The entire nucleotide sequences of two GB virus C/hepatitis G virus isolates of distinct genotypes from Japan

Hiroaki Okamoto,1 Haruhisa Nakao,2 Taisuke Inoue,3 Masako Fukuda,4 Junichi Kishimoto,4 Hisao Izuka,5 Fumio Tsuda,6 Yuzo Miyakawa7 and Makoto Mayumi1

1 Immunology Division, Jichi Medical School, Minamikawachi-Machi, Tochigi-Ken 329-04, Japan
2 First Department of Internal Medicine, Medical School, Nagoya City University, Aichi-Ken 467, Japan
3 First Department of Internal Medicine, Yamanashi Medical University, Yamanashi-Ken 409-38, Japan
4 Institute of Immunology, Tokyo 112, Japan
5 Japanese Red Cross Saitama Blood Center, Saitama-Ken 338, Japan
6 Department of Medical Sciences, Toshiba General Hospital, Tokyo 140, Japan
7 Miyakawa Memorial Research Foundation, Tokyo 107, Japan

Recently, putative viral agents responsible for human non-A to E hepatitis have been independently reported by two groups of investigators and designated GB virus C (GBV-C) and hepatitis G virus (HGV), respectively. The entire nucleotide sequences were determined for two viral genomes isolated from Japanese blood donors with GBV-C RNA. One of them (GT230) had a total genomic length of 9390 nucleotides (nt) with 5' and 3' untranslated regions of 551 and 313 nt, while the other (GT110) had genomic lengths of 9395, 281 and 315 nt, respectively. They both had a single long open reading frame, encoding 2842 amino acids (aa) in GT230 and 2933 aa in GT110. Surprisingly, they both lacked a clearly identifiable core gene, and possessed the E1/E2 gene with only four potential N-linked glycosylation sites. Pairwise comparison and phylogenetic analysis of the entire sequence indicated that the prototype GBV-C and two HGV isolates reported, as well as GT230 and GT110, are the same virus possibly of different genotypes. The five GBV-C/HGV isolates were variable up to 13.8% in the genomic nucleotide sequence, and contained deletions and insertions within the 5'–terminal 518–593 nt, which resulted in four different sizes of predicted polyproteins encoded by genomes of individual isolates. By contrast, the 3' untranslated region was well conserved. The high degree of sequence conservation within this region would favour it as a target for sensitive detection of GBV-C/HGV RNA.

Introduction

The discovery of hepatitis C virus (HCV) by Choo et al. (1989) has enabled the detection of HCV infection by serological assays and molecular biological techniques. It became clear that HCV is responsible for the majority of blood-borne non-A, non-B hepatitis cases worldwide. It is becoming increasingly evident, however, that there is still a residual risk of post-transfusion hepatitis in the recipients of blood units screened for antibody to HCV and hepatitis B surface antigen (Alter & Bradley, 1995). Likewise, not all patients with chronic non-A, non-B hepatitis are infected with HCV, and most patients with fulminant hepatitis of unknown aetiology do not have markers of infection with HCV or hepatitis A, B, D and E viruses (Alter & Bradley, 1995). Hence, there are agent(s) responsible for acute as well as chronic hepatitis categorized as non-A to E.

Recently, viruses possibly responsible for non-A to E hepatitis have been reported by two groups of investigators. Simons et al. (1995) recovered GB virus C (GBV-C) from patients with acute or chronic hepatitis or healthy individuals living in the United States, Canada and Africa, and Leary et al. (1996) determined the full-length sequence of a prototype GBV-C isolate from West Africa. Linnen et al. (1996) isolated...
one sample of hepatitis G virus (HGV) from an American patient with chronic hepatitis C and determined its entire nucleotide sequence (HGV-PNF2161), and another from an individual with a history of intermittently elevated liver enzymes (HGV-JC). Both GBV-C and HGV have a genomic organization resembling that of members of the \textit{Flaviviridae} family. Although they are similar in sequence to HCV, they are too divergent to be classified as genotypes of HCV (Leary \textit{et al.}, 1996; Schmidt \textit{et al.}, 1996). It is yet to be seen if GBV-C/HGV is hepatotrophic or can induce liver injury, or how it would account for non-A to E hepatitis in humans. The data obtained up to the present indicate that the virus can persistently infect patients on maintenance haemodialysis for 16 years or longer (Masuko \textit{et al.}, 1996), and is transmitted by transfusion and intravenous drugs (Aikawa \textit{et al.}, 1996; Linnen \textit{et al.}, 1996; Masuko \textit{et al.}, 1996; Schmidt \textit{et al.}, 1996). Remarkably, the infection with GBV-C/HGV is prevalent in carriers of HCV (Aikawa \textit{et al.}, 1996; Egawa \textit{et al.}, 1996; Linnen \textit{et al.}, 1996; Masuko \textit{et al.}, 1996), indicating a common route of transmission or a possible synergism of the two viruses. Although persistent infection with GBV-C does not seem to be associated with significant hepatic injuries (Masuko \textit{et al.}, 1996), GBV-C RNA has been reported in some patients with fulminant non-A to E hepatitis (Yoshida \textit{et al.}, 1995).

Further virological characterization is necessary to define GBV-C/HGV and their genetic variation, which would be prerequisite to the development of methods for their sensitive and specific detection in patients with non-A to E hepatitis and among the general population. Two viral isolates, designated GT110 and GT230, were propagated from Japanese blood donors with high titres of GBV-C RNA, and their sequences were determined in full. Sequence comparison among them and the three reported isolates of GBV-C/HGV revealed and characterized conserved as well as divergent regions. This information will help in designing primers to be used in RT–PCR for the detection of viral RNA.

### Methods

- **Preparation of nucleic acids from human plasma samples and synthesis of cDNA.** Plasma samples with high levels of alanine aminotransferase (ALT) of blood donors at regional Japanese Red Cross Blood Centers were screened for GBV-C RNA by RT–PCR with nested primers deduced from the non-structural (NS) 3 or helicase region (Yoshida \textit{et al.}, 1995). Two plasma samples with high titres were identified which were without markers of infection with HCV, hepatitis B virus or human immunodeficiency virus type 1. They differed substantially in the sequence of 5′-terminal nucleotides (see Results).

  Nucleic acids were extracted from 10 ml of each plasma by the method described previously (Okamoto \textit{et al.}, 1990), dissolved in RNAse-free distilled water and denatured at 70 °C for 1–3 min. They were used as templates for cDNA synthesis with antisense primers (20- to 25-mer) deduced from the helicase region of GBV-C (Simons \textit{et al.}, 1995), as well as from their own nucleotide sequences of the other genomic regions determined in this study, and a cloned reverse transcriptase of Moloney murine leukaemia virus (SuperScript II, GIBCO-BRL) following the method reported previously (Okamoto \textit{et al.}, 1994).

- **Amplification of cDNA by PCR.** cDNAs were denatured by heating at 94 °C for 15 min, and amplified by PCR with TaKaRa Ex Taq polymerase (TaKaRa Biochemicals) and TaqStart Antibody (CLONTECH Laboratories) for 35 cycles [94 °C, 30 s; 55 °C, 30 s; 72 °C, 45–150 s (8 min in the last cycle)]. When a second PCR was required, it was performed with 1–10 % of the products of the first PCR for 25–30 cycles with the same conditions as for the first PCR. The strategy for amplification of subgenomic regions for the two isolates is shown in Fig. 1; the isolates were recovered from two blood donors and named GT230 and GT110.

  For GT230, the nucleotide sequence was initially determined on region 11 amplified with helicase primers by PCR. Primers specific to GT230 were designed on the sequence of region 11 (Fig. 1a), and used to obtain sequences, indicated by closed bars with arrow heads, in both directions by rapid amplification of cDNA ends (RACE) using a Marathon cDNA amplification kit (CLONTECH Laboratories). By using antisense primers for an extension upstream and sense primers for that downstream, sequences were determined toward the 5′ and 3′ ends. Subgenomic regions represented by open bars were amplified by the conventional PCR with primers derived from GT230, and their sequences determined. For GT110, sequences represented by regions b to e (Fig. 1b) were amplified by the conventional PCR with primers deduced from GT230 or GT110 which became available as the sequencing proceeded.

  cDNAs covering the extreme 5′ end sequences, represented by region 1 and region a in Fig. 1, were tailed with dATP or dTTP homopolymer by a terminal deoxynucleotidyl transferase (Boehringer Mannheim), and amplified by single-sided PCR with 43-mer oligonucleotides containing (T)\textsubscript{17} or (A)\textsubscript{17}, following the method described previously (Okamoto \textit{et al.}, 1992). For the GT230 isolate, 27 clones were obtained by poly(T) tailing and 34 by poly(A) tailing. Of clones obtained by poly(T) tailing, 22 (81%) had the longest extension toward the 5′ end; they were presumed to represent the 5′ terminus of the GT230 isolate and their 5′ end was numbered nt 1. For the GT110 isolate, 22 clones were obtained from poly(T) tailing and 33 by poly(A) tailing: 18 (82%) of the clones with poly(T) tailing had the same length and were considered to represent the 5′ end.

  The amplification of the extreme 3′ end sequences, corresponding to region 22 and region p in Fig. 1, was attempted by single-sided PCR (Okamoto \textit{et al.}, 1992). However, no sequences were amplifiable by PCR with sense primers specific for GT230 or GT110 and the 20-mer primer (# 166), representing a part of the 43-mer oligonucleotide (Okamoto \textit{et al.}, 1992), on cDNA that had been reverse-transcribed from extracted RNAs with 43-mer oligonucleotide with either (T)\textsubscript{17} or (A)\textsubscript{17}. Therefore, the 3′ ends of extracted RNAs were tailed with poly(A) by poly(A) polymerase (TaKaRa Biochemicals), reverse-transcribed with a 43-mer oligonucleotide containing (T)\textsubscript{17}, and then amplified by the single-sided PCR. Using this procedure, 20 clones were obtained for GT230 and 17 for GT110.

- **Cloning and sequencing cDNAs.** PCR products were separated by gel electrophoresis and cloned into M13 phage vectors by the method described previously (Okamoto \textit{et al.}, 1992). Nucleotide sequences were determined for each cDNA clone on both plus and minus strands by ALF AutoRead DNA sequencing kits (Pharmacia) and Thermo Sequenase fluorescent-labelul primer cycle sequencing kit with 7-deaza-dGTP (Amersham). Three sequences were determined for each subgenomic area (Fig. 1), and the consensus sequence was adopted. A complete identity was obtained for consensus sequences overlapping neighbouring PCR amplicons spanning 26–370 nt (primer sequences excluded).
Genotypes of GB virus C/hepatitis G virus

**Fig. 1.** Genomic organization and strategy for sequencing GT230 and GT110 isolates. Shaded bars represent cDNAs for the extreme 5' and 3' ends of GBV-C genomes amplified by the single-sided PCR method. Sequences of 61 clones were determined for region 1, 20 clones for region 2, 55 clones for region 3, and 17 clones for region 4. Filled bars with arrowheads represent cDNAs obtained by the rapid amplification of cDNA ends method. The sequence of one clone each was determined for filled bars. Open bars represent cDNAs amplified by the conventional PCR with primers specific for GBV-C. For the GT230 isolate, sequences of up to three clones corresponding to open bars were determined, so as to make a total of three together with clones corresponding to closed bars: one clone for regions 5 and 10, two clones for regions 3, 8, 12, 14–16, and 21, and three clones for regions 11. For the GT110 isolate, three clones each for regions from a to p were obtained independently and sequenced. (a) Each region in GT230 isolate spanned: region 1, nt 1–150; 2, nt 8–539; 3, nt 32–539; 4, nt 218–2269; 5, nt 338–2233; 6, nt 549–2269; 7, nt 2177–4391; 8, nt 2184–3394; 9, nt 2924–4391; 10, nt 3185–4405; 11, nt 4278–4435; 12, nt 4408–7694; 13, nt 5109–5331; 14, nt 5770–6932; 15, nt 6738–7409; 16, nt 7386–8769; 17, nt 7386–8769; 18, nt 7386–8769; 19, nt 8701–9010; 20, nt 8701–9062; and 21, nt 8851–9390. (b) In GT110, regions spanned: a, nt 1–151; b, nt 33–542; c, nt 339–2236; d, nt 2177–2373; e, nt 2241–3377; f, nt 3188–4408; g, nt 4282–5180; h, nt 5112–5331; i, nt 5229–6367; j, nt 5958–6781; k, nt 6693–7429; l, nt 7382–8575; m, nt 8477–8761; n, nt 8711–9300; o, nt 9085–9355; and p, nt 9281–9395.

**Computer analysis.** This was performed using ODEN program version 1.1.1 (National Institute of Genetics, Mishima, Japan) and GENETYX version 8.0 (Software Development, Tokyo, Japan). Phylogenetic trees for the entire genomes of GBV-C/HGV were constructed by the unweighted pair-group method with arithmetic mean (Nei, 1987) and the neighbour-joining method (Saitou & Nei, 1987).

**Results**

**Sequences of GBV-C in Japanese blood donors**

Using RT–PCR with nested primers from the helicase region of GBV-C (Yoshida et al., 1995), a plasma sample (GT230) from a 46-year-old male blood donor with a high ALT level (528 IU/L) was identified, which contained a high titre of GBV-C RNA. The nucleotide sequence of the helicase region was determined for the GT230 isolate, and used as primers to determine its sequence toward both 5' and 3' ends (vide infra). Utilizing a sequence of putative 5' untranslated region (UTR) as primers, the 5' UTR sequences were amplified and determined in an additional 24 plasma samples from blood donors containing GBV-C RNA with normal or elevated ALT levels. A sequence of 316 bp spanning nt 64–379 (primer sequence excluded) of 23 samples showed a similarity > 95% to that of GT230 isolate, while one isolate (GT110), from a 33-year-old male donor with an elevated ALT level (108 IU/L), was similar to it only in 89%. Hence, the entire nucleotide sequence was determined for GT230 and GT110 isolates, which differed from each other in the sequence of 5' UTR.

**Full-length nucleotide sequences and deduced amino acid sequences of GT230 and GT110**

All the 5' end sequences of 22 and 18 clones from GT230 and GT110 isolates, respectively, obtained by amplification of poly(T)-tailed cDNAs were identical and started with 5'-TGACGTGGGG-. Likewise, the extreme 3' end sequences, which were apparently without any poly(A) or poly(U) stretches, were identical for all 20 clones from GT230 as well as all 17 clones from GT110 and terminated with -GGGTTCTACT-3'. Thus, these two isolates were considered to be sequenced up to the 5' and 3' termini.

GT230 isolate spanned 9390 nt with 5' UTR and 3' UTR of 551 and 313 nt, respectively. It has a single long open
12 Cs in 12. The 10 Cs, found in 57% of GT230 clones, and composed of 9 Cs in 2 clones, 10 Cs in 11, 11 Cs in 25 and GT230; there were 9 Cs in 2 clones, 10 Cs in 35, 11 Cs in 13 length of C-stretch was variable among 61 clones from
the 11 Cs stretching over nt 19–29 was absent in GT230. The
insertion of nucleotides in the 5 (9390 nt) and GT110 (9395 nt) was due to a deletion and all three clones of GT110.

Capsid protein (see Discussion). All three clones of GT230 had 105 aa in GT110 were not acceptable as representing the
defined in GT230 or GT110; it spanned only 14 amino acids
in HCV and the other viruses (Grakoui
been reported by Leary
(414 aa) and NS5b (563 aa). These subgenomic regions have
2933 aa. They both had the same length E1 (190 aa), E2
(387 aa), NS2 (281 aa), NS3 (677 aa), NS4 (316 aa), NS5a
(414 aa) and NS5b (563 aa). These subgenomic regions have
been reported by Leary et al (1996) in GBV-C and by
Muerhoff et al. (1995) in animal GB viruses, and are recognized
in HCV and the other viruses (Grakoui et al., 1993, von Heijne,
1986).

Like GBV-C and HGV, the putative C gene was not clearly
defined in GT230 or GT110; it spanned only 14 amino acids
(aa) in GT230, contrasting with 105 aa in GT110. Even the 105 aa in GT110 were not acceptable as representing the
capsid protein (see Discussion). All three clones of GT230 had the same starting site, which was different from that shared by
all three clones of GT110.

A difference of 5 nt in the genomic length between GT230
(9390 nt) and GT110 (9395 nt) was due to a deletion and insertion of nucleotides in the 5’- and 3’-terminal sequences.
Presuming a T at the extreme 5’ end of GT110 as nt 1, one of the 11 Cs stretching over nt 19–29 was absent in GT230. The
length of C-stretch was variable among 61 clones from GT230; there were 9 Cs in 2 clones, 10 Cs in 35, 11 Cs in 13 and 12 Cs in 8. Likewise, of 55 clones from GT110, it was composed of 9 Cs in 2 clones, 10 Cs in 11, 11 Cs in 25 and 12 Cs in 12. The 10 Cs, found in 57% of GT230 clones, and
11 Cs in 45% of GT110 clones created the difference of 1 nt within the C-stretch. GT230 was devoid of four other nucleotides at positions 415, 511, 9255 and 9261 in the sequence of GT110; they were unanimously observed in all three clones covering these respective positions in GT110.

GT230 and GT110 had a sequence similarity in the entire
genome of 87.7%, and the 2828 aa overlapping their ORFs were similar in 97.7% (Table 1). Therefore, they were both
considered to be the same virus. They varied in 10.5% within
591 nt including the putative 5’ UTR and an unidentifiable C
gene region of ambiguous 5’ end was excluded.

Comparison of the two Japanese isolates with two
American isolates and one African isolate within the
entire genome and subgenomic regions

Table 1 shows the results of two-by-two comparison, over
the entire genome and subgenomic regions, among the five
GBV-C/HGV isolates for which a full- or nearly full-length sequence has been determined. They included GT230 and GT110 described herein, and in addition, GBV-C (Leary et al., 1996). HGV-PNF2161 and HGV-JC (Linnen et al., 1996) with reported genomic lengths of 9125, 9392 and 9103 nt,

### Table 1. Genomic organization of the five GBV-C/HGV isolates and sequence similarity within the entire or subgenomic regions

<table>
<thead>
<tr>
<th>Region in comparison</th>
<th>Length nt (aa)</th>
<th>Percentage sequence similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GT230 vs GT110 nt (aa)</td>
</tr>
<tr>
<td>Total</td>
<td>9066–9392 (2828)</td>
<td>87.7 (97.7)</td>
</tr>
<tr>
<td>5’ UTR/C</td>
<td>518–593</td>
<td>89.5</td>
</tr>
<tr>
<td>E1</td>
<td>570 (190)</td>
<td>85.6 (97.9)</td>
</tr>
<tr>
<td>E2</td>
<td>1161 (387)</td>
<td>85.8 (95.1)</td>
</tr>
<tr>
<td>NS2</td>
<td>843 (281)</td>
<td>85.2 (97.5)</td>
</tr>
<tr>
<td>NS3</td>
<td>2031 (677)</td>
<td>86.9 (99.1)</td>
</tr>
<tr>
<td>NS4</td>
<td>948 (316)</td>
<td>86.0 (96.5)</td>
</tr>
<tr>
<td>NS5a</td>
<td>1242 (414)</td>
<td>88.3 (97.1)</td>
</tr>
<tr>
<td>NS5b</td>
<td>1689 (563)</td>
<td>90.3 (98.8)</td>
</tr>
<tr>
<td>3’ UTR§</td>
<td>313–315</td>
<td>97.1</td>
</tr>
</tbody>
</table>

* Comparison was made between any two from GT110, HGV-JC and HGV-PNF2161, making three combinations.
† Comparison was made between any two from GT110, HGV-JC and HGV-PNF2161, as well as between GT230 and GBV-C, totalling seven combinations.
§ The core gene region of ambiguous 5’ end was excluded.

Comparison of the five GBV-C/HGV isolates with a full- or nearly full-length sequence determined were compared disregarding insertion or deletion of nucleotides.
Genotypes of GB virus C/hepatitis G virus

Fig. 2. The 5'‑terminal sequences of the five GBV-C/HGV isolates. Sequences are shown for (1) GT110, (2) HGV-JC, (3) HGV-PNF2161, (4) GBV-C and (5) GT230 isolates. Four ATG codons are overlined with the nucleotide position marked. The first initiation codon in the ORF of each isolate is boxed, and deletions of nucleotides are indicated by slashes. Dashes indicate nucleotides identical to the top sequence (GT110).

respectively. Overall, the five isolates were 86.2–91.7% similar in nucleotide sequence and 96.1–98.8% in amino acid sequence within their overlapping regions, thereby indicating that they are all the same virus. The putative 5' UTR/C was longest in GT230 and GT110, and therefore, nt 1 was tentatively assigned to their 5' end (Fig. 2). They did not have a common starting site for the ORF, which was distributed in four different positions, although the putative starting position of the E1 region was the same for all five isolates. Hence, the 5' end of a single long ORF in the GBV-C/HGV genome is not to be determined at present; the 5' UTR and the capsid gene are not definable either.

By remarkable contrast, the sequence of the 3' UTR was very well preserved (Fig. 3). A 3'-terminal sequence up to the putative 3' end was determined only for three isolates, including GT230, GT110 and the previously reported HGV-PNF2161.

Fig. 3. The 3'-terminal sequences of three GBV-C/HGV isolates. Comparison were made between (1) GT110, (2) HGV-PNF2161 and (3) GT230 isolates for which sequences of 313–315 bp upstream of the putative 3' end are determined. In‑phase termination codons are overlined, and nt 9388 for HGV-PNF2161 of G or C is indicated by the ambiguity code (S). Dashes indicate nucleotides identical to the top sequence (GT110).
PNF2161 (Linnen et al., 1996). They showed a high similarity of 97.1–98.4%. Moreover, the sequence of 3′-terminal 140 nt was capable of making a characteristic secondary structure with stems and loops for both GT230 and GT110 (Fig. 4). For coding regions also, NS5b in the C-terminal had high similarity both in nucleotide and amino acid sequence (Table 1). However, the highest similarity in amino acid sequence was displayed by NS3 (98.1–99.7%) and NS5b regions (97.3–99.5%). In the NS3 region, a motif for trypsin-like serine proteinase (Bazan & Fletterick, 1989) borne by HGV (97%) was displayed by NS3 (98%–99%) represented by G1 for GBV-C; G2 for GT110, HGV-JC and HGV-PNF2161; and G3 for GT230.

Three genotypes of GBV-C/HGV were substantiated by two-by-two comparison within the entire genomic sequences and subgenomic regions (Table 1). Sequence similarities were higher for isolates of the same genotype than for those of distinct genotypes within the entire genome and subgenomic regions in comparison. The similarity in GBV-C/HGV isolates of the same genotype was highest within the 5′ UTR/C and 3′ UTR region.

**Phylogenetic analysis of the five GBV-C/HGV isolates**

A phylogenetic tree was constructed by the unweighted pair-group method with arithmetic mean (Nei, 1987) based on the entire nucleotide sequences of the five GBV-C/HGV isolates (Fig. 6). They were classified into at least three groups with evolutionary distance > 0.13. GT110 was closely related to both the HGV isolates. Although GT230 was closer to GBV-C, they were separated by an evolutionary distance of 0.134. A similar classification was obtained by the neighbour-joining method (Saitou & Nei, 1987). These three groups represent different genotypes of GBV-C/HGV, and they were tentatively designated G1 for GBV-C; G2 for GT110, HGV-JC and HGV-PNF2161; and G3 for GT230.

**Discussion**

The entire genomic sequence has been determined for two GBV-C/HGV isolates from Japanese blood donors with high serum ALT levels, and designated GT110 and GT230, respectively. They show a similarity in the entire nucleotide sequence of 87.2–91.7%, as well as 94.5–98.8% in deduced amino acid sequence, to the two HGV isolates (HGV-PNF2161 and HGV-JC) reported by Linnen et al. (1996), and those of 86.8–87.6% and 97.2–97.8%, respectively, to a GBV-C isolate reported by Leary et al. (1996). Therefore, the two Japanese isolates are considered to be the same virus as GBV-C and HGV. By phylogenetic analysis of the entire genomic sequence, GT110 was closely related to both HGV isolates, and probably of the same genotype. However, GT230 was different from either GBV-C or HGV, separated by a genetic distance > 0.13 (Fig. 6). Hence, there would be at least three genotypes of GBV-C/HGV. The genotype represented by GT230 is common in Japan since it accounts for 24 (96%) of 25 GBV-C/HGV isolates from Japanese blood donors as judged by the comparison of a 5′ UTR sequence of 316 bp (nt 64–379). The

**N-terminal amino acid sequence of GBV-C/HGV**

Fig. 5 compares N-terminal amino acid sequences of the five isolates. It starts at four different positions corresponding to Met₁, Met₁₋₂, Met₃₋₄ and Met₄₋₅ in the longest coding region represented by GT110. These variable starting positions made it difficult to define the capsid protein of GBV-C/HGV.

There were only four potential N-linked glycosylation sites in putative envelope proteins, one in the E1 region and three in the E2 region. They were possessed in common by the five GBV-C/HGV isolates. There were 11 cysteine residues in the E1 region and 18 in the E2 region. These 29 cysteine residues were shared by all five isolates, except C₃₈₈ that was replaced by R₃₈₈ in the HGV-JC isolate.

**Table 1. Comparison of nucleotide and amino acid sequence similarities**

<table>
<thead>
<tr>
<th>Genotype</th>
<th>NS3</th>
<th>NS5b</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBV-C</td>
<td>98.1%</td>
<td>98.1%</td>
</tr>
<tr>
<td>GT110</td>
<td>98%</td>
<td>98.1%</td>
</tr>
<tr>
<td>HGV-JC</td>
<td>97.3%</td>
<td>98%</td>
</tr>
<tr>
<td>HGV-PNF2161</td>
<td>97.3%</td>
<td>98%</td>
</tr>
</tbody>
</table>

*Fig. 4. Proposed model for the secondary structure constructed by the 3′-terminal 140 nt of the 3′-untranslated region of GBV-C/HGV. The same structure is constructed by GT110 and GT230 isolates. The minimal free energy of this model was calculated to be -81.20 Kcal/mol.*
Genotypes of GB virus C/hepatitis G virus

Fig. 5. N-terminal amino acid sequences of the viral protein encoded by the five GBV-C/HGV isolates. N-terminal amino acid sequences are shown for the predicted polyproteins encoded by the continuous long ORFs of (1) GT110, (2) HGV-JC, (3) HGV-PNF2161, (4) GBV-C and (5) GT230 isolates. Estimated starting positions of E1 and E2 regions are indicated by vertical bars with horizontal arrows. Conserved motifs of potential N-linked glycosylation sites (Asn-X-Thr/Ser) are boxed. Dashes indicate amino acids identical to the top sequence (GT110). The ORF starts variably at Met1, Met24, Met61 or Met92 in the five isolates.

Fig. 6. Proposed phylogenetic tree for the entire genomic sequences of the five GBV-C/HGV isolates. Comparison was made within nt 282–9080 (8799 bp), for which sequences of all isolates were determined and a phylogenetic tree was constructed by the unweighted pair-group method with arithmetic mean (Nei, 1987). Other two genotypes represented by GBV-C and both HGV isolates are likely to prevail in Africa and the United States, respectively, from where these isolates were reported, while the one represented by HGV occurs as a minor population (4%) in Japan.

These three genotypes of GBV-C/HGV were designated provisionally as G1 represented by GBV-C; G2 by GT110, HGV-JC and HGV-PNF2161; and G3 by GT230 (Fig. 6). The G in G1–G3 stands for either G, common in GBV-C and HGV, or genotype. The use of G would be appropriate, since GBV-C/HGV is the newest of hepatitis viruses labelled alphabetically for hepatitis A, B, C, D and E viruses, followed by a less common hepatitis F virus (Deka et al., 1994). G1–G3 are numbered in the order of description, and are defined by the full-genomic sequences. For these reasons and practical purposes, they would be used in the interim, adjusted as more sequence data accumulates on GBV-C/HGV and finalized by the nomenclature committee of Flaviviridae.

The divergence of > 12% in the entire genomic sequence as a criterion for genotypic differences is smaller than that for HCV at > 20% (Tokita et al., 1996), and compares with that for both poliovirus and hepatitis A virus at > 15% (Rico-Hesse et al., 1987; Robertson et al., 1992) and human immunodeficiency virus type 1 at > 14% (Lowagie et al., 1987; Robertson et al., 1992). Recently, Kao et al. (1996) proposed distinct genetic groups of GBV-C by the comparison of a helicase region of 117 bp. Muerhoff et al. (1996) found evidence for different genotypes of GBV-C based on the sequence heterogeneity within a 5' terminal sequence of 600 bp, and classified them into three major groups or types separated by an evolutionary distance of approximately 0.1, which further break down into subtypes (referred to as genotypes) separated by a distance of 0.034–0.056. In contrast to the report by Kao et al. (1996), genotypes defined by the 5' terminal sequence were not strongly supported by phylogenetic analysis of a helicase sequence of 135 bp (Muerhoff et al., 1996). It is to be hoped that genotypes of GBV-C/HGV would be defined by divergence in the full-length genomic sequence to avoid such a chaos as we have for HCV genotypes at present (Miyakawa et al. 1995).

The availability of full-length nucleotide sequences for the five GBV-C/HGV isolates (GT230, GT110, GBV-C, HGV-JC and HGV-PNF2161) provided an opportunity to further characterize the GBV-C/HGV genome. It has a genomic length of approximately 9.4 kb. Like HGV-PNF2161 (Linnen et al., 1996), neither GT230 nor GT110 had a poly(A) tail or a poly(U) stretch at the 3' end. Like members of the Flaviviridae
family, the genome of GBV-C/HGV had a single long ORF flanked by a 5' UTR and a 3' UTR. The length of the ORF from the GBV-C/HGV genome was variable, however, starting at four different methionine codons at the 1st, 24th, 61st and 92nd position in the GT110 isolate that possessed the longest ORF (Fig. 2). Even GBV-C/HGV isolates of the same genotype (G2) had different lengths of ORF; they were distinct from those of the isolates of other genotypes (G1 and G3). Differences within isolates of the same genotype would be attributable to point mutations, deletions and insertions of nucleotides in the 5' UTR, which would be characteristic of each GBV-C/HGV isolate, and make it difficult to clearly identify the core gene. The lack of a discernible core gene has been reported for GBV-C/HGV (Muerhoff et al., 1996), as well as its animal counterpart designated GBV-A (Muerhoff et al., 1995).

Table 2 compares nucleotide substitutions of the 1st, 2nd and 3rd letter in codons. Mutations were by far the commonest in the 3rd letter throughout the coding regions from E1 to NS5b (accounting for 72–86%), but they were equally frequent over the three letters in the region between the first methionine codon in GT110 (Fig. 2) and the start of E1 region spanning 315 nt (nt 282–596). Such an impartial distribution of point mutations in the three letters is observed also in an extreme 315 nt (nt 282–596). Such an impartial distribution of point codon in GT110 (Fig. 2) and the start of E1 region spanning over the three letters in the region between the first methionine NS5b (accounting for 72–86%), but they were equally frequent in the 3rd letter throughout the coding regions from E1 to and 3rd letter in codons. Mutations were by far the commonest animal counterpart designated GBV-A (Muerhoff et al., 1995).

Comparison was made only for GT110, GT230 and HGV-PNF2161 isolates for which 3' UTR sequences of 313–315 nt were available.

### Table 2. Substitution of nucleotides within various genomic regions among the five GBV-C/HGV isolates

Deletions and insertions were counted as substituted nucleotides.

<table>
<thead>
<tr>
<th>Regions</th>
<th>Nucleotide position</th>
<th>Length (nt)</th>
<th>Mutated nucleotides</th>
<th>Codon positions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>First</td>
</tr>
<tr>
<td>5' UTR/C*</td>
<td>76–281</td>
<td>206</td>
<td>39 (19%)</td>
<td>16 (41%)</td>
</tr>
<tr>
<td></td>
<td>282–596</td>
<td>315</td>
<td>63 (20%)</td>
<td>19 (30%)</td>
</tr>
<tr>
<td>E1</td>
<td>597–1166</td>
<td>570</td>
<td>130 (23%)</td>
<td>25 (19%)</td>
</tr>
<tr>
<td>E2</td>
<td>1167–2327</td>
<td>1161</td>
<td>285 (25%)</td>
<td>55 (19%)</td>
</tr>
<tr>
<td>NS2</td>
<td>2328–3170</td>
<td>843</td>
<td>204 (24%)</td>
<td>42 (21%)</td>
</tr>
<tr>
<td>NS3</td>
<td>3171–5201</td>
<td>2031</td>
<td>462 (23%)</td>
<td>56 (12%)</td>
</tr>
<tr>
<td>NS4</td>
<td>5202–6149</td>
<td>948</td>
<td>224 (24%)</td>
<td>35 (15%)</td>
</tr>
<tr>
<td>NS5a</td>
<td>6150–7391</td>
<td>1242</td>
<td>247 (20%)</td>
<td>38 (15%)</td>
</tr>
<tr>
<td>NS5b</td>
<td>7392–9080</td>
<td>1689</td>
<td>315 (19%)</td>
<td>38 (12%)</td>
</tr>
<tr>
<td>3' UTR†</td>
<td>9081–9395</td>
<td>315</td>
<td>9 (3%)</td>
<td>4 (44%)</td>
</tr>
</tbody>
</table>

* Comparison was made for a region starting from nt 76 for which the sequence was determined for all the five isolates. It was divided into a region from nt 76–281 that initiated the ORF of the GT110 isolate and thereafter up to nt 596.
† Comparison was made only for GT110, GT230 and HGV-PNF2161 isolates for which 3' UTR sequences of 313–315 nt were available.

It is to be hoped that primers deduced from such highly conserved sequences, identified by the comparison of five full-genomic sequences of GBV-C/HGV, will increase the sensitivity in detecting viral RNA in clinical studies and epidemiological surveys. Also, conserved amino acid sequences, highlighted in the NS3 and NS5b regions (Table 1), may help in developing serological assays for the infection with GBV-C/HGV.
We thank Japanese Red Cross Blood Centers for plasma samples containing GB virus C/hepatitis G virus. This work was supported in part by grants from the Ministry of Health and Welfare and the Ministry of Education, Science and Culture of Japan.

References


Received 28 June 1996; Accepted 17 December 1996

Genotypes of GB virus C/hepatitis G virus