Overlapping structure of hepatitis B virus (HBV) genome and immune selection pressure are critical forces modulating HBV evolution

Valeria Cento,1 Carmen Mirabelli,1 Salvatore Dimonte,1 Romina Salpini,1 Yue Han,2 Pascale Trimoulet,3 Ada Bertoli,4 Valeria Micheli,5 Guido Gubertini,5 Giuseppina Cappiello,6 Alberto Spanò,6 Roberta Longo,6 Martina Bernassola,6 Francesco Mazzotta,7 Giuseppe Maria De Sanctis,8 Xin Xin Zhang,2 Jens Verheyen,9 Antonella D’Arminio Monforte,10† Francesca Ceccherini-Silberstein,1 Carlo Federico Perno1 and Valentina Svicher1

1Department of Experimental Medicine and Surgery, University of Rome ‘Tor Vergata’, Rome 00133, Italy
2Department of Infectious Diseases, Rui Jin Hospital, Shanghai Jiao Tong University, Shanghai 200025, PR China
3Bordeaux University Hospital, University Victor Segalen, Bordeaux 33076, France
4University Hospital ‘Tor Vergata’, Rome 00133, Italy
5’Sacco’ Hospital, Milan 20157, Italy
6’S. Pertini’ Hospital, Rome 00157, Italy
7Hospital ‘Santa Maria Annunziata’, Florence 50126, Italy
8University ‘La Sapienza’, Rome 00185, Italy
9University of Cologne, Cologne 50923, Germany
10Infectious Diseases Unit, University of Milan, Milan 20122, Italy

How the overlap between the hepatitis B virus (HBV) reverse transcriptase (RT) and HBV S antigen (HBsAg) genes modulates the extent of HBV genetic variability is still an open question, and was investigated here. The rate of nucleotide conservation (≤ 1 % variability) followed an atypical pattern in the RT gene, due to an overlap between RT and HBsAg (69.9 % nucleotide conservation in the overlapping region vs 41.2 % in the non-overlapping region; P<0.001), with a consequently lower rate of synonymous substitution within the overlapping region [median(interquartile range)dS = 3.1(1.5–7.4) vs 20.1(10.6–30.0); P=3.249×10^{-22}]. The most conserved RT regions were located within the YMDD motif and the N-terminal parts of the palm and finger domains, critical for RT functionality. These regions also corresponded to highly conserved HBsAg domains that are critical for HBsAg secretion. Conversely, the genomic region encoding the HBsAg antigenic loop (where immune-escape mutations are localized) showed a sharp decrease in the extent of conservation (40.6 %), which was less pronounced in the setting of human immunodeficiency virus (HIV)-driven immune suppression (48.8 % in HIV–HBV co-infection vs 21.5 % in mono-infected patients; P=0.020). In conclusion, the overlapping reading frame and the immune system appear to have shaped the patterns of RT and HBsAg genetic variability. Highly conserved regions in RT and HBsAg may deserve further attention as novel therapeutic targets.

About 2 billion people worldwide are infected with hepatitis B virus (HBV), and about 350 million presently live with chronic HBV infection (WHO, 2012), thus making this viral infection a global health concern.
HBV is one of the smallest DNA viruses infecting humans, and its genome is a relaxed circular, partially double-stranded DNA of around 3200 bp. The genome contains four partially overlapping ORFs encoding the P (polymerase), C (core), S (surface) and X proteins, organized in order, resulting in about two-thirds of the viral genome encoding multiple proteins. From an evolutionary point of view, this genomic organization has a striking importance, as a synonymous nucleotide substitution in one ORF can potentially result in a non-synonymous nucleotide substitution in the overlapping ORF. In this way, it is believed that HBV genome evolution is constrained in order to maintain essential protein functions (Mizokami et al., 1997; Yang, 2007).

The replication cycle of HBV comprises an error-prone reverse-transcription step, made possible by a virus-encoded reverse transcriptase (RT, 344 aa). As a consequence, despite the evolutionary constraints acting on HBV evolution, its variation rate is much higher than that found for other DNA viruses and is more similar to the rate observed in RNA viruses with comparatively slow evolutionary rates (Mizokami & Orito, 1999).

On the basis of genome divergence, HBV is classified into ten genotypes (A–J), with >8% sequence diversity and a different geographical distribution (McMahon, 2009). Therefore, few regions of the HBV genome are expected to be conserved across all HBV genotypes, especially at the level of the RT and S proteins (targets of the currently available anti-HBV drugs and host’s neutralizing antibodies, respectively; Locarnini & Yuen, 2010; Lok & McMahon, 2009; Torresi et al., 2002).

Evaluation, at a population level, of HBV variability patterns along these two viral proteins may have remarkable importance in the definition of critical regions for structural integrity and function of HBV proteins. Such insight could provide directions for rational development of new antiviral strategies, involving non-nucleoside RT-directed inhibitors, entry inhibitors or small interfering RNAs. This is important because all clinically available anti-HBV drugs are nucleoside analogues acting as chain terminators, which exclusively target the RT protein and show largely overlapping resistance profiles.

In this light, the goal of this work was to provide a deep insight into HBV genetic variability at the level of RT and HBV S antigen (HBsAg) by analysing 3484 HBV RT/HBsAg nucleotide and amino acid sequences, isolated in Europe and Asia, from 555 drug-naive patients and 2927 antiviral-experienced patients infected with HBV genotypes A (n=542), B (n=607), C (n=1332), D (n=689), E (n=198), F (n=74) and G (n=42). The 555 sequences from drug-naive patients analysed in this study were collected from Italy (n=147, 26.5%), France (n=76, 13.7%), Germany (n=27, 4.9%) and China (n=305, 55.0%). The 2927 sequences from antiviral-experienced patients were retrieved from our clinical centre in Italy (n=192) and from GenBank (n=2735). All new sequences analysed have been submitted to GenBank (accession numbers JX849204–JX849647). Sequences containing undetermined nucleotides, premature stop codons and/or gaps were excluded. Amino acids with ≤1% prevalence in our set of sequences were defined as conserved.

The RT and HBsAg ORFs overlap at RT aa 8–236, with the HBsAg ORF shifted downstream by 1 nt (Fig. 1a). How this genomic organization can drive HBV evolution has been a matter of study for many molecular biologists (Maman et al., 2011; Mizokami & Orito, 1999; Mizokami et al., 1997; Yang et al., 1995). Indeed, the third nucleotide of any RT codon (p3) always corresponds to the second nucleotide of the S codon (s2). Similarly, the second nucleotide of a P codon (p2) always matches the first nucleotide of the S codon (s1), while the first nucleotide of a P codon (p1) corresponds to the third nucleotide of the S codon at position −1 (s3). Therefore, following the rules of the degenerate genetic code, nucleotide substitutions at position p2/s1 would affect amino acids in both RT and HBsAg. This situation can be disadvantageous for virus evolution and, indeed, the p2/s1 position showed the highest degree of nucleotide conservation (82.5%) in our population. Conversely, the p1/s3 and p3/s2 positions showed a significantly lower degree of conservation than p2/s1 (62.7% conserved nucleotides at p1/s3 and 64.5% at p3/s2; P<0.001 for both comparisons). As nucleotide substitutions at the third codon position generally do not change the encoded amino acid, nucleotide substitutions at p1/s3 and p3/s2 would indeed affect only one protein.

As proposed by Zaaijer et al. (2007), this may allow adaptive and independent evolution of HBV RT and HBsAg, despite the constraint imposed by the overlapping ORFs. Indeed, in our analysis, the RT region overlapping the HBsAg gene showed a significantly higher degree of nucleotide conservation than the non-overlapping region (69.9 vs 41.2%, respectively; P<0.001 by Wilcoxon test), and also at the level of each single nucleotide position of RT codons (62.7 vs 50.0% for p1; 82.5 vs 61.2% for p2; 64.5 vs 12.1% for p3; P<0.001 for all comparisons). Furthermore, analysis of the dN–dS ratio, estimated by the random-effect likelihood (REL) algorithm implemented in Hy Phy (Pond et al., 2005), showed a significantly higher rate of synonymous substitution in the non-overlapping than in the overlapping gene segment. Indeed, the median [interquartile range (IQR)] dS was 20.1 (10.6–30.0) in the non-overlapping region and 3.1 (1.5–7.4) in the overlapping region (P=3.249×10−22), while the median (IQR) dN–dS ratio was −4.41 (−6.74 to −1.06) vs −0.37 (−1.48 to 0.18), respectively (P=2.29×10−16) (Fig. 1b). Taken together, these results indicate that, in the RT region overlapping the HBsAg gene, classical neutral synonymous evolution is severely limited, due to the possibility of impairing the function of the HBsAg protein.

The extent of HBV genetic conservation was then analysed at the amino acid level. Analysis of 3482 full-length sequences of the RT enzyme showed conservation (<1% variability) in 223 of 344 (64.8%) amino acids, with no
Fig. 1. (a) Schematic representation of the HBV genome and its ORFs. The HBV genome comprises a 3.2 kb circular positive strand complementary to a negative shorter incomplete strand (1700–2800 nt). The RT and HBsAg ORFs overlap from RT aa 8–236, with the HBsAg ORF shifted downstream by 1 nt. In this study, the full-length sequences of RT (aa 1–344) and HBsAg (aa 1–226) were analysed. (b) Difference of synonymous and non-synonymous substitutions throughout the RT nucleotide sequence. (c) Percentage of amino acid conservation throughout the RT protein sequence. RT domains were assigned following Kohlstaedt et al. (1992). The mapping of HLA class I and II epitopes throughout the RT protein is also shown; the list of HLA epitopes in RT was retrieved from Desmond et al. (2008). Epitopes are shaded according to the degree of conservation: black indicates that all the residues within the epitope are conserved; dark grey, the percentage of conserved residues within the epitope is between 90 and 100%; light grey, the percentage of conserved residues within the epitope is between 80 and 90%; dashed lines, the percentage of conserved residues within the epitope is <80%.
significant difference in drug-naive versus drug-experienced patients. Among the 223 RT codons conserved at the amino acid level, 113 (50.7 %) were conserved even at the nucleotide level (variability ≤1% for each nucleotide of the codon). Notably, among amino acid residues that, because of the degeneration of the genetic code, could be encoded by more than four codons, the same codon usage occurred in 63 of 146 (43.2 %) residues in the overlapping region versus four of 77 (5.2 %) in the non-overlapping one, again underlying the higher freedom in synonymous variability within the non-overlapping portion of the RT ORF.

As expected, all three catalytically essential aspartate residues at RT positions 83, 205 and 206 were fully conserved (≤1% of variability), along with the tyrosine at RT position 203 in the YMDD motif. Only the YMMD methionine at position 204, well-known to be associated with drug resistance, showed variability (9.1 %). Other conserved areas of the RT enzyme were the N-terminal regions of the palm (aa 55–96, 82.5 % conserved amino acids) and finger (aa 1–54, 70.4 % conserved amino acids) domains (Figs 1c and 2). Within these domains, the regions encompassing aa 2–6, 24–34, 39–52, 58–75, 86–90 and 92–96 also showed a nucleotide variability <0.5 %, suggesting that these regions may contain residues critical for the interaction with the primer template and with the incoming dNTP substrates (Das et al., 2001). Furthermore, these regions correspond to specific HBsAg domains (such as the first transmembrane domain and the first cytosolic loop), containing residues critical for proper HBsAg secretion and maturation of virus particles (Blanchet & Sureau, 2006).

This means that specific HBV genomic areas encode, in alternative ORFs, critical domains for both RT and HBsAg function. Therefore, these highly conserved regions may deserve further investigation as a target for novel therapeutic approaches based on the design of non-nucleoside RT inhibitors and of new RNA interference (RNAi). Small interfering RNAs have been shown to be a potential new and attractive tool for HBV therapy in preclinical in vitro or in vivo studies (Zhang et al., 2010). However, this therapeutic approach is a long way from clinical trials in large cohort of patients. So far, only a phase 1b (first in humans) study has shown the safety and tolerability of the small interfering RNA NUC B1000, in three subjects with mild to moderate chronic hepatitis B (Gish et al., 2011). Thus, further efficacy and safety studies are needed to acquire a clear picture of both the therapeutic potential and risks of RNAi-based therapy for chronic hepatitis B in humans.

The above-mentioned highly conserved RT regions also contain the HLA class I epitope encompassing RT positions 42–51 and the HLA class II epitope encompassing RT positions 39–52 (Fig. 1c). The HBV-specific CD8 T-cell response is known to play a fundamental role in virus clearance and in the pathogenesis of liver disease (Chisari et al., 2010). Thus, these highly conserved HLA epitopes might deserve further investigation as novel potential vaccine targets based on eliciting a cytotoxic response. Indeed, the currently available HBV vaccine (based on recombinant HBsAg) is highly effective. However, HBV infections may occur despite vaccination, as reported in the Gambia, Alaska and Taiwan (Chang, 2010; McMahon, 2004; van der Sande et al., 2006). A recent study in Taiwan (Lai et al., 2012) showed an increased prevalence of HBV infection by HBsAg mutants in individuals older than 18 years after neonatal vaccination, thus suggesting that new preventative strategies may be needed for adults. As suggested recently (Chang, 2010), new vaccine targets might also be useful for vaccine therapies against chronic hepatitis infection, particularly during the immune-tolerance phase characterized by poor immune responses against HBV.

A decrease in the extent of genetic conservation was observed in the non-overlapping region of RT, such as the thumb domain and the C-term part of the palm domain (68.6 and 65.5 % conserved amino acids, respectively).

Finally, an interesting scenario was observed for the C-terminal part of the finger domain (aa 121–175). Indeed, the region encompassing aa 154–175, containing residues critical for RT activity (Das et al., 2001), was much more conserved than the region encompassing aa 121–153 (77.3 vs 33.3 % conserved amino acids). Up to 42.2 % of residues within the aa 121–153 region showed variability >25%. Interestingly, this region is lacking in the RT belonging to members of the family Retroviridae (Das et al., 2001), and corresponds to the first loop of the major B-cell epitope of the HBsAg, known as the ‘a-determinant’, where immune-escape mutations are localized. This suggests that the organization of the HBV genome in overlapping ORFs has shaped the structural

---

**Fig. 2.** Conservation pattern of HBV RT in relation to the three-dimensional structure of the RT protein, modelled for homology upon HIV-1 RT. RT domains are indicated in the protein structure (Kohlastaedt et al., 1992). RT residues involved in the overlap between RT and HBsAg (residues 8–236) are visualized as atom-bond spheres, while the non-overlapping RT region (residues 1–7 and 237–344) is represented as molecular surface.
organization of the RT and HBsAg proteins by imposing a compartmentalization of genomic areas that can tolerate a higher degree of genetic variability. This is critical in order to balance the need to preserve their correct function and to allow HBV to react to endogenous selective pressure imposed by the immune system. Beyond the mechanism of virus evolution via the p1/s3 and p3/s2 positions proposed by Zaaijer et al. (2007), this structural constraint may drive HBV adaptive evolution under positive selection pressure.

This is one of the first studies aimed at providing a map of conserved and variable regions throughout the entire HBsAg protein. Analysis of 3482 full-length HBsAg sequences showed 126 of 226 (55.8 %) conserved amino acids. Overall, the most conserved HBsAg regions were those between aa 25 and 39 (86.7 % conserved amino acids), aa 77 and 108 (71.9 % conserved amino acids, with the exception of position 85, with >12 % amino acid variability) and aa 144 and 157 (85.7 % conserved amino acids) (Fig. 3a). As reported above, residues 25–39 have been shown to be critical for HBsAg secretion or proper virus particle maturation (Blanchet & Sureau, 2006), thus supporting the high level of conservation in both the HBsAg and RT ORFs. Similarly, residues 77–108 contain amino acids fundamental for HBsAg secretion (Blanchet & Sureau, 2006) and the whole II transmembrane domain. Lastly, the region of aa 144–157 contains one of the two HBsAg N-linked glycosylation sites. The loss of this N-linked

Fig. 3. HBsAg sequence variability according to the distribution of amino acid variations along the RT protein in the overall population (a) and according to HIV–HBV co-infection status (b). TDM, Transmembrane domain; CYL, cytosolic domain. The mapping of HLA class I and II epitopes is also shown; the list of HLA epitopes was retrieved from Desmond et al. (2008). Epitopes are shaded according to the degree of conservation: black indicates that all the residues within the epitope are conserved; dark grey, the percentage of conserved residues within the epitope is between 90 and 100 %; light grey, the percentage of conserved residues within the epitope is between 80 and 90 %; dashed lines, the percentage of conserved residues within the epitope is <80 %. Solid line, HIV–HBV co-infected patients; dotted line, HBV mono-infected patients.
glycosylation site hampers HBV entry into hepatocytes. Indeed, in our population, this site is fully conserved. Conversely, the other glycosylation site at position 59, not critical for HBV entry, showed a substantial decrease in the extent of conservation (19.1% amino acid variability).

Recently, carbohydrate-binding agents targeting the glycans present on viral envelope glycoproteins have been proposed as an innovative therapeutic approach (Balzarini, 2007; François & Balzarini, 2012). In this light, the fully conserved N-linked glycosylation site at position 146 could represent a hot spot for targeting this new class of entry inhibitors.

A blunting in HBsAg conservation was observed in the antigenic loop containing the α-determinant (40.6% of conserved residues), in the HLA class I epitopes at HBsAg-positions 41–49 and 207–216 (44.4% and 40.0% conserved residues, respectively) and the HLA class II epitope at HBsAg positions 124–137 (35.7% conserved residues). This supports the central role of HBsAg in the immune control of HBV infection and, as mutations in this region may alter the antigenicity of the HBsAg protein, the presence of a strong positive selecting pressure. This concept was investigated further by analysing a subset of sequences derived from human immunodeficiency virus (HIV)–HBV co-infected and immune-compromised patients (n=108) (Table S1, available in JGV Online) versus a control population of mono-infected patients (n=577). Differences in the degree of amino acid conservation were specifically observed at the level of the ‘α’-determinant (aa 116–156) (Fig. 3b). In particular, co-infection status was associated with significantly higher amino acid conservation, with 20 of 41 (48.8%) of conserved ‘α’-determinant amino acids in HIV–HBV co-infected patients versus nine of 41 (21.5%) in HBV mono-infected (P=0.020 by Fisher’s exact test). This different pattern of amino acid conservation can be explained by a higher selecting pressure exerted by a competent immune system on HBV evolution, which selects a virus population with a better escape from a neutralizing immune response. Indeed, HBsAg positions 118, 120, 123, 126, 130, 133, 141 and 144, associated with the development of vaccine/immune-escape mutations, were more conserved in co-infected than in mono-infected patients.

Thus, overall findings support the hypothesis that the selective pressure imposed by the immune system (both neutralizing antibodies and cytotoxic responses) plays an important role in driving HBsAg genetic variability, and may thus explain the lower degree of conservation observed in HBsAg compared with the RT protein [126 of 226 (55.8%) conserved amino acids in HBsAg versus 223 of 344 (64.8%) in RT].

In this study, we provided a map of conserved regions throughout the RT and HBsAg proteins. However, we cannot exclude the possibility that mutations in these regions may occur under the pressure of novel drugs with a different mechanism of action. In addition, in our dataset, the number of RT and HBsAg sequences from patients infected with HBV genotypes E–G was lower than that from patients infected with HBV genotypes A–D. It is thus conceivable that, in these genotypes, the patterns of genetic conservation may be slightly different.

In conclusion, the overlapping reading frame and the immune system are major forces in shaping the patterns of RT and HBsAg genetic variability. During the course of evolution, the overlapping reading frame has shaped the structural organization of both the RT and HBsAg proteins in order to maximize the function of each protein and to allow the virus to react to the pressure imposed by the immune system. Specific genomic regions, critical for the function of both the RT and the HBsAg proteins, remain highly conserved at both the nucleotide and amino acid levels, and may thus be considered as novel targets for therapeutic strategies, involving the design of non-nucleoside RT inhibitors or carbohydrate-binding agents, or RNAi approaches.

Acknowledgements

We thank Michela Pollicita, Marianna Aragri and Claudia Alteri for their support in HBV sequencing, Massimiliano Bruni for data management, and all other participants and members of the ICoNA study group (Italy: A. Giacometti, A. Costantini, A. Riva (Ancona); G. Angaran, L. Monno, C. Carrisa, (Bari); F. Maggiolo, G. Lazzari (Bergamo); P. L. Viale, M. Borderi, G. Verucchi (Bologna); F. Castelli, C. Torti, C. Minardi (Brescia); T. Quirino, C. Abeli (Busto Arsizio); P. E. Manconi, P. Piano (Cagliari); J. Vecchiet, K. Falasca (Chieti); L. Sighinolfi, D. Segala (Ferrara); F. Mazzotta, S. Lo Caputo (Firenze); G. Cassola, G. Viscoli, A. Alessandri, R. Piscopo, G. Mazzarello (Genova); C. Mastroianni, V. Belvisi (Latina); P. Bonfanti, I. Caramma (Lecco); A. Chiodera, P. Castelli (Macerata); M. Galli, A. Lazzarin, G. Rizzardini, M. Puoti, A. d’Arminio Monforte, A. L. Ridolfi, R. Piolini, A. Castagna, S. Salpietro, A. Galli, A. Bigoloni, V. Spagnuolo, L. Careznì, P. Zucchi, M. C. Moioli, R. Rossotti, P. Cicconi, T. Formenti (Milano); C. Musmisi, L. Bisio (Modena); A. Gori, G. Lapadula (Monza), N. Abrescia, A. Chirianni, M. G. Guida, M. Gargiulo (Napoli); F. Baldelli, B. Belfiori (Perugia); G. Parruti, T. Ursini (Pescara); G. Magnani, M. A. Ursitti (Reggio Emilia); R. Cauda, M. Andreoni, A. Antinori, V. Tozzi, V. Vullo, A. De Luca, A. d’Avino, M. Zaccarelli, L. Gallo, E. Nicastro, R. Acinapura, M. Capozzi, R. Libertone, M. Lichtner, G. Tekano, (Roma); M. S. Mura, G. Madeddu (Sassari); P. Carmello, G. Di Perri, G. C. Orofino, M. Scandiera (Torino); G. Pellizzier, V. Manfrin (Vicenza). This work was supported by an grant from the Italian Ministry of Instruction, University & Research (MIUR) (RBAP11Y57K_001), Interomics project P. B. P05 coordinated by CNR, and the Aviralia Foundation.

References


