Identification of genotypes of hepatitis C virus by sequence comparisons in the core, E1 and NS-5 regions

P. Simmonds,1* D. B. Smith,1 F. McOmish,2 P. L. Yap,2 J. Kolberg,3 M. S. Urdea3 and E. C. Holmes4

1 Department of Medical Microbiology, 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 Chiron Corporation, 4560 Horton Street, Emeryville, California 94608, U.S.A. and 4 Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3RE, U.K.

Isolates of hepatitis C virus (HCV) show considerable nucleotide sequence variability throughout the genome. Comparisons of complete genome sequences have been used as the basis of classification of HCV into a number of genotypes that show 67 to 77% sequence similarity. In order to investigate whether sequence relationships between genotypes are equivalent in different regions of the genome, we have carried out formal sequence analysis of variants in the 5' non-coding region (5'NCR) and in the genes encoding the core protein, an envelope protein (E1) and a non-structural protein (NS-5). In the E1 region, variants grouped into a series of six major genotypes and a series of subtypes that could be matched to the phylogenetic groupings previously observed for the NS-5 region. Furthermore, core and E1 sequences showed three non-overlapping ranges of sequence similarity corresponding to those between different genotypes, subtypes and isolates previously described in NS-5. Each major genotype could also be reliably identified by sequence comparisons of part of NS-5 in the well conserved 5'NCR, although many subtypes, such as 1a/1b, 2a/2c and some of those of type 4, could not be reliably distinguished from each other in this region. These data indicate that subgenomic regions such as E1 and NS-5 contain sufficient phylogenetic information for the identification of each of the 11 or 12 known types and subtypes of HCV. No evidence was found for variants of HCV that had sequences of one genotype in the 5'NCR but of a different one in the E1 or NS-5 region. This suggests that recombination between different HCV types is rare or non-existent and does not currently pose a problem in the use of subgenomic regions in classification.

Introduction

Since the cloning and characterization of hepatitis C virus (HCV) as the main causative agent of post-transfusion non-A, non-B hepatitis (Choo et al., 1989), considerable sequence diversity has been found between variants infecting different individuals. For example, the complete genome sequence of the prototype virus (HCV-1; Choo et al., 1991) isolated in the U.S.A. shows only 69 to 85% sequence similarity with variants obtained from infected individuals in Japan (Kato et al., 1990; Takamizawa et al., 1991; Okamoto et al., 1991, 1992; Chen et al., 1992). Variants can be grouped into a number of different virus types by sequence comparisons of subgenomic regions of the virus such as the regions encoding the non-structural protein NS-5 (Enomoto et al., 1990; Mori et al., 1992; Chan et al., 1992; Cha et al., 1992; Simmonds et al., 1993a; Chayama et al., 1993), the core protein (Chan et al., 1992; Cha et al., 1992; Simmonds et al., 1993b), the envelope protein E1 (Cha et al., 1992; Stuyver et al., 1993b; Bukh et al., 1993), NS-3 (Tsukiyama Kohara et al., 1991; Chan et al., 1992) and the 5' non-coding region (5'NCR; Chan et al., 1992; Cha et al., 1992; Bukh et al., 1992; Stuyver et al., 1993a). Several different classification schemes have been proposed (Enomoto et al., 1990; Houghton et al., 1991; Okamoto et al., 1992; Chan et al., 1992; Mori et al., 1992; Cha et al., 1992; Simmonds et al., 1993a), and discussions have recently taken place over the most useful way to identify and assign genotypes of HCV. In particular it is not clear currently whether comparison of subgenomic regions is sufficient, as recently proposed on the basis of sequence comparisons of part of NS-5 (Simmonds et al., 1993a), or whether it is necessary to obtain complete genomic sequences before definitive classification may be carried out (Okamoto et al., 1992; Bukh et al., 1993).

The nucleotide sequence data reported in this paper have been submitted to GenBank and assigned accession numbers L29459 to L29467.
We have previously obtained evidence for six major genotypes of HCV in a worldwide collection of 76 samples from infected individuals by sequence comparisons in a 222 base fragment of the NS-5 gene (Simmonds et al., 1993a). Many of these HCV types comprise a number of more closely related subtypes, leading to a total of 11 genetically distinct virus populations. To address the issue of whether sequence comparisons in this region accurately model interrelationships between variants over the complete genome, we compared the phylogenetic tree produced from this fragment of NS-5 with that of the complete genomic sequence of types 1a, 1b, 2a and 2b. The evolutionary relationships between the complete sequence closely matched those obtained from analysis of NS-5, confirming that, for at least these genotypes, this region is representative (Holmes et al., 1994). We have also carried out similar sequence comparisons for several different coding regions of a limited number of viruses of types 1a, 1b, 2a, 2b, 2c, 3a and 4a, and again found equivalent phylogenetic relationships in each region studied (Chan et al., 1992; Cha et al., 1992).

Recently, a large amount of comparative sequence data for the E1 region of HCV variants was published, and on the basis of amino acid similarity between the 51 sequences, the existence of possibly as many as 12 genotypes was inferred (Bukh et al., 1993). We describe here a comparison of the phylogenetic relationships between nucleotide sequences in the E1 region with those which exist between the previously published NS-5 sequences. This provides a more stringent test of the validity of our proposed classification of HCV, as it allows a direct comparison of the evolutionary relationships between most of the genotypes in two widely spaced regions of the virus genome. Although the two sets of data include sequences from different infected individuals, we have been able to link them in two different ways. Firstly, the corresponding fragments of published complete genomic sequences were included in our analyses of E1 and NS-5, allowing several phylogenetic groupings in the two regions to be matched. Secondly, as the sequences in the 5'NCR of most of the variants published for E1 have been previously described (Bukh et al., 1992) we were able to match them to the corresponding 5'NCR sequences of the variants that we have previously classified in NS-5. In these ways, we have been able to establish equivalences between most of the genotypes for the two regions.

**Methods**

**Nucleotide sequences**

(i) NS-5. Sequences correspond to those obtained in our previous analysis of this region and include 35 sequences published elsewhere.

(ii) E1. Sequence data between positions 574 and 1149 (Choo et al., 1991) were derived from two sources. The first source was fragments of complete genomic sequences from several previous studies [HCV-1, Choo et al., 1991 (1); HCV-H, Inchauspe et al., 1991 (6); HCV-I, Kato et al., 1990 (17); HCV-BK, Takamizawa et al., 1991 (18); T, Chen et al., 1992 (19); HPCGENOM, Bi et al., GenBank accession number L02836 (26); HPCJTA and HPCJTB, Tanaka et al., 1992 (27 and 28); HCVJKIG, Honda et al., GenBank accession number X61596 (31); HC-J6, Okamoto et al., 1991 (36); HC-J8, Okamoto et al., 1992 (49)]. The numbers in parentheses refer to those allocated to each sequence by Simmonds et al. (1993a). The second source was previously published E1 sequences from 51 infected individuals (Bukh et al., 1993). The latter sequences have been assigned numbers 1 to 51 as described in Table 1. Partial sequences from E1 (nucleotide positions 987 to 1022) were also obtained from four of the samples that were previously sequenced in the NS-5 region (numbers 52, 54, 56 and 57; Cha et al., 1992).

(iii) Core. Sequence comparisons were made between positions 15 and 384 in the core gene for the 13 sequences listed in a previous analysis (Chan et al., 1992) and the corresponding fragments of a further seven complete genomic sequences (T, HPCGENOM, HPCJTA and HPCJTB, HCVJKIG and HC-J8).

(iv) 5'NCR. Sequences in the 5'NCR between positions –245 and –72 were obtained for 35 samples for which sequence data in the NS-5 region are available (Simmonds et al., 1993a), and for 35 samples where the E1 sequence was determined (Bukh et al., 1992). These sequences were compared with the corresponding fragments of the 11 previously published complete genomic sequences listed above. Restriction fragment length polymorphisms (RFLPs) between different 5'NCR sequences were calculated for previously described combinations of restriction enzymes as described in the legend to Fig. 1.

Table 1. Identification of samples sequenced in E1

<table>
<thead>
<tr>
<th>Isolate</th>
<th>No.</th>
<th>Isolate</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e</td>
<td>DK7</td>
<td>27c</td>
<td>US10</td>
</tr>
<tr>
<td>2e</td>
<td>U11</td>
<td>28e</td>
<td>T9</td>
</tr>
<tr>
<td>3e</td>
<td>DR4</td>
<td>29e</td>
<td>T2</td>
</tr>
<tr>
<td>4e</td>
<td>DR1</td>
<td>30e</td>
<td>DK11</td>
</tr>
<tr>
<td>5e</td>
<td>DK9</td>
<td>31c</td>
<td>SW3</td>
</tr>
<tr>
<td>6e</td>
<td>SW1</td>
<td>32e</td>
<td>DK8</td>
</tr>
<tr>
<td>7e</td>
<td>S14</td>
<td>33e</td>
<td>T8</td>
</tr>
<tr>
<td>8e</td>
<td>S18</td>
<td>34e</td>
<td>S38</td>
</tr>
<tr>
<td>9e</td>
<td>ND8</td>
<td>35e</td>
<td>DK12</td>
</tr>
<tr>
<td>10e</td>
<td>ND5</td>
<td>36e</td>
<td>HK10</td>
</tr>
<tr>
<td>11e</td>
<td>SW2</td>
<td>37e</td>
<td>S2</td>
</tr>
<tr>
<td>12e</td>
<td>HK3</td>
<td>38e</td>
<td>S54</td>
</tr>
<tr>
<td>13e</td>
<td>HK8</td>
<td>39e</td>
<td>S52</td>
</tr>
<tr>
<td>14e</td>
<td>S45</td>
<td>40e</td>
<td>Z4</td>
</tr>
<tr>
<td>15e</td>
<td>D3</td>
<td>41e</td>
<td>Z1</td>
</tr>
<tr>
<td>16e</td>
<td>T3</td>
<td>42e</td>
<td>Z6</td>
</tr>
<tr>
<td>17e</td>
<td>HK5</td>
<td>43e</td>
<td>Z7</td>
</tr>
<tr>
<td>18e</td>
<td>HK4</td>
<td>44e</td>
<td>DK13</td>
</tr>
<tr>
<td>19e</td>
<td>US6</td>
<td>45e</td>
<td>S55</td>
</tr>
<tr>
<td>20e</td>
<td>P10</td>
<td>46e</td>
<td>SA7</td>
</tr>
<tr>
<td>21e</td>
<td>SA10</td>
<td>47e</td>
<td>SA4</td>
</tr>
<tr>
<td>22e</td>
<td>T10</td>
<td>48e</td>
<td>SA1</td>
</tr>
<tr>
<td>23e</td>
<td>DK1</td>
<td>49e</td>
<td>SA6</td>
</tr>
<tr>
<td>24e</td>
<td>S9</td>
<td>50e</td>
<td>SA13</td>
</tr>
<tr>
<td>25e</td>
<td>D1</td>
<td>51e</td>
<td>HK2</td>
</tr>
<tr>
<td>26e</td>
<td>T4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Patterns of bands were coded as previously described (McOmish et al., 1993b), with the addition of a new pattern (E) for the Hinfl–MvaI digestion of sequence number 45 (no cuts with either enzyme; one band of 251 base pairs).

Nucleotide sequence comparisons. Distances between pairs of sequences were estimated using the DNADIST program of the PHYLIP package (version 3.5c) kindly provided by Dr J. Felsenstein (Felsenstein, 1993) using a model that allows for different rates of transition and transversion and different frequencies of the four bases. Phylogenetic trees were constructed using the neighbour-joining algorithm on the previous set of pairwise distances (Saitou & Nei, 1987) using the PHYLIP program, NEIGHBOR. Statistical analysis of the frequencies of sequence distances within four regions of the HCV genome (core, E1, NS-4 and NS-5) was carried out with in-house software.

Results

Comparison of sequences in the 5′ NCR

Although the 5′NCR is the most conserved region in the HCV genome, it is possible to identify many of the different genotypes from sequence comparisons in this region. Indeed, 5′NCR sequences amplified from plasma of infected individuals have been used in several different typing assays based upon RFLP analysis (Nakao et al., 1991; McOmish et al., 1993a, b) or hybridization with type-specific primers (Stuyver et al., 1993a).

To match genotypes identified in the NS-5 region with those in E1, the corresponding sequences in the 5′NCR from each were compared (Fig. 1). The sequences may be readily assigned into at least seven groups of identical or nearly identical sequences (A to G), which show a series of conserved differences from members of other groups. Groups differed from each other by between seven (between groups A and F) and 25 substitutions (groups C and D). In many cases within a group there were perfect matches between the 5′NCR sequences of variants previously sequenced in NS-5 (Simmonds et al., 1993a) with those in E1 (Bukh et al., 1993), although several of the 5′NCR groups contained sequences that differed by one to three nucleotide substitutions from the group consensus.

Sequences from Zaire (variants 40e to 43e) were heterogeneous and in some comparisons showed as many as four differences from each other and from DK13 (44e) and EG-7, EG-13 and EG-19 (variants 69 to 71). For the purposes of the analysis described below, Z6, Z7 and DK13 (42e to 44e) were grouped together with EG-7 and EG-13 (69 and 70), as they differ from each other at only one nucleotide position and with EG-19 (71), which differs at two (5′NCR group E). The other Zairean sequences (40e and 41e) differ from the others by four or five substitutions and were therefore left unassigned.

Phylogenetic analysis of E1

Published nucleotide sequences from E1 and the corresponding regions of 11 complete genomic sequences were compared using neighbour-joining and the resulting phylogenetic tree was plotted in unrooted form (Fig. 2a). A total of 12 clearly defined phylogenetic clusters were obtained, in agreement with a previous analysis of amino acid sequence similarity (Bukh et al., 1993). Overall, the topology and branching order of the E1 phylogenetic tree is remarkably similar to that of NS-5 (Fig. 2b). In particular, the phylogenetic tree of E1 variants splits into six major branches, some of which are further divided into more closely related sequences. Overall, the phylogenetic tree shows the same hierarchical division into types and subtypes that was previously observed for NS-5.

Comparison of genotype identification in E1 and NS-5

The groupings for E1 were matched with those for NS-5 in two ways. First, the complete genomic sequences that have been previously designated types 1a (1 and 6), type 1b (variants 17, 18, 19, 26, 27, 28 and 31), type 2a (36) and type 2b (variant 49) were identified in the two phylogenetic trees. These sequences were found in four separate groups in both trees. In both cases the sequences in the groups containing the type 1a and 1b variants were all more closely related to each other than they were to types 2a and 2b, showing that a hierarchy of types and subtypes also occurs in the E1 region.

The second way to match E1 and NS-5 groupings was to compare the distribution of 5′NCR sequences between the E1 and NS-5 trees (Fig. 2). Variants with group A 5′NCR sequences include those previously assigned as types 1a and 1b in the NS-5 region (numbers 1 to 31; Fig. 2b) and types 1a and 1b in the E1 region (numbers 1e to 25e; Fig. 2a). In this case, two phylogenetic groupings that are distinct in the coding region of the virus genome may contain viruses with similar or identical sequences in the 5′NCR. For example, sample 3e (type 1a) has a 5′NCR sequence identical to those of the type 1b variants 15, 10e, 11e, 14e, 15e, 17e, 19e, 20e, 22e and 24e (Fig. 1).

Group B 5′NCR sequences include the five previously designated type 2a variants (35 to 39) and the three type 2c variants (numbers 50 to 53; Fig. 1; Simmonds et al., 1993a). Four variants that were sequenced in E1 also contained similar sequences in the 5′NCR (26e to 28e and 34e). The complete genome sequence of HC-J6 (number 36) clusters with variants 26e to 28e in the E1 region identifying these variants as type 2a. Similarly, group C 5′NCR matches variants 40 to 48 (type 2b in NS-5) with 30e to 33e in E1 (Fig. 2), and this is confirmed by the clustering of the complete genome sequence of
Fig. 1. Comparison of sequences in the 5' NCR of HCV variants previously sequenced in the NS-5 region (column 3; Simmonds et al., 1993a), previously sequenced in E1 (column 4; Bukh et al., 1993) and variants for which the complete genomic sequence has been determined (column 2; Simmonds et al., 1993a). Sequence groups A to G (column 1) are individually boxed. Only variable sites are shown and the numbers refer to nucleotide positions -245 to -72, numbered as in Choo et al. (1991). * Indicates inserted bases present.
HC-J8 (type 2b; number 49) in these two phylogenetic groups.

With the information presented so far, it is not clear whether variants assigned to the third subtype of genotype 2 (type 2c; numbers 50 to 53) that all contain group B sequences in the 5'NCR are the same genotype as the third subtype of type 2 found upon phylogenetic analysis of the E1 region (34e). To investigate this, we have compared the sequence of a fragment of the E1 sequence of one of the type 2c variants (52) between nucleotide positions 987 and 1022 (Cha et al., 1992) with the equivalent region of the published type 2 sequences in E1 (Fig. 3). The nucleotide sequence of 34e differed from type 2c variant 52 at only two nucleotide positions, compared with at least eight differences when compared to other subtypes of type 2 and 15 to 17 differences upon comparison with type 1, indicating that 34e and 52 correspond to the same subtype.

Variants assigned as type 3a upon analysis of NS-5 all contain sequences in the 5'NCR corresponding to group D, a group that also contains E1 variants 35e to 39e (Fig. 1). To confirm that these two groups are equivalent, we have compared fragments of the E1 sequence of the type 3a variants 54, 56 and 57 with corresponding regions of the E1 sequences 35e to 39e (Fig. 3). Sequences within these two groups differed from each other by only one to five nucleotide substitutions compared with more than 15 differences with other genotypes. This is similar to the difference observed within subtypes of other genotypes, providing evidence that they belong to the same subtype of type 3. Phylogenetic analysis of NS-5 sequences reveals a second subtype of type 3 (3b) whose 5'NCR sequence is unknown. This subtype is probably not represented in the E1 dataset because the type 3 E1 sequences are monophyletic (Fig. 2a) and because it has only thus far been found in Thailand; this geographical region was not included in the E1 dataset.

The 5'NCR sequence of Z6, Z7 and DK13 (42e to 44e) was similar to the sequences of variants 69 to 71 (group E) allowing these two sets to be collectively identified as type 4. In this case, two separate subtypes of type 4 in E1 are identical in sequence in the 5'NCR (Z6, Z7 and DK13). Because of this, it is not possible to decide whether one of the two subtypes in E1 corresponds to the single subtype found previously in NS-5 (69, 70 and 71). The two other Zairean variants, Z4 and Z1 (numbered 40e and 41e), appear as two further subtypes of type 4 upon phylogenetic analysis of the E1 region, and each showed distinct sequences in the 5'NCR.

The final assignments can be made between the type 5a sequences in the NS-5 region (numbers 72 to 75) and sequences of 45e to 50e in E1 (group F in the 5'NCR), and between sequence 76 and sequence 51e (both in group G in the 5'NCR). In both cases the origins of these two genotypes were the same, type 5a from South Africa patients and type 6a from Hong Kong.

In summary, we have been able to match all six of the major genotypes and five of the nine subtypes in the two datasets. The E1 dataset does not contain sequences corresponding to types 1c or 3b and at least three subtypes of type 4 are not represented amongst the NS-5 sequences. Although in a few cases it was not possible definitively to match individual subtypes in the two regions, none of the sequence data in any of the three regions conflict with any other.

**Distribution of type and subtype sequence distances**

Pairwise comparison between each E1 nucleotide sequence (64 sequences, 2016 comparisons) produces three discrete distributions of evolutionary distance (Fig. 4), similar to that found for NS-5 (Simmonds et al., 1993a). Distances in the range 0.43 to 0.85 correspond to those between the major genotypes. This range of distances differs from those found between the subtypes (0.26 to 0.45), which in turn is separated from the range of variation found upon comparison of individual variants within a subtype (0.009 to 0.13). As with NS-5, we observed no overlap between these ranges, apart from a slightly closer than expected distance of 0.43 between the type 4 DK13 sequence and the type 1a sequence DR9 that lies just within the range of sequence distances between subtypes.

**Provisional assignment of new types and subtypes**

Given the restriction in the range of sequence diversity that exists between type and subtype in NS-5, we have previously suggested that the percentage sequence similarity with existing NS-5 sequences might be used provisionally to assign new genotypes and subtypes (Simmonds et al., 1993a). We chose NS-5 because this region is variable and could be readily amplified from plasma of infected individuals, and because there was already a considerable quantity of published sequence data in this region. The above analysis suggests that percentage similarity with the E1 sequences available from GenBank (Bukh et al., 1993) can also be used for

---

in some 5'NCR sequences between nucleotides -145 and -144 (two bases) and between -139 and -138 (one base). The right-hand three columns indicate the predicted RFLP patterns with three combinations of restriction enzymes, HaeIII/RsaI (5), HinfI/SacFI (6) and HinfI/MvaI (7), according to the previously described code (McOmish et al., 1993b).
Fig. 2. (a) Phylogenetic analysis of 64 sequences in the E1 region. Numbered sequences correspond to the E1 sequences of previously published complete genomes; 1, HCV-1; 6, HCV-H; 17, HCV-J; 18, HCV-BK; 19, T; 26, HPCGENOM; 27, HPCJTA; 28, HPCJTB; 31, HCVJKIG; 36, HC-J6; 49, HC-J8; see Methods for references. Genotypes that have been positively identified in E1 have been labelled (see text). Genotypes are labelled as previously proposed (Simmonds et al., 1993a). Small solid dots denote sequences for which 5'NCR sequence was not determined.

Fig. 3. Comparison of partial E1 nucleotide sequences (positions 987 to 1022; Bukh et al., 1993) with those of variants previously sequenced in NS-5 (column 2; Simmonds et al., 1993a) and assigned as type 2c and 3a. Differences from the prototype HCV-1 sequence are shown. Representative sequences from types la, lb, 2a and 2b are included for comparative purposes; column 1, sequence derived from complete genome sequences; column 3, sequence derived from previous E1 data.

Fig. 4. Distribution of evolutionary differences on pairwise comparison of the 64 E1 sequences (2016 comparisons). Number of calculated evolutionary distance measurements (in increments of 0.02) from 0.00 to 0.86 shown on y-axis.
Table 2. Sequence similarity between type, subtype and isolate in core, E1 and NS-5 regions

<table>
<thead>
<tr>
<th></th>
<th>Core</th>
<th>E1</th>
<th>NS-5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Start</strong></td>
<td>15</td>
<td>574</td>
<td>7975</td>
</tr>
<tr>
<td><strong>End</strong></td>
<td>384</td>
<td>1149</td>
<td>8196</td>
</tr>
<tr>
<td><strong>Length</strong></td>
<td>370</td>
<td>576</td>
<td>222</td>
</tr>
<tr>
<td><strong>No. seqs†</strong></td>
<td>20</td>
<td>64</td>
<td>76</td>
</tr>
<tr>
<td><strong>Isolate Min‡</strong></td>
<td>97·3 (OKA4core/J491)</td>
<td>99·1 (S54/S52)</td>
<td>99·5 (J483/J491)</td>
</tr>
<tr>
<td><strong>Max‡</strong></td>
<td>93·8 (JTA/B/GENANTI)</td>
<td>88·0 (T2/T9)</td>
<td>87·8 (I10/T983)</td>
</tr>
<tr>
<td><strong>Subtype Min</strong></td>
<td>92·8 (HCV-J/HCV-H)</td>
<td>78·6 (DK13/Z6)</td>
<td>86·0 (2TY4/GM2)</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>87·8 (HC-J6/HC-J8)</td>
<td>67·5 (S83/DK8)</td>
<td>74·8 (K1-3/2TY4)</td>
</tr>
<tr>
<td><strong>Type Min</strong></td>
<td>87·8 (E-b1/OKA1core)</td>
<td>69·4 (DK13/DR9)</td>
<td>72·1 (T10/K1)</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>81·2 (HC-J8/HPC-JTA/B)</td>
<td>53·5 (SA6/T8)</td>
<td>56·2 (EG-7/K2a)</td>
</tr>
</tbody>
</table>

* Sequences numbered as in Choo et al. (1991).
† Number of sequences compared.
‡ Minimum and maximum percentage sequence similarities. Individual sequence comparisons at the extremes of the range are indicated.

the identification of most of the known genotypes and the assignment of new virus types.

We have also carried out equivalent analyses of other regions of the genome such as the core region for which there are published sequence data for several genotypes (Chan et al., 1992; Simmonds et al., 1993b). The upper and lower limits of the ranges of sequence divergence that exist between type, subtype and isolate in each genomic region are compared in Table 2. The core region is the most conserved, with intertypic sequence similarities between major genotypes ranging from 81 to 88% and between subtypes from 88 to 93%. In the E1 region, the range of sequence similarity between types is similar to that of NS-5 (54 to 69% compared with 56 to 72%), whereas subtypes are less closely related to each other in E1 (range 68 to 79% compared with 75 to 86% for NS-5). Using these percentage similarities it would be possible to assign the correct genotype for all 96 of the sequences in the core and NS-5 regions, and all but one of the 64 sequences in the E1 region (DK13; see above). However, the identification of DK13 as a subtype of type 4 is apparent from the phylogenetic tree for E1 (Fig. 2), emphasizing the value of this type of analysis to confirm genotype identifications.

**Discussion**

The phylogenetic trees and the range of distances found upon pairwise comparisons of nucleotide sequences were equivalent for the E1 and NS-5 regions. Both regions showed a hierarchical division of sequences into six major groupings (designated HCV types in the NS-5 region), with further division into more closely related subtypes (Enomoto et al., 1990; Chan et al., 1992; Simmonds et al., 1993a). Furthermore, all variants that were assigned a particular genotype in one region of the virus genome always contained sequences corresponding to the same genotype elsewhere in the genome. These include the 13 complete genome sequences, and the 70 variants described in this study for which either E1 and 5'NCR or NS-5 and 5'NCR sequences were available. Together, these data suggest that relatively short subgenomic regions of the virus contain sufficient sequence information to allow reliable identification of HCV genotypes, irrespective of whether the sequence encodes structural or non-structural proteins, and irrespective of the degree of sequence variability within the coding region.

A potential objection to the use of subgenomic regions for virus identification is that recombination between different virus genotypes infecting the same cell might produce hybrid viruses comprising sequences corresponding to more than one genotype. We have found no evidence for recombination between genotypes in this study of 13 complete genome sequences and 70 HCV variants that were simultaneously sequenced in the E1 or NS-5 region and 5'NCR. This indicates that recombination is rare in vivo or that hybrids produced between different genotypes are not viable. In either case, recombination does not currently pose a problem for virus classification. However, until more HCV variants have been characterized, we have previously recommended that new genotypes should be assigned only after sequence comparisons are made with existing variants in at least two separate coding regions of the genome (Simmonds et al., 1993a).

The sequence comparisons described above between the E1 and NS-5 regions were based upon previous observations of an association between sequences in the 5'NCR and genotype. Although all six major HCV types are clearly distinct, this region is so highly conserved that, in many cases, variants corresponding to different subtypes contain indistinguishable sequences (Fig. 1). This does not invalidate the proposed type/subtype discrimination.
nomenclature of HCV, but serves to emphasize the difference in sequence diversity between type and subtype.

Although 5'NCR sequences of types 2a and 2b are consistently distinct from each other, the absence of corresponding differences between other subtypes (for example types 1a and 1b; types 2a and 2c; type 4 Z6 or Z7 and DK13) appears to preclude any typing assay based on 5'NCR sequences for identification of all HCV types and subtypes. RFLP analysis of this region can only differentiate the six major genotypes and type 2b from 2a and 2c (Fig. 1; McOmish et al., 1993b). Similar constraint applies to assays based upon hybridization of genotype-specific probes to amplified 5'NCR sequences (Stuyver et al., 1993a).

Simple methods for culture of HCV in vitro have not been reported, and animal models of infection are expensive and impractical for the majority of laboratories. There have been no attempts to classify variants of HCV by conventional serotyping reagents, a method which has been used for other genera in the Flaviviridae family, such as dengue virus, and within the Picornaviridae (poliovirus and coxsackie virus). However, because virus RNA may be readily amplified from the plasma of infected individuals by PCR and then sequenced, it has proved relatively easy to identify different genotypes of HCV on the basis of nucleotide sequence comparisons, without any information on the biological significance of such classification schemes. It is not currently known whether the major genotypes or subtypes would correspond to the serotypes of other RNA virus groups. Little is known about whether infection with different genotypes is associated with differences in the course and severity of liver or extrahepatic disease, although several studies have documented a poorer response to treatment with interferon of patients infected with type 1b virus (Kanai et al., 1992; Yoshioka et al., 1992; Takada et al., 1992).

To avoid confusion in the assignment of HCV genotypes in the future, it will be necessary for different laboratories to use a standardized nomenclature and methodology for assigning new genotypes. The system described in this and a previous paper (Simmonds et al., 1993a) has found wide acceptance, and is strengthened by the finding that the same type/subtype relationships between variants in NS-5 upon which the proposals were based are also found in E1. Furthermore, with the exception of the 5'NCR and certain hypervariable regions of the E2 gene, we believe that any region of the genome can be used as the basis for virus genotypic identification, provided sequences from each of the currently recognized genotypes are available. Sequences in the core region are particularly suitable for the purpose as this region is relatively easy to amplify from the plasma of infected individuals and sequences of all six genotypes and several subtypes are published or about to become available. One of the major tasks of any committee set up to co-ordinate the assignment of new genotypes is the collection and dissemination of representative sequences from the currently recognized genotypes in as many regions of the genome as possible.

The authors are grateful to Dr E. Follett and B. C. Dow of the Scottish National Blood Transfusion Service for providing samples for sequence analysis, and for technical assistance provided by N. Linskail. This work was funded in part by the MRC (grant number G9026615CA) and by the Scottish National Blood Transfusion Service.

References


(Received 15 November 1993; Accepted 21 December 1993)