Journal of General Virology (2001), **82**, 713–722. Printed in Great Britain

Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3′ end of the hepatitis G virus/GB-virus C genome

N. M. Cuceanu,† A. Tuplin and P. Simmonds

Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh EH9 1QH, UK

Hepatitis G virus (HGV)/GB virus C (GBV-C) causes persistent, non-pathogenic infection in a large proportion of the human population. Epidemiological and genetic evidence indicates a long-term association between HGV/GBV-C and related viruses and a range of primate species, and the co-speciation of these viruses with their hosts during primate evolution. Using a combination of covariance scanning and analysis of variability at synonymous sites, we previously demonstrated that the coding regions of HGV/GBV-C may contain extensive secondary structure of undefined function (Simmonds & Smith, *Journal of Virology* 73, 5787–5794, 1999). In this study we have carried out a detailed comparison of the structure of the 3′untranslated region (3′UTR) of HGV/GBV-C with that of the upstream NS5B coding sequence. By investigation of free energies on folding, secondary structure predictive algorithms and analysis of covariance between HGV/GBV-C genotypes 1–4 and the more distantly related HGV/GBV-C chimpanzee variant, we obtained evidence for extensive RNA secondary structure formation in both regions. In particular, the NS5B region contained long stem–loop structures of up to 38 internally paired nucleotides which were evolutionarily conserved between human and chimpanzee HGV/GBV-C variants. The prediction of similar structures in the same region of hepatitis C virus may allow the functions of these structures to be determined with a more tractable experimental model.

Introduction

A new member of the flavivirus family, variously described as hepatitis G virus (HGV) or GB virus C (GBV-C) has recently been characterized (Linnen *et al.*, 1996; Leary *et al.*, 1996). HGV/GBV-C is widely distributed in human populations, with frequencies of active or past infection ranging from 5 to 15%. Infection is frequently persistent and associated with high levels of circulating viraemia, although there is currently no evidence to link HGV/GBV-C infection to any identifiable hepatic or non-hepatic disease. HGV/GBV-C shows limited genetic heterogeneity, but with marked geographical differences in distribution of the four or five currently classified genotypes. It has been suggested that the presence and subsequent diversification of HGV/GBV-C in humans as they migrated out of Africa 100000 years ago accounts for the current association of different HGV/GBV-C genotypes with particular racial groups (Gonzalez-Perez *et al.*, 1997; Tanaka *et al.*, 1998; Katayama *et al.*, 1997). Supporting this conjecture, sequences from the Far East are almost invariably genotype 3, and this genotype is otherwise only found in native inhabitants of North and South America. In contrast, Caucasian and other populations from India westwards including those of Northern Africa are infected with genotype 2. Genotype 1 is confined to sub-Saharan Africa, and shows the greatest overall sequence diversity (Smith *et al.*, 1997; Muerhoff *et al.*, 1997). As further evidence for a very long-term association of this virus with humans and other primates, viruses closely related to HGV/GBV-C have been found in a variety of Old and New World monkey species, and their phylogenetic relationships mirror those of their hosts. Variants of HGV/GBV-C more divergent than human genotypes can be detected in wild-caught chimpanzees from Central and West Africa (Birkenmeyer *et al.*, 1998; Adams *et al.*, 1998). Even more distantly related viruses, collectively described as GBV-A, have been recovered from

Author for correspondence: Peter Simmonds.
Fax +44 131 650 7965. e-mail Peter.Simmonds@ed.ac.uk

† **Current address:** Centre of Infectious Diseases, Department of Medical Microbiology, Molecular Virology Laboratory, Leiden University Medical Centre, Leiden, The Netherlands.

0001-7355

Table 1. Free energy on folding the NS5B region and 3′UTR of GBV-C/HGV: comparison with sequence order-randomized controls

		NS5B region			3′UTR		
Accession no.	Genotype	ΔG/b*	$\Delta G/b_{sc}$†	% diff.	ΔG/b*	$\Delta G/b_{sc}$†	% diff.
U36380	1	−1·44	−1·19	17·33	−1·58	−1·33	15·91
AB013500	1	−1·51	−1·21	19·87	−1·55	−1·30	16·13
AB013501	2	−1·62	−1·22	24·95	−1·56	−1·29	13·21
U63715	2	−1·54	−1·22	20·78	−1·58	−1·26	20·25
U44402	2	−1·41	−1·20	14·89	−1·50	−1·27	15·33
AB008342	3	−1·41	−1·19	15·60	−1·52	−1·29	15·13
D90601	3	−1·51	−1·20	20·44	−1·51	−1·29	12·81
D87263	3	−1·51	−1·21	19·87	−1·56	−1·33	14·74
AB021287	4	−1·43	−1·17	18·82	−1·48	−1·20	18·92
AB018667	4	−1·43	−1·14	20·28	−1·52	−1·21	20·39
Mean‡		**−1·48 ± 0·07**	**−1·20 ± 0·04**	**18·85 ± 3·37**	**−1·54 ± 0·03**	**−1·26 ± 0·07**	**17·25 ± 3·98**
AF070476	CPZ	−1·47	−1·12	23·80	−1·29	−1·10	14·73
U22303	GBV-A	−1·46	−1·14	21·92	−1·46	−1·26	13·70

* Free energy on folding indicated in kJ/mol per base.
† sc, Sequence order-randomized control sequences (50 for sequences U36380, AB013501, D90601, three for other sequences analysed; mean value shown).
‡ Mean and standard deviation of free energy on folding (kJ/mol per base) or difference from sequence order-randomized controls.

several species of New World monkeys (Bukh & Apgar, 1997; Erker *et al.*, 1998; Leary *et al.*, 1997).

The great genetic stability of HGV/GBV-C and related viruses implied by the evidence for co-evolution with primates is difficult to reconcile with their observed rapid sequence change in individuals over short observation periods (Nakao *et al.*, 1997). We have previously found evidence for constraints on sequence change at many sites in the coding sequence, even at those where substitutions would be synonymous (Simmonds & Smith, 1999). Evidence that RNA secondary structure formation through internal base-pairing limits sequence variability at these sites was provided by the finding of multiple covariant sites spatially associated with potential stem–loop structures amongst HGV/GBV-C sequences of different genotypes. Furthermore, these occurred at positions in the genome that showed reductions in synonymous variability. In that study we excluded non-random nucleotide composition and biased codon usage as compounding factors in the use of RNA folding prediction algorithms and calculation of free energies.

In the current study we have used a variety of phylogenetic and free energy-based predictive algorithms to compare the extent and conservation of RNA secondary structure formation in the 3′ untranslated region (3′UTR) with upstream coding sequences from NS5B, a region encoding the viral RNA polymerase. Our findings indicate that the part of the RNA genome containing the coding sequences is more extensively structured than the 3′UTR, and shows better conservation between variants of HGV/GBV-C infecting different primates. These findings imply an important functional role(s) for the observed secondary structure.

Methods

■ **Sequences.** Currently available complete genomic sequences of HGV/GBV-C genotypes 1–4 (GenBank/EMBL accession numbers in parentheses) which included full-length or near full-length 3′UTR sequences were the genotype 1 sequences GBV-C (U36380) and AB013500; type 2a sequences PNF2161 (U44402), T55875 (AF031827), HGV-1517 (AF31828), HGV-1539 (AF031829), AF121950, HGV-Iw (D87255) and GT110 (D90600); the type 2b sequence GBV-C(EA) (U63715); type 3 sequences GT230 (D90601), HGV-IM71 (AB008342), GSI85 (D87262), D87708–D87714 (Katayama *et al.*, 1998); and type 4 sequences AB018667 and AB021287. Sequences were numbered from the start of the coding region after alignment. The chimpanzee homologue, HGV/GBV-C_{CPZ}, and GBV-A from the New World primate *Sanguinis mystax*, bore the accession numbers AF70476 and U22303. Sequences from the NS5 region of hepatitis C virus (HCV) were obtained from the following published sequences: genotype 1a (M62321), 2a (D00944), 3a (D17763), 4a (Y11604), 5a (Y13184) and 6a (Y12083).

Viroid sequences analysed were citrus exocortis viroid (accession no. X53715), potato spindle tuber viroid (U23058), chrysanthemum chlorotic mottle viroid (AJ247123), Mexican papita viroid (L78463) and potato spindle tuber viroid (X76846). Delta virus sequences of different genotypes were obtained from the following entries: AF098261, AJ000558, M21012, D01075 and M28267. Coding sequences of serum albumin were obtained from the following mammalian species: cat (X84842), cow (Y17729), gerbil (AB006197), horse (X74045), macaque (M90463) and rat (U01222). α-globin coding sequences were obtained from the following mammalian species: baboon (X05289), orang-utan (M12158), duck (X02008), marsupial cat (M17083) and human (V00493).

■ **Additional HGV/GBV-C 3′UTR sequences.** 3′UTR sequences

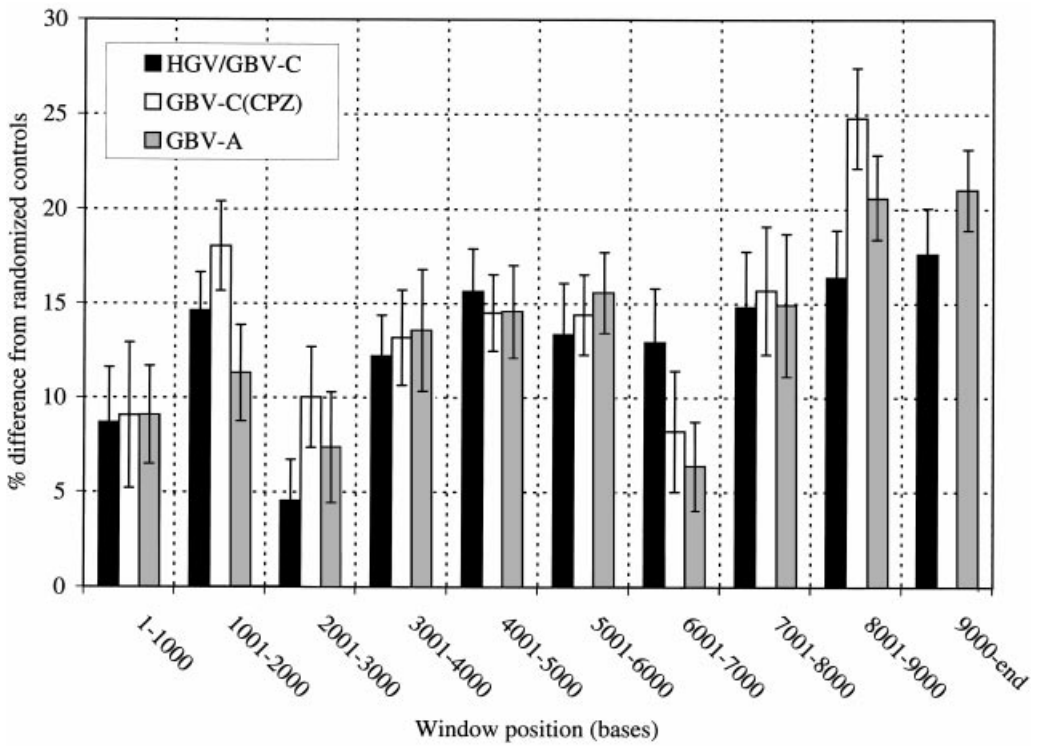

Fig. 1. Free energy on folding consecutive 1000 base fragments of HGV/GBV-C (AB013500), HGV/GBV-C_{CPZ} (AF070476) and GBV-A (U22303) expressed as percentage differences from sequence order-randomized controls. Bars show 95% confidence intervals calculated from multiple sequence randomizations. Bases numbered from start of genome.

Table 2. Free energy on folding of RNA sequences with defined secondary structure and controls

Sequence	*n**	ΔG/b†	$\Delta G/b_{sc}$‡	% diff.
HGV/GBV-C 5′UTR	4	−1·42 ± 0·05	−1·27 ± 0·05	10·27 ± 2·36
Delta virus	5	−1·61 ± 0·04	−1·29 ± 0·01	19·95 ± 1·93
Plant viroid	5	−1·45 ± 0·03	−1·17 ± 0·04	19·47 ± 2·37
α-Globin	5	−1·05 ± 0·14	−1·09 ± 0·12	−2·94 ± 3·22
Albumin	6	−0·80 ± 0·03	−0·82 ± 0·04	−2·30 ± 3·87
HCV NS5B	6	−1·33 ± 0·08	−1·02 ± 0·06	22·55 ± 6·44

* No. of sequences analysed.
† Mean and standard deviation of free energy on folding (kJ/mol per base).
‡ sc, Sequence order-randomized controls (three randomizations per sequence analysed).

were obtained from 17 samples whose genotype had been deduced from sequence comparisons of the 5′UTR and E2 regions (Smith *et al.*, 1997, 2000). RNA was extracted using proteinase K–SDS and phenol–chloroform as described previously (Jarvis *et al.*, 1994). Purified RNA was then reverse-transcribed and amplified by hemi-nested RT–PCR using primers derived from conserved regions of the HGV/GBV-C genome at the carboxyl end of the NS5B gene and the extreme 3′-end: Z3580 – outer sense (positions 8829–8848, 5′ GGTGGTNCATCAATTGGATT 3′, where N = A, C, G, T); Z3581 – inner sense (positions 8881–8900, 5′ GGTTCTTAGCCCTGCTCATC 3′); and Z3582 – outer and inner antisense (positions 9212–9231, 5′ AGTAGAACCCGGCCTTTGGG 3′). Reverse transcription was carried out at 42 °C for 30 min using avian myeloblastosis virus reverse transcriptase (Promega). The conditions for the first round of PCR were hot start at 80 °C for 2 min followed by 30 cycles of 94 °C for 18 s, 58 °C for 21 s and 72 °C for 90 s. At the end of the last cycle, samples were heated to 72 °C for 5 min to allow termination of incomplete strands. The second round of PCR was performed using 1 μl of the primary PCR product for the same number of cycles and conditions. The amplified PCR products were cloned into pGEM-T vector (Promega), and sequenced with both sense and antisense

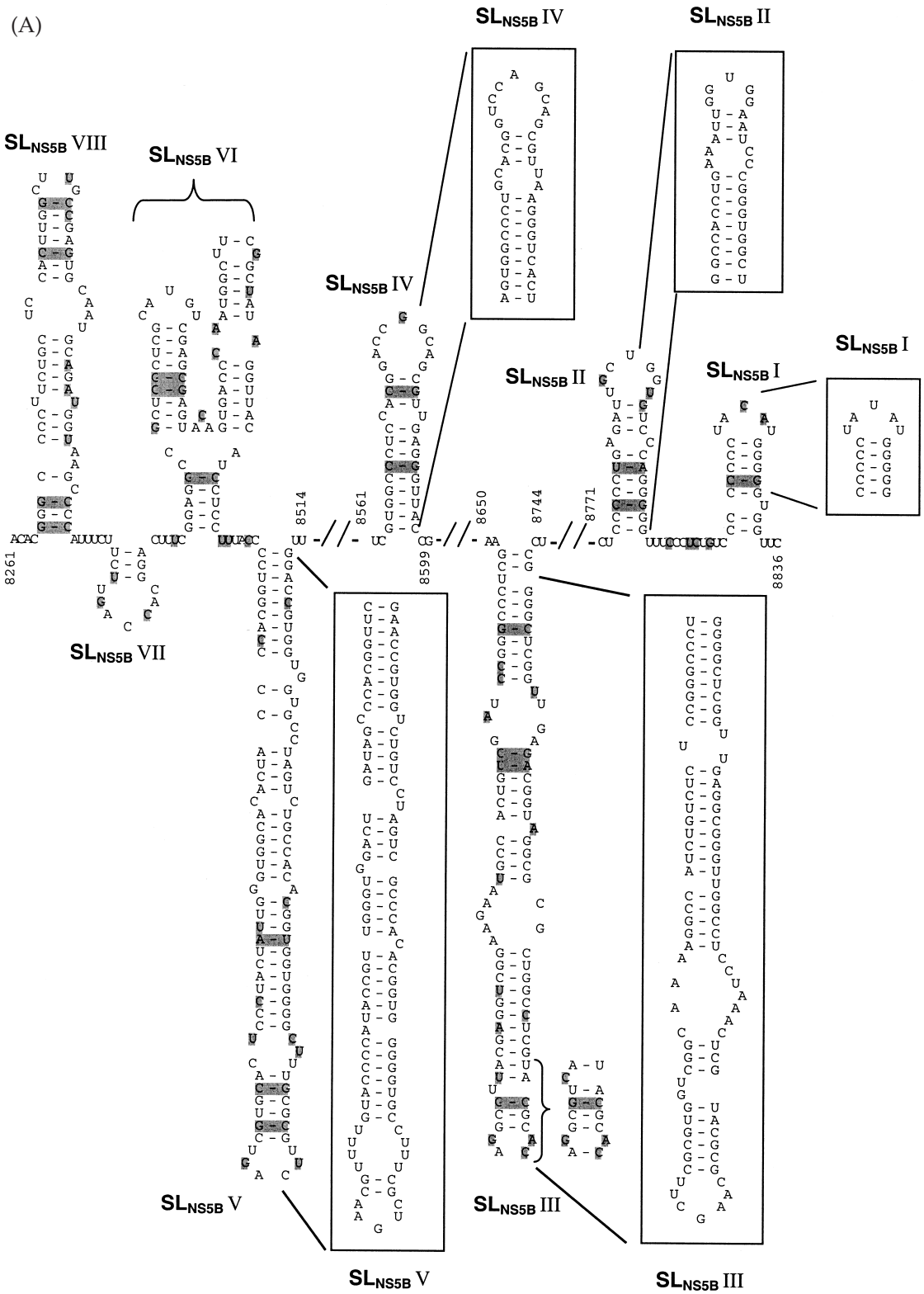

Fig. 2. For legend see facing page.

plasmid primers using T7 DNA polymerase (Sequenase, USB). The consensus sequence of one to three clones for each sample was used for phylogenetic analysis.

■ **Free energy on RNA folding.** The last 1000 bases of aligned HGV/GBV-C complete genome sequences were split into NS5B and 3′UTR regions. The NS5B sequence included the whole sequence

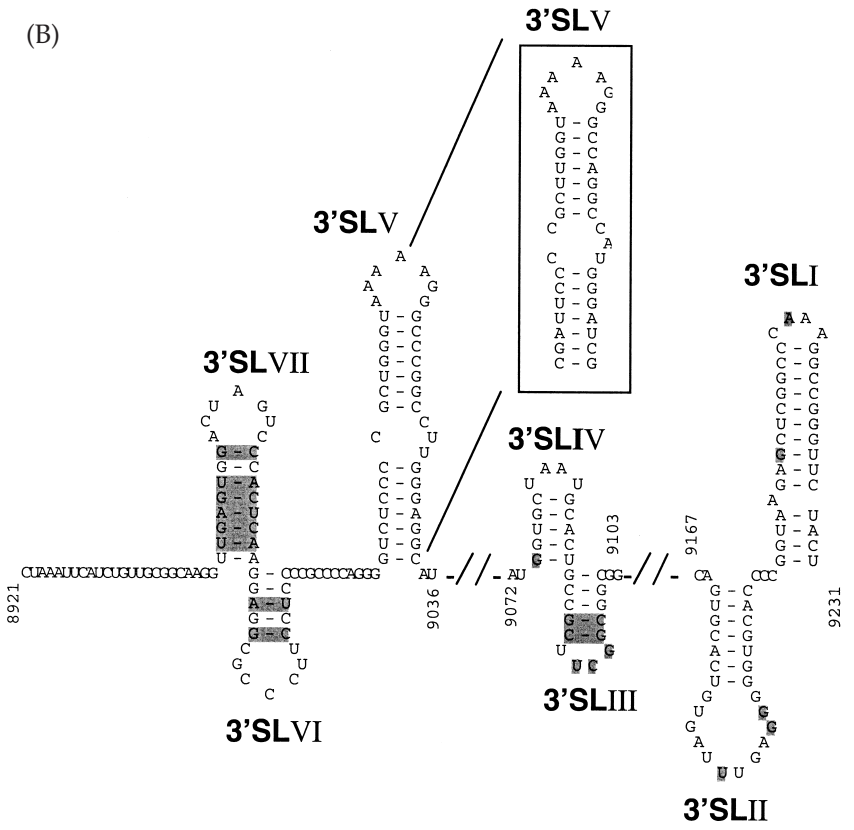

Fig. 2. Predicted secondary structure of (A) the NS5B region and (B) the 3′UTR of HGV/GBV-C. Bases numbered from start of the sequence AB013500. Variable sites between HGV/GBV-C genotypes 1–4 shown in grey boxes. Symbols: '–', canonical Watson–Crick base pairing or GU pairing; '//', intervening sequence without secondary structure (not shown). Stem–loops formed by HGV/GBV-C_{CPZ} that differ in structure shown in boxes.

upstream of and including the stop codon (670 bases in the sequence U36380), while the 3′UTR included the whole sequence downstream from this position (321 bases). The free energy of folding was calculated with the programs RNADraw v1.1 or MFOLD using default settings. The contribution of nucleotide order to free energy of folding was estimated by comparison of free energy with the mean value of sequences generated by independent sequence order randomizations. The variability in free energy on folding 50 sequence order randomizations of three representative HGV/GBV-C sequences of genotype 1 (U36380), 2 (AB013501) and 3 (D90601) was comparable to the combined variability shown by three sequence order randomizations of the seven HGV/GBV-C sequences shown in Table 1 [NS5 region: ±0·044 (±2·7%), ±0·041 (±2·8%) and ±0·040 (±2·8%); 3′UTR: ±0·043 (±4·0%), ±0·041 (±4·3%) and ±0·04 (±3·7%) for the three sequences with 50 randomizations].

This method was also used to analyse whole virus genomes of HGV/GBV-C, HGV/GBV-C_{CPZ} and GBV-A using a sliding window of 1000 bases. The free energies on folding each fragment of HGV/GBV-C, HGV/GBV-C_{CPZ} and GBV-A sequences were compared with those of 15 sequence order randomizations. Variability in the free energies of folding the latter sequences was used to calculate a standard error for the free energy difference estimate.

Free energies in the HGV/GBV-C NS5B and 3′UTR were compared with the mean values of sets of four 5′UTR sequences (U44402, U63715, AB008335 and D87263), five plant viroid sequences, five delta virus sequences, and five albumin and five α-globin sequences from a range of vertebrate species.

■ **Sequence software.** All randomization, free energy calculations and secondary structure predictions were made with the programs RNADraw v1.1 and MFOLD using default settings. Sequence alignments and distance measurements were performed with the Simmonic 2000 package, which is available from the authors.

Results

Free energy on RNA folding

The free energy of folding in different parts of the HGV/GBV-C genome was measured in consecutive 1000 base fragments spanning the genome (Fig. 1). Free energy of folding of the HGV/GBV-C sequence was compared with that of independently generated control sequences whose nucleotide sequence order had been randomized. Although each segment of the HGV/GBV-C genome showed a greater free energy than the control sequences, the difference was most marked at the 3′ end of the genome, with the fragments from 7000–8000, 8000–9000 and 9000–end showing differences ranging from 16 to 18%. To investigate whether the effect of sequence ordering on free energy was conserved amongst viruses related to HGV/GBV-C, similar analyses were carried out on sequences from HGV/GBV-C_{CPZ} and GBV-A. Although these viruses display only 75% and ≈ 40% sequence similarity to HGV/GBV-C, they both showed remarkably similar free

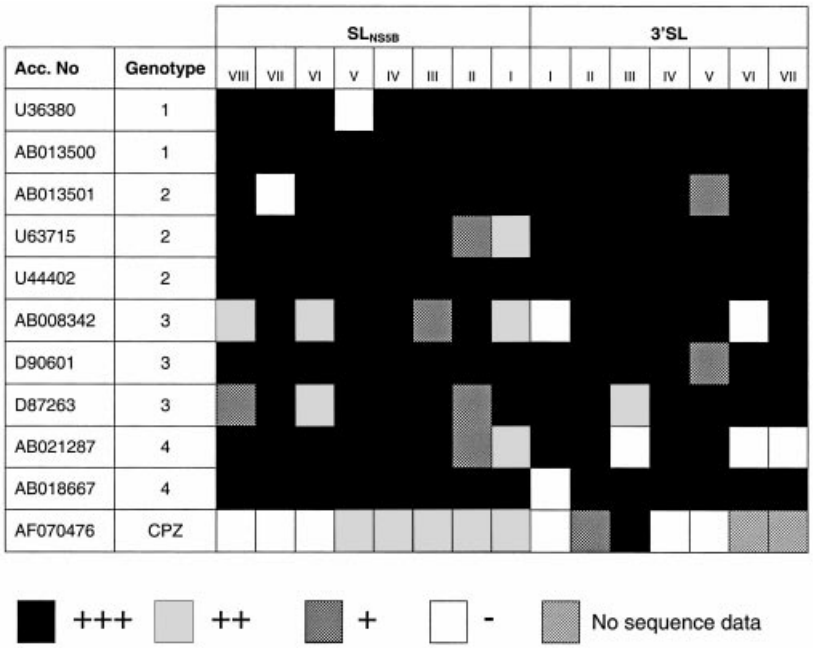

Fig. 3. Structure conservation of predicted stem–loops in the NS5B region (SL_{NS5B}I–SL_{NS5B}VIII) and 3′UTR (3′SLI–3′SLVII) between representative sequences of HGV/GBV-C genotypes depending on the degree of similarity to the most common structure in HGV/GBV-C sequences as follows: ' + + + ', stem–loop structurally identical; ' + + ', minor differences in base-pairing but conservation of overall size and shape of stem–loop; ' + ', different structure in the same region; '-', no secondary structure detected.

energy profiles, with particularly large differences in the 3′-terminal regions. The large free energies on folding and hence the possibility of secondary structure formation in genomic RNA are therefore a conserved evolutionary feature of this subgenus of viruses.

In this study we have concentrated our analysis on the 3′-terminal region of HGV/GBV-C (Table 1). Fragments of the coding NS5B gene sequences, and the non-coding 3′UTR of different HGV/GBV-C genotypes, showed differences in free energy of 15–23% (mean 18·9 ± 3·4%) and 15–20% (mean 17·3 ± 4·0%) from sequence order-randomized controls. The significance of these free energy differences is indicated by a comparison with sequences of other viruses or virus-like agents with previously documented RNA secondary structure (Table 2). Firstly, the 5′UTR of HGV/GBV-C, whose conserved secondary structure is believed to function as an internal ribosomal entry site (Simons *et al.*, 1996), shows a free energy on folding of 1·4 kJ/base, 10% higher than randomized controls. Secondly, plant viroids contain covalently closed single-stranded RNA genomes with extensive stem–loop structures required for various RNA-catalysed replicative functions. These sequences showed a mean free energy on folding of −1·17 kJ/base (range −1·10 to −1·20), and a difference in free energy of 19·5% (16–23%). Finally, the non-coding region of the single-stranded RNA delta virus genome also contains a well-characterized RNA structure, in which RNA is folded into a ribosome-like domain. The mean free energies on folding [1·29 kJ/base (range −1·27 to −1·32), 20·0% difference from controls (range 15–23%)] of five delta virus variants were also remarkably similar to those calculated for the NS5B region of HGV/GBV-C (Table 2). As negative controls, sequences of the coding regions of several different mammalian and non-mammalian α-globin and albumin genes (which would not be expected to form secondary structures) showed no significant differences in free energy on folding from sequence order-randomized controls. These sequences showed markedly different codon biases; albumin shows a low G + C content (45·4% G + C overall, 34·2% G + C at 3rd base positions), while α-globin has a high G + C content (62·0% overall, 83·0% at 3rd base positions). Since these sequences show no difference from randomized controls in free energy on folding, it is unlikely that biased codon usage per se would influence the values calculated for HGV/GBV-C, as its G + C content lies within these two extremes (63·3% overall, 65·7% at 3rd base positions).

Prediction of RNA secondary structure

Two RNA structure prediction methods (RNADraw and MFOLD) were used to compare the folding of different variants of HGV/GBV-C and the more divergent HGV/GBV-C_{CPZ} (Fig. 2). Results from the two methods were equivalent. Both predicted seven potential stem–loops formed by RNA folding in the 3′UTR (provisionally assigned as 3′SLI–3′SLVII), and eight in the NS5B region (labelled SL_{NS5B}I–SL_{NS5B}VIII). Most of the loops in both regions were conserved between all HGV/GBV-C genotypes, and many were also found in HGV/GBV-C_{CPZ}. A scoring system was adopted to indicate the degree of structure conservation between RNA sequences (Fig. 3), differentiating between identity of RNA structure (+ + +), similar folding but with minor differences in the positions of bulges and/or minor slippage (+ +) and folding the same region but with a different structure (+). (Examples of structural differences scored as + + and + between human and chimpanzee HGV/GBV-C secondary structures are shown

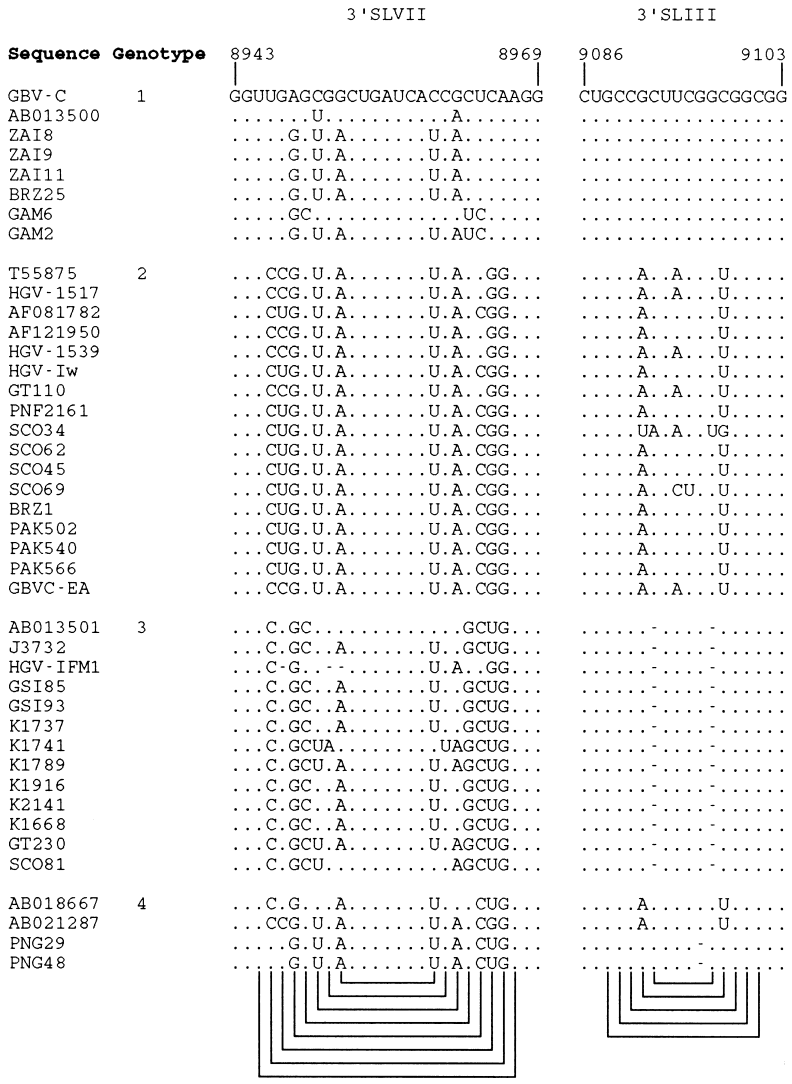

Fig. 4. Covariant substitutions in stem–loops 3′SLVII and 3′SLIII in the 3′UTR between variants of HGV/GBV-C. Bases identical to that of GBV-C indicated as '.'. Proposed base pairings in stem–loops indicated underneath alignment.

in Fig. 2). By these comparisons, substantial structural similarity was found between all sequences in the NS5B region, with similar or identical stem–loops $SL_{NS5B}I$ to $SL_{NS5B}V$ being predicted for both human and chimpanzee HGV/GBV-C sequences. In contrast, only stem–loop 3′SLIII was completely conserved in the 3′UTR, the predicted structure of the HGV/GBV-C_{CPZ} sequence differing considerably elsewhere in this region. The different prediction for this variant may, however, have resulted at least in part from the likely incompleteness of the published HGV/GBV-C_{CPZ} sequence. Alignment of homologous bases indicated that the terminal 55 bases were missing, including the regions forming stem–loops 3′SLII and 3′SLI.

Analysis of covariance

A large number of fully covariant and semi-covariant (C ↔ U opposite G, A ↔ G opposite U, and G ↔ U opposite A) sites were identified in predicted base-paired regions, particularly between human and chimpanzee sequences (Fig. 2). These accommodated the sequence variability between genotypes in structurally conserved regions. Although sequence variability between HGV/GBV-C genotypes was distributed throughout the NS5B and 3′UTR regions, a lower frequency of substitutions were found in regions predicted to be base-paired (18 from 273 sites in NS5B, 2 from 116 sites in the 3′UTR) than in regions predicted to be non-base-paired (57/267 and 21/188; $P < 10^{-7}$ and $P = < 0{\cdot}002$ respectively by Fisher's Exact Test).

One of the problems of verifying structure predictions for the 3′UTR using covariance was lack of comparative sequence information of two of the four HGV/GBV-C genotypes. In GenBank, there are currently only 25 complete or near-complete 3′UTR sequences of which only two are from type 1 and two of type 4. We obtained additional 3′UTR sequences of genotypes 1–4 from HGV/GBV-C-infected individuals from various geographical locations. Sequence variability in the 3′UTR was largely confined to the predicted stem–loop 3′SLVII (Fig. 4), which demonstrated a large number of

covariant substitutions both between and within genotypes. Covariant sites, and in genotype 3 a paired deletion, were also found in stem–loop 3′SLIII. The remaining predicted structures were in regions of the 3′UTR invariant between HGV/GBV-C genotypes, although covariant changes and some structural differences with the HGV/GBV-C_{CPZ} sequence were detected in loops 3′SLVI and 3′SLIV (data not shown).

Sequence variation in the NS5B region was also marked by multiple covariant sites, generally between paired, usually synonymous, 3rd codon positions in predicted stem–loops (data not shown). There are also examples of minor structural differences between HGV/GBV-C genotypes, such as the terminal region of stem–loop SL_{NS5B}III. The existence of multiple covariant sites in each of the predicted stem–loops, and their conservation of the majority of structures in the more divergent HGV/GBV-C_{CPZ} sequence, provides strong supporting evidence for the secondary structure predicted by free energy calculations.

Discussion

Structure of the 3′UTR

The secondary structure of the 3′UTR was inferred by folding algorithms based on calculations of free energy, structural conservation between HGV/GBV-C genotypes and the occurrence of covariant changes in base-paired regions. Our prediction differs from that of a previous study which was based on an analysis of the last 140 nucleotides of the 3′UTR from three HGV/GBV-C sequences (Okamoto *et al*., 1997). The more extensive analysis of 42 sequences in this study identified several substitutions that disrupted the previously proposed base-pairings. For example, the proposed pairing of positions 9098 (U) and 9229 (A) would be disrupted in the majority of sequences because of U → C substitution at position 9098, and in one group 2 sequence (SCO34) because of a U → G substitution; another U–A pair between 9109 and 9117 is affected in three sequences (SCO34, PNG29, PNG48) because of a U → C substitution. Other positions at which substitutions produce mismatches in regions predicted to be base-paired by Okamoto *et al*. are: C–C at positions 9140 and 9193 in four group 3 and two unassigned sequences; G–G at positions 9150 and 9189 in three genotype 3 and two genotype 4 sequences; A–G at positions 9130 and 9201 in one genotype 2 sequence. In addition, the Okamoto *et al*. model is not supported by any covariant substitutions and does not include the fragment upstream of position 9092 in which other structures were found.

Computer-predicted folding patterns and RNase cleavage experiments have previously demonstrated the existence of a long stable hairpin structure (3′-LSH) within the distal part of the 3′UTR of several different flaviviruses (Proutski *et al*., 1997; Brinton *et al*., 1986; Rice *et al*., 1985), some positive-strand RNA plant viruses (Strauss & Strauss, 1983), HCV (Blight & Rice, 1997; Kolykhalov *et al*., 1996), GBV-B (Rijnbrand *et al*., 2000) and pestiviruses (Yu *et al*., 1999; Deng & Brock, 1993). Other studies have provided evidence for a specific interaction between the 3′LSH of flaviviruses and host cellular proteins, components of the virus replication complex or have demonstrated a specific binding of cellular proteins to the 3′-terminal 98 nucleotides of the HCV RNA (Ito & Lai, 1997) and determined which regions of the HCV 3′-UTR are critical for *in vivo* virus replication (Lefrere *et al*., 1999).

The configurations of the predicted terminal loops 3′SLII and 3′SLI in the 3′UTR of HGV/GBV-C (Fig. 2) closely resemble those predicted for HCV (Tanaka *et al*., 1996; Kolykhalov *et al*., 1996) and GBV-B (Rijnbrand *et al*., 2000). However, there was no evidence for a conserved third loop (3′SLIII) nor primary sequence similarity between HGV/GBV-C and HCV or GBV-B. Additionally, the terminal loop is shorter than the HCV and GBV-B homologues (14 base pairs in the stem instead of 19 or 20), although the predicted free energy for formation of the terminal loop (−73 kJ, −2·2 kJ/b) is similar to that of HCV (−110 kJ, −2·4 kJ/b), and indicates a high probability of its formation *in vivo*.

The 3′UTR sequences of human HGV/GBV-C genotypes were highly conserved, with mean pairwise distances between genotypes ranging from 3·8 to 6·6%, compared with 12·8 to 13·4% over the rest of the genome. The HGV/GBV-C_{CPZ} sequence, however, did not display the same differential in divergence, with 23–26% sequence divergence from human genotypes in the 3′UTR, only slightly lower than observed upon comparison of coding sequences (30%). Structurally, only loop 3′SLV was found in both human and chimpanzee variants, although the alignment indicated that the region containing the two terminal loops was missing from the published HGV/GBV-C_{CPZ} sequence. The lack of structural conservation between HGV/GBV-C sequences is not unexpected, given the lack of similarity between other members of the hepaciviruses and pestiviruses in this region. For example, only the terminal three loops are conserved between HCV and GBV-B, and there is also considerable sequence variability between HCV genotypes in non-coding regions 5′ to this, including the great variability in length of the poly(U) tract. Clearly there are varying constraints on sequence change between different regions of the 3′UTR. However, apart from the involvement of the terminal loops in transcription initiation of HCV (Lohmann *et al*., 1999) and potentially other hepaciviruses, it remains unclear what other functional roles secondary structure in the 3′UTR may play.

Secondary structure in NS5B

The prediction methods used to determine the structure of the 3′UTR were also used to analyse the coding sequence of NS5B. Surprisingly, this region showed even greater free energy on folding than the 3′UTR, several large stem–loops such as those numbered SL_{NS5B}III and SL_{NS5B}V and, in contrast to the 3′UTR, substantial structural similarity between

human and chimpanzee HGV/GBV-C variants. Generally, secondary structures were either identical between HGV/GBV-C variants or, particularly on comparison with the HGV/GBV-C_{CPZ} sequence, showed some differences in the identity of the bases involved in base-pairing, but retained conservation of the overall shape and size of the stem–loop. The finding of secondary structure in this region and differences in free energy with sequence order-randomized sequences throughout the genome (Fig. 1) confirms and extends our previous predictions for extensive structure of the HGV/GBV-C genome based on a novel method of covariance scanning and analysis of the distribution of variability at synonymous sites (Simmonds & Smith, 1999). The involvement of such a high proportion of bases in internal base-pairing in the NS5B region, and by implication elsewhere in the genome, suggests that the RNA molecule may be extensively folded through local and possible longer-range interactions to form a 'tertiary' RNA structure.

The conservation in structure and the large number of covariant substitutions suggest a functional role(s) for the RNA structures. Particularly striking was the similarity in free energy on folding the NS5B sequences with the free energies observed for plant viroids and the non-coding region of delta viruses. For these agents, the secondary structure is essential for the replication of the genome, where specific domains may catalyse RNA cleavage, ligation and editing of the genomic RNA sequence. For HGV/GBV-C, amongst several possibilities, RNA folding may be required for packaging of the HGV/GBV-C genome into virus particles, or to protect the genome from RNA-degrading enzymes, particularly as HGV/GBV-C and related viruses do not appear to encode a conventional nucleocapsid protein. In other RNA viruses, secondary structures such as the *cis*-acting replication element in picornaviruses may play a role in initiation of RNA synthesis through long-range interactions with the 3′-terminal region of the genome (Goodfellow *et al.*, 2000; McKnight & Lemon, 1998). The interactions between different genomic regions implied by these observations suggest that HGV/GBV-C might also have an organized overall structure of the RNA genome, in which stem–loop structures may play a role in virus replication.

Enzymatic and chemical methods have been used to provide evidence of secondary structures independently of sequence analysis. While we have considered this approach for the further investigation of the HGV/GBV-C described in this and our previous study (Simmonds & Smith, 1999), the problem with the analysis of sequences such as the NS5B is that they are too long to be easily resolvable by conventional methods. Although it may be possible to separately analyse shorter lengths of sequence in this region, splitting sequences in this way could disrupt the longer-range interactions such as the base-pairing in the base of stem–loops SL_{NS5B}III and V. Direct visualization by electron microscopy of RNA folded in physiological conditions is potentially a better method to determine secondary structure of longer sequences of RNA, particularly if combined with hybridization with gold-labelled probes to identify specific sequences within the observed structures. We are currently carrying out this analysis with RNA transcripts from the two regions analysed in this study. Additionally, experimental manipulation of the recently described infectious clone of HGV/GBV-C and methods to culture the virus *in vitro* (Xiang *et al.*, 2000) may allow a direct investigation of the functional significance of RNA folding in this region of the genome.

Using methods described in this and our previous study, we have also commenced secondary structure analyses of other members of the flavivirus family. HCV shows an even greater excess of free energy on folding the NS5B region (23%, Table 2), several stem–loop structures conserved between HCV genotypes (which are much more divergent in nucleotide sequence than between the HGV/GBV-C sequences analysed in this study), and the occurrence of multiple covariant sites in each of the predicted stem–loops (data not shown). The availability of a replicating clone of HCV (Lohmann *et al.*, 1999) may allow the role of such structures to be experimentally investigated.

References

Adams, N. J., Prescott, L. E., Jarvis, L. M., Lewis, J. C. M., McClure, M. O., Smith, D. B. & Simmonds, P. (1998). Detection of a novel flavivirus related to hepatitis G virus/GB virus C in chimpanzees. *Journal of General Virology* **79**, 1871–1877.

Birkenmeyer, L. G., Desai, S. M., Muerhoff, A. S., Leary, T. P., Simons, J. N., Montes, C. C. & Mushahwar, I. K. (1998). Isolation of a GB virus-related genome from a chimpanzee. *Journal of Medical Virology* **56**, 44–51.

Blight, K. J. & Rice, C. M. (1997). Secondary structure determination of the conserved 98-base sequence at the 3′ terminus of hepatitis C virus genome RNA. *Journal of Virology* **71**, 7345–7352.

Brinton, M. A., Fernandez, A. V. & Dispoto, J. H. (1986). The 3′-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology* **153**, 113–121.

Bukh, J. & Apgar, C. L. (1997). Five new or recently discovered (GBV-A) virus species are indigenous to New World monkeys and may constitute a separate genus of the Flaviviridae. *Virology* **229**, 429–436.

Deng, R. T. & Brock, K. V. (1993). 5′ and 3′ untranslated regions of pestivirus genome – primary and secondary structure analyses. *Nucleic Acids Research* **21**, 1949–1957.

Erker, J. C., Desai, S. M., Leary, T. P., Chalmers, M. L., Montes, C. C. & Mushahwar, I. K. (1998). Genomic analysis of two GB virus A variants isolated from captive monkeys. *Journal of General Virology* **79**, 41–45.

Gonzalez-Perez, M. A., Norder, H., Bergstrom, A., Lopez, E., Visona, K. A. & Magnius, L. O. (1997). High prevalence of GB virus C strains genetically related to strains with Asian origin in Nicaraguan hemophiliacs. *Journal of Medical Virology* **52**, 149–155.

Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J. W., Barclay, W. & Evans, D. J. (2000). Identification of a cis-acting replication element within the poliovirus coding region. *Journal of Virology* **74**, 4590–4600.

Ito, T. & Lai, M. M. C. (1997). Determination of the secondary structure of and cellular protein binding to the 3′-untranslated region of the hepatitis C virus RNA genome. *Journal of Virology* **71**, 8698–8706.

Jarvis, L. M., Watson, H. G., McOmish, F., Peutherer, J. F., Ludlam, C. A. & Simmonds, P. (1994). Frequent reinfection and reactivation of hepatitis C virus genotypes in multitransfused hemophiliacs. *Journal of Infectious Diseases* **170**, 1018–1022.

Katayama, Y., Apichartpiyakul, C., Handajani, R., Ishido, S. & Hotta, H. (1997). GB virus C hepatitis G virus (GBV-C/HGV) infection in Chiang Mai, Thailand, and identification of variants on the basis of 5′-untranslated region sequences. *Archives of Virology* **142**, 2433–2445.

Katayama, K., Kageyama, T., Fukushi, S., Hoshino, F. B., Kurihara, C., Ishiyama, N., Okamura, H. & Oya, A. (1998). Full-length GBV-C/HGV genomes from nine Japanese isolates: characterization by comparative analyses. *Archives of Virology* **143**, 1063–1075.

Kolykhalov, A. A., Feinstone, S. M. & Rice, C. M. (1996). Identification of a highly conserved sequence element at the 3′ terminus of hepatitis C virus genome RNA. *Journal of Virology* **70**, 3363–3371.

Leary, T. P., Muerhoff, A. S., Simons, J. N., Pilot-Matias, T. J., Erker, J. C., Chalmers, M. L., Schlauder, G. S., Dawson, G. J., Desai, S. M. & Mushahwar, I. K. (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.

Leary, T. P., Desai, S. M., Erker, J. C. & Mushahwar, I. K. (1997). The sequence and genomic organization of a GB virus A variant isolated from captive tamarins. *Journal of General Virology* **78**, 2307–2313.

Lefrere, J. J., Roudotthoraval, F., Morandjoubert, L., Brossard, Y., Parnetmathieu, F., Mariotti, M., Agis, F., Rouet, G., Lerable, J., Lefevre, G., Girot, R. & Loiseau, P. (1999). Prevalence of GB virus type C hepatitis G virus RNA and of anti-E2 in individuals at high or low risk for blood-borne or sexually transmitted viruses: evidence of sexual and parenteral transmission. *Transfusion* **39**, 83–94.

Linnen, J., Wages, J., Zhangkeck, 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. K., Young, L., Piatak, M., Hoover, C., Fernandez, J., Chen, S., Zou, J. C., Morris, T., Hyams, K. C., Ismay, S., Lifson, J. D., Hess, G., Foung, S. K. H., 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.

Lohmann, V., Korner, F., Koch, J. O., Herian, U., Theilmann, L. & Bartenschlager, R. (1999). Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science* **285**, 110–113.

McKnight, K. L. & Lemon, S. M. (1998). The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA* **4**, 1569–1584.

Muerhoff, A. S., Smith, D. B., Leary, T. P., Erker, J. C., Desai, S. M. & Mushahwar, I. K. (1997). Identification of GB virus C variants by phylogenetic analysis of 5′-untranslated and coding region sequences. *Journal of Virology* **71**, 6501–6508.

Nakao, H., Okamoto, H., Fukuda, M., Tsuda, F., Mitsui, T., Masuko, K., Lizuka, H., Miyakawa, Y. & Mayumi, M. (1997). Mutation rate of GB virus C hepatitis G virus over the entire genome and in subgenomic regions. *Virology* **233**, 43–50.

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.

Proutski, V., Gould, E. A. & Holmes, E. C. (1997). Secondary structure of the 3′ untranslated region of flaviviruses: similarities and differences. *Nucleic Acids Research* **25**, 1194–1202.

Rice, C. M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L. & Strauss, J. H. (1985). Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. *Science* **229**, 726–733.

Rijnbrand, R., Abell, G. & Lemon, S. M. (2000). Mutational analysis of the GB virus B internal ribosome entry site. *Journal of Virology* **74**, 773–783.

Simmonds, P. & Smith, D. B. (1999). Structural constraints on RNA virus evolution. *Journal of Virology* **73**, 5787–5794.

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

Smith, D. B., Cuceanu, N., Davidson, F., Jarvis, L. M., Mokili, J. L. K., Hamid, S., Ludlam, C. A. & Simmonds, P. (1997). Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5′ non-coding region. *Journal of General Virology* **78**, 1533–1542.

Smith, D. B., Basaras, M., Frost, S., Haydon, D., Cuceanu, N., Prescott, L., Kamenka, C., Millband, D., Sathar, M. A. & Simmonds, P. (2000). Phylogenetic analysis of GBV-C/hepatitis G virus. *Journal of General Virology* **81**, 769–780.

Strauss, E. G. & Strauss, J. H. (1983). Replication strategies of the single stranded RNA viruses of eukaryotes. *Current Topics in Microbiology and Immunology* **105**, 1–98.

Tanaka, T., Kato, N., Cho, M. J., Sugiyama, K. & Shimotohno, K. (1996). Structure of the 3′ terminus of the hepatitis C virus genome. *Journal of Virology* **70**, 3307–3312.

Tanaka, Y., Mizokami, M., Orito, E., Ohba, K., Kato, T., Kondo, Y., Mboudjeka, I., Zekeng, L., Kaptue, L., Bikandou, B., Mpele, P., Takehisa, J., Hayami, M., Suzuki, Y. & Gojobori, T. (1998). African origin of GB virus C hepatitis G virus. *FEBS Letters* **423**, 143–148.

Xiang, J., Wunschmann, S., Schmidt, W., Shao, J. & Stapleton, J. T. (2000). Full-length GB virus C (hepatitis G virus) RNA transcripts are infectious in primary CD4-positive T cells. *Journal of Virology* **74**, 9125–9133.

Yu, H. Y., Grassmann, C. W. & Behrens, S. E. (1999). Sequence and structural elements at the 3′ terminus of bovine viral diarrhea virus genomic RNA: functional role during RNA replication. *Journal of Virology* **73**, 3638–3648.

Received 17 August 2000; Accepted 10 January 2001