 - 0.19 | 2.41 - 2.90 |
| 1863 | 2 | 0 | 4.35 | 0.04 | 4.14 - 4.49 | 0.00 - 0.11 |
| 1896 | 12 | 11 | 22.77 | 18.33 | 20.89 - 25.11 | 16.30 - 20.59 |
| 1913 | 3 | 0 | 7.03 | 0.44 | 6.29 - 7.68 | 0.23 - 0.80 |
| 1915 | 3 | 0 | 7.72 | 0.53 | 6.97 - 8.56 | 0.29 - 0.87 |
| 1934 | 2 | 0 | 918 | 1.22 | 8.52 - 10.41 | 0.82 - 1.86 |
| 1942 | 0 | 2 | 0.29 | 5.14 | 0.13 - 0.47 | 4.65 - 5.68 |
| 1951 | 2 | 0 | 9.2 | 1.24 | 8.09 - 10.04 | 0.79 - 2.08 |
| 1960 | 1 | 0 | 1.78 | 0.00 | 1.70 - 1.84 | - |

[a] Evolutionary change in the Maximum Likelihood tree.

[b] Evolutionary change among 100 bootstrap trees.

[c] Mean bias-corrected and accelerated 95% bootstrap (n=999) confidence intervals for the mean.

**References.**


1. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer Associates, Inc., Sunderlands, Massachusetts) p 664.

2. Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* 20:406-416.

manuscript with changes highlighted

Click here to download Additional Material for Reviewer: Jones.et.al.R1-track-changes.docx

Sede et al paper

Click here to download Additional Material for Reviewer: Sede_2014.CHB.family.pdf