Molecular Evolution of Hepatitis B Virus over 25 Years

Carla Osiowy,1* Elizabeth Giles,1 Yasuhiro Tanaka,2 Masashi Mizokami,2 and Gerald Y. Minuk3

National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada1; Department of Clinical Molecular Informative Medicine, Nagoya City University Graduate School of Medical Sciences, Kawasaki, Mizuho, Nagoya, Japan2; and Section of Hepatology, Department of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada3

Received 15 May 2006/Accepted 3 August 2006

Determining the longitudinal molecular evolution of hepatitis B virus (HBV) is difficult due to HBV’s genomic complexity and the need to study paired samples collected over long periods of time. In this study, serial samples were collected from eight hepatitis B virus e antigen-negative asymptomatic carriers of HBV genotype B in 1979 and 2004, thus providing a 25-year period to document the long-term molecular evolution of HBV. The rate, nature, and distribution of mutations that emerged over 25 years were determined by phylogenetic and linear regression analysis of full-length HBV genome sequences. Nucleotide hypervariability was observed within the polymerase and pre-S/S overlap region and within the core gene. The calculated mean number of nucleotide substitutions/site/year (7.9 × 10⁻⁵) was slightly higher than published estimates (1.5 × 10⁻⁵ to 5 × 10⁻⁵). Nucleotide changes in the quasispecies population did not significantly alter the molecular evolutionary rate, based on linear regression analysis of evolutionary distances among serial clone pre-S region sequences. Therefore, the directly amplified or dominant sequence was sufficient to estimate the putative molecular evolutionary rate for these long-term serial samples. On average, the ratio of synonymous (dS) to nonsynonymous (dN) substitutions was highest for the polymerase-coding region and lowest for the core-coding region. The low dN/dS ratios observed within the core suggest that selection occurs within this gene region, possibly as an immune evasion strategy. The results of this study suggest that HBV sequence divergence may occur more rapidly than previously estimated, in a host immune phase-dependent manner.

MATERIALS AND METHODS

Patients. The eight hepatitis B virus surface antigen (HBsAg)-positive subjects described in this study were identified as a result of their participation in a large seroepidemiologic survey of HBV infection within their community conducted in 1979 (20). Twenty-one additional subjects were also found to be HBsAg positive during that survey, but these individuals had either died or left the settlement or were out “on the land” hunting when investigators returned in 2004 to repeat clinical evaluations and obtain follow-up serum samples. None of the eight participating subjects had received HBV treatment or HBV immunoprophylaxis or had been treated with immunosuppressant drugs for other medical disorders during the intervening 25 years. Informed consent was obtained from each subject on both occasions. The two study protocols (1979 and 2004) were approved by the University of Manitoba Conjoint Ethics Committee for Human Experimentation.

DNA extraction. DNA was extracted from 150 μl of serum by the proteinase K-sodium dodecyl sulfate lysis and phenol-chloroform extraction methods, as described previously (25), and was resuspended (final volume, 30 μl) in sterile, nuclease-free water. The extracted DNA was stored at −20°C.

Viral load determination. HBV DNA quantitation was performed by real-time PCR analysis using a RealArt HBV PCR kit (Artus Biotech, Qiagen, Mississauga, Ontario, Canada) with an ABI Prism 7500 sequence detection system. Briefly, 2.5 μl of DNA extract was added to 7.5 μl sterile water and 15 μl kit master mix consisting of buffers, enzyme, primers, and probe for the specific amplification of a 134-bp region of the HBV genome. One microliter of kit internal control was also added per reaction to identify possible PCR inhibition.

The DNA quantity (international units [IU]/ml) was determined by comparison to a standard curve obtained by serial dilutions of a positive control DNA sample. The PCR products were separated on a 2% agarose gel, and purity was verified by an internal control. The DNA was quantitated as the 10 IU/ml standards were chosen and used to obtain the final quantitation (range, 10 IU/ml to 1 × 10⁸ IU/ml). Real-time PCR qPCR data were analyzed using the Ct value obtained from the reaction, with the ΔCt value determined as the mean of samples divided by the mean of controls. The ΔCt value correlated with the log10 IU/ml ranges for each sample and was used to determine the number of copies per milliliter of serum for each sample.

The DNA was extracted from 150 μl of serum by the proteinase K-sodium dodecyl sulfate lysis and phenol-chloroform extraction methods, as described previously (25), and was resuspended (final volume, 30 μl) in sterile, nuclease-free water. The extracted DNA was stored at −20°C.

Viral load determination. HBV DNA quantitation was performed by real-time PCR analysis using a RealArt HBV PCR kit (Artus Biotech, Qiagen, Mississauga, Ontario, Canada) with an ABI Prism 7500 sequence detection system. Briefly, 2.5 μl of DNA extract was added to 7.5 μl sterile water and 15 μl kit master mix consisting of buffers, enzyme, primers, and probe for the specific amplification of a 134-bp region of the HBV genome. One microliter of kit internal control was also added per reaction to identify possible PCR inhibition.

The DNA quantity (international units [IU]/ml) was determined by comparison to a standard curve obtained by serial dilutions of a positive control DNA sample. The PCR products were separated on a 2% agarose gel, and purity was verified by an internal control. The DNA was quantitated as the 10 IU/ml standards were chosen and used to obtain the final quantitation (range, 10 IU/ml to 1 × 10⁸ IU/ml). Real-time PCR qPCR data were analyzed using the Ct value obtained from the reaction, with the ΔCt value determined as the mean of samples divided by the mean of controls. The ΔCt value correlated with the log10 IU/ml ranges for each sample and was used to determine the number of copies per milliliter of serum for each sample.

The DNA was extracted from 150 μl of serum by the proteinase K-sodium dodecyl sulfate lysis and phenol-chloroform extraction methods, as described previously (25), and was resuspended (final volume, 30 μl) in sterile, nuclease-free water. The extracted DNA was stored at −20°C.

Viral load determination. HBV DNA quantitation was performed by real-time PCR analysis using a RealArt HBV PCR kit (Artus Biotech, Qiagen, Mississauga, Ontario, Canada) with an ABI Prism 7500 sequence detection system. Briefly, 2.5 μl of DNA extract was added to 7.5 μl sterile water and 15 μl kit master mix consisting of buffers, enzyme, primers, and probe for the specific amplification of a 134-bp region of the HBV genome. One microliter of kit internal control was also added per reaction to identify possible PCR inhibition.

The DNA quantity (international units [IU]/ml) was determined by comparison to a standard curve obtained by serial dilutions of a positive control DNA sample. The PCR products were separated on a 2% agarose gel, and purity was verified by an internal control. The DNA was quantitated as the 10 IU/ml standards were chosen and used to obtain the final quantitation (range, 10 IU/ml to 1 × 10⁸ IU/ml). Real-time PCR qPCR data were analyzed using the Ct value obtained from the reaction, with the ΔCt value determined as the mean of samples divided by the mean of controls. The ΔCt value correlated with the log10 IU/ml ranges for each sample and was used to determine the number of copies per milliliter of serum for each sample.
cycling parameters and result interpretation were carried out according to the manufacturer’s protocol.

**PCR amplification.** Full-length genome sequencing of HBV DNA was performed by nested PCR with a full-length amplicon obtained using the primers and thermocycling conditions described by Günther et al. (10). Thereafter, several nested PCR steps were performed in order to increase the sensitivity of detection and produce a sufficient amount of amplicon for sequence analysis. The sequences and annealing temperatures of the nested PCR primer sets used are shown in Table 1. PCR was performed using an ultra-high-fidelity polymerase (AccuPrime Pfx DNA polymerase; Invitrogen Life Technologies, Burlington, Ontario, Canada) to ensure low to nil error rates during amplification (5). Reaction tubes for PCR contained 5 μl DNA extract or 2 μl of the first-stage PCR product, AccuPrime Pfx reaction mix (Invitrogen Life Technologies), a 0.5 μM concentration of each primer, and 1 U AccuPrime Pfx DNA polymerase. Thermal cycling parameters for each set of primers were those suggested by the manufacturer (Invitrogen Life Technologies) for three-step cycling using the annealing temperatures listed in Table 1.

**Full-genome sequence analysis.** Nested PCR products were gel purified prior to cycle sequencing with an ABI Prism 3100 genetic analyzer (Applied Biosystems, Foster City, California), using BigDye v3.1 Terminator chemistry. All sequences were assembled using SeqMan II software (DNASTar Inc., Madison, Wisconsin). Full-length genome sequence alignments were performed using ClustalX v1.8 (32). Sequence identity and divergence were calculated based on the number of nucleotide changes per total number of nucleotides analyzed (3,215 bp). The number of nucleotide substitutions per site per year, based on the number of nucleotide changes per total number of nucleotides analyzed (3,215 bp), was calculated using the equation of Gojobori and Yokoyama (8, 24). Phylogenetic tree analysis was performed using the Tamura-Nei model of evolutionary distance, and the topology was evaluated by bootstrap analysis (1,000 replicates) using the neighbor-joining method. Linear regression analyses based on Tamura-Nei model evolutionary distances were performed to calculate a mean evolutionary rate of quasispecies sequences over the 25-year period.

**Nucleotide sequence accession numbers.** The full-length genome sequences obtained from the patients at each time point were submitted to the National Center for Biotechnology Information GenBank database under accession numbers DQ463787 to DQ463802.

## RESULTS

### Patient clinical and virological data.
As shown in Table 2, the mean (± standard deviation) age of the eight HBsAg-positive subjects in 2004 was 69.8 ± 12.6 years, and 7/8 (88%) subjects were male. All subjects were HBBeAg negative and anti-HBe positive on both occasions (1979 and 2004). Seven subjects were positive for antibody to hepatitis A virus (anti-HAV) and negative for anti-HCV in both 1979 and 2004. The remaining individual was anti-HAV negative in 1979 but positive in 2004. This individual was also anti-HCV negative on both occasions. Liver biochemistry (serum alanine aminotransferase, alkaline phosphatase, total bilirubin, and albumin) results for these patients are shown in Table 2.

### Statistical analysis

The full-length genome sequences were assembled using SeqMan II software (DNAStar Inc., Madison, Wisconsin), using BigDye v3.1 Terminator chemistry. All sequences were assembled using SeqMan II software (DNAStar Inc., Madison, Wisconsin). Full-length genome sequence alignments were performed using ClustalX v1.8 (32). Sequence identity and divergence were calculated based on the number of nucleotide changes per total number of nucleotides analyzed (3,215 bp). The number of nucleotide substitutions per site per year, based on the number of nucleotide changes per total number of nucleotides analyzed (3,215 bp), was calculated using the equation of Gojobori and Yokoyama (8, 24). Phylogenetic tree analysis was performed using the Tamura-Nei model of evolutionary distance, and the topology was evaluated by bootstrap analysis (1,000 replicates) using the neighbor-joining method. Linear regression analyses based on Tamura-Nei model evolutionary distances were performed to calculate a mean evolutionary rate of quasispecies sequences over the 25-year period.

### Quasispecies analysis.
Viral quasispecies were investigated by clonal analysis of the pre-S gene regions from four sample pairs. A 479-bp amplicon from the pre-S1/pre-S2 gene region was obtained, using primers P1 and P2 as described previously (17). Amplicons were gel purified and cloned into a pCR2.1-TOPO plasmid vector (Invitrogen Life Technologies) according to the manufacturer’s instructions. Ligated products were transformed into *Escherichia coli* TOP10F cells (Invitrogen Life Technologies), at least 10 individual colonies were picked, and the plasmid DNA inserts were sequenced.

Mean genetic distances of all synonymous and nonsynonymous positions were calculated using the Pamilo-Bianchi-Li (P-B-L) model (MEGA v.2.1). Linear regression analyses based on Tamura-Nei model evolutionary distances were performed to calculate a mean evolutionary rate of quasispecies sequences over the 25-year period.

### Viral load analysis.

Viral load analysis was performed on all eight samples (1979 and 2004) using the real-time PCR assay. The results are shown in Table 2.

### Summary of findings

In summary, the results of this study indicate that there was no significant change in the viral load over the 25-year period. The mean viral load was 10^6 IU/ml in 1979 and 10^6 IU/ml in 2004 for the eight subjects. The mean viral load was 10^6 IU/ml in 1979 and 10^6 IU/ml in 2004 for the eight subjects. The mean viral load was 10^6 IU/ml in 1979 and 10^6 IU/ml in 2004 for the eight subjects. The mean viral load was 10^6 IU/ml in 1979 and 10^6 IU/ml in 2004 for the eight subjects. The mean viral load was 10^6 IU/ml in 1979 and 10^6 IU/ml in 2004 for the eight subjects.
Ferace, alkaline phosphatase, total bilirubin, and albumin levels) tests were normal for all individuals on each occasion. HBV DNA viral loads were variable for most patients at the two time points measured, ranging over approximately 1 log. Overall, the DNA levels remained very low at both time points for all patients (<10^4 IU/ml).

**Sequencing.** All sera collected in 2004 were extracted, amplified, and sequenced prior to analysis of the 1979 sample set to avoid the possibility of contamination between matched samples. Following sequencing and assembly of each genome sequence, a 3,215-bp genome was obtained for each sample. Each genome sequence was genotyped using the NCBI genotyping tool (26), and all were determined to be genotype B. One sample pair (from patient 462-4) had a premature stop codon within HBcAg (182 of 184 codons) and lacked a start codon for the pre-S2 gene for the viral genomes from both time points. The precore stop mutation at nucleotide 1896 of the HBeAg gene was observed in all genome sequences; however, none of the genomes contained mutations at nucleotides 1762 and 1764 within the core promoter region.

**Phylogenetic analysis.** The 16 full-length HBV genome sequences were aligned, and a histogram was prepared to visually demonstrate regions of hypervariability and relative conservation along the length of the genome (Fig. 1). Nucleotide changes occurred throughout the entire genome and in each coding region. Regions of hypervariability were observed within the core gene, the 3' end region of the S gene, and other regions of overlap between the polymerase and pre-S/S genes. Conversely, the X gene and the overlap region encompassing the S major hydrophilic region and reverse transcriptase domains B and C within the polymerase gene were observed to have fewer nucleotide substitutions among the 16 sequences.

Phylogenetic analysis showed that most matched samples formed unique clusters within the tree, while two sample pairs did not cluster significantly, suggesting that the matched samples from these pairs diverged independently (Fig. 2). In order to avoid a false overestimation of the mean molecular evolutionary rate, sequences from these two sample pairs were not included in rate calculations, except for quasispecies analysis (for patient 234-6 only). The range of sequence divergence
among the eight matched pairs was 0.3% (patient 462-4) to 1.9% (patient 234-6).

**Molecular evolutionary rate analysis.** Values for the number of nucleotide substitutions per site per year, based on direct comparison between the six sample pair sequences that uniquely clustered, ranged from $6.23 \times 10^{-5}$ (patient 462-4) to $3.59 \times 10^{-4}$ (patient 539-16), with a mean value of $1.9 \times 10^{-4}$.

To more accurately determine the molecular evolutionary rate based on a putative ancestral sequence, linear regression analysis was performed (Fig. 3). Using this method of analysis, the mean evolutionary rate among six sample pairs was determined to be $7.9 \times 10^{-5}$ nucleotide substitutions/site/year.

**Quasispecies analysis.** The contribution of viral quasispecies to the molecular evolutionary rate of sample pairs over the 25-year period was investigated by clonal analysis of four sample pairs. Sample pairs were chosen based on their divergence over the study period, with two pairs demonstrating the most sequence divergence over time (234-6 and 539-16 [1.9% and 1.8% divergence, respectively]) and two...
pairs demonstrating the least sequence divergence over time (462-4 and 650-15 [0.3% and 0.4% divergence, respectively]). The pre-S region was selected for quasispecies analysis because this region is recognized for its hypervariability within the HBV genome (13).

The mean overall genetic distances (including synonymous and nonsynonymous sites) between quasispecies sequences from 1979 and 2004 were determined (Table 3). A significant reduction in the overall genetic distances in 2004 compared to those in 1979 was observed (except for samples from patient 650-16), suggesting that selective pressure occurred within the pre-S region. Following nonsynonymous versus synonymous substitution analysis of quasispecies, only sample pair 234-6 quasispecies showed greater nonsynonymous than synonymous changes over time (ratio, 1.247), indicating positive selection. The other three sample pairs showed no positive selection over the 25-year period (ratios of nonsynonymous to synonymous substitutions, 0.438 [462-4], 0.492 [539-16], and 0.232 [650-16]). Sample pair 234-6 quasispecies sequences also showed a shift within the pre-S2 region, with a three-codon deletion in 6 of 11 clones from 2004 that was not observed in any clones from 1979 (n/H11005 11).

The mean molecular evolutionary rate based on pre-S quasispecies sequences from the four sample pairs (8.3 × 10⁻⁵/site/year) (Fig. 4A) was compared to the calculated rate based on the pre-S region from each dominant sequence of the six full-length sample pairs as well as the samples from patient 234-6 (7.2 × 10⁻⁵/site/year) (Fig. 4B). No significant difference in mean rate was observed (P > 0.05), indicating that quasispecies variation does not contribute considerably to the overall mean evolutionary rate during long-term follow-up over 25 years.

**Synonymous versus nonsynonymous substitutions of full-length genome coding regions.** The ratio (d_S/d_N) of synonymous (d_S) to nonsynonymous (d_N) substitutions was calculated for the HBsAg, polymerase, core, and X coding regions for the six sample pairs showing unique phylogenetic clustering (Table 4). This ratio determines the extent of natural selection, such that a ratio of <1 indicates positive selection within the gene. Ratios of <1 were observed for several coding regions from several of the patients, but the majority of coding regions had ratios of >1, indicative of sequence stability or negative selection over time. In general, the lowest ratios were observed for the core coding region. Conversely, the polymerase coding region demonstrated the highest d_S/d_N ratios among all patients.

**DISCUSSION**

The current study investigates the molecular evolution of HBV in individuals over a 25-year period. To our knowledge, this study is one of the first to analyze longitudinal HBV evolution in individuals over a long period of time.

![Comparison of evolutionary rate calculations based on quasispecies versus dominant strain sequences.](http://jvi.asm.org/)

**FIG. 4.** Comparison of evolutionary rate calculations based on quasispecies versus dominant strain sequences. Evolutionary rates were calculated by regression analyses (Tamura-Nei model) using either (A) quasispecies (cloned) pre-S sequences (n = 114) from four sample pairs or (B) dominant strain (directly amplified) pre-S sequences (n = 14) from seven sample pairs. The differences between the evolutionary rates in panels A and B are not significant (P > 0.05).
TABLE 4. Ratio of synonymous to nonsynonymous substitutions for coding regions among full-length genomes

<table>
<thead>
<tr>
<th>Sample pair</th>
<th>Coding region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HBsAg</td>
</tr>
<tr>
<td></td>
<td>Sd/a</td>
</tr>
<tr>
<td>621-14</td>
<td>2.00</td>
</tr>
<tr>
<td>524-3</td>
<td>1.00</td>
</tr>
<tr>
<td>650-15</td>
<td>0.00</td>
</tr>
<tr>
<td>539-16</td>
<td>5.00</td>
</tr>
<tr>
<td>462-4</td>
<td>1.00</td>
</tr>
<tr>
<td>928-9</td>
<td>2.00</td>
</tr>
</tbody>
</table>

a Number of observed synonymous substitutions.
b Number of observed nonsynonymous substitutions.
c Ratio of synonymous to nonsynonymous substitutions, calculated using the Jukes-Cantor correction for multiple hits of pS and pN (23).
d NA, not available (Sd/Sn was not calculated if either Sd or Sn equaled 0).

The observation of independent evolution in these two study subjects led us to investigate the contribution of quasispecies to the HBV molecular evolutionary rate. Immune selective pressure coupled with the lack of proofreading activity by the HBV polymerase likely contributes to the development of quasispecies complexity and diversity during infection (27). Viral quasispecies have very closely related genomes but exist in an environment of mutation, selection, and competition, thus creating a dynamic and changing population over time (7). Based on phylogenetic and linear regression analysis of pre-S1/S2 sequence quasispecies from four sample pairs, it was determined that the mean molecular evolutionary rate did not diverge significantly from the rate calculated using pre-S sequences derived from the directly sequenced or dominant strain. This result suggests that the quasispecies populations from HBeAg-
negative, asymptomatic chronic HBV carriers during long-term follow-up over 25 years did not contribute significantly to the putative overall evolutionary rate and, therefore, that the dominant sequence is sufficient for rate estimation.

Although the sample pair quasispecies evolutionary rate was not significantly different, the observation that two sample pairs showed independent evolution may be related to quasispecies competition. Replacement of the dominant strain observed in 1979 with a minority quasispecies strain may have occurred due to a selective advantage of the minority variant during the quiescent phase of chronic infection in the study participants (9). For example, the three-codon deletion observed in the pre-S2 region from the majority of 234-6 quasispecies clones from 2004 may contribute a selective advantage to the virus to allow it to become the dominant strain sometime in the future. Such mutations within the pre-S2 region (deletions and start codon mutations) are characteristic of genomes from the HBcAg-negative phase of infection (9). Another explanation for the observed independent evolution may be re-infection with a different HBV genotype B strain during the follow-up period.

The mean nucleotide substitution rate observed in this study was slightly higher than previously estimated rates based on HBcAg-positive carriers, as transmission is assumed to occur predominantly through HBcAg-positive donors (1, 13, 24, 29). However, transmission from carriers negative for HBcAg has been documented (4, 6, 31). The observed evolutionary rate validates previous statements that viruses lacking HBcAg evolve more rapidly, possibly as a function of increased immune pressure during the immune clearance phase of infection (1, 3, 11, 19, 28). The chronic infection phase for all patients investigated in this study was typically quiescent or asymptomatic and HBcAg negative, suggesting a reduction in host immune activity following seroconversion. Therefore, the significantly higher evolutionary rate observed in this study, despite less selection pressure, may be related to the seroconversion event driving quasispecies complexity and diversification. The more diversified quasispecies pool would then undergo competition during the follow-up period to obtain the most "fit," and thus dominant, genome. Furthermore, the reduced selection pressure during the HBcAg-negative chronic phase may allow the accumulation of mutations due to error-prone reverse transcription during replication. Indeed, since overall more synonymous mutations were observed in the coding regions of the study sequences, it is likely that selective outgrowth of sequences having a structure/function advantage for the virus occurred throughout the follow-up period (27).

In conclusion, further analysis of HBV evolutionary patterns should include both HBcAg-positive and -negative symptomatic and asymptomatic patients representing different HBV genotypes to truly characterize HBV sequence divergence over time. In this manner, estimating the molecular clock and origins of HBV may be done more accurately.

REFERENCES


