Journal of General Virology (1997), **78**, 2771–2777. Printed in Great Britain

Sequence variation and phylogenetic analysis of envelope glycoprotein of hepatitis G virus

M. Y. Lim, K. Fry, A. Yun, S. Chong, J. Linnen,† K. Fung and J. P. Kim

Genelabs Technologies, Inc., 505 Penobscot Drive, Redwood City, California 94063, USA

A transfusion-transmissible agent provisionally designated hepatitis G virus (HGV) was recently identified. In this study, we examined the variability of the HGV genome by analysing sequences in the putative envelope region from 72 isolates obtained from diverse geographical sources. The 1561 nucleotide sequence of the E1/E2/NS2a region of HGV was determined from 12 isolates, and compared with three published sequences. The most variability was observed in 400 nucleotides at the N terminus of E2. We next analysed this 400 nucleotide envelope variable region (EV) from an additional 60 HGV isolates. This sequence varied considerably among the 75 isolates, with overall identity ranging from 79·3% to 99·5% at the nucleotide level, and from 83·5% to 100% at the amino acid level. However, hypervariable regions were not identified. Phylogenetic analyses indicated that the 75 HGV isolates belong to a single genotype. A single-tier distribution of evolutionary distances was observed among the 15 E1/E2/NS2a sequences and the 75 EV sequences. In contrast, 11 isolates of HCV were analysed and showed a three-tiered distribution, representing genotypes, subtypes, and isolates. The 75 isolates of HGV fell into four clusters on the phylogenetic tree. Tight geographical clustering was observed among the HGV isolates from Japan and Korea.

Introduction

Recently a novel virus, provisionally designated hepatitis G virus (HGV), was identified from the plasma of a patient (PNF2161, GenBank accession number U44402) with chronic post-transfusion hepatitis C (Linnen *et al.*, 1996). A second HGV isolate, R10291 (GenBank accession number U45966), was isolated from an individual with a history of elevated liver enzyme but no hepatitis A–E infection (Linnen *et al.*, 1996). HGV is transfusion transmissible, prevalent among blood donors and causes persistent infection. However, its aetiological role in liver disease remains to be defined.

Author for correspondence: Moon Young Lim.
Fax +1 650 368 0709. e-mail moonl@genelabs.com

†**Present address:** Johnson & Johnson Clinical Diagnostics Rochester, New York, USA.

The GenBank accession numbers of the sequences reported in this paper are: U87653–U87664 for E2-1, E2-11, E2-10, E2-12, E2-2, E2-3, E2-4, E2-5, E2-6, E2-7, E2- and E2-9, respectively, and U87665–U87724 for A11, A12, A3, A6, A7, B1, F9, G1, G12, G3, G6, D9, E1, E11, E3, E4, E5, E6, E8, F11, F12, F4, F6, E10, A9, F1, F10, F5, F8, G4, G5, G7, G9, H1, H2, A10, G8, B6, C11, D1, A8, B5, F2, C12, C5, D11, G10, B10, B11, B2, B4, B8, B9, C3, C6, C9, D12, D5, D6 and D8, respectively.

HGV has a positive-sense RNA genome of approximately 9·4 kb. Phylogenetically, HGV is related to other *Flaviviridae* such as hepatitis C virus (HCV), GB virus A (GBV-A; Simons *et al.*, 1995), GB virus B (GBV-B; Simons *et al.*, 1995) and GB virus C (GBV-C; Simons *et al.*, 1996; Leary *et al.*, 1996). Amino acid sequence comparison shows that HGV (PNF2161) is distantly related to HCV with 25% global identity, but that it is more closely related to GBV-A and GBV-B with 43·8% and 28·4% identity, respectively. HGV isolates PNF2161, R10291 and GBV-C (GenBank accession number U36380) have 96–97·5% identity at the amino acid level indicating that they belong to the same species.

Like other *Flaviviridae*, the prototypic HGV isolate PNF2161 contains a continuous open reading frame (ORF) which encodes a 3000 amino acid polyprotein. The genomic structures of HGV and HCV are similar in that the 5′ untranslated region (UTR) is followed by the structural and nonstructural genes which are then followed by the 3′ UTR. The core region of HGV (GBV-C isolate) has not been identified (Simons *et al.*, 1996). Analogous to HCV, the presence of amino acid sequence motifs imply functions for the nonstructural region of the HGV genome. These include a zinc

0001-4879 © 1997 SGM

protease in NS2b (Reed *et al*., 1995), a chymotrypsin-like serine protease and a helicase in NS3 and an RNA-dependent RNA polymerase in NS5b. Among the *Flaviviridae*, the zinc protease motif is found only in HCV, HGV, GBV-A and GBV-B (Linnen *et al*., 1996).

HCV isolates display substantial sequence variation throughout the viral genome (Okamoto *et al*., 1991) with the envelope glycoprotein coding region being the most variable (Hijikata *et al*., 1991; Weiner *et al*., 1991), and the 5′ UTR being the most conserved (Okamoto *et al*., 1991; Han *et al*., 1991; Bukh *et al*., 1992). Based on this variation, most HCV isolates can be classified into six major genotypes and a number of subtypes (Simmonds *et al*., 1993*a*). Sequence variations of viral proteins can alter their antigenic properties. For example type-specific epitopes in the NS4 protein of HCV display different antigenicities (Simmonds *et al*., 1993*b*; Stuyver *et al*., 1993), whereas changes in the sequence of lentivirus envelope proteins alter the susceptibility of the virus to neutralizing antibodies (Burns & Desrosiers, 1994). The sequence variation of the virus may thus have critical implications in designing an antibody assay or vaccine.

Analogous to HCV, the 5′ UTR sequence of HGV was found to be highly conserved among 34 isolates collected from different regions of the world (Linnen *et al*., 1997). To further analyse HGV sequence variation among isolates, a 1561 bp sequence from the E1/E2/NS2a region was compared among 15 isolates. A 400 bp sequence from the N-terminal region of E2 was found to be particularly variable. Here, we report the sequence comparison between the most variable region of E2 (EV) and the phylogenetic relationships among 60 additional worldwide HGV isolates.

Methods

Samples. Plasma or serum samples were obtained from 72 HGV-infected blood donors, patients with chronic hepatitis, haemophiliacs, intravenous drug users or transfusion patients from different geographical locations. Specimens D1 and D5–D6 were from Australia; E2-5, E2-6, E2-11 and E2-12 from Egypt; A6–A10, B1–B2, B4–B6, B8–B11, C3, C5–C6, C9, C11–C12, D8 and E2-10 from England; A11–A12, E2-1, E2-2, G12 and H1 from Greece; E2-7 and E2-8 from Jamaica; D9, D11–D12, E1, E3–E6, E8, E10–E11 and F1–F2 from Japan; G9–G10 from Korea; A3 from Peru; E2-3, E2-4, E2-9, F4–F6, F8–F12, G1, G3–G8 and H2 from USA. The sequences of three HGV isolates, namely, HGV PNF2161 (USA), HGV R10291 (USA) and GBV-C (West Africa) were used as references. HGV-infected specimens were identified by RT–PCR using the NS5a primer combination of 77F and 211R followed by solution hybridization or Southern blot analysis using probe 152F (Linnen *et al*., 1996).

RNA extraction and RT–PCR amplification. Viral RNA was extracted from 50 μl of HGV-positive serum, and reverse transcribed using random hexamers as described previously (Linnen *et al*., 1996). The E1–NS2a region was amplified by semi-nested PCR using the primers E2F1 (5′ GAGGACATHGGSTTCTGCCTIGA 3′, position +663) and E2R2 (5′ GTGTCCACCTCGAATGTRRCATCGAAG 3′, position −2575), followed by inner primers E2F2 (5′ GTGGGAAAGTGAG-TTTTGGAGATGGACTG 3′, position +959) and E2R2. Degenerate primers E2F1 and E2F2 were designed from the relatively well conserved regions of the HGV sequence based on the HGV 5′ end sequence information obtained from a large number of HGV isolates (Linnen *et al*., 1997). This amplification product contained a portion of E1 and complete E2 and NS2a sequences. PCR amplification of the amino terminus of the E2 region (EV) was performed using primers EVF (5′ GGGATG-TCGCAAGGCGCYCC 3′, position +1149) and EVR (5′ TCCCKCACA CAGGTGCCGCA 3′, position −1660). Nucleotide positions are numbered as for PNF2161 (Linnen *et al*., 1996). Mixed base code H was used for the mixture of A, T and C; R for A and G; Y for C and T; K for G and T.

Sequencing. The 1617 bp RT–PCR products from the E1–NS2a region were subcloned into the pCRII vector (Invitrogen) and sequenced. At least three clones for each isolate were used to generate a consensus sequence. Double-stranded DNA sequencing was performed by the dideoxy chain termination method and automated analysis (Applied Biosystems 373A DNA sequencer). The 512 bp PCR products of the amino-terminal end of the E2 region were purified through MicroSpin S-300 HR columns (Pharmacia) and subjected to direct sequencing. These methods allowed comparison of the 1561 bp fragment of E1/E2/NS2a DNA, homologous to positions 988–2548, and the 400 bp EV fragment DNA, homologous to positions 1209–1608 (position numbered as in the PNF2161; Linnen *et al*., 1996).

Sequence comparison and phylogenetic analysis. Nucleotide sequences and their translated amino acid sequences were aligned using the CLUSTAL V program (Higgins *et al*., 1992). Nucleotide sequences were compared using the DNAPARSIMONY program and amino acid sequences were compared using PROTEINPARSIMONY program in the PHYLIP package version 3.5 (Felsenstein, 1993). Sets of pairwise distances were generated using the program CLUSTAL V. These distances were further interpreted by the programs CLUSTAL V (Neighbor-Joining setting) and DRAWTREE in the PHYLIP package to construct the phylogenetic trees.

Results and Discussion

Sequence variability in the E1/E2/NS2a regions of HGV and HCV

RNA was extracted from 12 HGV-infected serum or plasma samples (E2-1–E2-12). Viral RNA was amplified using semi-nested RT–PCR and sequenced. The 1561 nucleotides (E1/E2/NS2a) were analysed. These 12 nucleotide and their deduced amino acid sequences were aligned together with the three published HGV sequences (PNF2161, R10291 and GBV-C) (data not shown). Nucleotide changes were observed throughout the sequence. The overall sequence identity among the 15 isolates ranged from 84·1% to 92·6%. The identity at the amino acid level was higher, ranging from 91·9% to 98·8%. Both constant and variable regions were found within the 519 amino acid E1/E2/NS2a sequences (amino acid position 188–706 in PNF2161). The first 70 amino acids at the carboxyl end of the putative E1 glycoprotein were relatively well conserved. The next 140 amino acids at the amino terminus of the putative E2 glycoprotein (EV region) displayed more variability. The 160 amino acids, situated in the middle of the E2 protein, showed a higher degree of conservation than those

at the amino terminal end. The 60 amino acids at the carboxyl end of the E2 protein sequence were identical among the 15 isolates. The NS2a protein sequences were also highly conserved.

Pairwise comparisons of the 15 HGV E1/E2/NS2a amino acid sequences showed a relatively narrow range of evolutionary distances (data not shown), ranging from 0·0116 to 0·0809. The evolutionary distances were subjected to the neighbour-joining algorithm (Saitou & Nei, 1987) to determine the phylogenetic relationship between the 15 isolates. The phylogenetic tree exhibited four clusters of more closely related isolates (data not shown) but could not be used alone to classify the 15 isolates into genotypes or subtypes.

In HCV, there is a high degree of sequence variability between strains. This variation is dependent on the regions being compared. For example, the 5′ UTR region is most conserved (Han *et al*, 1991; Okamoto *et al*., 1991), whereas the envelope glycoprotein region is highly variable (Weiner *et al*., 1991; Hijikata *et al*., 1991). We used the amino acid sequence of the HCV E1/E2/NS2a region (amino acids 169–795 of HCV-1), which corresponds to the 519 amino acid E1/E2/NS2a region of HGV for the analysis. Eleven complete HCV E1/E2/NS2a sequences were selected from the GenBank database. Their accession numbers are X76918, M67463, L02836, D17763, D10749, D49374, D28917, D10988, D11168, M62321 and D00944. These represent genotypes 1, 2 and 3 and their subtypes. Complete E1, E2 and NS2a region of genotypes 4, 5 and 6 were not available in the GenBank database at the time of analysis. Pairwise comparison of the E1/E2/NS2a amino acid sequences between the 11 variants indicated identities ranging from 62·1% to 94·7% (data not shown). This variation was much broader than the 91·9% to 98·8% variation observed in HGV. This indicated that the HGV E1/E2/NS2a sequences were more conserved than the corresponding HCV sequences. A phylogenetic tree, constructed for sequences from the 11 HCV isolates, clearly showed three genotypes with a number of subtypes and isolates (data not shown). Pairwise analysis of evolutionary distances among the 11 HCV E1/E2/NS2a sequences (Fig. 1 *c*) showed a wide evolutionary distance between them (0·05–0·37). Furthermore, three levels of non-overlapping distributions were observed representing types (0·29–0·37), subtypes (0·19–0·22) and isolates (0·05–0·10). Here, the E1/E2/NS2a region of HCV was successfully used to classify the HCV isolates. This was in agreement with the use of other regions, e.g. NS5a and core, by Simmonds *et al*. (1993).

Because the HCV E1/E2/NS2a analysis produced the same type and subtype classification as previously shown, we performed a comparable analysis with HGV E1/E2/NS2a sequences, and looked for tiers of HGV classification. Distribution of pairwise evolutionary distances between the 15 HGV E1/E2/NS2a amino acid sequences was studied. A total of 105 comparisons revealed a single level of narrow evolutionary distances ranging from 0·01 to 0·08 (Fig. 1 *a*), indicating that all 15 isolates belong to the same genotype. This differs from the variability observed within the comparable region of HCV. However, we cannot conclusively classify HGV based on this analysis alone, as the number of sequences examined was limited. More distant variants of

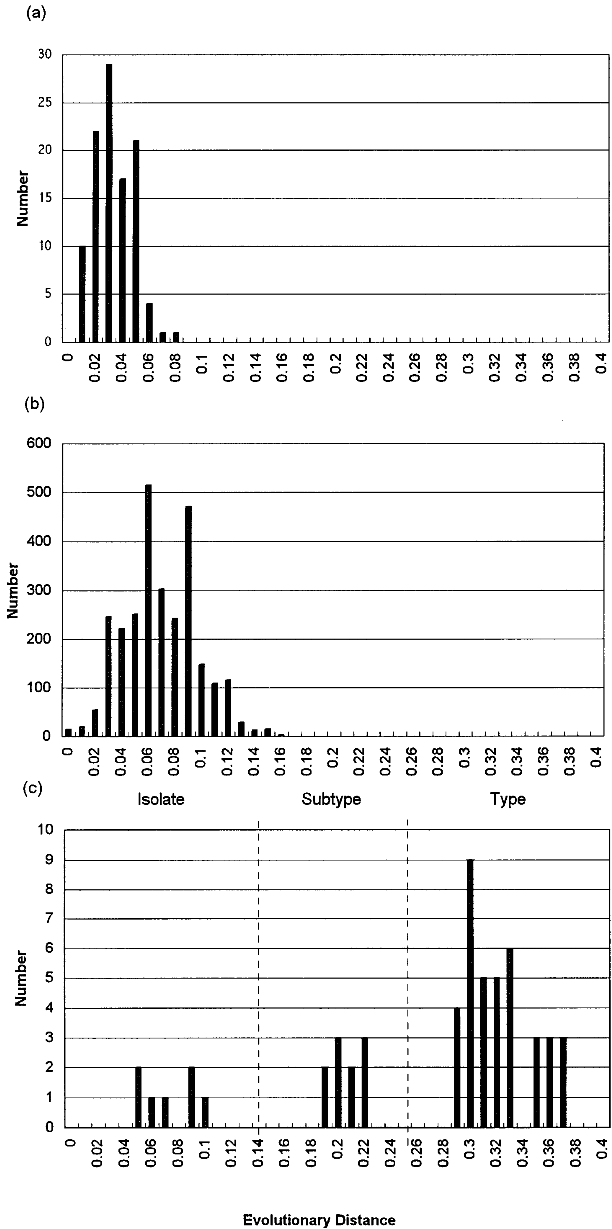

Fig. 1. Distribution of pairwise evolutionary distances of HGV and HCV. (*a*) Pairwise comparison of evolutionary distances among 15 HGV isolates using the 519 amino acid E1/E2/NS2a region (amino acid position 188–706 on PNF2161) of HGV (105 comparisons). (*b*) Pairwise comparison of evolutionary distances among 75 HGV isolates using the 133 amino acid EV region (amino acid position 261–393 on PNF2161) of HGV (2775 comparisons). (*c*) Pairwise comparison of evolutionary distances among 11 HCV isolates using the 617 amino acid E1/E2/NS2a region (amino acid position 169–795 on HCV-1) of HCV (55 comparisons). The number of calculated evolutionary distance measurements (in increments of 0·01) from 0 to 0·4 is recorded on the *y*-axis.

```
         261                                                                                            360    361                                   393
A10      TWQSCSCRAN GSRIPTGEKV WDRGNVTLLC DCPNGPWVWL PAFCQAIGWG DPITHWSHGQ NQWPLSCPQY VYGSASVTCV WGSASWFATT GGRDSKIDVW   SLVPVGSASC TIAALGSSDR DTVVELSEWG VPC
A11      ..KA...... .......... .E........ .......... .......... ....Y..... .......... ....V..... ...V....S. ..........   .......... .......... .......... ...
A12      ..K....... ...V...Q.. .E........ .......... .......... .......... .R........ ....V..... ...V....S. .....R....   ....T..... .......... .......... ...
A3       ..K.....S. ...V....R. .......... .......... ..V....... .......... .......... ...AV..... ...V....S. ..........   .......... ......T... .......... I..
A6       .......... ...V...... .......... .......... ..L...V... ....Y..... .R........ ....V..... ...V....S. D....R....   .......... .......... .......K.. ...
A9       .......... .......... .......... .......... ......V... .......... .......... ....V..... ........S. ..........   .......... .......... .......... ...
PNF2161  ...T...... ...FS..... ........Q. .......... .......... ....Y..... .......... .....T.... ........S. S.........   ........T. .......... ...PG..... I..
B1       .....T.K.. .......... .......... .........M .......... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
F8       .......... ...V...... .......... .......... .......... .......... .........F ....V..... ...V....S. ..........   .......... .......... .......... ...
F9       ...P...... .......... .......... .......... ..L...V... ......G... .R........ ....V..... ........S. .A........   ........T. .......... .......... I..
G1       .......... ...V...... .......... .......... .......... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
G12      ...A...... .......... .E........ .......... .......... ....Y..... .......... ....V..... ...V....S. ..........   .......... .......... .......... ...
G3       ........S. ...LI..... .......... Y......... .......... .......... .......... ....V..... ........S. ..........   .......... .......... .......... ...
G4       ........S. ....I..... .......... .......... ......V... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
G5       .......... .......... .......... .......... ......V... .......... .R........ ....V..... ........S. ..........   .......... .....R.... .......... ...
G6       .......... .......... .......... .......... ......V... E......... .......... ....V..... ...V....S. ..........   .......... .......... .......... ...
D9       .......... ...V....R. .......... .........V ..L....... .......... .........F ...AV..... ...V....S. ..........   .......... .......... .......... I..
E1       ........S. ...V....R. ......S... .......... ..I....... .......... ....M..... ...AV..... ...V....S. .A........   N......... .......... ...A...... T..
E10      .......... ...K....W. .......... .......... ....R..... .......... .......... .H.AV..... ...V....S. ..........   .......... .......... .......... I..
E3       .......... ...V....R. .......... .......... ..V....... .......... .......... ...AV..... ...V....S. ..........   ....I..... .......... ...AA..... I..
E4       .......... ...V...HR. .......... .......... ..V....... N...Y..... .......... ...AV..... ...V....S. R.........   .......... .......... .......... I..
E5       .......... ...V....R. .......... .......... .......... .......... .........F ...AV..... ...V....S. R.........   .......... .......... .......... I..
E6       .......... ...V....R. .......... .......... ..V....... ......G... .......... ...AV..... ...V....S. ..........   ....I..... ......T... .......... I..
E2-4     ........S. ........R. .......... .......... ..V....... .......... .......... ...AV..... ...V....S. ..........   .......... ......T... .......... I..
E2-11    .......... .......... ...R...... .......... .......... .......... .......... ....V..... ...V....S. ..........   ....T..... .......... .......... ...
E2-12    .......... ...V...... .......... .......... .......... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... I..
E8       .......... ........R. .H........ .......... ..L.K..... ....Y..... .......... ...AV..... ...V....S. ....A.....   .......... ........N. ...P...... I..
F1       .......K.. ...V....W. .......... .......... ..V....... .......... .........F ...AV..... ...V....S. ..........   .......... .......... ...P...... I..
F10      .......... ..Q....... .......... .......... ..L.R.V... .......... .R.......F ....V..... ...V....S. ..H.......   ....T..... .......... .......... ...
F11      .......... .......... ...R...... ...I...... .......... E......... .......... ....V..... ........S. ..........   ........T. .......... .......... ...
F4       .......K.. .......... .......... .......... ......V... .......... .R........ ....V..... .......... R....R....   ........T. .......... .......... E..
F5       .......... .......... .......... .......... .......... .......... .......... ....V..... ...V....S. .......E..   .......... ..T..S.... .V........ ...
F6       ..K.....S. ...Y...... .......... .......... ......V... ......G... .......... ....V..... ...V....S. ..........   ....T..... .......... .......... ...
E2-10    .......... .......... .......... .......... .......... .......... .......... ....V..... ...V....S. .....R....   G......... .......... .......... ...
A8       ...T...... ........R. .......... .......I.. ..L...V... .......... .R........ ....V..... ........S. ..........   ........T. .......... .......... I..
R10291   .......... ...YT..... .......... .......... .......... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
GBV-C    ........S. ...V....R. .E........ .........V ..L....... .......... .........F ...AV..... ...V....S. ......V...   .......... .......... .......... I..
E2-1     .......... ...TT..... .......... .......... ..L...V... .......... .......... ....V..... ...V....S. .....R....   .......... .......... .......... ...
E2-2     .......K.. ...V...... .......... .......... ..L.R.V... .......... .R.......F ....V..... ...V....S. ..H.......   N...T..... .......... .......... ...
E2-7     ...T...K.. ...T...... .......... .......... ......V... .......... .......... ...PV..... ........S. ..........   ........T. ......T... .......G.. I..
E2-5     .......... .......... .......... .......... .......... .......... .......... ....V..... ...V....S. ..........   .......... ......T... .......... ...
E2-6     .......... ...V...... .......... .......... ......V... .......... .......... ....V..... ...V....S. ..........   .......... .......... ...A...... I..
E2-8     ........S. ...LI..... .......... .......... .......... .......... .......... ....V..... ........S. R.........   .......... .......... .......... ...
E2-3     ...T...... .......... .......... .......... .......... .......... .......... ....V..... ........S. ..........   ........T. .......... .......... I..
E2-9     ...A...... ...V...... .......... .......... .......... .......... .R........ ....V..... ........S. S...A.....   .......... .......... .......... ...
G7       .......... .......... .......... .......... ......V... .......... .......... ....V..... ...V....S. ..........   .......... .......... .......... ...
G8       ........S. ...V....R. .H........ .......... .......... .......... .......... ...AV..... ...V....S. ..........   .......... .......... ...P...... ...
G9       .......K.. ...V....R. .......... .......F.. ..V....... .......... .......... ...AV..... ...V....S. ..........   .......... .......... .V........ I..
H1       ........S. .......... .......... .......... .......... .......... ....M..... ....V..... ...V....S. R.........   ....T..... .......PGP ......F... ...
H2       ........S. ...V...... .E....S... .........Q .......... .......... .........F ...AV..... ...V....S. ..........   .......... .......P.. .V.D.R.... ...
E11      .......K.. ........R. .......... .......... ..V..S.... .......... .......... ...AV....E ...V....S. R.........   .......... ......I... .......... I..
F12      .....A.... ...F...... .......... .......... .......... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
B5       .......... .......... .......... .......... ......V... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
F2       ........S. ...L....R. .......... .......... ..V....... .......... .......... ...AV..... ...V....S. ..........   .......... .......... ...R...... I..
C12      .......... ...V...... .......... .......... .......... .......... .......... .......... ........S. ..........   ........T. .......... .......... ...
C5       ...T...K.. ...V...... .......... .......... ..L....... ....Y.G... .R........ ....V..... ........S. HS........   ........T. .......... .V........ ...
D11      ........S. ...L....R. .......... .......... ..V....... N......... .......... ...AV..... ...V....S. R.........   .......... D......... ...R....V. I..
G10      ...A...... ...V...... .......... .......... .......... .......... .......... ...AV..... ...V....S. ..........   ....I..... .......... .......... ...
A7       ........S. ..NV....R. .N....S... .......... ..V....... ....Y.G... .......... ...AV..... ...V....S. .R..RR....   .......... .......... .......... I..
B10      ........S. ....T..... .......... .......... .......... .......... .......... ....V..... ...V....S. ..........   .......... .......... .......... ...
B11      ........S. ....I..... .......... .......... ......V... .......... .......... ....V..... ........S. ..........   .......... .......... .......... I..
B2       .....T.K.. .......... .......... .......... .......... ......G... .R........ .......... ........S. ..........   ........T. .......... .......... I..
B4       ........S. ...V...... .E........ .......... .......... ......G... .......... ....V..... ...V....S. ..........   .......... .......... ...P...... ...
B6       .......... ...V...... .......... .......... ..L...V... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .V........ ...
B8       ........S. ...LI..... .......... .......... .......... .......... .R........ ....V..... ...V....S. ..........   ....I..... .......... .......... ...
B9       .......... .......... .......... .......... ......V... .......... .......... ....V..... ...V....S. ..........   .......... .......... .......... ...
C11      ...T...... ...V...... .......... .......... .......... .......... .R........ ....V..... ........S. ......LE..   ........T. ..P....... .S........ I..
C3       .......... .......... .......... .......... .......... .......... .......... ....V..... ........S. .....R....   ........T. .......... .......... I..
C6       .......... .......... .......... .......... ......V... .......... .R........ ....V..... ...V....S. ..........   .......... .......... .......... ...
C9       .......... .......... .......... .......... .......... .......... .......... ....V..... ...V....S. .....R....   G......... .......... .......... ...
D1       ...T...... ........R. .......... .......I.. ......V... .......... .R........ ....V..... ........S. ..........   ........T. .......... .......... I..
D12      .......... ........R. .......... .......... ..V....... .......... .......... ...AV..... ...V....S. ..........   .......... ......T... .......... I..
D5       ........S. ...RI....L .......... .......... ..L...V... .......... ........H. ....V..... ........S. ..........   .......... .......... .P.A...... ...
D6       ...A.....D .......... .E........ .......... .......... .......... .......... ...AV..N.. ...V....S. .....R....   .......... .......... ...P...... ...
D8       ........S. ...SI..... .......... .......... ..L...V... .......... .........F .H..V..... ........S. R.........   .......... .......... .......... ...
```

Fig. 2. Comparative alignment of the deduced amino acid sequences of EV from 75 HGV isolates. Sequence numbers based on PNF2161 coordinate as in Linnen *et al.* (1996). Dots (.) represent identity with the first sequence.

HGV, which need to be included for a more conclusive classification, may have been overlooked by the initial selection of HGV specimens, which used the NS5 region primers 77F and 211R.

Sequence variability in the E2 variable (EV) region of HGV

The amino terminus of the E2 region from HGV displayed a higher degree of variability compared to the neighbouring regions and was thus selected for further sequence comparison studies. To amplify this E2 variable (EV) region, two primers (EVF and EVR) were designed from well conserved regions within the E1/E2/NS2a sequences which flanked the EV region. RNA was amplified from 60 additional HGV-positive serum or plasma samples from geographically diverse regions. EV nucleotide sequences from these 60 HGV-positive specimens and the same region from the previously analysed 15 were compared. At the nucleotide level, sequence identity ranged from 79·3% to 99·5% (data not shown). Deduced amino acid sequences derived from 75 EV nucleotide sequences (amino acid 261–393 in PNF2161) were aligned (Fig. 2). The overall amino acid sequence identity among the 75 isolates ranged from 83·5% to 100% (data not shown). Although there were some regions which displayed more variability than others, there was no apparent hypervariable region. In HCV, the hypervariable domain of approximately 28 amino acids is located in the E2 envelope glycoprotein region, which is

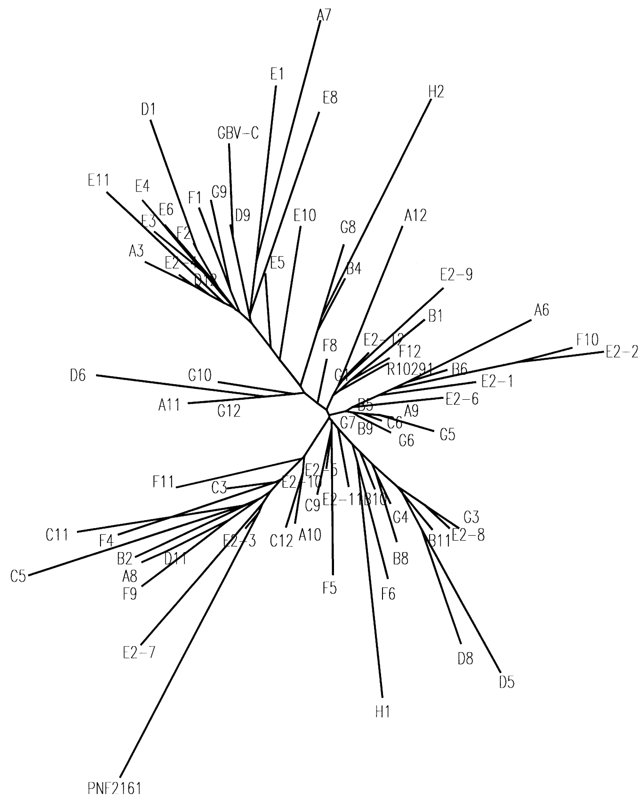

Fig. 3. Phylogenetic analysis of EV sequences from 75 isolates of HGV.

directly downstream of a putative signal peptide of E2/NS1 (Weiner *et al*., 1991). This hypervariable region, which represents 8% of the complete envelope nucleotide sequence, accounts for between 40% and 63% of the observed amino acid variation in the total putative envelope region of the HCV isolates. Lack of high sequence variability in the HGV E2 region suggests that this region is not subjected to strong host immune pressure and may be important for viability of the virus.

Pairwise comparison showed a relatively narrow range of evolutionary distances among the 75 HGV EV sequences. The evolutionary distances ranged from 0 to 0·1654. The distribution of the pairwise evolutionary distances of 75 HGV EV sequences exhibited a single peak (Fig. 1*b*). Again, this suggested that the HGV isolates were more closely related to each other than the HCV isolates. Although our analysis classified the 75 HGV isolates in the same genotype, the existence of a more distant HGV variant cannot be excluded. As discussed earlier, initial selection of the HGV specimen by primers used in this study could have limited the amplification of distant HGV isolate sequences. Despite this, several lines of evidence support our finding that HGV isolates are more closely related than the HCV isolates. Firstly, there was a close correlation in detecting HGV/GBV-C by RT–PCR, using primer sets derived from NS5a and NS3 region sequences (J. W. Mosley and others; L. J. Jeffers and others: personal communications). Also, the detection of HGV using the primer set derived from the HGV 5′ UTR region was highly concordant with the results obtained from the NS5a (J. Linnen and others, personal communication) or the NS3 primer sets (Bhardwaj *et al*., 1997). Thus, the detection of HGV by the primer sets derived from three different regions of HGV was highly concordant. Secondly, a high degree of sequence conservation was observed among a large number of isolates from around the world. Finally, the sequence of GBV-C from West Africa, obtained by an independent laboratory (Leary *et al*., 1996), showed high sequence similarity with the HGV isolates analysed in this study.

No different genotypes were deduced by analysing the constructed phylogenetic tree from the 75 HGV EV amino acid sequences. However, the isolates were closely clustered in four groups (Fig. 3). The first group contained PNF2161 (USA), the second group specimen D5 (Australia), the third group specimen E2-2 (Greece) and the fourth group specimen GBV-C (West Africa). Moreover, geographical clustering was noticeable among isolates of HGV from Japan and Korea (specimens D9, D12, E1, E3–E6, E8, E10–E11, F1–F2 and G9–G10) in a group represented by GBV-C. The significance of this geographical distribution is not clear at this point as the epidemiology of HGV spread is not understood. Sequence variation of HGV may play a role in clinical or serological differences among different isolates.

Several groups have examined the genetic variation in HGV by comparing sequences from other segments of HGV genome. In agreement with our findings from the envelope region of HGV, limited genetic variation in HGV has been reported. Studies with NS3, core and NS5a regions showed a single distribution of evolutionary distances between the isolates (Pickering *et al*., 1997; Viazov *et al*., 1997). Despite this other groups have suggested that HGV can be classified into genotypes or subtypes. Okamoto *et al*. (1997) proposed three genotypes based on the divergence of $>$ 12% in the entire genomic sequence. The three genotypes were represented by PNF2161, GBV-C and a Japanese isolate GT230. In our studies PNF2161 and GBV-C were separated into different groups. Furthermore, analysis of these isolates using the 400 bp envelope region sequence suggests that the definition of three HGV genotypes based on the entire genome might be similar to our grouping. Other groups have also suggested similar classification of HGV into two to five genotypes, subtypes or groups based on their analyses of 5′ UTR region (Muerhoff *et al*., 1996; Fukushi *et al*., 1996; Ding *et al*., 1997; Hsieh *et al*., 1997; Mukaide *et al*., 1997) or NS3 helicase region sequences (Kao *et al*.,1996). Furthermore, Mukaide *et al*. (1997) also identified ‘genotype’-specific restriction sites, *Scr*FI and *Bsm*FI, and developed restriction polymorphism analysis. It would be interesting to see whether the classification of HGV based on the restriction analysis of the 5′ UTR region coincides with grouping based on sequences from other regions. Regardless of the genomic regions used for phylogenetic analyses, it is

clear that HGV isolates are genetically more closely related to each other compared to isolates of HCV. The formal classification of the virus awaits further analysis and discussion.

We are indebted to Professors Steven Foung, Stephanos Hadziyannis, Howard Thomas, Georg Hess, Stephen Locarnini, Kendo Kiyosawa and Kwan Soo Byun, and Doctors Peter Karayiannis, Kenneth Hyams, Douglas Watts and Andrew Corwin for providing valuable serum specimens. We are grateful to Doctors Nancy Alexi and Elaine Goodman for careful reading of the manuscript.

References

Bhardwaj, B., Qian, K., Detmer, J., Mizokami, M., Kolberg, J. A., Urdea, M. S., Schlauder, G., Linnen, J. M., Kim, J. P., Davis, G. L. & Lau, J. Y. N. (1997). Detection of GB virus-C/hepatitis G virus RNA in serum by reverse transcription polymerase chain reaction. *Journal of Medical Virology* **52**, 92–96.

Bukh, J., Purcell, R. H. & Miller, R. H. (1992). Sequence analysis of the 5′ noncoding region of hepatitis C virus. *Proceedings of the National Academy of Sciences, USA* **89**, 4942–4946.

Burns, D. P. & Desrosiers, R. C. (1994). Envelope sequence variation, neutralizing antibodies, and primate lentivirus persistence. *Current Topics in Microbiology and Immunology* **188**, 185–219.

Ding, M., Yuwen, H., Mitchell, F., Biswas, R., Ndimbie, O. K. & Farshid, M. (1997). Sequence characterization of the 5′ noncoding region of GB virus C/hepatitis G virus. *Biochemical and Biophysical Research Communications* **231**, 606–609.

Felsenstein, J. (1993). PHYLIP inference package version 3.5. Department of Genetics, University of Washington, Seattle, USA.

Fukushi, S., Kurihara, C., Ishiyama, N., Okamura, H., Hoshino, F. B., Oya, A. & Katayama, K. (1996). Nucleotide sequence of the 5′ noncoding region of hepatitis G virus isolated from Japanese patients: comparison with reported isolates. *Biochemical and Biophysical Research Communications* **226**, 314–318.

Han, J. H., Shyyamala, V., Richman, K. H., Brauer, M. J., Irvine, B., Urdea, M. S., Tekamp Olson, P., Kuo, G., Choo, Q. L. & Houghton, M. (1991). Characterization of the terminal regions of hepatitis C viral RNA: identification of conserved sequences in the 5′ untranslated region and poly(A) tails at the 3′ end. *Proceedings of the National Academy of Sciences, USA* **88**, 1711–1715.

Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Computer Applications in the Biosciences* **8**, 189–191.

Hijikata, M., Kato, N., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S. & Shimotohno, K. (1991). Hypervariable regions in the putative glycoprotein of hepatitis C virus. *Biochemical and Biophysical Research Communications* **175**, 220–228.

Hsieh, S. Y., Yang, P. Y., Chen, H. C. & Liaw, Y. F. (1997). Cloning and characterization of the extreme 5′-terminal sequences of the RNA genomes of GB virus C/hepatitis G virus. *Proceedings of the National Academy of Sciences, USA* **94**, 3206–3210.

Kao, J. H., Chen, P. J., Hsiang, S. C., Chen, W. & Chen, D. S. (1996). Phylogenetic analysis of GB virus C: comparison of isolates from Africa, North America, and Taiwan. *Journal of Infectious Diseases* **174**, 410–413.

Leary, T. P., Muerhoff, A. S., Simons, J. N., Pilot-Matias, T. J., Erker, J. C., Chalmers, M. L., Schlauder, G. G., Dawson, G. J., Desai, S. M. & Mushahwar, I. (1996). Sequence and Genomic organization of GBV-C: a novel member of the *Flaviviridae* associated with human non-A–E hepatitis. *Journal of Medical Virology* **48**, 60–67.

Linnen, J., Wages, J., Jr, Zhang-Keck, Z. Y., Fry, K. E., Krawczynski, K. Z., Alter, H., Koonin, E., Gallagher, M., Alter, M., Hadziyannis, S., Karayiannis, P., Fung, K., Nakatsuji, Y., Shih, J. W., Young, L., Piatak, M., Jr, Hoover, C., Fernandez, J., Chen, S., Zou, J. C., Morris, T., Hyams, K. C., Ismay, S., Lifson, J. D., Hess, G., Foung, S. K. J., Thomas, H., Bradley, D., Margolis, H. & Kim, J. P. (1996). Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. *Science* **271**, 505–508.

Linnen, J., Fung, K., Fry, K. E., Mizokami, M., Ohba, K., Wages, J. M., Jr, Zhang-Keck, Z. Y., Song, K., & Kim, J. P. (1997). Sequence variation and phylogenetic analysis of the 5′ terminus of hepatitis G virus. *Journal of Viral Hepatitis* (in press).

Muerhoff, A. S., Simons, J. N., Leary, T. P., Erker, J. C., Chalmers, M. L., Pilot-Matias, T. J., Dawson, G. J., Desai, S. M. & Mushahwar, I. K. (1996). Sequence heterogeneity within the 5′-terminal region of the hepatitis GB virus C genome and evidence for genotypes. *Journal of Hepatology* **25**, 379–384.

Mukaide, M., Mizokami, M., Orito, E., Ohba, K., Nakano, T., Ueda, R., Hikiji, K., Iino, S., Shapiro, S., Lahat, N., Park, Y. M., Kim, B. S., Oyunsuren, T., Rezieg, M., Al-Ahdal, M. N. & Lau, J. Y. (1997). Three different GB virus C/hepatitis G virus genotypes. Phylogenetic analysis and a genotyping assay based on restriction fragment length polymorphism. *FEBS Letter* **407**, 51–58.

Okamoto, H., Okata, S., Sugiyama, Y., Kurai, K., Iizuka, H., Machida, A., Miyakawa, Y. & Mayumi, M. (1991). Nucleotide sequence of the genomic RNA of hepatitis C virus isolated from a human carrier: comparison with reported isolates for conserved and divergent regions. *Journal of General Virology* **72**, 2697–2704.

Okamoto, H., Nakao, H., Inoue, T., Fukuda, M., Kishimoto, J., Iizuka, H., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1997). The entire nucleotide sequences of two GB virus C/hepatitis G virus isolates of distinct genotypes from Japan. *Journal of General Virology* **78**, 737–745.

Pickering, J. M., Thomas, H. C. & Karayiannis, P. (1997). Predicted secondary structure of the hepatitis G virus and GB virus-A 5′ untranslated regions consistent with an internal ribosome entry site. *Journal of Viral Hepatitis* **4**, 175–184.

Reed, K. E., Grakoui, A. & Rice, C. M. (1995). Hepatitis C virus-encoded NS2-3 protease: cleavage-site mutagenesis and requirements for biomolecular cleavage. *Journal of Virology* **69**, 4127–4136.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.

Simmonds, P., Holmes, E. C., Cha, T.-A., Chan, S.-W., McOmish, F., Irvine, B., Beall, E., Yap, P. L., Kolberg, J. & Urdea, M. S. (1993*a*). Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *Journal of General Virology* **74**, 2391–2399.

Simmonds, P., Rose, K. A., Graham, S., Chan, S. W., McOmish, F., Dow, B. C., Follett, E. A. C., Yap, P. L. & Marsden, H. (1993*b*). Mapping of serotype-specific, immunodominant epitopes in the NS-4 region of hepatitis C virus (HCV)-use of type-specific peptides to serologically differentiate infections with HCV type 1, type 2, and type 3. *Journal of Clinical Microbiology* **31**, 1493–1503.

Simons, J. N., Leary, T. P., Dawson, G. J., Pilot-Matias, T. J., Muerhoff, A. S., Schlauder, G. G., Desai, S. M. & Mushahwar, I. K. (1995). Isolation of novel virus-like sequences associated with human hepatitis. *Nature Medicine* **1**, 564–569.

Simons, J. N., Desai, S. M., Schultz, D. E., Lemon, S. M. & Mushahwar, I. (1996). Translation initiation in GB virus A and C: evidence for internal ribosome entry and implications for genome organization. *Journal of Virology* **70**, 6126–6135.

Stuyver, L., Van Arnhem, W., Wyseur, A., DeLeys, R. & Maertens, G. (1993). Analysis of the putative E1 envelope and NS4a epitope regions of HCV type 3. *Biochemical and Biophysical Research Communications* **192**, 635–641.

Viazov, S., Riffelmann, M., Khoudyakov, Y., Fields, H., Varenholz, C. & Roggendorf, M. (1997). Genetic heterogeneity of hepatitis G virus isolates from different parts of the world. *Journal of General Virology* **78**, 577–581.

Weiner, A. J., Brauer, M. J., Rosenblatt, J., Richman, K. H., Tung, J., Crawford, K., Bonino, F., Saracco, G., Choo, Q. L., Houghton, M. & Han, J. H. (1991). Variable and hypervariable domains are found in the regions of HCV corresponding to the flavivirus envelope and NS1 proteins and the pestivirus envelope glycoproteins. *Virology* **180**, 842–848.

Received 15 April 1997; Accepted 17 July 1997