Analysis of a new hepatitis C virus type and its phylogenetic relationship to existing variants

S.-W. Chan, 1 F. McOmish, 2 E. C. Holmes, 3 B. Dow, 4 J. F. Peutherer, 1 E. Follett, 4 P. L. Yap 2 and P. Simmonds 1.*

1Department of Medical Microbiology, Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, 2Edinburgh and South East Scotland Blood Transfusion Service, Royal Infirmary of Edinburgh, Lauriston Place, Edinburgh EH3 9HB, 3Division of Biological Sciences, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JN and 4SNBTS Microbiology Reference Unit, Regional Virus Laboratory, Ruchill Hospital, Glasgow G20 9NB, U.K.

Sequences obtained in the 5' non-coding region (5'NCR) of hepatitis C virus (HCV) were obtained from Scottish blood donors and compared with previously published HCV sequences. Phylogenetic analysis revealed the existence of three distinct groups of sequences; two of these corresponded to the recently described HCV types 1 and 2 variants, while viral sequences detected in around a third of the blood donors formed a separate phylogenetic group that probably represents infection with a novel virus species. Nucleotide sequences of this latter group differed from all previously published 5'NCR sequence variants by at least 9%. This new virus type also differed considerably from previously published variants in other regions of the viral genome (core, NS-3 and NS-5), with corrected nucleotide distances of 15, 43 and 49% respectively from the prototype HCV-1 sequence. Formal phylogenetic analysis of each of the coding regions confirmed that HCV type 1 variants could be clearly differentiated into regional variants (Far East and U.S.A./European), in contrast to the clearly overlapping geographical distributions of the main HCV types in U.K. blood donors. We discuss the evidence for and against the hypothesis that the three main phylogenetic groups identified in this study represent separate species of HCV.

Introduction

The aetiological agent responsible for most cases of post-transfusion non-A, non-B (NANB) hepatitis has been cloned and characterized (Choo et al., 1989, 1991). This virus, now termed hepatitis C (HCV), is a positive-strand RNA virus distantly related to the pestiviruses and flaviviruses (Miller & Purcell, 1990; Koonin, 1991). Its genome consists of an approximately 332 nucleotide 5' non-coding region (5'NCR) followed by a continuous single open reading frame encoding a polypeptide of around 3010 amino acids and then a short 3' untranslated region (Kato et al., 1990; Takamizawa et al., 1991; Choo et al., 1991). By analogy with flaviviruses, this polypeptide has been divided into a 5' structural region consisting of putative core and envelope proteins and a 3' region corresponding to non-structural (NS-1 to NS-5) proteins. Recombinant proteins cloned from the prototype virus and synthetic peptides based on the viral sequence have been used to detect HCV antibodies (Kuo et al., 1989; Muraiso et al., 1990; Hosein et al., 1991), and screening of blood donors has been initiated in several countries to prevent post-transfusional NANB hepatitis. However, there remain donations that transmit hepatitis C virus but which are seronegative or 'indeterminate' in commercial serological tests (Esteban et al., 1990; van der Poel et al., 1991; Japanese Red Cross Non-A, Non-B Research Group, 1991). It is possible that some of these false negative serological results may be the result of infection by extreme sequence variants of HCV that elicit an antibody response that has limited or no cross-reactivity with the peptide antigens used in serological assays. Supporting this hypothesis is the recent discovery of HCV variants (Enomoto et al., 1990; Nakao et al., 1991; Okamoto et al., 1991) that differ markedly in sequence from the original prototype HCV (HCV-1; Choo et al., 1991) and others found in Japanese patients (Kato et al., 1990; Takamizawa et al., 1991).

To investigate whether sequence heterogeneity might influence the effectiveness of serological screening for HCV in blood donors, we initiated a study to examine nucleotide sequence diversity of HCV in naturally
infected individuals. Available for the study were anti-HCV positive blood donors, intravenous drug users (IVDUs) and haemophiliacs exposed previously to non-heat-treated factor VIII, and who had biochemical evidence of liver disease (Simmonds et al., 1990b). It is hoped that such studies will also assist in the development of type-specific and type-common antigens for serological diagnosis, to allow the detection and typing of HCV variants by polymerase chain reaction (PCR), and may ultimately assist in vaccine research.

**Methods**

**Samples.** Plasma from 18 different blood donors (E-b1 to E-b18), that were repeatedly reactive on screening by Abbott second generation enzyme immunoassay (EIA), and confirmed or indeterminate by a recombinant immunoblot assay (RIBA, Ortho; Chan et al., 1991) were the principal samples used in this study. In the NS-3 region from five anti-HCV positive IVDUs (abbreviated as i1 to i5 in Simmonds et al., 1990b), five haemophiliacs who had received non-heat-treated clotting concentrate, and who were also anti-HCV-positive (h1 to h5), three pools of 1000 donations collected in 1983 (p1 to p3), and five separate batches of commercially available non-heat-treated factor VIII (f1 to f5) correspond to those described previously (Simmonds et al., 1990b).

**Primers.** The primers used for cDNA synthesis and PCR are listed in Table 1. They were synthesized by Oswel DNA Service, Department of Chemistry, University of Edinburgh, U.K.

**RNA extraction and PCR.** HCV virions in 0.2 to 1.0 ml volumes of plasma were pelleted from plasma by ultracentrifugation at 100000 g for 2 h at 4 °C. RNA was extracted from the pellet as previously described (Chomczynski & Sacchi, 1987; Simmonds et al., 1990b). First strand cDNA was synthesized from 3 μl of RNA sample at 42 °C for 30 min with 7 units of avian myeloblastosis virus reverse transcriptase (Promega) in 20 μl buffer containing 50 mM-Tris-HCl pH 8.0, 5 mM-MgCl₂, 5 mM-DTT, 50 mM-KCl, 0.05 μg/μl BSA, 15% DMSO, 600 μM each of dATP, dCTP, dGTP and dTTP and 0.5 μM of each of the outer nested primers. One μl of the reaction mixture was then transferred to a second tube containing the same medium but with the inner pair of nested primers, and a further 25 heat cycles were carried out with the same programme. The PCR products were subjected to electrophoresis in 3% low melting point agarose gel (IBI) and the fragments were detected by ethidium bromide staining and u.v. illumination. For sequence analysis, single molecules of cDNA were obtained at a suitable limiting dilution at which a Poisson distribution of positive and negative results was obtained (Simmonds et al., 1990a).

**Direct sequencing of PCR products.** The PCR products were purified by glass-milk extraction (GeneClean; Bio101). One-quarter of the purified products was used in sequencing reactions with T7 DNA polymerase (Sequenase; United States Biologicals) performed according to the manufacturer's instructions except that the reactions were carried out in 10% DMSO and the template DNA was heat-denatured before primer annealing.

**Phylogenetic methods.** The sequences were compiled by version 2.0 of the programs of Staden (1984) and analysed by programs available in the University of Wisconsin Genetics Computer Group sequence analysis package, version 7.0 (Devereux et al., 1984). Phylogenetic trees were inferred using two different programs available in the PHYLIP package of Felsenstein (version 3.4, June 1991; Felsenstein, 1988). The program DNAML finds the tree of the highest likelihood (the maximum likelihood tree) given a particular stochastic model of molecular evolution and has been shown to perform well in simulation studies (Saitou & Imanishi, 1989). In the analyses performed here the global (G) option was used as this searches a greater proportion of all possible trees. The second program used was NEIGHBOR which clusters (following the algorithm of Saitou & Nei, 1987) a matrix of nucleotide distances previously estimated using the program DNADIST (which itself was set, using the D option, to use the same stochastic model as underlies DNAML in order to estimate distances corrected for the probabilities of multiple substitution). In all cases the maximum likelihood and neighbour joining procedures produced congruent trees and thus only the former have been presented here.

**Table 1. Sequences and sources of primers used for amplification of HCV genome**

<table>
<thead>
<tr>
<th>Name</th>
<th>Region</th>
<th>Position of 5' base</th>
<th>Polarity</th>
<th>Sequences 5' to 3'</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>209</td>
<td>5'NCR</td>
<td>8</td>
<td>-</td>
<td>ATATCTCGGAGGCAGCTAGCTACGGACGAGACCT</td>
<td>Garson et al. (1990)</td>
</tr>
<tr>
<td>211</td>
<td>5'NCR</td>
<td>-29</td>
<td>-</td>
<td>CACTCTCGAGACCCCTATCGAGGAGT</td>
<td>Garson et al. (1990)</td>
</tr>
<tr>
<td>939</td>
<td>5'NCR</td>
<td>-297</td>
<td>+</td>
<td>CTGTGAAGAACTACGTCTTT</td>
<td>Okamoto et al. (1990a)</td>
</tr>
<tr>
<td>940</td>
<td>5'NCR</td>
<td>-279</td>
<td>+</td>
<td>TTCAGCAGAAGTGCTAG</td>
<td>Okamoto et al. (1990a)</td>
</tr>
<tr>
<td>410</td>
<td>Core</td>
<td>410</td>
<td>-</td>
<td>ATGTACCCCATGAGGTCGCC</td>
<td>Okamoto et al. (1990a)</td>
</tr>
<tr>
<td>406</td>
<td>Core</td>
<td>-21</td>
<td>+</td>
<td>ACGTCTCTAGACCGCTCGAC</td>
<td>Simmonds et al. (1990b)</td>
</tr>
<tr>
<td>288</td>
<td>NS-3</td>
<td>4951</td>
<td>-</td>
<td>CCGGCATGATGTCATGATGAT</td>
<td>Simmonds et al. (1990b)</td>
</tr>
<tr>
<td>290</td>
<td>NS-3</td>
<td>4933</td>
<td>-</td>
<td>GTAATGTTGATGCTTGGCCGTC</td>
<td>Simmonds et al. (1990b)</td>
</tr>
<tr>
<td>207</td>
<td>NS-3</td>
<td>4662</td>
<td>+</td>
<td>TCTTTGAAATTTTGGGAGGGCGTC</td>
<td>Simmonds et al. (1990b)</td>
</tr>
<tr>
<td>207</td>
<td>NS-3</td>
<td>4699</td>
<td>+</td>
<td>CATATAGTGCCACTTCTATC</td>
<td>Simmonds et al. (1990b)</td>
</tr>
<tr>
<td>242</td>
<td>NS-5</td>
<td>8304</td>
<td>-</td>
<td>GCGGGGAATCTCCTGCTAGGCTGCGTG</td>
<td>Enomoto et al. (1990)</td>
</tr>
<tr>
<td>555</td>
<td>NS-5</td>
<td>8227</td>
<td>-</td>
<td>CGCAGACTGATCATCGGCT</td>
<td>Enomoto et al. (1990)</td>
</tr>
<tr>
<td>243</td>
<td>NS-5</td>
<td>7904</td>
<td>+</td>
<td>TGGGATGCTCTGATGATTACCCCGTGCTT</td>
<td>Enomoto et al. (1990)</td>
</tr>
<tr>
<td>554</td>
<td>NS-5</td>
<td>7935</td>
<td>+</td>
<td>CTCAACCGCTGACTGAAAGGACAT</td>
<td>Enomoto et al. (1990)</td>
</tr>
</tbody>
</table>

* Position of 5' base relative to HCV genomic sequence in Choo et al. (1991).
† Orientation of primer sequence (+, sense; −, antisense).
To establish the interrelationships of the major types of HCV, we have separately analysed several regions of the viral genome that differ in sequence variability and evolutionary constraint. Thus the conclusions drawn from the sequence comparisons are not subject to spurious evolutionary phenomena that may affect a particular region. However, one problem with the analysis presented here was the absence of a viral sequence that was sufficiently distantly related to HCV to serve as an out-group. Thus, although we describe the interrelationships of different sequence variants of HCV, it should be stressed that we have no means of deciding which sequence is ancestral to the others. The trees are thus drawn in the less familiar unrooted form to indicate this. All sequences reported in this publication have been submitted to GenBank (accession numbers D10113 to D10134).

Results

Analysis of the 5'NCR

Samples were obtained from 18 blood donors that were repeatedly reactive in the Abbott second generation EIA and which were confirmed or indeterminate in the Chiron 4-RIBA (E-b1 to E-b18; Follett et al., 1991). HCV sequences present in stored plasma samples from each donor were amplified with primers corresponding to sites in the 5'NCR (Garson et al., 1990; Okamoto et al., 1990a) that are well conserved between all known HCV variants (Table 2, sequences 1 to 15, 28). Sequencing of the PCR product, after limiting dilution to isolate single molecules of cDNA before amplification, allowed approximately 190 bp in the centre of the region to be compared with equivalent published sequences (Fig. 1).

Within the sequences, constant as well as variable regions can be found. Six sequences from donors E-b13 to E-b18 closely resembled HCV-1 (sequence no. 1) and the other published sequences 2 to 15 (Table 2), whereas others (E-b9 to E-b12) resembled the recently reported highly divergent K2 and HC-J6 sequences (nos. 24, 27, 28). However, eight sequences (E-b1 to E-b8) appear quite distinct from the others. Division of the sequences into three groups is supported by formal phylogenetic analysis using the maximum likelihood (Fig. 2) and neighbour joining algorithms (data not shown) of the blood donor sequences along with previously published sequences (identified in Table 2). The group labelled 1 contains sequences of HCV with a world-wide distribution (sequences 1 to 15; Table 2), and group 2 contains K2 and J6 sequences (nos. 24, 27, 28). Sequence variability within the three groups is in each case considerably less than that which separates them, and no sequence intermediate between the three groups was found. This tree shows that the third group is equally distinct from group 1 as is group 2. Using the DNAML model, the corrected distances between sequences within each group were in each case less than 3%. Between groups, they ranged from 9% (between groups 1 and 3, and between groups 1 and 2) to 13% (between groups 2 and 3) (Table 3).

Analysis of the NS-5 region

The nucleotide sequence of the NS-5 region has been found to vary significantly between the previously

Table 2. Source and citation of previously published HCV sequences used in this study

<table>
<thead>
<tr>
<th>No.</th>
<th>Type*</th>
<th>Abbreviation</th>
<th>Geographical source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (I)</td>
<td>HCV-1</td>
<td>U.S.A.</td>
<td>Choo et al. (1991)</td>
</tr>
<tr>
<td>2</td>
<td>1 (I)</td>
<td>P1</td>
<td>Japan</td>
<td>Nakao et al. (1991)</td>
</tr>
<tr>
<td>3, 4</td>
<td>1 (I)</td>
<td>H77, H90</td>
<td>U.S.A.</td>
<td>Enomoto et al. (1990)</td>
</tr>
<tr>
<td>5, 6</td>
<td>1 (I)</td>
<td>GM-1, GM-2</td>
<td>Germany</td>
<td>Ogata et al. (1991)</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>J1</td>
<td>Japan</td>
<td>Han et al. (1991)</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>A1</td>
<td>Australia</td>
<td>Han et al. (1991)</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>S1</td>
<td>S. Africa</td>
<td>Han et al. (1991)</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>T1</td>
<td>Taiwan</td>
<td>Han et al. (1991)</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>U18/124</td>
<td>U.S.A./Italy</td>
<td>Han et al. (1991)</td>
</tr>
<tr>
<td>12</td>
<td>1 (II)</td>
<td>HCV-J</td>
<td>Japan</td>
<td>Kato et al. (1990)</td>
</tr>
<tr>
<td>13</td>
<td>1 (II)</td>
<td>HCV-BK</td>
<td>Japan</td>
<td>Takamizawa et al. (1991)</td>
</tr>
<tr>
<td>14, 15</td>
<td>1, 1 (II)</td>
<td>HC-J1, -14</td>
<td>Japan</td>
<td>Okamoto et al. (1990b)</td>
</tr>
<tr>
<td>16-20</td>
<td>1 (II)</td>
<td>K1, K1-1-K1-4</td>
<td>Japan</td>
<td>Enomoto et al. (1990)</td>
</tr>
<tr>
<td>21</td>
<td>1 (II)</td>
<td>JH</td>
<td>Japan</td>
<td>Kubo et al. (1989)</td>
</tr>
<tr>
<td>22</td>
<td>1 (II)</td>
<td>J7</td>
<td>Japan</td>
<td>Takeuchi et al. (1990)</td>
</tr>
<tr>
<td>23</td>
<td>1 (II)</td>
<td>T3</td>
<td>Taiwan</td>
<td>Chen et al. (1991)</td>
</tr>
<tr>
<td>24-27</td>
<td>2 (III)</td>
<td>K2a, K2a-1</td>
<td>Japan</td>
<td>Nakao et al. (1991)</td>
</tr>
<tr>
<td>28</td>
<td>2 (III)</td>
<td>HC-J6</td>
<td>Japan</td>
<td>Okamoto et al. (1991)</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>Clone A</td>
<td>Japan</td>
<td>Tsukiyama-Kohara et al. (1991)</td>
</tr>
</tbody>
</table>

* Designation of sequences follows the classification described by Enomoto et al. (1990). The alternative classification (Houghton et al., 1991) is shown in parentheses (see Discussion).
Fig. 1. Comparison of nucleotide sequences in the 5'NCR from British blood donors (E-b1 to E-b18) with previously published HCV sequences. Dots indicate identity with sequence of E-b1 (top line); nucleotide substitutions are indicated. Nucleotide numbering corresponds to that described for the prototype HCV-1 sequence (Choo et al., 1991). Source and citation of published sequences are shown in Table 2; phylogenetic groups are indicated in the left column.
described K1 and K2 variants of HCV (Enomoto et al., 1990). To investigate whether the previously identified new group of sequences were equally distant from the other two in this region as well as in the 5'NCR, we compared sequences from four blood donors with group 3 sequences (E-b1, E-b2, E-b3 and E-b7) and one in group 2 (E-b12) with previously published sequences (Fig. 3). As NS-5 is a coding region, the inferred amino acid sequences are presented to indicate the degree of phenotypic variation between the different viral sequences. Phylogenetic analysis of the nucleotide sequences is shown in Fig. 4; corrected nucleotide distances between the different groups are shown in Table 3.

A remarkable variation was observed between sequences of the three groups in this region. Again, sequences falling into the third group clustered separately from the others in this region. However, unlike the 5'NCR, there appear to be subdivisions within the other groups. Group 1 sequences are split between those found in Japanese infected individuals (e.g. HCV-J; HCV-BK; sequence numbers 12, 13, 16 to 20 in Table 2) and those of U.S.A. origin (HCV-1, Pt-1, H77, H90; sequence numbers 1 to 4; Fig. 4). There is also some evidence for a split between group 2 sequences, with K2a and HC-J6 (nos. 24, 25 and 28) appearing distinct from type K2b sequences (nos. 26, 27) and the Scottish blood donor, E-b12.

Table 3 shows that the average nucleotide distances between the two clusters of HCV group 1 sequences is 25% [indicated here as type la/I (U.S.A.) and type lb/II (Japanese)], with variation of only 4 to 7% within each. However, this distance is considerably less than those which exist between group 1 and group 2 sequences (52 to 62%), and group 3 (48 to 49%), and the distance between group 2 and group 3 sequences (53 to 61%).

Table 3. Nucleotide distances between the three HCV groups in four regions of the genome

<table>
<thead>
<tr>
<th>Region</th>
<th>Group*</th>
<th>No.†</th>
<th>1a (I)</th>
<th>1b (II)</th>
<th>2a (III)</th>
<th>2b (IV)</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'NCR</td>
<td>1</td>
<td>21</td>
<td>0.0163†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>0.0867</td>
<td>0.0214</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8</td>
<td>0.0948</td>
<td>0.1331</td>
<td>0.0123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core</td>
<td>1a (I)</td>
<td>6</td>
<td>0.0227</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b (II)</td>
<td>5</td>
<td>0.0855</td>
<td>0.0359</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2a (III)</td>
<td>1</td>
<td>0.2226</td>
<td>0.2051</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>0.1511</td>
<td>0.1802</td>
<td>0.2188</td>
<td>ND§</td>
<td>0.0000</td>
</tr>
<tr>
<td>NS-3</td>
<td>1a (I)</td>
<td>34</td>
<td>0.0536</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b (II)</td>
<td>4</td>
<td>0.2310</td>
<td>0.0823</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2a (III)</td>
<td>1</td>
<td>0.3416</td>
<td>0.4072</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2b (IV)</td>
<td></td>
<td>1</td>
<td>0.4337</td>
<td>0.3447</td>
<td>0.2776</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>0.4281</td>
<td>0.3747</td>
<td>0.4885</td>
<td>0.3552</td>
<td>0.0460</td>
</tr>
<tr>
<td>NS-5</td>
<td>1a (I)</td>
<td>4</td>
<td>0.0372</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b (II)</td>
<td>7</td>
<td>0.2478</td>
<td>0.0743</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2a (III)</td>
<td>3</td>
<td>0.6170</td>
<td>0.5920</td>
<td>0.0789</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2b (IV)</td>
<td>3</td>
<td>0.5732</td>
<td>0.5215</td>
<td>0.2328</td>
<td>0.0655</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>0.4890</td>
<td>0.4755</td>
<td>0.6051</td>
<td>0.5300</td>
<td>0.0322</td>
</tr>
</tbody>
</table>

* See footnote for Table 2.
† Number of sequences analysed.
‡ Groups 1 and 2 could not be subdivided in the 5'NCR, so distances are presented in an intermediate column.
§ ND, Not done.
∥ HCV clone A sequence shown as 2b/IV for presentation purposes only (see text).
portion of anti-C-100 positive sera from haemophiliacs, they were less effective with sera from IVDUs

Analysis of the NS-3 region

Amplification reactions were carried out using previously published primer sequences in the NS-3 region (Simmonds et al., 1990). Although these primers amplified HCV sequences from a high proportion of anti-C-100 positive sera from haemophiliacs, they were less effective with sera from IVDUs (Simmonds et al., 1990b) and with blood donor samples (three positive out of 15 tested; data not shown). Two conserved sites in the amplified fragment were identified by sequence analysis of the NS-3 region from the haemophiliac and IVDU patients, and two new primers corresponding to these were specified (207, 208; Table 1). The combination of 288 and 208 (first round) and 290 and 207 (second round) primers successfully amplified samples from four donors infected with HCV variants from group 3 (E-b1, E-b2, E-b6 and E-b7) but none of those infected with group 2 sequences (data not shown). This enabled a comparison of sequence group 3 with our own (Simmonds et al., 1990b) and previously published sequences (Fig. 5, 6; Table 3). For clarity, only seven of the group 1 sequences obtained in this and our previous study [E-b16, E-b17, i3, i4, h5, h3 and h1 (nos. 19 to 23, filled circles)] are shown in the tree. These sequences are representative of the range of variation found in this region in individuals infected in Britain; comparison of the tree previously published (Simmonds et al., 1990b) with Fig. 6 shows that the former forms a very small component of the overall tree obtained once published and group 3 sequences are added.

The maximum likelihood tree again shows three main groups of sequences. As was found in the NS-5 region, sequences in the group 1 5'NCR are split into geographical groups in NS-3. Sequences of Japanese and Taiwanese origin (HCV-J, HCV-BK, JH and T3, nos. 12, 13, 21, 23; Table 2) are distinct from the HCV-1 representative of the range of variation found in this study [E-bl6, E-bl7, il to i5] and group 3 sequences are added. The maximum likelihood tree again shows three main groups of sequences. As was found in the NS-5 region, sequences in the group 1 5'NCR are split into geographical groups in NS-3. Sequences of Japanese and Taiwanese origin (HCV-J, HCV-BK, JH and T3, nos. 12, 13, 21, 23; Table 2) are distinct from the HCV-1 representative of the range of variation found in this study [E-bl6, E-bl7, il to i5] and group 3 sequences are added.

The maximum likelihood tree again shows three main groups of sequences. As was found in the NS-5 region, sequences in the group 1 5'NCR are split into geographical groups in NS-3. Sequences of Japanese and Taiwanese origin (HCV-J, HCV-BK, JH and T3, nos. 12, 13, 21, 23; Table 2) are distinct from the HCV-1 representative of the range of variation found in this study [E-bl6, E-bl7, il to i5] and group 3 sequences are added.

The maximum likelihood tree again shows three main groups of sequences. As was found in the NS-5 region, sequences in the group 1 5'NCR are split into geographical groups in NS-3. Sequences of Japanese and Taiwanese origin (HCV-J, HCV-BK, JH and T3, nos. 12, 13, 21, 23; Table 2) are distinct from the HCV-1 representative of the range of variation found in this study [E-bl6, E-bl7, il to i5] and group 3 sequences are added.
Phylogenetic analysis of HCV

![Diagram]

Fig. 5. Comparison of deduced amino acid sequences in the NS-3 region of blood donors E-bl, E-b2, E-b6, E-b7 (type 3) with those previously published (Table 2). Group 1/1, amino acid sequence of fl, f3, f4, h2, h3, h4 (one), 1/2, i1, i3, i4, p1, p2; group 1/2, amino acid sequence of i5; group 1/3, amino acid sequence of h2, h3, h4 (one), h5, f2, p3, 1/4; group 1/4, amino acid sequence of h1 (one); group 1/5, amino acid sequence of h1 (one). Numbering, symbols and abbreviations are as described in the legend to Fig. 3.

![Diagram]

Fig. 6. Phylogenetic analysis of nucleotide sequences in the NS-3 region using the maximum likelihood algorithm, shown as an unrooted tree. Representative nucleotide sequences of the five groups of type 1 sequences shown in Fig. 5 are coded as follows: O i3; O i4; ~ h5; ~ h3; h1. Symbols are as described in the legend to Fig. 2.

within group 1 (23%) is lower than those which exist between group 1 and the four group 3 sequences (37 to 43%), and between the group 2 sequence, HC-J6, no. 28 (34 to 41%). As reported previously (Simmonds et al., 1990b), the majority of nucleotide substitutions that exist between type 1 sequences are silent (i.e. do not affect the encoded amino acid sequence), but numerous amino acid substitutions exist between sequences in the three main groups (Fig. 5). The analysis of the NS-3 region includes the sequence of clone A, labelled as no. 29 (Tsukiyama-Kohara et al., 1991), which was obtained from Japanese patients with NANB hepatitis, and which was reported to be distinct from existing HCV sequences. In Fig. 6, this sequence appears to cluster with HC-J6 (no. 28; 72% similarity, compared with 57 to 66% and 64% with group 1 and 3 sequences). Although sequence data for clone A are restricted, it is possible that it may be homologous with K2b/E-bl2, as the sequence difference between it and HC-J6 is comparable to that between K2b/E-bl2 and HC-J6 in the NS-5 region (Table 3).

Partial sequence of the putative core region of HCV

The region encoding the putative core protein of HCV is comparatively well conserved in its nucleotide sequence between previously published sequences. Sequences classified in the 5'NCR as group 1 show nucleotide and amino acid sequence similarities of 90 to 98% and 98 to 99% respectively (Fuchs et al., 1991; Ogata et al., 1991), whereas 81% (nucleotide) and 90% (amino acid) simi-
Fig. 7. Comparison of deduced amino acid sequences in the HCV core region of blood donor E-bl (type 3) with those previously published (Table 2). Numbering, symbols and abbreviations are as described in the legend to Fig. 3.

Fig. 8. Phylogenetic analysis of nucleotide sequences in the core region using the maximum likelihood algorithm, shown as an unrooted tree. Symbols are as described in the legend to Fig. 2.

Similarities are reported between HC-J6 (group 2) and HCV-1 (group 1; Okamoto et al., 1991). The core region from the blood donor E-bl (group 3), amplified with primers 410 and 406, was found to be distinct from both group 1 and group 2 (Fig. 7, 8; Table 3). Again there was a prominent subdivision of group 1 sequences into Japanese (HCV-J, HCV-BK, HC-J4, JH and J7; sequences 12, 13, 15, 21, 22) and U.S.A./European (HCV-1, H77, H90, GM1, GM2; nos. 1, 3 to 6, 14) sequences. As was found in NS-3, very little amino acid sequence variation is found between geographically separated variants of group 1 in the core regions; almost all of the nucleotide differences between the two groups are at 'silent' sites. By contrast, HC-J6 shows at least 10 and the E-bl (group 3) sequence shows at least eight amino acid substitutions in comparison with group 1 sequences.
Discussion

A major difficulty associated with the use of the PCR to assess sequence variation is the possibility that mismatches between the primers and the variant sequence will prevent amplification. In this study, we have used several strategies to overcome this problem. For initial virus detection, we used primers in the 5'NCR, which are reported to be highly conserved amongst previously published sequences (sequences 1 to 15, 28; Table 2). Analysis of the blood donor samples revealed the existence of three phylogenetically distinct groups of sequences. Sequences in the third group have not been previously described. Based on this grouping, we sought corroboration of our findings in other (coding) and more variable regions of the viral genome.

Analysis of the NS-5 region, which was based on several sequences of each of the three main variants (Fig. 3, 4; Table 3), revealed that group 3 sequences formed a relatively homogeneous group that was quite distinct from published sequences. Group 1 and 2 sequences that appeared homogeneous in the 5'NCR split into two distinct clusters. The proposed subdivision of K2 sequences (group 2; Enomoto et al., 1990) into K2a and K2b types is also supported by the phylogenetic analysis presented in this paper. We observed that one of the blood donor sequences obtained in this study appears most similar to K2b, and HC-J6 (Okamoto et al., 1991) falls into the K2a group. Differentiation of HCV type 1 sequences into two groups is also clearly shown in the core (Fig. 7) and NS-3 regions (Fig. 5), in both cases with the group 2 and 3 sequences appearing considerably more distant.

This issue of whether the distinct phylogenetic groupings of HCV constitute separate species cannot be clearly resolved at this stage, as the criteria by which HCV variants may be so differentiated have not yet been defined. Circumstantial evidence in favour of the distinct species hypothesis includes the very marked clustering of phylogenetically distinct groups, without intermediate type sequences. Secondly, we and others have found an overlapping distribution in a single geographical area, and indeed all three groups appear capable of mixed infection in multiply exposed individuals, such as haemophiliacs (Enomoto et al., 1990; Nakao et al., 1991; Chan et al., 1991; Pozzato et al., 1991). Such findings are quite contrary to the alternative hypothesis that the different groups represent epidemiologically separated and hence divergent variants of the same species.

Unlike the 5'NCR, where there are only three distinct groups, each of the coding regions shows prominent differentiation of group 1 sequences into two separate clusters. In this case, the sequences in the two subgroups are geographically separated between Far East (Japan, Taiwan) and U.S.A./European variants (Table 2). However, an exception to this is the HC-J1 sequence (no. 14; Okamoto et al., 1990b), which clusters with the U.S.A./European group in the core region (Fig. 8) but was obtained from a Japanese patient with NANB hepatitis. The Pt-1 sequence (no. 2) that also clusters away from the Far East group was obtained from a Japanese haemophiliac treated with imported factor VIII of U.S.A. origin (Enomoto et al., 1990; Nakao et al., 1991). There are insufficient sequence data to indicate whether the two proposed type 2 subtypes, K2a and K2b (Enomoto et al., 1990; Nakao et al., 1991), also represent geographically distinct variants.

There have been several attempts to classify HCV sequence variants into different types. Enomoto et al. (1990) classified sequences corresponding to the phylogenetic groups labelled 1 and 2 in this paper as K1/PT and K2, the first being subdivided along the geographical lines discussed previously, and K2 comprising the K2a and K2b subgroups. Houghton et al. (1991) classified sequences as type 1 (sequences of U.S.A./European origin in our phylogenetic group 1), type II (Far East group 1 sequences) and type III (sequences in our phylogenetic group 2), presumably with the option of extending the classification to type IV to differentiate the two clusters of group 2 sequences. However, this classification is uneven in that the difference between type I and II sequences in each protein coding region is consistently less than that between I and III and between I and the new group 3 sequences reported here (Table 3). It is also particularly noticeable that far fewer amino acid substitutions exist between type I and II sequences than between other variants, and type I and II sequences cannot be differentiated in the 5'NCR. For these reasons, we would favour classifying our new sequences in group 3 as type 3, although if the alternative system is accepted, we propose that they be provisionally termed type V.

The genomic organization of HCV corresponds to that of flaviviruses and pestiviruses, with a single open reading frame encoding a polyprotein that is subsequently cleaved into structural and non-structural proteins. Although the overall degree of sequence dissimilarity between the three main groups cannot be measured by comparison of the small regions of sequence analysed in this study, a rough estimate of the extent of divergence in protein coding regions is given by an examination of the divergence of the partial core sequences obtained here. This shows that the difference between HCV group 1, 2 and 3 core regions (approximately 10% amino acid sequence divergence) is comparable to that which exists between different serotypes of the flavivirus, tick-borne encephalitis virus (14%; Mandl
et al., 1988), but lower than that which is found between serotypes of a mosquito-borne flavivirus, dengue fever virus (67%), and the West Nile virus (WNV) subgroup (60%). The 5′NCR sequences of the different members of the WNV subgroup are also considerably more diverse (< 50% similarity) than those of the three major types of HCV, although within each of the members, e.g. Murray Valley encephalitis virus, the 5′NCR is extremely well conserved (> 95% similarity; Coelen & Mackenzie, 1990). On the basis of these analogies, we speculate that the major types of HCV could conceivably represent distinct serotypes, each capable of human infection irrespective of the immune response mounted against other HCV types.

The existence of different hepatitis C viral types opens up the possibility that the distinct clinical disease syndromes associated with HCV infection may reflect underlying differences in the pathogenicity of the different types of the virus. There is some evidence that infection with HCV group 2 variants (type III) leads to more severe disease than group 1, and is less susceptible to interferon treatment (Pozzato et al., 1991). There are, as yet, no data to link virus type with different sources of infection (particularly non-parenteral infection). Our own preliminary investigations have shown that infection with HCV type 3 is more strongly associated with previous intravenous drug misuse than types 1 and 2 (unpublished data).

The degree of sequence variability found between HCV types would be expected to affect profoundly the antigenicity of many of the putative proteins of HCV. We have previously shown that sera from blood donors infected with different HCV types show distinct differences in the pattern of reactivity to a range of structural and non-structural proteins in two commercial immuno-blot assays for HCV antibody (Ortho RIBA and Innogenetics LIA; Chan et al., 1991). In particular, no reactivity with C-100 and infrequent reactivity with C33c were observed in patients infected with HCV types 2 and 3, presumably reflecting the high degree of sequence variability in the NS-3 and NS-4 regions of the genome. Reactivity, however, was always found with the core protein, which is consistent with the degree of sequence conservation in this region (Table 3).

This then provides at least one explanation for the observation that blood donor screening with the original C-100-based immunoassay reduced the incidence but did not entirely prevent post-transfusional NANB hepatitis (Esteban et al., 1990; Japanese Red Cross Non-A, Non-B Hepatitis Research Group, 1991). The use of second generation tests that incorporate core proteins will undoubtedly increase the effectiveness of blood donor screening, although the most effective test for HCV infection would be an assay incorporating representative antigens from all three HCV types. Indeed, an immunoblot assay that included polypeptides corresponding to the C-100 protein of HCV types 1, 2 and 3 may serve to type infected individuals serologically, by virtue of the apparent type-specific serological reactivity to this variable protein.

The authors are indebted to staff at the Glasgow and West Scotland Blood Transfusion Service, the Edinburgh and South East Scotland Blood Transfusion Service for supplying blood donor samples, Dr H. Hart of the Scottish Protein Fractionation Centre for plasma pools, Dr G. Kembal-Cook, National Institute for Biological Standards and Control for non-heat-treated commercial factor VIII and Dr C. A. Ludlam, Department of Haematology, Royal Infirmary of Edinburgh for haemophiliac blood samples used for sequence analysis. This work was supported by the Medical Research Council (grant number G9020615CA) and by the Scottish National Blood Transfusion Service.

References


S' untranslated region and poly(A) tails at the 3' end. Proceedings of the National Academy of Sciences, U.S.A. 88, 1711–1715.


(Received 5 December 1991; Accepted 31 January 1992)