Phylogenetic analysis of GBV-C/hepatitis G virus

Donald B. Smith,¹ Miren Basaras,² Simon Frost,³ Dan Haydon,⁴ Narcisa Cuceanu,¹ Linda Prescott,¹ Cara Kamenka,¹ David Millband,¹ Mahomed A. Sathar⁵ and Peter Simmonds¹

¹Department of Medical Microbiology, University of Edinburgh, Summerhall, Edinburgh EH9 1QH, UK
²Department of Microbiology, School of Medicine, Universidad del Pais Vasco, 48080 Bilbao, Spain
³Centre for HIV Research, University of Edinburgh, Waddington Building, King’s Buildings, West Mains Road, Edinburgh EH9 3JN, UK
⁴Centre for Tropical Veterinary Medicine, University of Edinburgh, Easter Bush, Roslin EH25 9RG, UK
⁵Department of Medicine, University of Natal, Congella, South Africa 4013

Comparison of 33 epidemiologically distinct GBV-C/hepatitis G virus complete genome sequences suggests the existence of four major phylogenetic groupings that are equally divergent from the chimpanzee isolate GBV-C_virus and have distinct geographical distributions. These four groupings are not consistently reproduced by analysis of the virus 5′-noncoding region (5′-NCR), or of individual genes or subgenomic fragments with the exception of the E2 gene as a whole or of 200–600 nucleotide fragments from its 3′ half. This region is upstream of a proposed anti-sense reading frame and contains conserved potential RNA secondary structures that may be capable of directing the internal initiation of translation. Phylogenetic analysis of this region from certain South African isolates is consistent with previous analysis of the 5′-NCR suggesting that these belong to a fifth group. The geographical distribution of virus variants is consistent with a long evolutionary history that may parallel that of pre-historic human migrations, implying that the long-term evolution of this RNA virus is extremely slow.

Introduction

The similarity in genome organization between GB virus-C/hepatitis G virus (GBV-C/HGV) and hepatitis C virus (HCV) has led to the naïve expectation that variation of these closely related and persistent flaviviruses might also be similar. However, our limited understanding of the causal reasons for virus variability is underscored by the increasing evidence that these viruses vary in quite different ways. Although both viruses have a similar rate of nucleotide substitution during persistent infection [0.4–1.9 x 10⁻³] for HCV (Major et al., 1999; Smith et al., 1997a; Okamoto et al., 1992; Ogata et al., 1991), 0.4–2.4 x 10⁻³ for HGV (Khudyakov et al., 1997; Nakao et al., 1997)], GBV-C/HGV lacks a hypervariable region comparable to that present at the NH₂ terminus of the HCV E protein (Takahashi et al., 1997a; Nakao et al., 1997) and observed ratios of synonymous to nonsynonymous substitution are higher for GBV-C/HGV (30:1) than within HCV subtypes (9:1) although the latter are less divergent (Muerhoff et al., 1997; Smith et al., 1997b). In addition, while different genotypes of HCV differ by more than 30%, the most extreme GBV-C/HGV variants differ by only 14%. Previous studies have identified three (Suzuki et al., 1999; Okamoto et al., 1997), four (Charrel et al., 1999) or five (Takahashi et al., 1997b) phylogenetic groupings of GBV-C/HGV, although some of these groupings are weak and inconsistent between different studies. However, whereas HCV genotypes can be distinguished by phylogenetic analysis of a variety of subgenomic regions as small as 222 nt, variants of GBV-C/HGV cannot be reliably identified in this way. Systematic analysis of six complete GBV-C/HGV genome sequences revealed that congruent phylogenetic relationships were obtained for only a minority of 300, 600 and 1200 nt fragments, and that the optimal region was all or part of the 5′-noncoding region (5′-NCR) (Muerhoff et al., 1997; Smith et al., 1997b).

At present 33 epidemiologically unrelated GBV-C HGV

Author for correspondence: Donald Smith. Present address: Garden Cottage, Clerkington, Haddington, East Lothian EH41 4NJ, UK. e-mail Donald.B.Smith@gardencottage.screaming.net

The GenBank accession numbers of the sequences reported here are AF181977–AF181981.
complete virus genome sequences are available, including three of group 1 and several isolates of uncertain relationship to previously defined groupings. In addition, the recent discovery of closely related chimpanzee viruses (Adams et al., 1998) and the availability of a complete genome sequence (GBV-C\textsubscript{tr}) (Birknenmeyer et al., 1998) allows phylogenetic trees to be constructed using a more appropriate outgroup. We have therefore undertaken a re-analysis of the phylogenetic relationships of GBV-C/HGV complete genome sequences, the extent to which these can be reproduced by analysis of subgenomic regions and the implications of virus geographical variation for theories about its evolutionary history.

**Methods**

- **Nucleotide sequences.** Nucleotide sequences were obtained from GenBank and manipulated and aligned using Simmicon 2000 software (P. Simmonds, unpublished). Nucleotide positions of aligned sequences are numbered relative to the AUG codon at the beginning of the E1 gene of the prototype isolate GBV-C (U36380; Leary et al., 1996b). The 32 other epidemiologically unrelated complete genome sequences were AB003288–AB003293 (Takahashi et al., 1997b), AB013500 (Saio et al., 1999), AF104403 (Charrel et al., 1999), AB013501 (Konomi et al., 1999), AF031829 (Bukh et al., 1998), D87255 (Shao et al., 1996), D90600, D90601 (Okamoto et al., 1997), U44402, U45966 (Linnen et al., 1996), U63715 (Erker et al., 1996), AB008342 (Kaneko et al., 1998), AF006500 (L. Lu et al., unpublished), D87265 (Nakao et al., 1997), D87708–D87715 (Katayama et al., 1998), U75356 (Y. S. Zhou and H. T. Wang unpublished), U94695 (Wang et al., 1997), AB018667 and AB021287 (H. Naito & K. Abe, unpublished), together with the outgroup AF007046 (GBV-C\textsubscript{tr}; Birknenmeyer et al., 1998). The partial sequence (positions 396 to 6118) of an additional isolate from Thailand (K-10) was made available by Sirirug Songsivilai (personal communication). The boundaries between genes were as given in Charrel et al. (1999).

- **Serum samples.** Sera were collected from individuals from Papua New Guinea, Sudan and The Gambia as part of population-based serum samples. “New Guinea, Sudan and The Gambia” as part of population-based Serum samples.

- **PCR amplification of GBV-C/HGV genomes.** RNA was extracted from plasma using a proteinase K–SDS lysis buffer (Jarvis et al., 1999). The 5'-NCR was amplified for 5 outer primers 5'-TGGCAATCCGGCGCTGGTGCAGT 3' (443 to 426), and the inner primers 5'-GGCGTGGTGTGGTGATG 3' (positions 21 to 41). Amplification reactions were 30 cycles for each primer pair containing 12 cycles for 5°C for 18 s, 52 °C for 21 s and 72 °C for 90 s. Set 1 consisted of outer primers 5'-TACGATGTGACGCACTG 3' (positions 979–998), and inner primers 5'-ACCCCAAATGCTGGTGGG 3' (1375–1353) and 5'-GTGNYGACCTCTGACGTG 3' (1027–1046).

The region of NSSA where duplications are observed in some isolates (Tanaka et al., 1998) was amplified for 5'-NCR-PCR positive samples from the Democratic Republic of Congo (n = 7). The Gambia (n = 4), Sudan (n = 2) and Papua New Guinea (n = 8) using outer primers 5'-CATAATGCTGTAGTTGCTCG 3' (positions 6736–6714) and 5'-CCATCGCCCGCACTWATCTCGG 3' (positions 6409–6430), and inner primers 5'-TACRAGARCCGCCCACTTGAAGAC 3' (positions 6573–6550) and 5'-ACGNAGAGACCTGACTG 3' (positions 6433–6453). Amplifications were started at 80 °C followed by 30 cycles of 94 °C 18 s, 52 °C 21 s and 72 °C 90 s. The size of amplified DNA fragments was assessed by electrophoresis through 4% Metaphor agarose gels, with expected sizes of 140 or 180 bp for fragments with and without a duplication.

- **Nucleotide sequencing and phylogenetic analysis.** Nucleotide sequences of amplified fragments were obtained by direct sequencing of amplified genome regions using Thermosequenase (Amersham) in reactions containing 33P-labelled deoxynucleotides. Potential RNA secondary structures were investigated using RNADraw version 1.0 (Matzura, 1995).

Completed coding region sequences were analysed by three different methods. (1) Distances were estimated using an F84 model of substitution (Felsenstein, 1984), which allows for unequal transition/transversion rates and unequal base frequencies. 100 bootstrapped datasets, distance matrices, neighbour joining trees and a consensus tree were produced using SEQBOOT, DNADIST, NEIGHBOR and CONSENSE, all part of the PHYLIP package (Felsenstein, 1993). (2) Synonymous and non-synonymous distances were estimated using Method I of Nei & Gojobori (1984), which allows for unequal transition rates between purines and pyrimidines, together with rate heterogeneity modelled as a proportion of invariable sites plus eight rates taken from a discrete gamma distribution using parameters estimated from the data. 100 bootstrapped datasets were produced using SEQBOOT, and distance matrices calculated with the programs PUZZLE version 4.0.2 (Strimmer & von Haeseler, 1996) and PUZZLEBOOT version 1.01 (Holder & Roger, 1999) using parameters estimated from the dataset. Trees with the lowest least-squares deviation were produced using the program FITCH applying global search, and a consensus tree produced using CONSENSE. Phylogenetic trees of subgenomic regions were produced with MEGA (Kumar et al., 1993) using the Kimura 2-parameter model of substitution on datasets of 100 bootstrap replicates.

In order to determine which GBV-C/HGV group was closest to the outgroup GBV-C\textsubscript{tr} (AF070476), we used a reduced dataset consisting of three sequences from each group (Group 1, HGU36380, AB003291, AB013500; Group 2, AB013501, U44402, U63715; Group 3, AB003288, D90601, U94695; Group 4, AB003292, AB018667, AB021287). We estimated the likelihood of the outgroup clustering with each of the four groups under four different models of nucleotide substitution: (i) an F84
model; (ii) an F84 model allowing for rate variation modelled using a discrete gamma distribution with eight categories; (iii) as for (ii), but allowing for different transition/transversion ratios, different base frequencies, different rates and different levels of rate heterogeneity for each gene; and (iv), as for (ii) but allowing parameters to differ for each codon position. These models were fitted and the bootstrap support for each cluster compared using the fast approximate bootstrap procedure of Kishino & Hasegawa (1989) as implemented in the PAML version 2.0 package (Yang, 1999). Similar analyses of amino acid sequences were carried out using Phylip 3.57c (Kimura substitution matrix), or RELL (Jones et al. (1992) substitution matrix] assuming gamma rate heterogeneity.

**Results**

**Complete genome sequences**

Phylogenetic analysis of the coding region of 33 epidemiologically unrelated complete GBV-C/HGV genome sequences yielded evidence for four distinct phylogenetic groupings each of which was supported by high levels of bootstrap support (Fig. 1). These groupings correspond to the three groupings previously identified from the analysis of complete genome sequences (Okamoto et al., 1997) together with an additional group consisting of an unclassified Japanese sequence (Takahashi et al., 1997b) and three sequences from Thailand (K-10; Sirirrug Songtivilai, unpublished results), Vietnam (AB018667) and Myanmar (AB021287). Similar groupings and levels of bootstrap support were obtained when the analysis was confined to either synonymous or non-synonymous sites, or to amino acid substitutions, if comparisons were extended to include the entire virus genome, or if the chimpanzee isolate GBV-Ctro was used as an outgroup (data not shown).

Although subgroupings of groups 1, 2 and 3 have been distinguished by phylogenetic analysis of subgenomic regions (Muerhoff et al., 1996; Takahashi et al., 1997a; Muerhoff et al., 1997) there was only limited evidence from the analysis of complete genome sequences for subgroupings. Three group 3 isolates from China (U75356, U94695 and AF006500) grouped together with high bootstrap support, while the isolates AB013501 and U63715 grouped separately from the other

---

**Fig. 1.** Phylogenetic tree of GBV-C/HGV complete coding region sequences. Maximum likelihood distances between sequences were calculated with Phylip DNADIST (ts:tv = 2, assuming no rate variation between sites), and used to produce a neighbour joining tree with Phylip NEIGHBOR. Bootstrap values (100 replicates) were obtained with the SEQBOOT and CONSENSE options of Phylip. The tree was produced using Treeview version 1.5.
group 2 isolates (group 2a) and isolate AB003292 from Japan was relatively divergent from other group 4 isolates. However, analysis of pairwise distances between GBV-C/HGV sequences revealed a distribution with only two overlapping peaks (Fig. 2) corresponding to comparisons between sequences belonging to different groups (distances of 0.076–0.131 or within one of the four phylogenetic groups 0.131–0.171) or within one of the four phylogenetic groups (0.137–0.171) or within one of the four phylogenetic groups (0.171–0.200). The overlap between these two distributions comprised 18/528 pairwise comparisons (3.5%), more than half of which involved a divergent group 2 isolate from South America (AB013501).

A previous study suggested that certain group 1 isolates containing a duplicated region within NS5A (e.g. isolate AB003291) were ancestral to other GBV-C/HGV isolates relative to GBV-A (Tanaka et al., 1998). However, this was not confirmed when the analysis was extended to complete genome sequences (Suzuki et al., 1999), or when we investigated the relative likelihood with which the more closely related GBV-C_{tr} isolate clustered with three representatives from each group (Table 1). Under four different models of substitution, clustering with group 1 could be rejected at the 5% level. Although clustering with group 2 gave the best fit to the data for the models which included rate heterogeneity, differences between genes in transition/transversion ratios, rate heterogeneity and base frequencies were relatively small (results not shown), and the bootstrap support for this clustering was not high enough to be conclusive. The best fit to the data were obtained with the model in which parameters could vary according to position within codons, reflecting the strong bias against non-synonymous substitution. Bootstrap resampling analysis using simpler models of substitution at synonymous or non-synonymous sites supported group 4 as being ancestral, while analysis of amino acid sequences supported either groups 1 or 3 as being ancestral depending on the substitution matrix used. Hence, our analysis does not consistently place any one of the four human GBV-C/HGV groups as being ancestral.

Furthermore, the duplicated region may be derivative rather than ancestral since this requires fewer evolutionary steps (Suzuki et al., 1999). Most isolates that lack the duplication have instead an 8 nt direct repeat (ACCCCGTTC, positions 6457–6464 and 6487–6494) flanking a sequence with the potential to form a structure with an 11 nt stem and a 3 nt loop. The formation of this stem–loop during RNA synthesis could lead to slippage and mispairing between the direct repeats resulting in the duplication of the hairpin loop and the direct repeat. We did not detect the duplication in NS5A amongst other isolates from Africa (n = 13) or Papua New Guinea (n = 8) or other NS5A sequences available in GenBank (n = 94).

### Subgenomic regions

Next, we investigated the extent to which subgenomic regions of GBV-C/HGV bear the same phylogenetic relationships as do entire virus genomes. Analysis of individual virus genes failed to produce congruent phylogenetic trees with the sole exception of the E2 gene (Fig. 3). Similar analysis of subgenomic fragments of 2000, 1000 or 500 nt produced congruent trees only for fragments including the COOH-

---

**Table 1. Log likelihood and approximate bootstrap support for the clustering of the outgroup GBV-C_{tr} with groups 1–4 under different models of substitution**

<table>
<thead>
<tr>
<th>Model of substitution</th>
<th>Log likelihood for clustering (bootstrap support)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1</td>
</tr>
<tr>
<td>F84</td>
<td>–54053:387 (0-44%)</td>
</tr>
<tr>
<td>F84 with rate variation</td>
<td>–49041:827 (1-58%)</td>
</tr>
<tr>
<td>F84 with rate variation gene specific</td>
<td>–49018:110 (1-52%)</td>
</tr>
<tr>
<td>F84 with rate variation codon specific</td>
<td>–47139:032 (0-5%)</td>
</tr>
</tbody>
</table>
terminal region of the E2 gene. Analysis of the entire 3’-terminal half of the virus genome or any of its subfragments failed to produce congruent trees.

In most cases, sequences that grouped aberrantly were not consistently associated with a particular group. The only exception was the group 4 sequence AB0021287 that grouped with group 2 sequences for subgenomic fragments encompassing the NH\textsubscript{2}-terminus of the NS2 gene or the COOH-terminal half of the NS5B gene. Even if the assumption is made that this isolate is a recombinant between isolates of group 4 and 2 [the major GBV-C/HGV groups found in this geographical region (Naito et al., 1999)], the only additional subgenomic region that then produces a congruent phylogenetic tree is a 1000 nt fragment encompassing the COOH terminus of the E2 gene and most of the NS2 gene.

Since a 500 nt fragment of the E2 gene could reproduce the phylogenetic relationships of the complete genome sequences, we next analysed this region in more detail (Figs 4 and 5). Congruent phylogenetic trees were produced using a region as small as 200 nt (positions 1344–1543). The shortest region that gave a congruent tree with more than 98% bootstrap support for each group was the 600 nt region from positions 994–1594. This region also produced a congruent tree when analysed at synonymous sites (> 85% support) but not at nonsynonymous sites.

Previous studies have suggested that the phylogenetic relationships of GBV-C/HGV complete genome sequences can be reproduced by analysis of the whole or part of the 5’-NCR (Muerhoff et al., 1996; Smith et al., 1997b; Muerhoff et al., 1997). The extreme 5’-terminal sequence is available for only a single group 1 and none of the group 4 isolates, so an adequate analysis can only be carried out from positions –388 to –1. Analysis of this region and of various subfragments failed to produce congruent phylogenetic trees (Fig. 6), with
two sequences responsible for the majority of aberrant groupings AB013500, an extreme group 2 isolate from Bolivia, groups with group 3 isolates for the 5’-NCR and some subgenomic fragments in the 3’ half of the genome, although the 5’-NCR is atypical. Similarly, U75356, an unpublished group 3 isolate from China, has a 5’-NCR sequence similar to but distinct from those of group 2 isolates. This sequence also has a frameshift within NS5A and an unusual sequence at the 3’ terminus (Smith et al., 1997b). Hence, it is uncertain whether these isolates represent recombinants (An et al., 1997) or divergent variants with unique 5’-NCR sequences. If these two sequences are excluded from the dataset, a congruent tree is obtained when the region fi388 to fi1 is analysed. However, subfragments of this region, including those previously identified as reproducing the phylogenetic relationships of group 1 and 2 isolates (Smith et al., 1997b), provide < 70% bootstrap support for either groups 3 or 4.

Evidence for additional groupings from the analysis of subgenomic regions

Having identified a region within the E2 gene that consistently reproduces the phylogenetic relationships of complete genome sequences, we compared published sequences of this region for evidence of additional GBV-C/HGV groups. Of 16 additional sequences for the region 994–1594 (accession nos U87653–U87664, AF063827, AF063828, AF063830, AF017533), 14 (from USA, Jamaica, Greece, UK and Egypt) grouped with group 2, and one each with groups 1 (USA) and 3 (Hong Kong) (> 99% support for all groups; data not shown). Similarly, nine additional sequences of the region 1414–1693 from Italian patients (AF015842–AF015862) clustered with group 2, although as expected from the analysis of complete genome sequences (Fig. 4), the bootstrap support was low (44%). In contrast, E2

---

**Fig. 4.** Congruence of phylogenetic relationships produced using fragments of the E2 gene. The extent to which phylogenetic groupings observed from analysis of complete genome sequences are supported by comparison of fragments of the E2 gene is indicated by the level of shading. The number of groups supported by more than 70% of bootstrap replicates is indicated. The minimum degree of bootstrap support is given for selected examples. Also indicated are the position of an anti-genome reading frame (Kondo et al., 1998), a region where synonymous substitutions are suppressed (Simmonds & Smith, 1999) and regions where the presence of covariant nucleotide substitutions suggests the presence of RNA secondary structures (Simmonds & Smith, 1999).
sequences for the 350 nt region 1146–1495 of five GBV-C variants from South Africa with unusual 5′-NCR sequences (Sathar et al., 1999) formed an additional group (Fig. 5), while other South African isolates with 5′-NCR sequences that clustered with groups 1 (n = 3) or 2 (n = 2) clustered with groups 1 or 2 respectively. Isolates from Papua New Guinea had E2 sequences that clustered with group 2 (n = 2) or group 4 (n = 2), while an isolate from Sudan (n = 1) clustered with group 2.

Similar analysis of 5′-NCR sequences is complicated by the large number of sequences available within this region (1409 accessions) and the inconsistent phylogenetic relationships of all but the largest fragments (Fig. 6). Visual examination for motifs distinct from those typical of groups 1, 2, 3 and 4 revealed a small number of unusual sequences, but of these only the variants from South Africa (Sathar et al., 1999) described above, and isolates from Spain (Lopez-Alcorocho et al., 1999) grouped separately from the 5′-NCR sequences of
D. B. Smith and others

Fig. 6. Congruence of phylogenetic relationships produced using fragments of the 5′-NCR. The extent to which phylogenetic groupings observed from analysis of complete genome sequences are supported by comparison of fragments of the 5′-NCR is indicated by the level of shading. Groupings were considered to be supported if they were observed in > 70% of bootstrap replicates.

Discussion

Identification of phylogenetic groups

This analysis of complete genome sequences of GBV-C/HGV identifies four main phylogenetic groupings (Fig. 1). The existence of the fourth group had been suggested from the analysis of 5′-NCR sequences of Southeast Asian isolates (Naito et al., 1999). A previous phylogenetic analysis of complete genome sequences also proposed four groups in which groups 1 and 4 were combined and group 3 was split into two (Charrel et al., 1999), but the published bootstrap analysis did not support these groupings. Similarly, an analysis based upon genetic distances identified five groups with group 1 split into two (Takahashi et al., 1999), but our analysis suggests that phylogenetic groups cannot be defined using pairwise distances (Fig. 2).

An unexpected finding of this study is that the phylogenetic relationships of complete GBV-C/HGV sequences (but not subgroupings) can be reproduced by short COOH-terminal fragments of the E2 gene (Fig. 4). Previous studies of smaller numbers of sequences have shown that phylogenetic relationships are inconsistent when comparisons are made for individual genes (Takahashi et al., 1997b) or subgenomic fragments (Smith et al., 1997b). Although three (Cong et al., 1999) or four (Lim et al., 1997) phylogenetic groupings are observed when fragments larger than 1500 nucleotides including the E2 gene are compared, interpretation of these studies is complicated by idiosyncratic labelling, the absence of group 4 isolates and the failure to assess the robustness of groupings. We show here that analysis of a 200 nt fragment from the centre of the E2 gene (positions 1344–1543) provides > 75% bootstrap support for all four phylogenetic groupings, while a 600 nt region (positions 994–1594) provided > 98% support. This is in stark contrast to other coding regions: all subgenomic fragments of 2000 nt or less that did not contain this region of E2 failed to produce congruent trees with the sole exception of a 2000 nt fragment encompassing the remainder of the E2 gene, NS2 and the NH₂-terminal half of complete genome sequences for the region −388 to −1. These later isolates could represent recombinants since they have motifs typical of group 3 sequences between positions −489 and −459 but the remainder of the 5′-NCR is similar to, although distinct from, that of group 2 isolates.
NS3 (Fig. 3). An important unresolved question is the extent to
which some complete genome sequences represent recombi-
nants between different groups. Inconsistent relationships
were observed for some sequences between the 5′-NCR and
coding regions or between different coding regions. However,
the aberrant groupings were typically weak and inconsistent
suggesting that these do not represent simple recombinants.

Our analysis helps to clarify previous conflicting studies on
the extent to which subgenomic regions can be used to
identify GBV-C/HGV phylogenetic groupings. Most early
studies of virus diversity concentrated on a 118–135 nt
fragment within NS3 that was the first part of the genome to
be sequenced (Simons et al., 1995; Masuko et al., 1996;
Heringlake et al., 1996; Kao et al., 1996; Berg et al., 1996;
Schreier et al., 1996; Tsuda et al., 1996; Schmidt et al., 1996;
Muerhoff et al., 1997; Pickering et al., 1997; Ibanez et al., 1998).
However, the phylogenetic conclusions of these studies appear
to be unreliable since analysis of even the entire NS3 gene fails
to produce congruent groupings. Inconsistent phylogenetic
groupings have also been observed for the NH$_2$ terminus of E2
(Lim et al., 1997; Muerhoff et al., 1997), 354 nt of NS5A/NS5B
(Viazov et al., 1997) or 279 nt of NS5B (Muerhoff et al., 1997).
Finally, although analysis of a 2600 nt fragment containing
both NS5A and NS5B produced a congruent tree for 12
isolates from groups 1, 2 and 3 (Khudyakov et al., 1997),
comparison of this region from the 33 complete genome
sequences gave aberrant groupings of some isolates with only
marginal support for groups 1 and 4 (data not shown).

The inconsistent phylogenetic relationships of GBV-
C/HGV subgenomic coding regions contrasts with that
observed for HCV where analysis of a variety of subgenomic
regions reproduces the phylogenetic relationships of complete
genomes (Simmonds et al., 1994; Tokita et al., 1998). This
difference could arise if different regions of the GBV-C/HGV
genome were subject to different evolutionary processes in
which case combining these regions could produce a less
reliable reconstruction than the analysis of individual regions
(Bull et al., 1993). However, the validity of the analysis based
on complete genome sequences is supported by the correlation
between their phylogenetic relationships and their geographi-
cal origin. An alternative explanation is that the amino acid
sequence of the GBV-C/HGV polyprotein is well conserved
($d_{AA}$: $d_{E2}$, 0.033, divergence < 11%) relative to HCV subtypes
($d_{AA}$: $d_{E2}$, 0.094, divergence < 10%). Phylogenetic groupings of
GBV-C/HGV therefore rely more on variation at synonymous
sites, many of which are invariant or saturated, possibly
because of the presence of extensive RNA secondary structures
within the GBV-C/HGV genome (Simmonds & Smith, 1999).
Consequently, GBV-C/HGV isolates with a common evolu-
tionary origin may share only a small number of poly-
morphisms and analysis of subgenomic regions could often fail
to produce congruent phylogenetic trees.

There are several potential explanations for our observation
that analysis of the E2 gene or specific subfragments produces
phylogenetic trees congruent with those observed for com-
plete genome sequences. First, since E2 is the most variable
part of the GBV-C/HGV polyprotein (Katayama et al., 1998;
Erker et al., 1996), the phylogenetic relationships of complete
genome sequences could depend entirely on substitutions in
E2. However, identical groupings are observed if E2 is excluded
and most amino acid substitutions occur at the NH$_2$ terminus
of E2 (Lim et al., 1997; Katayama et al., 1998; Cong et al., 1999),
whereas it is only the central and 3′ regions that produce
congruent phylogenetic trees, and then only for synonymous
rather than nonsynonymous substitutions. A second potential
explanation is that this region of the genome encodes an open
reading frame on the anti-sense strand (nucleotides 870–1226,
Fig. 4), possibly encoding a nucleocapsid protein (Kondo et al.,
1998). This would constrain the accumulation of substitutions
and so retain evidence of phylogenetic relationships. However,
although synonymous substitutions are suppressed in this part
of the genome (Muerhoff et al., 1997; Simmonds & Smith,
1999), phylogenetic analysis of the anti-sense reading frame
fails to produce a congruent tree (Fig. 4). Another potential
source of constraint in this region of the genome would be if
translation of the anti-sense reading frame was dependent on
an upstream internal ribosome entry site. Indeed, substitutions
in this region of the genome are frequently covariant and
associated with potential RNA secondary structures
(Simmonds & Smith, 1999) (Fig. 7), while RNA folding
predictions for representatives of groups 1–3 identify
structures with free energies one standard deviation below
those of random sequences of the same base composition.
However, these RNA structures were not observed in
sequences from group 4 isolates, and the anti-sense reading
frame is 10 residues longer in some group 3 isolates and lacks
the initial AUG codon in GBV-C$_{trio}$.
Geographical distribution and origin of GBV-C HGV isolates

The distinct geographical distribution of GBV-C/HGV variants is consistent with their co-evolution with humans during pre-historic migrations. Group 1 and 5 isolates are African and have relatively diverse 5'NCR sequences (Muerhoff et al., 1997; Smith et al., 1997b; Naito et al., 1999; Sathar et al., 1999), while the remaining groups correspond to the three main waves of human migration from Africa to Europe (group 2), northern Asia (group 3) and southern Asia (group 4). In addition, the presence of group 3 variants amongst native populations in South America (Pujol et al., 1998; Gonzalez-Perez et al., 1997) is consistent with the first colonization of this continent from northern Asia via the Bering Strait. Although Japanese isolates have been found belonging to groups 1–4 (Takahashi et al., 1997b; Okamoto et al., 1997), most are group 3 suggesting that the presence of other groups represents recent introductions.

Finally, since a virus related to GBV-C/HGV is present in chimpanzees (Adams et al., 1998; Birkenmeyer et al., 1998), while New World monkeys harbour more distantly related but species-specific variants (Bukh & Apgar, 1997; Leary et al., 1996a) it is possible that GBV-C/HGV has been continuously present in human populations since speciation. In this case the virus appears to have evolved at an overall rate of less than 10^{-8} per site per year (Suzuki et al., 1999; Simmonds & Smith, 1999), although its rate of sequence evolution measured in longitudinal studies is similar to that of other RNA viruses. Further sequence analysis of the E2 region from GBV-C/HGV variants isolated from different geographical regions may help to clarify the origins of this unusual virus.

We are grateful to John Mokili, David Pritchard, David Arnot, Eleanor Riley and Brian Greenwood for providing serum samples, and to David Leach for advice on the mechanisms of RNA rearrangements. M.B. is supported by a grant from the Sociedad Espanola de Quimioterapia. D.B.S. was supported by a grant from the Welcome Trust. P.S. is a Darwin Trust Fellow.

References


Simmonds, P., Smith, D. B., McOmish, F., Yap, P. L., Kolberg, J.,


Received 17 August 1999; Accepted 30 November 1999