Sequencing Variation within a Nonstructural Region of the Hepatitis G Virus Genome

YURY E. KHUDYAKOV,1* MIAN-ER CONG,1,2 MARIA-Teresa BONAFONTE,1 SULAIMAN ABDULMALEK,1,3 BARBARA L. NICHOLS,1 STEPHEN LAMBERT,1 MIRIAM J. ALTER,1 AND HOWARD A. FIELDS1

Hepatitis Branch, Division of Viral and Rickettsial Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia1; Institute of Virology, Chinese Academy for Preventive Medicine, Beijing, People’s Republic of China2; and Specialty Laboratories, Inc., Santa Monica, California3

Received 6 December 1996/Accepted 10 June 1997

Nine sets of nested PCR primers from a 2.6-kb region of the hepatitis G virus (HGV) genome at nucleotide positions 5829 to 8421 were designed and used to analyze serum specimens obtained from patients with community-acquired non-A, non-B hepatitis who were HGV RNA positive. One set of primers was found to be most efficient in detecting HGV and was subsequently used to test 162 HCV-positive and 11 HCV-negative plasma units obtained from individual paid donors. HGV RNA was detected in 30 (17.3%) plasma units, 2 of which were found among the 11 HCV-negative specimens. A complete set of nine PCR fragments was obtained from two patients with community-acquired acute non-A, non-B hepatitis and from four paid donors. All PCR fragments were sequenced and were shown to have a nucleotide similarity of 85.9 to 92.3% and a derived amino acid similarity of 96.0 to 99.0%. The majority of nucleotide changes occurred in the third position of codons. The HGV nucleotide and protein sequences obtained in this study were compared with HCV sequences. Based on this analysis the 2.6-kb fragment was predicted to encode the C-terminal part of the putative NS4b, the entire NS5a, and almost the complete NS5b proteins. Putative protease cleavage sites separating these proteins were also predicted. In serial specimens obtained from the two HGV-infected patients, no significant variations were found in the HGV nucleotide and derived amino acid sequences over time. The HGV sequences obtained from one patient showed no changes over 6 months, whereas more than 99.0% homology was observed for sequences from the second patient over 2.5 years. Heterogeneity analysis performed on 10 sequences obtained in this study and corresponding regions from 6 known full-size sequences of the HGV genomes demonstrated notable discrete heterogeneity consistent with the existence of HGV genetic groups or types.
different HGV isolates. This observation adds credence to the idea that distinct HGV genetic groups exist.

MATERIALS AND METHODS

Serum specimens. Ten serum specimens obtained from 10 patients that were HGV RNA positive by PCR (8a) were identified through the Centers for Disease Control and Prevention Sentinel Counties study of community-acquired viral hepatitis (2).

Serum samples (n = 173) from paid blood donors were randomly selected from a collection of anti-HCV-positive units repositioned at Boston Biomedica Inc. (West Bridgewater, Mass.). Initially, each plasma unit obtained for this study was tested for anti-HCV positive by a screening assay, which was performed at the donation site. However, upon repeat testing of these units with a second generation anti-HCV enzyme immunoassay (Abbott Laboratories, Chicago, Ill.) and supplemental testing with a semiautomated dot blot immunoassay (MATRIX; Abbott Laboratories) or recombinant immunoblot assay (RIBA-3; Ortho Diagnostic Systems, Inc., Raritan, N.J.), the anti-HCV status could not be confirmed for 11 specimens, while the status of 3 specimens was identified as indeterminate because of immunoreactivity with a single HCV protein.

All plasma units were tested by nested reverse transcription-PCR for the presence of the HCV genome. The protocol for HCV RNA extraction from serum and cDNA preparation has been described (4).

The protocol for nested PCR was 30 cycles of amplification with a pair of external primers 5'-CTGTTAGGGGCTGGTATG 3' and 5'-GGTGCTTACATCACCGCCAC 3'.

For the identification of the HGV genome (see Materials and Methods), the selected 2.6-kb HCV region for this study was predicted to encode the C-terminal NS4b region, was found to be the most efficient primer derived from the NS5 region, was found to be HGV positive. Nevertheless, YK875 was modified to 5'-CAGAACATCAATGCTTGGTAC (YK1183) to avoid this mismatch. Although no detectable increase in sensitivity was observed, all further experiments used YK1183 in place of YK875. No single specimen tested with these nine sets of primers yielded all nine PCR fragments, and additional primers were designed to obtain missing PCR fragments. The design of these additional primers was based upon new sequence information obtained from sequencing PCR fragments for each individual specimen.

Sequencing. The primary structure of PCR fragments was determined with an automated sequencer (373 DNA sequence; Applied Biosystems, Foster City, Calif.) according to the manufacturer’s protocol. In all cases, internal PCR primers were used as sequencing primers.

Computer-assisted sequence analysis. Nucleotide and amino acid sequence comparisons were performed with the MegAlign program from the Lasergene software package (DNAStar Inc., Madison, Wis.) and the PILEUP program from the Wisconsin Sequence Analysis Package (Genetics Computer Group, Madison, Wis.). Evolutionary distances between pairs of sequences were determined by the DNADIST program from the PHYLIP package (J. Felsenstein, University of Washington, Seattle). The computed distances were used to construct phylogenetic trees by the program NEIGHBOR. Additionally, the program MegAlign was used to construct phylogenetic trees.

Nucleotide sequence accession numbers. GenBank accession numbers for the HGV full-size genome sequences used in this study are as follows: GBV-C from West Africa, U36380; HGV PNF2161 from North America, U44402; HGV R10291 from North America, U43796; HGV D79255 from Japan, D87255; HGV U63715 from East Africa, U63715; and HGV U75356 from China, U75356. Sequences of the nonstructural region of HGV variants were deposited in GenBank with accession no. U96121 for PRCS118, U96119 for JFA2113, U96123 for HGV-14, U96124 for HGV-16, U96126 for HGV-31, and U96125 for HGV-26. A follow-up specimen corresponding to PRCS118 was sequenced and given GenBank accession no. U96122. For JFA2113, three additional isolates were obtained and deposited under GenBank no. U96117, U96118, and U96120 (see below for details).

RESULTS AND DISCUSSION

Identification of HGV-positive specimens. Using the PNF2161 HGV prototype sequence (13) we designed nine sets of nested PCR primers. All primers were designed to amplify overlapping PCR fragments of approximately 400 bp that collectively span the 2.6-kb nonstructural region at nucleotide positions 5829 to 8421. Based on a comparison of HGV and HCV nucleotide sequences, the selected 2.6-kb HGV region for this study was predicted to encode the C-terminal NS4b region, the entire NS5a region, and a significant part of NS5b. Ten of all 11 PCR sets of primers were applied to 11 serum specimens previously identified as HGV RNA positive by PCR. One set of primers, amplifying a 400-bp fragment derived from the NS5 region, was found to be the most efficient for the identification of the HGV genome (see Materials and Methods). This set (set 6) amplified a fragment from all HGV-positive serum specimens tested, whereas other sets often failed.

Set 6 was used to test the 173 plasma units obtained from individual paid donors for HGV RNA. A total of 30 units were found to be HGV PCR positive: 27 (16.9%) of the 159 HCV-positive units, 2 (18.2%) of the 11 HCV-negative units, and 1 of 2 units that yielded anti-HCV-indeterminate and HCV PCR-negative results.

Sequences of the nonstructural regions from six HGV variants. Two specimens (PRCS118 and JFA2113, drawn 6 and 4.5 years, respectively, after onset of illness) from patients with chronic-Aquatic acute non-A, non-B hepatitis and four specimens from plasma donors were used for sequencing. Specimen PRCS118 was positive for HCV RNA by PCR and for anti-HCV antibody by both the screening and the supplemental assays, while specimen JFA2113 lacked any markers of HCV infection. Three specimens (HGV-14, HGV-16, and
HGV-31) from paid plasma donors were positive for both HCV RNA and antibody, and one specimen (HGV-26) was negative for all markers of HCV infection. All six sequences were compared to the sequences of the corresponding region from the known HGV full-size genomes (PNF2161, R10291, D87255, U63715, and U75356) and to GBV-C sequences (8, 12, 13, 21, 27). The results of this analysis are presented in Fig. 1. The nucleotide sequences of these six new HGV variants were 86.4 to 91.5% similar to each other. When the prototype HGV strains (8, 13, 21, 27) and GBV-C (12) sequences were used in the homology analysis, the similarity varied from 85.9 to 92.3%, with the greatest divergence being observed between GBV-C and PNF2161 or R10291 HGV and between HGV D63715 from East Africa and U75356 from China. The highest similarity was observed between PNF2161 and HGV-26. Among the new variants identified in this study, HGV-14 and JFA2113 were the least homologous (86.2%). The deduced amino acid sequences were 94.5 to 99.5% similar. The highest protein similarity was observed for PRC5118 and Japanese HGV D87255 (2). The lowest amino acid sequence similarity was found for HGV-16 and Chinese HGV U75356 (27).

In general, all of the derived HGV sequences from within the 2.6-kb nonstructural region demonstrated relatively low variability. This observation may be explained by significant conservation of the primary structure because of important, or even essential, functions provided by these proteins in the HGV replication cycle. Alternatively, the low variability may be due to the origin of the isolates, since all of them came from North America. Various genotypes of HCV predominate in different parts of the world. It is important to note that experimental conditions can also affect the interpretation of sequence comparisons, and since HGV variants were identified by PCR, the design of the primers may have also introduced a selection bias. Because the complete nucleotide sequence for the entire 2.6-kb region was obtained only for those serum specimens that were amplified by several sets of primers, an additional selection bias may have been introduced.

Nucleotide and amino acid sequence variations within different regions. Analysis of the sequence variations demonstrated that nucleotide changes are almost uniformly distributed across the 2.6-kb region with only a few relatively conserved sites identified. The amino acid sequence showed relatively low variability with four regions demonstrating a high degree of variation (Fig. 1). It is interesting that one of the most variable protein regions is encoded by a relatively conserved RNA region identified as region I (Fig. 1). The opposite is true for variable regions II and IV, which demonstrate a high degree of conservation at the amino acid level. Region III, however, is very conserved at both the nucleotide and amino acid level. This region encodes a putative cleavage site between NS5a and NS5b (12), which may explain this observation. RNA regions IV and V are located at the 5' terminus of the predicted NS5b gene. Both regions maintain almost identical degrees of variability, with nucleotide changes affecting almost
30% of all nucleotide positions among nine HGV isolates. However, protein regions IV and V demonstrate two very different patterns: region IV is highly conserved, whereas region V is rather variable. Another important feature of this entire area is that RNA regions IV and V are flanked by highly conserved sequences that appear to be particularly suitable for designing PCR primers. This observation was confirmed by the finding that the most sensitive set of primers found in this study was derived from these conserved regions. Because of these properties, regions IV and V may represent an ideal location to perform genotype and/or phylogenetic analysis for HGV.

**Synonymous mutations.** A comparison of HGV nucleotide sequences revealed a substantial number of silent mutations of codons, which may explain the high degree of homology between the predicted protein sequences. Region IV (Fig. 1) provides the most notable examples of such mutations. As discussed above, the nucleotide sequence of this region and the encoded amino acid sequence are constrained differently: the nucleotide sequence demonstrates significant changes, while the amino acid sequence is highly conserved. Another interesting example of synonymous mutations was found within the JFA2113 sequence between nucleotide positions 8277 and 8306 (data not shown). This sequence contains a region in which 9 of 10 consecutive codons are affected by mutations, although none resulted in amino acid changes.

**Predicted protein cleavage sites.** The 2.6-kb region of the HGV genome analyzed in this study encodes the C-terminal part of NS4b, all of NS5a, and a significant part of NS5b, as predicted by comparison of HCV and HGV sequences. Alignment of sequences revealed the identity of RNA polymerase domains within the NS5b protein encoded by both the PRC2161 (13) and GBV-C (12) strains. Alignment analysis also allowed the prediction of two protease cleavage sites. One site was predicted at amino acids 1893 to 1896, separating NS4b from NS5a, and another at aa 2307 to 2313, separating NS5a from NS5b (Fig. 2). Neither of these predicted sites contained cysteine residues, while both sites contained negatively charged amino acids common to HCV and HGV. The HGV cleavage site separating NS5a and NS5b had more similarity with the HCV site than did the site between NS4b and NS5a. This observation allowed the prediction of the boundary between NS5a and NS5b more accurately at positions 2309 and 2310, with alanine as the C-terminal residue for NS5a and serine as the N-terminal residue for NS5b. This observation strongly confirms previous prediction of the NS5a-NS5b cleavage site made for the GBV-C polyprotein alone (12). It is interesting that when the entire 2.6-kb region was aligned with HCV sequences, a deletion of two codons, encoding both predicted protease cleavage sites, within the HGV sequence was identified (Fig. 2b).

**HGV sequences from serial serum specimens.** In addition to sequence variations between HGV isolates from different infected persons, another important characteristic of HGV genome stability is sequence variation over time for HGV strains.

---

**FIG. 2.** (a) Amino acid homology between the HCV NS4b-NS5a and NS5a-NS5b protein cleavage sites and the corresponding protein regions within the HGV polyprotein. (b) Nucleotide homology between the HCV genome regions encoding the NS4b/NS5a and NS5a/NS5b protein cleavage sites (9) and corresponding regions of the HGV genome. The HCV cleavage sites (a) and the HCV genome regions encoding these cleavage sites (b) are indicated by downward arrows. The hypothetical HGV protein cleavage sites are shown with brackets and upward arrows. HCV sequences used for the comparison are HCV1 (6), HCVJ6 (17), HCVJ8 (18), and HCVG9 (19).
obtained from individual patients. To analyze this sequence variation, we selected additional specimens from patients from whom PRC5118 and JFA2113 were isolated. For the former, a 6-month follow-up specimen was subsequently sequenced over the entire 2.6-kb region (GenBank accession number is U96122). For the latter patient, three additional specimens were identified, two samples that were collected 1 and 2 years before the specimen (see above) already sequenced (GenBank accession no. U96117 and U96118, respectively) and a third specimen (GenBank accession no. U96120) that was collected 6 months after the sequenced specimen. Analysis of these sequences showed that only minor nucleotide changes occurred over 30 months for the patient from whom JFA2113 was obtained, and no changes were found over the 6-month period for the patient from whom PRC5118 was obtained (data not shown). Percent pairwise homology of the HGV nucleotide sequences varied from 99.3 to 99.7% for all four specimens tested from the former patient, with derived amino acid sequence homology varying between 99.2 and 99.8%. Nucleotide and amino acid substitutions were uniformly scattered across the entire region. Thus, for the analyzed 2.6-kb region of the HGV genome, the nucleotide sequence is very stable and does not undergo considerable change over the course of infection in individual patients.

**Heterogeneity of HGV.** All 10 sequences from the HGV 2.6-kb nonstructural region obtained in this study were compared with the corresponding regions of five HGV sequences (8, 13, 21, 27) and one GBV-C sequence (12). Phylogenetic analysis of all 16 sequences was performed, and the results are presented in Fig. 3 and 4. The phylogenetic tree presented in Fig. 3 contains three major branches, 1, 2 and 3, which may represent three different genetic groups. The evolutionary distances were not normally distributed when all 16 sequences were analyzed (Fig. 4). Three distinct peaks, A, B, and C, were identified. The deviation from a normal distribution may be attributed to the small number of observations included in this analysis and/or may reflect discrete HGV heterogeneity suggesting the existence of distinct genetic groups. Peak A corresponds to sequence comparisons between HGV isolates from serial specimens obtained from individual patients and, thus, presents variations within a single HGV strain. Peak B corresponds to sequence homologies between related HGV strains obtained from different patients, while peak C, indicating sequences with highest evolutionary distances, implies the pres-
ence of distinct HGV groups. However, when sequences from only group 1, which were obtained from different individuals, were compared a normal distribution of evolutionary distances was observed (Fig. 4). Groups 2 and 3 do not have enough members for such an analysis. Also, a normal distribution was observed when sequences from different groups were compared to each other (Fig. 4). Thus, groups 1, 2, and 3 (Fig. 3) may represent three different HGV genotypes. However, this phylogenetic analysis failed to show full separation between genetic groups identified by peak C and subgroups identified by peak B in Fig. 4. This may pose a significant problem for establishing a firm genetic classification of different HGV variants into genotypes and subtypes.

Recently, evidence for the existence of different HGV genotypes based on analysis of the sequence heterogeneity of the 5′-terminal region of the HGV genome was reported (15). In a more recent publication, this observation was confirmed when full-length HGV genomic sequences were compared (20). Despite concordance in the number of genotypes found, it is difficult to determine whether the three genotypes found in this study and in the above-mentioned publications are identical (15, 20). It is interesting that phylogenetic analysis performed on a small NS5 region derived from HGV isolates obtained from many parts of the world failed to identify HGV genetic types (25). This discrepancy suggests that not every region of the HGV genome contains a sufficient amount of genetic information to discriminate between different HGV genotypes. Thus, larger regions or carefully selected regions should be used for the HGV phylogenetic analysis.

In conclusion, although only a small number of sequences were available to perform a complex heterogeneity analysis of the HGV genome, making definitive conclusions on the extent of HGV heterogeneity difficult, the data obtained in this study indicate that the HGV genome is discretely heterogeneous, consistent with the grouping of HGV into subtypes and putative genotypes.

REFERENCES

9. Gallagher, M. Personal communication.