Journal of General Virology (1993), **74**, 661–668. *Printed in Great Britain*

Sequence variability in the 5′ non-coding region of hepatitis C virus: identification of a new virus type and restrictions on sequence diversity

P. Simmonds,[1]* F. McOmish,[2] P. L. Yap,[2] S.-W. Chan,[1] C. K. Lin,[3] G. Dusheiko,[4] A. A. Saeed[5] and E. C. Holmes[6]

[1]*Department of Medical Microbiology, Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG,* [2]*Edinburgh and South East Scotland Blood Transfusion Service, Royal Infirmary of Edinburgh, Lauriston Place, Edinburgh EH3 9HB, U.K.,* [3]*Hong Kong Red Cross Blood Transfusion Service, 15 King's Park Rise, Yaumatei, Kowloon, Hong Kong,* [4]*Department of Medicine, Royal Free Hospital, London NW3 2PS, U.K.,* [5]*Riyadh Armed Forces Hospital, Riyadh 11159, Saudi Arabia and* [6]*Division of Biological Sciences, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JN, U.K.*

We have analysed the pattern of nucleotide sequence variability in the 5′ non-coding region (5′ NCR) of geographically dispersed variants of hepatitis C virus (HCV). Phylogenetic analysis of sequences in this region indicated the existence of a new virus type, provisionally termed type 4, the identity of which was confirmed by further analysis of the more variable part of the HCV core protein coding region. The geographical distribution of HCV type 4 was distinct from that of other HCV types, it being particularly widespread in Africa and absent or rare in Europe and the Far East. Much of the variability in the 5′ NCR appears to be constrained by a requirement for specific secondary structures in the viral RNA. In one of the most variable regions of the 5′ NCR (positions −169 to −114), most of the nucleotide changes that are characteristic of different HCV types were covariant, with complementary substitutions at other positions. According to the proposed secondary structure of the 5′ NCR, such changes preserved base pairing within a stem–loop structure, whereas the nucleotide insertions found in a proportion of 5′ NCR sequences, including those of type 4, localized exclusively to the non-base-paired terminal loop. The specific nucleotide substitutions in the 5′ NCR that differentiate each of the four HCV types can be detected by restriction enzyme cleavage, providing a rapid and reliable method for virus typing.

Introduction

The introduction of serological tests to detect antibody to hepatitis C virus (HCV) in blood donors is now the principal measure to prevent post-transfusional non-A, non-B hepatitis (NANBH). The effectiveness of such assays is dependent in part upon the extent of sequence diversity between different isolates of HCV in different geographical regions. Published sequences of different HCV variants can be classified into a number of different virus types on the basis of overall sequence similarity in both coding and non-coding parts of the viral genome. Phylogenetic analysis of the 5′ non-coding region (5′ NCR) led us to propose a classification of HCV into three major types, 1, 2 and 3 (Chan *et al.*, 1992*b*). Analysis of the more variable coding regions of the viral genome (core, NS-3, NS-5) indicated that each major type was composed of two or three distinct subtypes, termed 1a, 1b, 2a etc. (Chan *et al.*, 1992*a*, *b*). An alternative typing scheme (Houghton *et al.*, 1991) is similar to this except that types 1a (represented by the prototype virus HCV-1; Choo *et al.*, 1989) and 1b (exemplified by HCV-J and HCV-BK; Kato *et al.*, 1990; Takamizawa *et al.*, 1991) are called types I and II respectively, while types 2a and 2b (Enomoto *et al.*, 1990; Okamoto *et al.*, 1991, 1992*a*) are called type III. More recently, sequences originally described as type 3 (Chan *et al.*, 1992*b*) have been reported and termed group IV (Cha *et al.*, 1992) whereas a previously undescribed group of variants, so far found only in South Africa, is described as group V. A third system (Okamoto *et al.*, 1992*a*; Mori *et al.*, 1992) divides type 2 (or type III) sequences into types III and IV, and the two subtypes of type 3 (group IV) are described as types V and VI (Mori *et al.*, 1992). To avoid confusion in the

The nucleotide sequence data reported have been submitted to the GenBank database and assigned the accession numbers L08136 to L08164.

current communication, HCV variants will be described as previously proposed (Chan *et al.*, 1992*b*).

The degree of sequence variability differs throughout the genome. At one extreme, the core protein demonstrates approximately 90% sequence similarity between different HCV types, whereas high degrees of sequence divergence are found in certain non-structural proteins (e.g. NS-4) and in the region of the genome encoding putative envelope proteins, with greater than 50% sequence divergence in certain 'hypervariable' regions (Hijikata *et al.*, 1991; Weiner *et al.*, 1991; Kato *et al.*, 1992). We have previously shown that infection with HCV variants other than type 1 (whose sequences are used for the antigens used in screening assays) may elicit an antibody response that is not cross-reactive with the NS-4-encoded antigens used in first generation screening assays (5-1-1, c100-3; McOmish *et al.*, 1993). There is now substantial data showing the increased sensitivity of second generation serological assays that include the well conserved core protein in addition to non-structural proteins (Craxi *et al.*, 1991; van der Poel *et al.*, 1990, 1992; Chan *et al.*, 1991; Watson *et al.*, 1992; Lelie *et al.*, 1992), and it can be reasonably anticipated that such tests will be more effective than the original anti-C100-based assays for prevention of transfusion-transmitted HCV infection. Nevertheless, the serological response to HCV-encoded antigens is often narrow in specificity and is generally of extremely low titre. Acute infection with HCV following transfusion of infectious blood or blood products may fail to elicit antibody for several months (Lelie *et al.*, 1992; Alter *et al.*, 1989; van der Poel *et al.*, 1992). Furthermore, individuals who are only marginally immunosuppressed (such as renal dialysis patients, the elderly, haemophiliacs and neonates) generally show very restricted and idiosyncratic patterns of serological reactivity, often to only one (or possibly none) of the four antigens used in current screening assays (Lelie *et al.*, 1992; Watson *et al.*, 1992; Allain *et al.*, 1991; Lam *et al.*, 1993).

We have previously argued that serological tests for anti-HCV should be equally sensitive for all variants of HCV, not just type 1 (McOmish *et al.*, 1993). In order to develop such tests, we have explored the distribution of different HCV types world-wide, and carried out a search for new variants outside the existing classification (F. McOmish *et al.*, unpublished results). In this study we have carried out sequence analysis of the 5′ NCR amplified from plasma of anti-HCV-positive blood donors or NANBH patients from Scotland, Finland, Holland, Australia, Hong Kong, Japan, Turkey, Egypt and other countries in the Middle East. The distribution of the three established HCV types (1 to 3, as defined by Chan *et al.*, 1992*b*) is the subject of a separate study (F. McOmish *et al.*, unpublished results); here we analyse those sequences which differ substantially from those previously described, to investigate whether they constitute further HCV types. Our analysis of sequence variability in the 5′ NCR has led to the identification of a new HCV type, and has indicated possible constraints that limit the degree of sequence variability in this untranslated region.

Methods

Samples. RNA was extracted from plasma samples from blood donors and patients with NANBH that were repeatedly reactive on second generation screening assays for HCV, and which were either confirmed (significant reactivity with two or more antigens in the Chiron recombinant immunoblot assay; Chiron Corporation, Emeryville, Ca., U.S.A.) or indeterminate (reactivity with only one antigen) upon supplementary testing. Most of the samples containing sequences that differed substantially from known HCV types came from individuals of Egyptian origin (EG-1 to EG-96). Variant sequences were also detected in several Hong Kong blood donors (HK-1 to HK-4), in a Dutch blood donor who was likely to have been originally infected in Indonesia (IN-26; T. Cuypers, personal communication), and a patient with NANBH from Iraq (IQ-48).

Sequence determination. HCV sequences were reverse-transcribed and amplified by PCR using *Taq* polymerase (Northumbria) and primers matching conserved regions in the 5′ NCR as previously described (Chan *et al.*, 1992*b*). For analysis of the core region, RNA was reverse-transcribed using a primer of sequence CA(T/C)GT-(A/G)AGGGTATCGATGAC (5′ base: 383; numbered as in Choo *et al.*, 1991). cDNA was amplified using this primer and a primer in the 5′ NCR of sequence ACTGCCTGATAGGGTGCTTGCGAG (5′ base: −53). The second PCR used primers of sequences AGG-TCTCGTAGACCGTGCATCATG (5′ base: −20) and TTGCG-(G/T/C)GACCT(A/T)CGCCGGGGGTC (5′ base: 353). Amplified DNA in both regions was directly sequenced as previously described (Chan *et al.*, 1992*b*).

Sequence analysis. Sequences were aligned using the ClustalV program (Higgins *et al.*, 1992) as implemented in the GDE (version 2.0) sequence analysis package kindly provided by the Harvard Genome Laboratory. Phylogenetic trees were reconstructed by a number of programs taken from the PHYLIP package (version 3.4) of Felsenstein (1991). The principal method utilized was maximum likelihood (program DNAML) although neighbour-joining trees (program NEIGHBOR) were also reconstructed to instil more confidence in the results. Global branch-swapping was used in multiple runs of DNAML and nucleotide sequence distances for the neighbour-joining analysis were estimated using the same evolutionary model as is used in the maximum likelihood method (program DNADIST). Furthermore, in order to assess the reliability of specific groupings of sequences found in the neighbour-joining analysis, 500 bootstrap replications of the data were performed (programs SEQBOOT and CONSENSE) and groupings were taken to be significantly supported when they were found in at least 95% of bootstrap replicates (Felsenstein, 1991). Estimation of mean nucleotide distances within and between HCV types 1, 2, 3 and 4 in the 5′ NCR and core regions was carried out on non-identical sequences using DNADIST.

RNA secondary structures in the 5′ NCR of five representative HCV variants (Choo *et al.*, 1989; Kato *et al.*, 1990; Takamizawa *et al.*, 1991; Okamoto *et al*, 1991, 1992*a*) were predicted using the program FOLD (Devereux *et al.*, 1984). Three predictions were made from each sequence between nucleotides −318 to 0, −318 to +281, and −318 to +881 to allow for possible long-range interactions. Comparison of the predicted conformations for each sequence over the different lengths

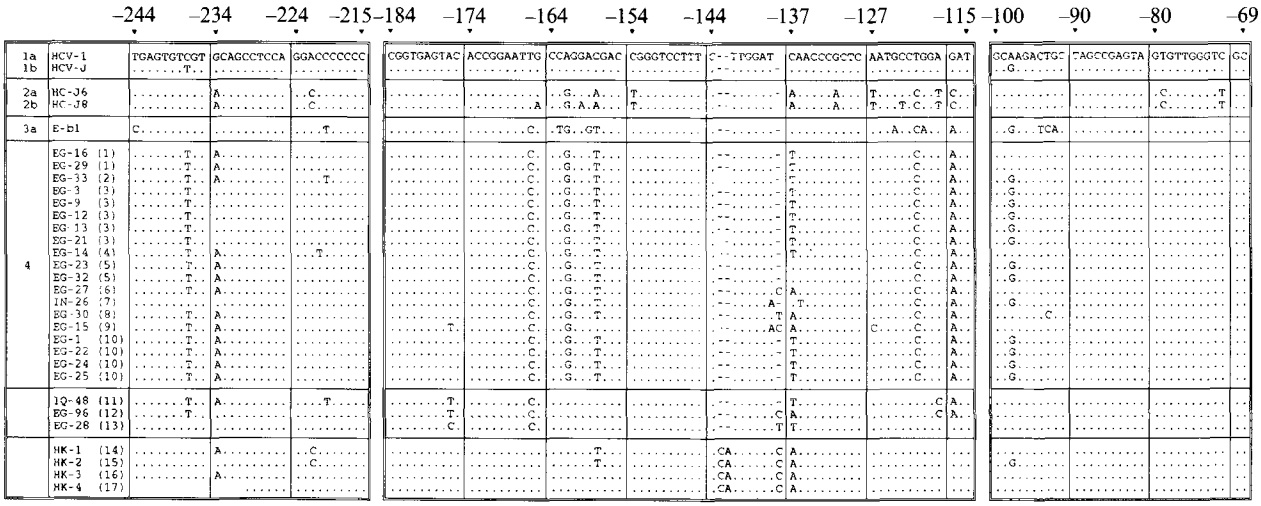

Fig. 1. Comparison of divergent HCV sequences with representative type 1, 2 and 3 sequences in variable regions of the 5′ NCR. Sequences from −254 to −245, −214 to −185, −114 to −101 and −68 to −61 identical to prototype sequence, and not shown to save space. (.) Sequence identity with HCV-1; (−) gap introduced in sequences to preserve alignment; ( ) sequence not determined. Figures in parentheses identify each non-identical sequence used for phylogenetic analysis in Fig. 2.

showed that only relatively small-scale features, such as the stem and loop (analysed in Results) were at all conserved (data not shown).

Results

Divergent 5′ NCR sequences

Several sequences in the 5′ NCR detected in samples from blood donors originally infected in Egypt, Indonesia and Hong Kong, and from NANBH patients in Iraq and Egypt, differed substantially from those found in Scottish blood donors and those reported elsewhere (Fig. 1). Instead of showing the nucleotide substitutions that distinguish HCV types 1, 2 and 3 from each other, a new set of differences was observed that appeared to place them outside the existing system of virus classification. The evolutionary relationships between the sequences from bp −244 to −69 (numbered as in Choo *et al.*, 1991) were reconstructed using the maximum likelihood method (program DNAML), and are presented as an unrooted phylogenetic tree (Fig. 2).

Types 2 and 3 were found to be the most distinct on the basis of the maximum likelihood tree (Fig. 2), a finding confirmed by the neighbour-joining phylogenetic analysis (trees not shown) where more than 95% of bootstrap replicates supported the separation of types 2 and 3 from any of the other sequences. The maximum likelihood tree also suggested the existence of further clusters of variants, most notably the sequences previously described as type 1 and the new sequences presented here, although the phylogenetic position of these variants could not be established unequivocally. This was also found to be the case in a neighbour-joining analysis where these clusters could not be assigned with any statistical certainty by the bootstrap method.

Thus, there is some evidence from the phylogenetic analysis of the 5′ NCR that some of the new sequences presented here are not members of types previously designated as types 1, 2 or 3. We refer to these divergent variants (sequence numbers 1 to 10) as type 4, although this provisional designation may change as further variants are reported by other research groups. Using this classification, mean distances within type 4, and between type 4 and the other HCV types, in the 5′ NCR were comparable to those previously described for types 1 to 3 (Table 1 *a*). It is evident from this phylogenetic tree that the majority of variants identified in a previous world-wide survey (Bukh *et al.*, 1992) can be readily identified as types 1, 2 or 3 (Fig. 2). HCV variants detected in Zaire, however, cluster closely with the type 4 sequences reported here.

Finally, the status of a number of other sequences appears to be ambiguous. Three sequences, IQ-48 (11), EG-96 (12) and EG-28 (13) are linked to the other sequences by long branch lengths (Fig. 2) and change positions between the maximum likelihood and neighbour-joining trees (data not shown). Three sequences obtained from South African patients (Bukh *et al.*, 1992) appear to be distinct from those of both type 1 and type 4, and may possibly represent another HCV type or subtype, as is the case for a single variant from Indonesia (IN-26). The status of variants detected in individuals from Hong Kong (HK-1 to -4) also remains ambiguous. The actual sequences in the 5′ NCR differ little from type 1, and cluster relatively closely to the main group of type 1 sequences in the phylogenetic tree (numbers 14 to 17; Fig. 2). However, all show insertions at two positions (between nucleotides −144 and −143, and between −138 and −137 in the HCV-1 sequence; see below).

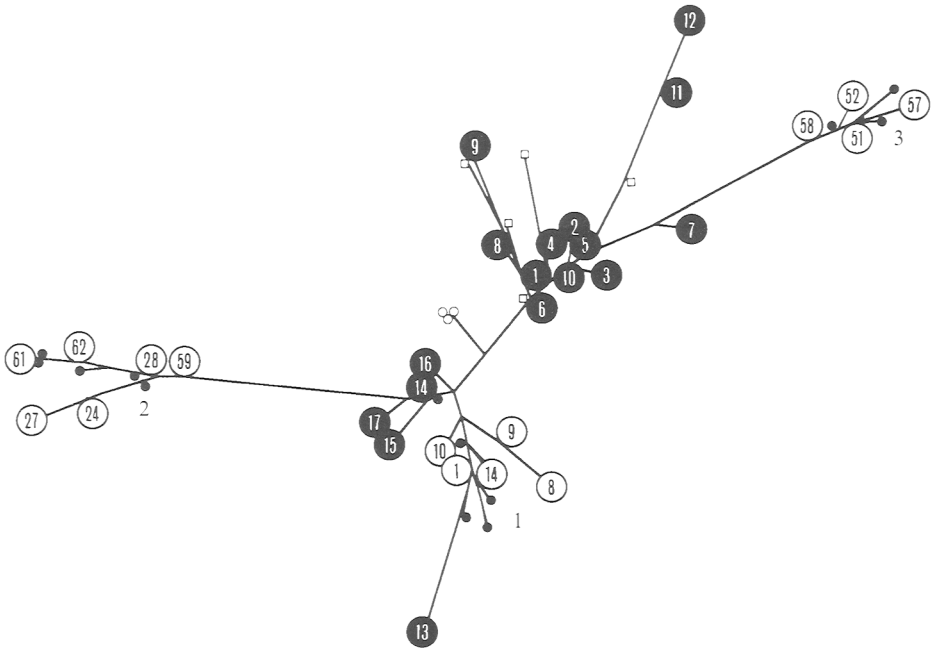

Fig. 2. Phylogenetic analysis of sequences between −244 and −69 in the 5′ NCR using DNAML, shown as an unrooted tree. ❶ to ⓱, sequences numbered as in Fig. 1; previously published sequences numbered as in Chan *et al.* (1992*b*); Scottish blood donor sequences E-b1 to E-b12 numbered ㊶ to ㊷. For clarity, only non-identical sequences are shown in the tree; e.g. sequence 1 corresponds to those found in samples EG-16 and EG-29 etc. (Fig. 1). (□) Published sequences (Bukh *et al.*, 1992) from Zaire; (○) sequences from South Africa; (●) sequences obtained elsewhere in the world. All branch lengths are shown to scale.

Table 1. *Mean nucleotide distances within and between HCV types 1 to 4 in* (a) *the 5′ NCR and* (b) *the core region*

			HCV type					
(*a*)		n*	1†	2	3	4		
	1	14	0·018					
	2	13	0·091	0·020				
	3	9	0·095	0·144	0·012			
	4	13	0·054	0·089	0·080	0·022		
(*b*)		n*	1a	1b	2a	2b	3a	4
	1a	6	0·023					
	1b	5	0·086	0·036				
	2a	1	0·223	0·205	0			
	2b	1	0·193	0·206	0·130	0		
	3a	1	0·151	0·180	0·219	0·220	0	
	4	3	0·161	0·153	0·222	0·195	0·227	0·042

* Numbers of non-identical sequences in each group.
† Sequences 14 to 17 (HK-1 to -4) that appear to group with type 1 sequences excluded from comparison because of uncertain status.

Insertion of a single nucleotide at the latter position is also frequently found amongst type 4 sequences. It is difficult to assess the significance of these insertions, and the issue of whether the variants detected in Hong Kong also represent a new type or subtype cannot be decided at present.

In order to assess further the phylogenetic position of type 4 sequences, RNA from three representative type 4 variants (EG-29, -33, -21; corresponding to 5′ NCR sequence numbers 1 to 3) was amplified using primers representing the core region of the HCV polyprotein. All three sequences differed at both the nucleotide and amino acid levels from HCV types 1, 2 and 3 (Fig. 3 *a* and *b* respectively). Phylogenetic analysis of HCV sequences between positions 13 and 380 in the core region further supports the proposition that the sequences designated type 4 do in fact constitute a new type of HCV (Fig. 4). The unrooted maximum likelihood tree clearly depicts the three Egyptian-derived sequences as an independent phylogenetic group being separated by long branch lengths from the other sequences used in this analysis. Estimated nucleotide distances between type 4 and types 1 to 3 sequences were comparable to the distances between the three known types of HCV (Table 1*b*). Although most of the nucleotide sequence changes were silent, there were between four and nine amino acid differences between the new variants and the other types.

Significance of sequence variability in the 5′ NCR

The reasons for this region being so conserved between different HCV types, to the extent that the sequence subtypes cannot be differentiated from each other, are currently unclear. However, as the 5′ NCR adopts a secondary structure with presumed regulatory roles in virus replication or intiation of translation, the requirement for internal base pairing may considerably

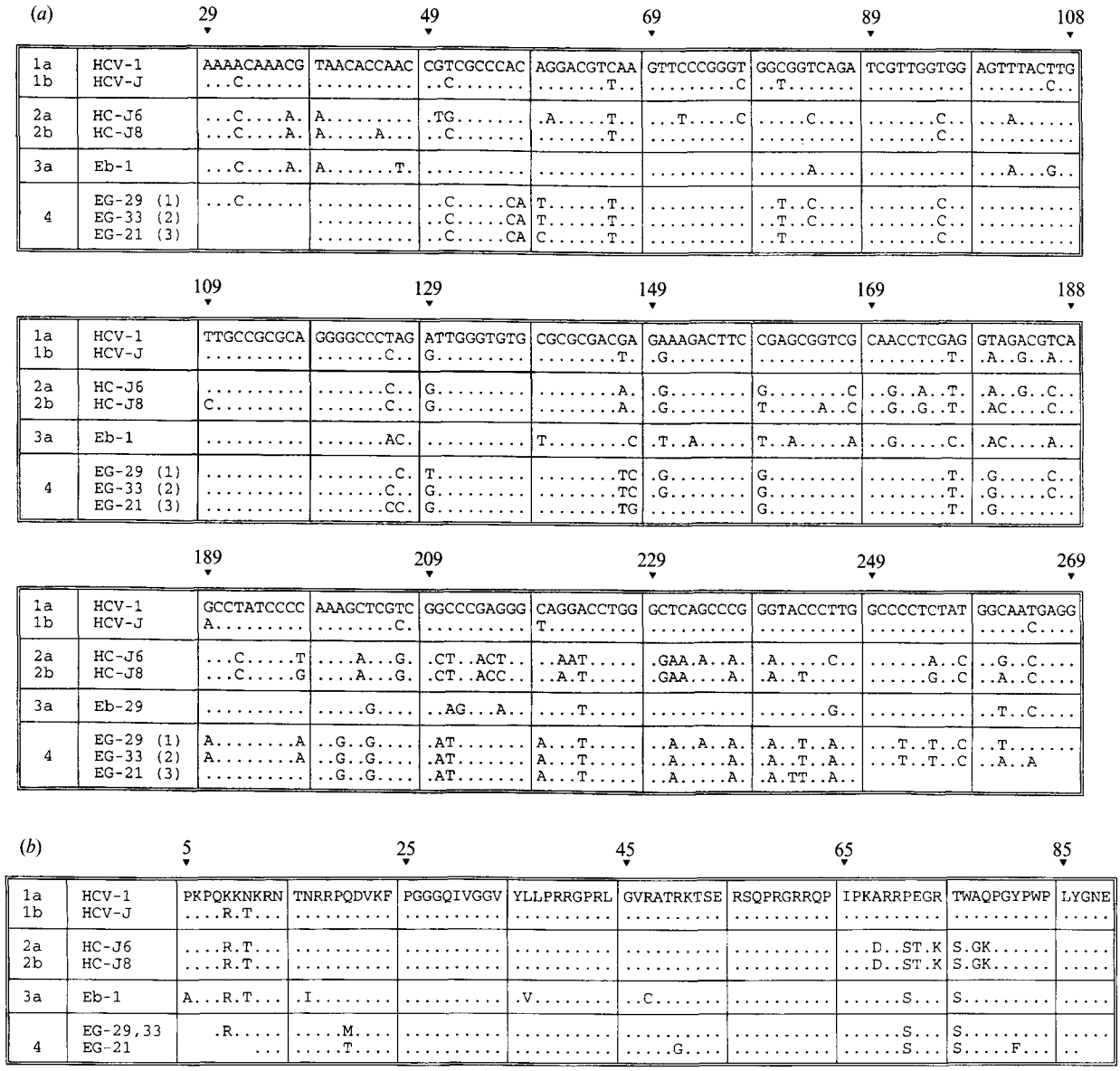

Fig. 3. Comparison of (*a*) nucleotide and (*b*) amino acid sequences in the core region of three type 4 variants with published sequences of HCV types 1, 2 and 3. Symbols as for Fig. 1. Single-letter amino acid codes are used in the sequences shown in (*b*).

restrict the degree of variability possible. Using the program FOLD with a range of input parameters, we have found that many of the small-scale secondary structures for HCV type 1 (Tsukiyama Kohara *et al.*, 1992; Brown *et al.*, 1992) are well conserved amongst all types for which there are sufficient sequence data for complete analysis (types 1a, 1b, 2a and 2b; see Methods section). However, as many of the predicted base pairings involve sequences outside the region described in this and our previous paper, we have concentrated our analysis on the predicted stem–loop structure formed by a highly variable region of the 5′ NCR (positions −169 to −114; Fig. 5).

The stem formed by all of the 5′ NCR sequences described here can be predicted to have considerable stability, with consecutive runs of mainly G and C nucleotides linked by Watson–Crick base pairings. Although there is some variability in the shape of the stem, all variants showed substantial negative free energies (predicted ΔG values ranged from −61·5 to −82·8 kJ/mol; Fig. 5). Further evidence that the stem–loop is present in different HCV types is provided by the pattern of sequence variability. Comparison of the eight HCV sequences in Fig. 5 reveals five covariant positions within the 17 base paired residues forming the stem. For example, position −162 in type 1 contains an A residue which can be predicted to bind to U at −121; in types 2, 3 and 4, G substitutes for A, and to maintain base pairing the U is substituted by a C. Other sites of covariance are found between type 1 and other HCV

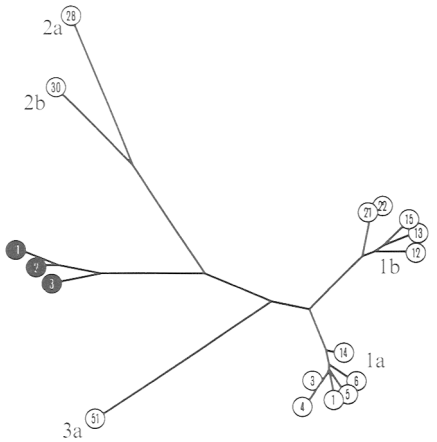

Fig. 4. Phylogenetic analysis of nucleotide sequences of part of the core region (positions 13 to 380) using DNAML, shown as an unrooted tree. Sequences numbered as in Fig. 2; sequence 30 corresponds to isolate HC-J8 (Okamoto *et al.*, 1992*a*). All branch lengths are shown to scale.

types at positions −163/−120, −160/−123, and −154/−131; between type 2 and type 3 and 4, there is a further covariant site at position −158/−127 (Fig. 5). This frequency of compensatory substitutions which maintain base pairing would be unlikely to have arisen by chance. For the two compensatory substitutions observed in the type 2a sequence, $P = 0{\cdot}002$, while for the three covariant sites in type 2b, $P = 0{\cdot}00015$.

Another significant feature of sequence variability in the 5′ NCR is the positions of the single or double nucleotide insertions. Whereas sequences of HCV types 1 to 3 published to date have been entirely collinear, three of the 19 sequences provisionally assigned to type 4, as well as EG-28 and -96 show a single base insertion between nucleotides −139 and −138. Furthermore, all four sequences from Hong Kong blood donors also showed a second two base insertion 5 bp upstream from the first site. Both nucleotide insertions localize to the non-base-paired terminal loop (Fig. 5) and would therefore have no effect on base pairing within the proposed stem structure. As with other HCV types, the loop is also a region of considerable variability amongst type 4 sequences, with frequent substitutions either side of the point of the single base insertion (Fig. 1).

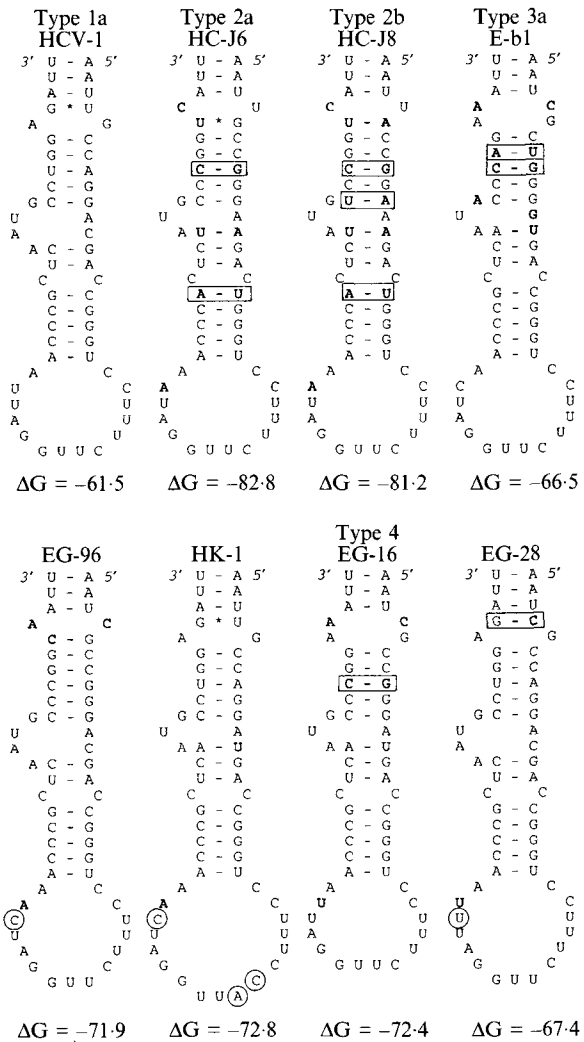

Fig. 5. Predicted secondary structure between nucleotides −169 and −114 of the 5′ NCR for representative sequences of HCV types 1 to 3 and selected sequences reported in this paper. (−) Watson–Crick base pairing; (*) G-T pairing tolerated in secondary structures. Differences from HCV-1 prototype sequence indicated in bold; covariance to maintain base pairing indicated in boxes; inserted nucleotides ringed. Free energy (ΔG) in kJ/mol indicated under each sequence.

Discussion

The 5′ untranslated region has been recognized as the most conserved between different HCV variants worldwide, and oligonucleotides corresponding to it in sequence are almost universally used for amplification by PCR. Greater sequence conservation in this region, compared with coding regions of the genome, is characteristic of several positive-stranded virus groups which produce and subsequently cleave a single, virus-encoded polyprotein (flaviviruses, picornaviruses, pestiviruses). One difficulty with this strategy of replication is the possibility of premature initiation and truncation of protein synthesis caused by the chance production of ATG triplets upstream from the correct reading frame during eukaryotic ribosomal scanning. Different HCV isolates demonstrate variability in both the number (one to four) and position of such triplets before the correct initiating methionine codon. It has recently been shown that sequences in the 5′ NCR that include the stem and loop structure (Fig. 5) play an essential role in the initiation of translation in cell-free expression systems (Tsukiyama Kohara *et al.*, 1992). By analogy with the structure of the equivalent region in the genome of picornaviruses, it is possible that this is achieved by secondary structure-dependent binding of the ribosome

complex, or other cellular factors, to the 5′ NCR, thereby enabling the efficient initiation of translation from the correct methionine codon.

Although computer algorithms can predict possible secondary structures in RNA sequences, there are often many radically different alternatives differing little in overall free energy. However, parts of the predicted secondary structure of HCV type 1 (Tsukiyama Kohara *et al.*, 1992) have recently been experimentally verified by nuclease digestion experiments (Brown *et al.*, 1992). In the current study, the detection of linked substitutions amongst related RNA sequences that maintain base pairing between opposed nucleotides (covariance) provides evidence for the proposed stem–loop structure in all HCV variants. These sites were found in the stem–loop structure in the centre of domain III (Brown *et al.*, 1992), the region implicated in promoting translation of the HCV polyprotein (Tsukiyama Kohara *et al.*, 1992). A similar degree of covariance has been described in stem and loop structures in the 5′ NCR of a wide range of picornaviruses (Rivera *et al.*, 1988; Skinner *et al.*, 1989).

On the basis of the evidence provided by the pattern of sequence variation in the 5′ NCR presented in this paper and elsewhere, we propose that secondary structure considerations may be a significant restriction on the degree of variability possible in the 5′ NCR, because of the requirement for simultaneous compensatory changes elsewhere. This is perhaps reflected in the absence of any consistent differentiation of HCV subtypes amongst the 5′ NCR sequences, in marked contrast to similar analyses of the coding regions (Chan *et al.*, 1992*b*). In evolutionary terms, these data are consistent with the hypothesis that many of the differences in the 5′ NCR between major HCV types developed before the diversification process that led to the appearance of the descendant subtypes of the virus.

The marked clustering of many of the 5′ NCR sequence variants into an apparently new phylogenetic group suggested the existence of at least one new HCV type. However, the relatively small numbers of substitutions that differentiated these sequences from those of types 1, 2 and 3 necessitated an analysis of more variable regions of the genome. Phylogenetic analysis of the core region showed that sequences assigned to type 4 were distinct from the previously identified HCV types, justifying the assignment of 'type' status to the new group. Type 4 sequences show a markedly different geographical distribution from the three previously described types of HCV, accounting for almost all HCV infections in Egypt. The finding that these sequences group with those from Zaire (Bukh *et al.*, 1992) suggests that type 4 has a broad distribution in Africa. In contrast, types 1 to 3 are more frequent elsewhere in the world, and account for all infections in Scotland (McOmish *et al.*, 1992), Finland, The Netherlands, Australia and Japan (F. McOmish *et al.*, unpublished; Okamoto *et al.*, 1992*b*). Although we were unable to obtain sequences in the NS-5 region from our own samples to compare with those published for the new South African type (group V; Bukh *et al.*, 1992; Cha *et al.*, 1992), the latter sequences demonstrate a number of differences from those of both type 1 and type 4 in the 5′ NCR, suggesting that they may represent another major type of HCV, currently with an exclusively African distribution. Considerable work remains to further characterize other variants from the Middle East (IQ-48, EG-96), Indonesia (IN-26) and Hong Kong because as it is important to determine whether there are a relatively restricted number of HCV types corresponding to those identified so far, or whether the total will continue to proliferate as HCV sequences from more countries are obtained.

The existence of relatively conserved patterns of substitutions in the 5′ NCR that are characteristic of different HCV types provides useful 'signature' sequences for the identification of HCV genotypes. Having compared large numbers of different HCV type 1, 2 and 3 sequences we developed a method that differentiated HCV types 1 to 3 by restriction endonuclease cleavage of amplified DNA (McOmish *et al.*, 1993). However, enzymatic cleavage of the 19 type 4 sequences obtained in this study would produce bands corresponding to electrophoretic types Aa and Ab, and would therefore be indistinguishable from type 1 sequences. All type 4 sequences, however, showed a change from U to C at position −166, which creates a novel *Hin*fI site absent in all type 1 (and type 2) sequences. In combination with *Scr*FI, and *Hae*III/*Rsa*I, it has now proved possible to identify provisionally the new type broadly distributed throughout most countries in the Middle East, as well as in Africa (P. Simmonds *et al.*, unpublished). In the future, it should be possible to introduce further modifications to the restriction fragment length polymorphism analysis to identify other HCV types as they are discovered.

Note added in proof. Nucleotide sequences of variants identified here as type 4 have subsequently been obtained from the NS-5 region. Phylogenetic analysis confirmed their provisional designation, arrived at from analysis of the core region, as a new HCV type. A system of virus classification that broadly follows the nomenclature proposed by Chan *et al.* (1992*b*) has been agreed between several laboratories working on HCV sequence variability. The new nomenclature confirms the designation of sequences EG-1 to EG-33 published here as type 4.

The authors are indebted to Dr Theo Cuypers of the Central Laboratory of the Netherlands Red Cross Blood Transfusion Service, to Drs S. Leong and C. Lai and staff of the Hong Kong Red Cross Blood Transfusion service, and to Dr A. M. Al-Rasheed, Debbie Rankin, Gillian Olewicz and staff at the Riyadh Armed Forces Hospital for collection and dispatch of samples analysed in this paper. Fiona McOmish was supported by a grant from the Scottish National Blood Transfusion Service and Shui-Wan Chan was supported by a project grant from the Medical Research Council, grant number G9020615CA.

References

ALLAIN, J., DAILEY, S. H., LAURIAN, Y., VALLARI, D. S., RAFOWICZ, A., DESAI, S. M. & DEVARE, S. G. (1991). Evidence for persistent hepatitis C virus (HCV) infection in hemophiliacs. *Journal of Clinical Investigation* **88**, 1672–1679.

ALTER, H. J., PURCELL, R. H., SHIH, J. W., MELPOLDER, J. C., HOUGHTON, M., CHOO, Q. L. & KUO, G. (1989). Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A, non-B hepatitis. *New England Journal of Medicine* **321**, 1494–1500.

BROWN, E. A., ZHANG, H., PING, H.-L. & LEMON, S. M. (1992). Secondary structure of the 5′ nontranslated region of hepatitis C virus and pestivirus genomic RNAs. *Nucleic Acids Research* **20**, 5041–5045.

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, U.S.A.* **89**, 4942–4946.

CHA, T. A., BEALL, E., IRVINE, B., KOLBERG, J., CHIEN, D., KUO, G. & URDEA, M. S. (1992). At least five related, but distinct, hepatitis C viral genotypes exist. *Proceedings of the National Academy of Sciences, U. S. A.* **89**, 7144–7148.

CHAN, S.-W., SIMMONDS, P., MCOMISH, F., YAP, P.-L., MITCHELL, R., DOW, B. & FOLLETT, E. (1991). Serological reactivity of blood donors infected with three different types of hepatitis C virus. *Lancet* **338**, 1391.

CHAN, S.-W., HOLMES, E. C., MCOMISH, F., FOLLETT, E., YAP, P. L. & SIMMONDS, P. (1992*a*). Phylogenetic analysis of a new, highly divergent HCV type (type 3): effect of sequence variability on serological responses to infection. *Hepatitis C Virus and Related Viruses: Molecular Virology and Pathogenesis. Venice.* Abstract D5, 73.

CHAN, S.-W., MCOMISH, F., HOLMES, E. C., DOW, B., PEUTHERER, J. F., FOLLETT, E., YAP, P. L. & SIMMONDS, P. (1992*b*). Analysis of a new hepatitis C virus type and its phylogenetic relationship to existing variants. *Journal of General Virology* **73**, 1131–1141.

CHOO, Q. L., KUO, G., WEINER, A. J., OVERBY, L. R., BRADLEY, D. W. & HOUGHTON, M. (1989). Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome. *Science* **244**, 359–362.

CHOO, Q. L., RICHMAN, K. H., HAN, J. H., BERGER, K., LEE, C., DONG, C., GALLEGOS, C., COIT, D., MEDINA SELBY, R., BARR, P. J., WEINER, A. J., BRADLEY, D. W., KUO, G. & HOUGHTON, M. (1991). Genetic organization and diversity of the hepatitis C virus. *Proceedings of the National Academy of Sciences, U.S.A.* **88**, 2451–2455.

CRAXI, A., FIORENTINO, G., DI MARCO, V., MARINO, L., MAGRIN, S., FABRIANO, C. & PAGLIARO, L. (1991). Second generation tests in diagnosis of chronic hepatitis C. *Lancet* **337**, 1354.

DEVEREUX, J., HAEBERLI, P. & SMITHIES, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

ENOMOTO, N., TAKADA, A., NAKAO, T. & DATE, T. (1990). There are two major types of hepatitis C virus in Japan. *Biochemical and Biophysical Research Communications* **170**, 1021–1025.

FELSENSTEIN, J. (1991). *PHYLIP Manual Version* 3.4. Berkeley: University Herbarium, University of California.

HIGGINS, D. G., BLEASBY, A. J. & FUCHS, R. (1992). Clustal V: improved software for multiple sequence alignments. *CABIOS* **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.

HOUGHTON, M., WEINER, A., HAN, J., KUO, G. & CHOO, Q. L. (1991). Molecular biology of the hepatitis C viruses: implications for diagnosis, development and control of viral disease. *Hepatology* **14**, 381–388.

KATO, N., HIJIKATA, M., OOTSUYAMA, Y., NAKAGAWA, M., OHKOSHI, S., SUGIMURA, T. & SHIMOTOHNO, K. (1990). Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. *Proceedings of the National Academy of Sciences, U.S.A.* **87**, 9524–9528.

KATO, N., OOTSUYAMA, Y., TANAKA, T., NAKAGAWA, M., NAKAZAWA, T., MURAISO, K., OHKOSHI, S., HIJIKATA, M. & SHIMOTOHNO, K. (1992). Marked sequence diversity in the putative envelope proteins of hepatitis C viruses. *Virus Research.* **22**, 107–123.

LAM, J. P. H., MCOMISH, F., BURNS, S. M., YAP, P. L., MOK, J. Y. Q. & SIMMONDS, P. (1993). Infrequent vertical transmission of hepatitis C virus. *Journal of Infectious Diseases* (in press).

LELIE, P. N., CUYPERS, H. T. M., REESINK, H. W., VAN DER POEL, C. L., WINKEL, I., BAKKER, E., EXEL OEHLERS, P. J., VALLARI, D., ALLAIN, J. P. & MIMMS, L. (1992). Patterns of serological markers in transfusion-transmitted hepatitis C virus infection using second generation HCV assays. *Journal of Medical Virology* **37**, 203–209.

MCOMISH, F., CHAN, S.-W., DOW, B. C., GILLON, J., FRAME, W. D., CRAWFORD, R. J., YAP, P. L., FOLLETT, E. A. C. & SIMMONDS, P. (1993). Detection of three types of hepatitis C virus in blood donors: investigation of type-specific differences in serological reactivity and rate of alanine aminotransferase abnormalities. *Transfusion* **33**, 7–14.

MORI, S., KATO, N., YAGYU, A., TANAKA, T., IKEDA, Y., PETCHCLAI, B., CHIEWSILP, P., KURIMURA, T. & SHIMOTOHNO, K. (1992). A new type of hepatitis C virus in patients in Thailand. *Biochemical and Biophysical Research Communications* **183**, 334–342.

OKAMOTO, H., OKADA, 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., KURAI, K., OKADA, S., YAMAMOTO, K., LIZUKA, H., TANAKA, T., FUKUDA, S., TSUDA, F. & MISHIRO, S. (1992*a*). Full-length sequence of a hepatitis C virus genome having poor homology to reported isolates: comparative study of four distinct genotypes. *Virology* **188**, 331–341.

OKAMOTO, H., SUGIYAMA, Y., OKADA, S., KURAI, K., AKAHANE, Y., SUGAI, Y., TANAKA, T., SATO, K., TSUDA, F., MIYAKAWA, Y. & MAYUMI, M. (1992*b*). Typing hepatitis C virus by polymerase chain reaction with type-specific primers: application to clinical surveys and tracing infectious sources. *Journal of General Virology* **73**, 673–679.

RIVERA, V. M., WELSH, J. D. & MAIZEL, J. V. (1988). Comparative sequence analysis of the 5′ noncoding region of the enteroviruses and rhinoviruses. *Virology* **165**, 42–50.

SKINNER, M. A., RACANIELLO, V. R., DUNN, G., COOPER, J., MINOR, P. D. & ALMOND, J. W. (1989). New model for the secondary structure of the 5′ noncoding RNA of poliovirus is supported by biochemical and genetic data that also show that RNA secondary structure is important in neurovirulence. *Journal of Molecular Biology* **207**, 379–392.

TAKAMIZAWA, A., MORI, C., FUKE, I., MANABE, S., MURAKAMI, S., FUJITA, J., ONISHI, E., ANDOH, T., YOSHIDA, I. & OKAYAMA, H. (1991). Structure and organization of the hepatitis C virus genome isolated from human carriers. *Journal of Virology* **65**, 1105–1113.

TSUKIYAMA KOHARA, K., IIZUKA, N., KOHARA, M. & NOMOTO, A. (1992). Internal ribosome entry site within hepatitis C virus RNA. *Journal of Virology* **66**, 1476–1483.

VAN DER POEL, C. L., REESINK, H. W., SCHAASBERG, W., LEENTVAAR KUYPERS, A., BAKKER, E., EXEL OEHLERS, P. J. & LELIE, P. N. (1990). Infectivity of blood seropositive for hepatitis C virus antibodies. *Lancet* **335**, 558–560.

VAN DER POEL, C. L., BRESTERS, D., REESINK, H. W., PLAISIER, A. A. D., SCHAASBERG, W., LEENTVAAR KUYPERS, A., CHOO, Q. L., QUAN, S., POLITO, A., HOUGHTON, M., KUO, G., LELIE, P. N. & CUYPERS, H. T. M. (1992). Early anti-hepatitis C virus response with 2nd generation C200/C22 ELISA. *Vox Sanguinis* **62**, 208–212.

WATSON, H. G., LUDLAM, C. A., REBUS, S., ZHANG, L. Q., PEUTHERER, J. F. & SIMMONDS, P. (1992). Use of several second generation assays to determine the true prevalence of hepatitis C infection in haemophiliacs treated with non-virus inactivated factor VIII and IX concentrates. *British Journal of Haematology* **80**, 514–518.

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 3 November 1992; Accepted 9 December 1992)