Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: implications for virus classification

J. Mellor, E. C. Holmes, L. M. Jarvis, P. L. Yap, P. Simmonds* and The International HCV Collaborative Study Group†

Genotypes of hepatitis C virus (HCV) present within 104 samples from HCV-infected individuals from Africa, the Middle East, the Indian subcontinent and South-East Asia were identified by sequence comparisons in the core and NS-5 regions. Relatively short sequences (such as the 222 bp fragment of NS-5) provided effective discrimination of types, subtypes and isolates, and produced equivalent relationships between genotypes as were found upon comparison of longer sequences of NS-5, of the core region, and by comparison of the limited number of complete genomic sequences currently available. Measurement of evolutionary distances in the core and NS-5 regions allowed 79 of the 104 samples to be identified as examples of known genotypes, while 17 of the remainder could be provisionally classified as new subtypes of types 1 (Nigeria), 2 (Gambia), 3 (India, Pakistan and Bangladesh) and 4 (Middle East) on the basis of sequence comparison in core and NS-5 (n = 9) or provisionally using core alone (n = 8). The remaining sequences from Thailand made up two groups showing no close similarity to any of the six major genotypes classified to date, although one corresponded to an as yet unclassified variant of HCV also found in Thailand. However, phylogenetic analysis of the core and NS-5 regions indicated a distant relationship between these sequences with variants found in Vietnam and with type 6a, and collectively they formed a diverse single phylogenetic group. The existence of great diversity within a single genotype was also found amongst type 3 sequences in the Indian subcontinent, amongst type 4 variants in Central Africa and the Middle East, and amongst type variants in Nigeria. These findings may provide clues for understanding the origins and mechanisms of transmission of HCV.

Introduction

Hepatitis C virus (HCV) is a positive-stranded RNA virus that causes persistent infection in humans, and is the principal cause of post-transfusion non-A, non-B hepatitis in many countries (Choo et al., 1990; Kuo et al., 1989). In common with other RNA viruses, HCV shows high levels of sequence variability, but the absence of simple methods for virus culture in vitro has precluded its classification into serotypes using neutralization assays. However, comparisons of nucleotide sequence data have allowed classification into genetically distinct groups, termed genotypes, that differ from each other by approximately 30% over the entire genome, and which may correspond to the serotypes of other RNA viruses. So far, a total of six major genotypes of HCV have been described (Simmonds et al., 1993a; Bukh et al., 1993), and these can be divided further into a series of more
closely related subtypes, that differ from each other by around 20\% (Enomoto et al., 1990; Chan et al., 1992; Simmonds et al., 1993a; Bukh et al., 1993; Stuyver et al., 1993).

Considerable attention is now focused on the possibility that biological differences exist between different HCV genotypes. There is some evidence that infection with different virus genotypes is associated with differences in the course of disease, and the likelihood of developing hepatocellular carcinoma. It is more clearly established that individuals infected with type 1 viruses are much less likely to respond to treatment with interferon than for other genotypes such as types 2 and 3 (Dusheiko et al., 1994). These differences have clinical implications for patient treatment selection, and in understanding the pathogenesis of HCV infection.

The ongoing international research effort in documenting HCV variability and its clinical implications requires an agreed framework by which variants are classified, and on the criteria by which novel genotypes may be assigned. Furthermore, it is necessary to develop a procedure by which novel genotypes may be named. The lack of such criteria has led to the present situation where the same labels have been used to describe different HCV genotypes. Examples include two variants described as type 1c (Okamoto et al., 1994; Simmonds et al., 1993a), two 4a (Simmonds et al., 1993b; Bukh et al., 1993); two each of types 4e and 4f (Bukh et al., 1994; Stuyver et al., 1994) and two type 3c (Tokita et al., 1994b; Greene et al., 1995). Alternatively, one genotype has been described by two different names: for example, a new subtype of type 1 has been described as both type 1c (Okamoto et al., 1994) and type 1d (Hotta et al., 1994a, b).

Another related problem is the use of different methods for classification of HCV variants. Complete genomic sequences have been obtained for the most common genotypes encountered clinically in Western and Far Eastern countries (Types 1a, 1b, 2a, 2b, 3a and 3b; see Methods for sources), and phylogenetic analysis reveals a clear distinction between virus type, subtype and isolate (Okamoto et al., 1992b; Sakamoto et al., 1994). However, classification of more geographically restricted HCV genotypes and subtypes has so far been dependent upon comparisons of subgenomic regions. Sequence comparisons of different regions of the genome such as core, E1, NS-3, NS-4 and NS-5 produce equivalent relationships between HCV variants which are supported by phylogenetic analysis (Chan et al., 1992; Simmonds et al., 1994b; Stuyver et al., 1994; Tokita et al., 1994a, b). However, whether this is always true, and whether certain methodologies or regions are more informative than others has not been entirely resolved. For example, it is known that different subtypes of HCV (such as types 1a and 1b) may show identical sequences in the 5' non-coding region (5'NCR); conversely, parts of the envelope gene (such as the 5' end of E2) may be too variable (Weiner et al., 1991).

In this study, we have addressed these questions by carrying out statistical analysis of sequence variability between previously recognized HCV genotypes in various parts of the genome, using fragments of different length. Criteria derived from this analysis were then used to classify sequences obtained from a series of novel variants found in HCV-infected individuals from Africa, Europe and South-East Asia within the existing framework of HCV types and subtypes. The considerable sequence diversity found within most major genotypes in certain distinct geographical regions provides evidence of their endemic origin, and a different epidemiology for the limited variability of types and subtypes found elsewhere.

Methods

Samples. Plasma or serum samples used in this study were referred from blood donors, patients with chronic hepatitis and haemophiliacs from the Middle East (Bahrain, Lebanon and Saudi Arabia) (n = 13), Kuwait (n = 4), Yemen (n = 5), South Africa (n = 9), Nigeria (n = 3), Gambia (n = 2), Egypt (n = 16), India (n = 2), Pakistan (n = 3), Bangladesh (n = 5), Thailand (n = 32) and Hong Kong (n = 4).

Three unusual samples collected from blood donors in the UK are included for comparative purposes. One had previously received a blood transfusion in South Africa and was suspected to carry type 5 virus based on restriction fragment length polymorphism (RFLP) analysis of sequences amplified from the 5'NCR (Davidson et al., 1995). The second donor was investigated because of discrepant results between the RFLP assay (type 3) and a serotyping assay (type 1) based upon the detection of antibody to NS-4 peptides (Table 4 in Simmonds et al., 1993b). A third sample (EUUK43) was obtained from an individual who may have been originally infected in Bangladesh. Samples were also analysed from patients from Italy and Greece with chronic HCV with type 4 infection (Dusheiko et al., 1994). None of these six samples are representative of their countries of origin.

RNA extraction and RT–PCR amplification. Virus RNA was extracted directly from 100 \mu l of serum using proteinase K/Sarkosyl, followed by phenol–chloroform extraction and precipitation in ethanol (Cuypers et al., 1992). RNA was reverse transcribed using outer antisense primers specific for the core (primer 410, 5' ATGTACC-CATGAGGTCGGC 3', position -410) and NS-5 (primer 1204, 5' GGAGGGGGGAATTGCATAGCGCTCGGTGAA 3', position -54) were used, followed by inner primers 951 (5' CAC/GAGTTACTCGATGAC 3', position -8250), and 122 (5' CTCAACCGTATGAGCAGAGC 3', position -8383). Amplification of the core and NS-5 regions was achieved by nested PCR using specific primers. For the core, the primers were 410 (above) and 954 (5' ACTGCGCTGATGAGCTGGGAGC 3', position -54) were used, followed by inner primers 951 (5' CAC/TGTA/GAGGGTATCGATGAC 3', -383) and 953 (5' AGGTCCT-GTAGACCGTGCATCG 3', -21) respectively. For NS-5, the primers were 1204 and 1203 (5' ATGGGGGTTCCTGATGACCC-GCTGCTGTTGACTC 3', +7903) followed by various combinations of inner primers: 123 (5' GCTGCTGCTGGTCGCGCTC 3', -8250), and 122 (5' CTCAACCGTATGAGCAGAGC 3', 554 (5' CTCAACCGTATGAGCAGAGC 3', 518 (5' CTCAACCGTATGAGCAGAGC 3', 517 (5' CTCAACCGTATGAGCAGAGC 3', all at position +7935 (nucleotide positions as in Kato et al., 1990). For some samples, supplementing the PCR buffer...
with 10% dimethylsulphoxide (DMSO) improved the specificity and yield of the amplification reaction. For each set of primers, reaction tubes were cycled 30 times through 95 °C for 0.6 min, 45 °C for 1.5 min and 68 °C for 3 min.

Direct sequencing of PCR products. Positive samples were re-amplified from primary product with one of the inner primers followed by alkaline treatment to release the unlabelled strand into the amplified from primary product with one of the inner primers Sequenase sequencing kit (United States Biochemical) according to the manufacturer's instructions, except that reactions contained 10% DMSO and the template DNA was heat-denatured before primer annealing.

Sequence comparisons. Sets of pairwise evolutionary distances for core and NS-5 nucleotide sequences were generated using the program DNADIST in the PHYLIP package (Felsenstein, 1993). Evolutionary distances were corrected for multiple substitution using a model which allows different rates of transition and transversion and different frequencies of the four nucleotides. Pairwise distances between complete genome sequences were calculated arithmetically. Sets of sequence distances were analysed with standard statistical software (SYSTAT) using files derived from DNADIST distance tables.

Phylogenetic analysis was carried out using a maximum likelihood method (program fastDNAML; Olsen et al., 1994) (see Fig. 5). Trees are shown unrooted with all branch lengths drawn to scale. The precision of the groupings was assessed using the bootstrap resampling procedure. Because of the limits on computation imposed by the maximum likelihood algorithm, 1000 replicate neighbour-joining trees were used in the bootstrap analysis (programs SEQBOOT, DNADIST, NEIGHBOR and CONSENSE in the PHYLIP package: Felsenstein, 1993). Because of the very large number of sequences available in the 222 bp NS-5 region, especially from genotype 1b, it was decided to use a subset of these in the phylogenetic analysis. A smaller dataset will also consider that the names la, lb, 2a, 2b, 3a-3f, 4b-4d, 5a and 6a are ambiguous. For other variants, we have added an initial in parentheses after the proposed name to indicate the source of the unambiguous. For other variants, we have added an initial in parentheses after the major genotype number while new major genotypes additional to the original scheme have ambiguous names because the same genotype has two different names, or more frequently, because different genotypes have the same name. In this paper, we have considered that the names 1a, 1b, 2a, 2b, 3a-3f, 4b-4d, 5a and 6a are unambiguous. For other variants, we have added an initial in parentheses after the proposed name to indicate the source of the sequence. Thus type 1a(0) refers to the subtype of type 1 found in Indonesia (Okamoto et al., 1994; Hotta et al., 1994a). Similarly, type 4a(E) from Egypt (Simmonds et al., 1993a) is distinct from type 4a(B) from Zaire (Bukh et al., 1993), while types 4e(B) and 4f(B) (Bukh et al., 1994) are distinct from 4e(S) and 4f(S) found in Gabon (Stuyver et al., 1994). In order to prevent further confusion we have given novel subtypes described in this study a temporary label of Roman numbers in parentheses after the major genotype number while new major genotypes are labelled NG. Novel variants of HCV from Thailand (Apichartpiyakul et al., 1994) were not labelled in the original publication, and are referred to here as H1 (BB9, D86/93, PE and PB) and H2 (B4/92, PC).
Results

Analysis of sequence variability in the core and NS-5 regions

In order to investigate the distribution of evolutionary distances between types, subtypes and isolates of HCV, we have compiled a reference dataset of 173 NS-5 sequences of length 222 bp and 93 core region sequences of 308 bp (see Methods). Each of these sequences has been classified into previously described genotypes [lα-lα(O), lα(E), 2a-2c, 3a-3f, 4a(E), 5a and 6a].

For the 222 bp fragment of NS-5, pairwise comparisons between sequences of different major genotypes produced a distribution with a median of 0.532 and in which 95% of the individual values fall within the range 0.418-0.655 (Fig. 1F). This distribution is approximately normal since the median was close to the mean (0.536) and the mean ±2 standard deviations (0.414-0.659) expected to contain 94.6% of the values was similar to the 95th percentile values. The distribution of values for both intra-subtype and intra-isolate pairwise comparisons also corresponded closely to normal distributions, with medians and 95th percentiles of 0.235 (0.191-0.300) and 0.052 (0.23-0.106) respectively (Fig. 1E, D). These results provide a basis for classifying new variants by comparison of percentage sequence identity values to sequences of known genotype. However, small overlaps in range occur between the type/subtype and subtype/isolate distributions, and the identity of variants with sequence distances falling within these regions would be ambiguous. For example, the smallest pairwise distance between subtypes was 0.149, lower than the largest pairwise distance between sequences considered to be of the same genotype (0:151). A similar overlap of values is found between subtype and type (0.349-0.317).

Similar comparisons of nucleotide sequences comprising 308 bp of the core region also produce three approximately normal distributions, with median/mean values of 0.192/0.188 for pairwise comparisons between major genotypes, 0.099/0.100 between subtypes and 0.033/0.033 between isolates (Fig. 1A-C). Although this region is less variable than other coding parts of the genome, most pairwise comparisons between virus subtypes and isolates are distinct, with no overlap in the pairwise values found between subtypes with those between major genotypes (Fig. 1B, C). The lowest intertypic distance is 0.084, while the highest distance between subtypes is 0.153, and there is also an overlap of the 95th percentile ranges. A considerable proportion of pairwise values would therefore be inconclusive without further comparisons in more variable regions of the genome, such as E1, NS-4 or NS-5.

Choice of region for HCV classification

A subject of current controversy is the length of sequence required to accurately identify HCV genotypes. Although genotypes 1-6 and many subtypes can be distinguished by phylogenetic analysis of a 222 bp fragment of NS-5 (Simmonds et al., 1993a), or of E1 (Bukh et al., 1993), other authors have suggested that precise classification requires the analysis of longer subgenomic fragments (Stuyver et al., 1994) or of complete genomic sequences (Tokita et al., 1994a). However, comparison of the discriminating power of different regions is confounded by the different datasets employed by different authors, in particular the number of sequences being compared.

In order to address the issue systematically, we calculated the range and median values of isolate, subtype and type distributions of evolutionary distances for collections of sequences over successively smaller genomic regions (Fig. 2). As previously described (Tokita et al., 1994a), the range of percentage sequence similarity values upon comparison of 16 complete genomic sequences was highly restricted; pairwise similarity values range from 65-7-68.8 for different genotypes (1, 2 and 3), from 76-9-80-11 between subtypes [lα/lβ/lα(O) and 2a/2b], and from 90-8-99% within subtypes (lα and lβ).

Surprisingly, analysis of subgenomic sequences of the same 16 HCV variants produced similarly restricted distributions of sequence similarity (Fig. 2A). Thus the range of intertypic, subtype and isolate variation for 222 bp of NS-5 are little different from the values obtained from comparison of complete genomic sequences. A similar analysis of evolutionary distances for different fragments of NS-5 was carried out for 43 HCV variants for which 1093 bp of sequence in NS-5 was available (Fig. 2B) (Tokita et al., 1994a). Little difference between the 1093, 329 or 222 bp fragments was observed in the ranges of pairwise values for types, subtypes and isolates, although these were greater than the values for the 16 genomic sequences. Differences in ranges of values were not related to the length of sequence being compared since the range over which 95% of pairwise values was found within a distribution (95th percentile) was similar irrespective of the region compared (Fig. 2B). Indeed, the distribution of distances for the 329 bp fragment was slightly more restricted than for the 1093 bp fragment. The small overlap in ranges
using the shortest sequence (222 bp) resulted from a single pairwise value between subtypes overlapping with two values between major genotypes.

Finally, it was possible to compare a large number of sequences using the 222 bp fragment of NS-5 (n = 205; 173 sequences from the original reference dataset plus 32 NS-5 sequences from Tokita et al., 1994a; 20910 pairwise comparisons) (Fig. 2C). Percentile ranges were unchanged or more restricted than for comparisons between 43 sequences and the increase in the range of values for 16, 43 and 205 sequences corresponds closely to that predicted for normally distributed data.

In conclusion, the major determinant of the range of values of sequence similarity or evolutionary distance between the three proposed categories (type, subtype and isolate) is the number of sequences being compared, rather than the length of the fragment. The analysis also indicates that variability within subgenomic regions of NS-5 is generally representative of the virus as a whole and this may equally apply to other subgenomic regions such as E1 or NS-4. Finally, it predicts that sequencing larger pieces of the genome may not necessarily resolve problems of HCV classification based upon sequence similarity; an overlap between the ranges of type and subtype values might also have been found if it had been possible to obtain the corresponding complete genome sequences for the 205 variants analysed in NS-5 (Fig. 2C).

Fig. 1. Distribution of pairwise evolutionary distances for core (A–C) and NS-5 (D–F) nucleotide sequences, separately analysed for major genotype (C, F), subtype (B, E) and isolate (A, D). Dotted lines indicate the range for each distribution containing 95% of all sequence comparisons (95th percentile). Arrows indicate the total range and median value. Abbreviations: No., number of values (pairwise comparisons) within distribution; Max, maximum value; 95th, upper 95th percentile value; Med, median value; 95th, lower 95th percentile value; Min, minimum value; Mn, mean value; SD, value of 1 standard deviation.
Fig. 2. Comparison of range, upper and lower percentile and median values for type, subtype and isolate in different regions of the HCV genome. (A) Comparison of the range and median of percentage sequence divergence (uncorrected values) of 16 complete genome sequences (Tokita et al., 1994a) with those of subgenomic fragments of NS-5 derived from the same variants. (B) Evolutionary distances between 43 HCV variants using NS-5 sequences of length 1093 bp (Tokita et al., 1994a). (C) Evolutionary distances between 205 HCV variants using NS-5 sequences of length 222 bp. One division on the scale corresponds to an evolutionary distance of 0.1 (B, C) or an uncorrected distance of 10% (A); distances are scaled on the median distance between major genotypes for each of the subgenomic fragments. Shading indicates ranges of pairwise distances outside the 95th percentile (B, C only).

Table 1. Distribution of previously classified HCV genotypes in surveyed areas

<table>
<thead>
<tr>
<th>County</th>
<th>Genotype</th>
<th>Number*</th>
<th>Core Mean evolutionary distance†</th>
<th>NS-5 Mean evolutionary distance†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh</td>
<td>3b</td>
<td>3(C), 1(N)</td>
<td>0.047 0.089</td>
<td></td>
</tr>
<tr>
<td>Egypt</td>
<td>4a</td>
<td>14</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>6a</td>
<td>4</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>3b</td>
<td>1</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>Italy/Greece</td>
<td>4a</td>
<td>2</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>Kuwait</td>
<td>4a</td>
<td>4</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>Middle East</td>
<td>4a</td>
<td>12</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td>3a</td>
<td>1</td>
<td>0.031 0.151</td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td>4c (Bukh)</td>
<td>1</td>
<td>0.021 NA</td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>1a</td>
<td>2</td>
<td>0.018 0.048</td>
<td></td>
</tr>
<tr>
<td>Yemen</td>
<td>4a</td>
<td>3</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4d (Bukh)</td>
<td>2</td>
<td>0.031 NA</td>
<td></td>
</tr>
</tbody>
</table>

* Number of samples within survey population; C, N, number of sequences obtained in core and NS-5 regions respectively.
† Mean pairwise distance between survey samples with previously classified sequences of indicated HCV genotype.
‡ Sequence of indicated genotype not available from the literature.

Classification of new HCV variants

HCV infected serum samples used in this study were referred from blood donors, patients with chronic hepatitis and haemophiliacs from the Middle East (Egypt, Bahrain, Lebanon, Saudi Arabia, Kuwait and Yemen), Africa (South Africa, Nigeria and Gambia), the Indian subcontinent (India, Pakistan and Bangladesh) and Thailand. HCV sequences were amplified from extracted RNA using primers for core and NS-5 regions, and sequenced directly (Simmonds et al., 1993a).

Sequences were obtained in the core region for 104 samples and compared with previously published sequences of genotypes 1a, 1b, 1c(O), 1c(E), 2a, 2b, 2c, 3a–3f, 4a(E), 4a(B)–4f(B), 5a and 6a (listed in Methods). A total of 79 could be classified as examples of previously described HCV genotypes (Table 1). Genotypes 3a and 3b were found in South-East Asia, type 4a(E) in the Middle East and type 5a in South Africa. Examples of types 4c(B)–4f(B) previously found in Zaire (Bukh et al., 1994) and Gabon (Stuyver et al., 1994) were also found in Italy, South Africa and Yemen.

Twenty-five samples contained virus with sequences in the core region with no clear similarity to known HCV genotypes and this was also true when comparisons were
Table 2. Provisional classification of HCV variants with no close sequence similarity to existing HCV genotypes

<table>
<thead>
<tr>
<th>Country</th>
<th>Variant</th>
<th>n*</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangladesh</td>
<td>EUB21</td>
<td>1</td>
<td>0.177</td>
<td>0.257</td>
<td>0.144</td>
<td>0.222</td>
<td>0.219</td>
<td>0.237</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(I)</td>
</tr>
<tr>
<td></td>
<td>EUB8</td>
<td>1</td>
<td>0.180</td>
<td>0.252</td>
<td>0.132</td>
<td>0.238</td>
<td>0.226</td>
<td>0.220</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(II)</td>
</tr>
<tr>
<td>Egypt</td>
<td>EUEG15</td>
<td>1</td>
<td>0.148</td>
<td>0.243</td>
<td>0.218</td>
<td>0.099</td>
<td>0.167</td>
<td>0.177</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4(I)</td>
</tr>
<tr>
<td></td>
<td>EUEG28</td>
<td>1</td>
<td>0.145</td>
<td>0.236</td>
<td>0.235</td>
<td>0.109</td>
<td>0.183</td>
<td>0.171</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4(II)</td>
</tr>
<tr>
<td>Gambia</td>
<td>EUGAM29</td>
<td>2</td>
<td>0.228</td>
<td>0.111</td>
<td>0.256</td>
<td>0.229</td>
<td>0.166</td>
<td>0.226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2(I)</td>
</tr>
<tr>
<td></td>
<td>EUGAM28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.513</td>
<td>0.303</td>
<td>0.624</td>
<td>0.583</td>
<td>0.527</td>
<td>0.566</td>
<td>2(I)</td>
</tr>
<tr>
<td>India</td>
<td>EUIND7</td>
<td>1</td>
<td>0.184</td>
<td>0.236</td>
<td>0.137</td>
<td>0.200</td>
<td>0.200</td>
<td>0.181</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(III)</td>
</tr>
<tr>
<td>Middle East</td>
<td>EUAH3485</td>
<td>1</td>
<td>0.144</td>
<td>0.223</td>
<td>0.215</td>
<td>0.101</td>
<td>0.167</td>
<td>0.187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4(I)</td>
</tr>
<tr>
<td>Nigeria</td>
<td>EUNIG13</td>
<td>1</td>
<td>0.109</td>
<td>0.209</td>
<td>0.193</td>
<td>0.160</td>
<td>0.173</td>
<td>0.199</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1(I)</td>
</tr>
<tr>
<td></td>
<td>EUNIG14</td>
<td>1</td>
<td>0.085</td>
<td>0.185</td>
<td>0.167</td>
<td>0.142</td>
<td>0.175</td>
<td>0.151</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1(II)</td>
</tr>
<tr>
<td></td>
<td>EUNIG45</td>
<td>1</td>
<td>0.117</td>
<td>0.222</td>
<td>0.196</td>
<td>0.170</td>
<td>0.179</td>
<td>0.196</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1(III)</td>
</tr>
<tr>
<td>Pakistan</td>
<td>EUPK6</td>
<td>2</td>
<td>0.220</td>
<td>0.270</td>
<td>0.158</td>
<td>0.231</td>
<td>0.253</td>
<td>0.239</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(IV)</td>
</tr>
<tr>
<td>Thailand</td>
<td>EUH46</td>
<td>1</td>
<td>0.173</td>
<td>0.239</td>
<td>0.128</td>
<td>0.216</td>
<td>0.211</td>
<td>0.224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(V)</td>
</tr>
<tr>
<td></td>
<td>EUH19</td>
<td>3</td>
<td>0.208</td>
<td>0.237</td>
<td>0.229</td>
<td>0.229</td>
<td>0.203</td>
<td>0.124</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(VI)</td>
</tr>
<tr>
<td></td>
<td>EUH49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.422</td>
<td>0.582</td>
<td>0.270</td>
<td>0.429</td>
<td>0.478</td>
<td>0.600</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>EUH86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.523</td>
<td>0.643</td>
<td>0.550</td>
<td>0.601</td>
<td>0.531</td>
<td>0.453</td>
<td>NG(I)</td>
</tr>
<tr>
<td></td>
<td>EUH88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.480</td>
<td>0.732</td>
<td>0.577</td>
<td>0.619</td>
<td>0.518</td>
<td>0.421</td>
<td>NG(II)</td>
</tr>
<tr>
<td></td>
<td>EUH86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.480</td>
<td>0.732</td>
<td>0.577</td>
<td>0.619</td>
<td>0.518</td>
<td>0.421</td>
<td>NG(II)</td>
</tr>
<tr>
<td></td>
<td>EUH86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.480</td>
<td>0.732</td>
<td>0.577</td>
<td>0.619</td>
<td>0.518</td>
<td>0.421</td>
<td>NG(II)</td>
</tr>
<tr>
<td></td>
<td>EUKUK2060</td>
<td>1</td>
<td>0.175</td>
<td>0.258</td>
<td>0.130</td>
<td>0.202</td>
<td>0.208</td>
<td>0.213</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(V)</td>
</tr>
<tr>
<td></td>
<td>EUUK43</td>
<td>1</td>
<td>0.191</td>
<td>0.248</td>
<td>0.157</td>
<td>0.226</td>
<td>0.226</td>
<td>0.207</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3(VI)</td>
</tr>
<tr>
<td></td>
<td>EUUK3</td>
<td>1</td>
<td>0.166</td>
<td>0.245</td>
<td>0.222</td>
<td>0.183</td>
<td>0.051</td>
<td>0.210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5(I)</td>
</tr>
</tbody>
</table>

* Number of sequences corresponding to the new genotype.
† Mean pairwise distance between variant with previously published sequences classified as HCV genotypes 1–6; mean weighted to represent each subtype equally, irrespective of number of sequences within each (see text).

Extended to a fragment of NS-5 (Table 2). Individual comparisons between new HCV variants with those previously classified sequences were plotted to investigate the relationship between novel variants of HCV with those previously classified (Fig. 3). For example, the plot of individual pairwise distances in the NS-5 region between a variant found in Nigeria (EUNIG13) and representative examples of genotypes 1–6 shows a tight scattering of values ranging from 0.184–0.26 with type 1 sequences, and between 0.42 and 0.70 against other genotypes (Fig. 3A). Similarly, for the core region and despite the overlap in values between type and subtype in the reference dataset (Fig. 1B, C), EUNIG13 also appears as a new subtype of type 1 (Fig. 3A). Other examples are a novel type 2 subtype from the Gambia (Fig. 3B; EUGAM28), and sequences from blood donors in Thailand with no clear similarity in any of the six major genotypes (Fig. 3C, D) and provisionally described as NG(I) and NG(II) (see Methods). In each case, individual evolutionary distances between new sequences and those of known genotype were distributed over relatively narrow ranges that generally fell within the 95th percentile limits previously defined (Fig. 1).

For each region, the mean pairwise distance with each of the six major genotypes of HCV was calculated (Table 2). Where a genotype contains several subtypes, equal weight was given to each, avoiding a bias towards common subtypes such as type 1a at the expense of rarer ones [e.g. type 1c(E)]. In general, sequence relationships between sequences obtained in this study with those previously classified were equivalent in the core and NS-5 regions. The only exceptions were EUH46, which is distinct from type 3b in the core region (evolutionary distance 0.084), but similar in NS-5 (0.099), and EUUK3 which is similar to type 5a in the core region (0.051; below the subtype limit of 0.072), but in the middle of the subtype range for NS-5 (0.228 compared with a median of 0.235).

We have attached provisional labels to novel sequences obtained in this study [e.g. type 3(I), 3(II) etc.] until formal procedures are available to name them without conflicting with other publications. Summarizing Table 2, we describe new subtypes labelled as 1(I), 1(II) and 1(III) from Nigeria, a new subtype of type 2 from Gambia [2(I)], subtypes 3(I)-3(VI) from India, Bangladesh and Pakistan, two new subtypes of type 4
Fig. 3. Distribution of individual pairwise distances (y-axis) in core and NS-5 between (A) EUNIG13; (B) EUGAM28; (C) EUTH86; (D) EUTH36 and sequences previously classified as types 1–6 (x-axis). Dotted lines show upper and lower 95th percentile limits for type, solid lines for subtype; arrows indicate the range for each; shading indicates the area of overlap between the upper 95th percentile for subtype and the lower 95th percentile for type calculated for sequences in the core region. Arrows indicate total range and the median is marked.
Fig. 4. Phylogenetic analysis of core (A–D) and NS-5 (E–H) nucleotide sequences for genotypes 1 (A, E), 2 (B, F), 3 (C, G) and 4 (D, H), using a type 2a (HC-J6) or 1a (HCV-PT) sequence as an outgroup. ●, New genotypes listed in Table 2; □, new variants of HCV available from GenBank (see Methods for accession numbers).
(Egypt and Bahrain), and a new subtype of type 5 from a UK blood donor, and two new genotypes from Thailand (NG1, II) unrelated to genotypes 1–6 of unknown source of infection.

The six new subtypes of type 3 were highly divergent when sequenced in the core region, with mean values between them and previously classified type 3 subtypes ranging from 0·132–0·158, in many cases greater than the 95th percentile value for subtypes of the standard dataset (0·139; Fig. 1). However, type 3 sequences previously classified as types 3a–3f (Tokita et al., 1994b) are more diverse compared with subtypes of other genotypes, in the NS-5 and core regions analysed here, and also in longer fragments of NS-5 and E1. Four of these six subtypes fall outside the 95th percentile values for the 222, 329 and 1093 bp fragments of NS-5, and for the 576 bp E1 series. The implications of this for HCV classification are discussed below.

**Phylogenetic analysis of core and NS-5 sequences**

Phylogenetic analysis of nucleotide sequences provides an alternative approach to HCV classification, but one that has produced equivalent results to date. In order to confirm our provisional assessments based on evolutionary distances, we have performed phylogenetic analysis of variants within genotypes 1–4, using type 2a as an outgroup (or type 1a for type 2 sequences; Fig. 4A–D, core; Fig. 4E–H, NS-5; Table 2). In both the core and NS-5 regions, proposed new subtypes formed distinct branches consistent with their assignment as separate subtypes. The new subtypes of type 1 [1a(I–III)] appear equally distinct from each other as they do from other subtypes of type 1 [1a, 1b, 1c(E) and 1c(O)], with longer branches than found between sequences of the same subtype [e.g. HCV-PT and HCV-H (type 1a)]. Similarly, the two samples from Gambia group together and are distinct from other subtypes of type 2, although the core and NS-5 trees differ in how this genotype is placed relative to other subtypes (Fig. 4B, F). For type 4 sequences, proposed subtypes 4(I) and 4(II) branch separately from variants reported from Central Africa, and from the original type 4a(E) sequence found in Egypt (represented by sequences EUME3657 and EUME3662; Fig. 4D, H). The tree indicates the close relationship between variants in the Yemen and Italy (EUYEM3706, EUYEM2 and EUIT3835) with the previously described type 4d variants found in Denmark, and between EUSA7 (South Africa) and type 4c. The variants described as types 3(I)–3(VI) on the basis of evolutionary distances in core and NS-5 regions all grouped with type 3 sequences (Fig. 4C, G).

For each genotype, there were only minor differences in the branching order between sequences in the core and NS-5 region, indicating that two different and relatively short regions of the HCV genome can produce equivalent results. The inclusion of published sequences for genotypes 1–4 reveals the great diversity of HCV subtypes present, with evidence for the existence of 14 subtypes of type 3 and 11 subtypes of type 4 (Fig. 4C, D) and possibly further type 4 subtypes [4f(S), 4g(S) and 4h(S)] which have sequences available only in the NS-5 region (Fig. 4H). Generally, the greatest subtype diversity originates from restricted geographical regions. For example, almost all of the type 3 variants originate from the Indian subcontinent (including Nepal). The three samples obtained from Nigeria each provided evidence for the existence of novel subtypes of type 1.

**Classification of NG(I) and NG(II)**

The identity of the variants from Thailand that lacked any close relationships with types 1–6 was also investigated by phylogenetic analysis with examples of known HCV genotypes (Fig. 5). NG(I) is similar to a novel genotype previously found in Thailand (Apichartpiyakul et al., 1994), referred to here as H1 (see Methods), and more distantly to a related Thai variant, H2, with pairwise evolutionary distances in the NS-5 region of 0·206 and 0·268 respectively. Together these sequences cluster with a recently described series of variants found in Vietnam that have been proposed to represent separate major genotypes 6, 7, 8 and 9 on the basis of pairwise distances in core/E1 and NS-5 regions (Tokita et al., 1994a). However, it is difficult to classify NG(I) and NG(II), or other variants from Thailand (Hotta et al., 1994b) in a consistent way. For example, the mean pairwise distance between the two Thailand variants PC and B4/92 (termed type H2) with the five type H1/NG(I) sequences is 0·2493 in the NS-5 region, suggesting that they are subtypes. The H1/NG(I) sequences show mean distances of 0·2792 and 0·2766 with variants described as 8a(O) and 8b(O) from Vietnam, again within the subtype range. However, the other subtype, type H2 shows mean distances of 0·3384 and 0·3531 with types 8a(O) and 8b(O), values that are ambiguous but closer to the lower percentile of the type range, than the upper percentile for subtypes. Furthermore, although type 8a/b(O) and type 9a(O) were described as separate major genotypes, type H2 shows a similar mean distance to type 9a(O) (0·3428) as it does to the type 8(O) sequence distances. Similar inconsistencies arise upon attempts to incorporate EUTH36 [NG(II)] in the part of the tree containing type 7(O) sequences, with mean distances of 0·3108 with type 7b(O), but 0·3526 with type 7a(O).

A close relationship was observed between core and NS-5 sequences obtained from Vietnam and Thailand with those previously classified as type 6a (originally
Fig. 5. Unrooted maximum likelihood trees of (A) 576 bp of the E1 region; (B) 1093 bp of NS-5; (C) 222 bp of NS-5. All branch lengths are shown to scale. The percentage of bootstrap replications supporting the branches leading to each clade is indicated. H1 and H2 are temporary labels of HCV genotypes found in Thailand (Apichartpiyakul et al., 1994). For clarity, analysis is confined to at most two sequences for each subtype as listed in Methods.
found in Hong Kong and Macau) (Fig. 5). Trees based upon E1 (Fig. 5A), 1093 bp of NS-5 (Fig. 5B) and upon 222 bp of NS-5 (Fig. 5C) are congruent, supporting our conclusion presented earlier in this paper that even the relatively short 222 bp fragment of NS-5 is generally sufficient for HCV classification. Each phylogenetic tree reveals a shared common ancestor for Vietnamese and Thailand variants with type 6a, where the division of HCV into only six clades is, in each region, supported by large numbers of bootstrap replications. However, despite their common origin, variants within the clade labelled ‘6’ are diverse in sequence, showing greater variability than found amongst variants of other clades. Because of the noise caused by excessive multiple substitutions at synonymous sites, neighbour-joining trees were also reconstructed using non-synonymous sites only. In all cases (NS-5, 222 bp, 1093 bp and E1), similar relations between type and subtype were found, including a common ancestor for type 6 and variants from Vietnam and Thailand (data not shown). Whether HCV should be classified by evolutionary relationships or simply by sequence similarity is discussed further below.

Discussion

Identification of HCV genotype using subgenomic sequences

This study has allowed us to evaluate different methods of HCV classification, and to compare the discriminating power of different genomic regions and of different lengths of sequence information. Statistical analysis of the distribution of pairwise evolutionary distances in the core and NS-5 regions for standard datasets indicated that distances between different major genotypes, different subtypes of the same type and between different variants of the same subtype (isolates) corresponded closely to normal distributions. The spread of values within each category was principally determined by the number of sequence comparisons, and was hardly affected by the length of sequence compared.

Consequently, a better indicator of the usefulness of different subgenomic regions for classification is provided by comparison of percentile values, such as the 95% range, for type, subtype and isolate since these, unlike the total range, are not affected by the number of component values. Relatively short regions of NS-5, such as the 329 and 222 bp fragments produced 95% percentile ranges that were similar to those obtained for the 1093 bp fragment of NS-5 (Fig. 2) or the 576 bp fragment of E1 (data not shown), and therefore are equivalent for the purpose of classification using similarity values. This conclusion is consistent with observations that classifications of HCV based upon subgenomic regions such as E1, NS-3, NS-4 and NS-5 accurately reflect those obtained for the complete genome sequence (Chan et al., 1992; Simmonds et al., 1994b; Stuyver et al., 1994; Tokita et al., 1994a, b).

For the core region, the 95th percentile ranges between subtype and isolate were distinct but the type and subtype percentiles overlapped (Fig. 1). Despite this, there were only two discrepancies between the core and NS-5 region when evolutionary distances were used to classify novel HCV variants obtained in this study. Furthermore, phylogenetic analysis of core and NS-5 sequences for subtypes of HCV types 1–4 produced consistent results (Fig. 4), and confirmed the identification of the novel genotypes listed in Table 2. The use of two regions has allowed us to add several new HCV subtypes to types 1–5 within the framework proposed for HCV classification (Simmonds et al., 1994a).

Whatever the method of analysis, the relationship of NG(I) and NG(II) with other HCV variants remains problematic. Using pairwise evolutionary distances, NG(I) appears closely related to isolates of type H1, which are themselves most closely related to types 8a(O), 8b(O) and 9a(O) from Vietnam, although not closely enough to consider them as subtypes of each other. Similarly, NG(II) is more closely related to variants from Vietnam described as type 7a(O) and 7b(O), although sequence distances of 0.35 and 0.31 in the NS-5 region placed it outside the expected subtype range (upper percentile). The variants from Vietnam, NG(I), NG(II) and the other variants from Thailand showed a greater than expected similarity with type 6a sequences, and grouped together with them in phylogenetic trees based upon NS-5 and E1 sequences (Fig. 5). Within this group as a whole, there is no clear evidence for clustering into the type, subtype and isolate categories apparent elsewhere in the tree and it appears that variability amongst these isolates is structured differently from that of the other genotypes. The close evolutionary relationship between type 6 and variants classified as types 7, 8 and 9 from Vietnam was partially obscured using the unweighted pair-group method with arithmetic mean (UPGMA) method for phylogenetic analysis (Tokita et al., 1994a). This method does not take into account variation in rate of sequence change in different lineages, and minimizes the size of the branch leading to the common ancestor of the ‘type 6’ clade.

In previous studies confined to HCV genotypes 1–6, phylogenetic analysis produced an equivalent classification to that derived by simply comparing pairwise distances between new sequences with those previously classified. Given the discrepancies now found between methods it is not clear whether a virus should be classified on the basis of evolutionary relationships,
which would place types 6, 7, 8 and 9 into a single genotype, or on the basis of sequence similarity, which would place them as separate genotypes, as previously proposed (Tokita et al., 1994a).

It is accepted in most areas of systematics that the best classifications are those that reflect the underlying phylogenetic relationships of the organisms (or sequences) in question. The theoretical basis for this position is that there is only one true phylogeny and that using this is both objective and natural (Ridley, 1986). However, it can be argued that this approach is only as good as our confidence in the phylogeny itself and that if the phylogeny is uncertain, so will be the classification, particularly if it is based on a very small and often biased sample of extant lineages. For example, distinct clusters of sequences (genotypes) could simply arise from intermediate sequences being under-represented in the sample. A more useful classification may be one which reflects some aspect of function, such as serological relationships, which is of more relevance to those developing vaccines.

On the basis of the data presented here we suggest a pragmatic approach to the classification of HCV, involving measures of phylogenetic relationship and genetic distance. In most cases, these will be equivalent. However, variation in rate of sequence change may explain the difficulties described above for the classification of variants found in Vietnam and Thailand, as the clade labelled ‘6’ contains HCV variants much more divergent in sequence from each other than found amongst the other HCV genotypes.

Geographical patterns of HCV sequence diversity

The pattern of HCV variation in certain African and South-East Asian regions appears to be distinct from that in Europe and other Western countries. In the latter (pattern ‘A’), a restricted number of HCV genotypes occur, of which types 1a, 1b and 3a are the most frequent. The occurrence of additional genotypes in these countries is often attributable to recent travel or immigration; for example, a Canadian blood donor infected with type 6 was found to be a recent immigrant from Vietnam (Murphy et al., 1994). In Thailand, Nepal, India and Central Africa (pattern ‘B’), HCV infection within a population is predominantly of a single major genotype, but a wide range of subtypes are found. For example, six subtypes of type 3 were found in ten hepatitis C patients from Nepal (Tokita et al., 1994b), and six subtypes from eleven individuals in Bangladesh, India and Pakistan (Tables 1 and 2). Although types 3a and 3b are relatively widespread, the other novel subtypes of type 3 from Nepal (3c–3f) were represented by single examples (Tokita et al., 1994b), as were those described in this study from neighbouring countries.

An analogous situation has been described in Central Africa, where nine subtypes of type 4 have been obtained from a relatively small number of HCV-infected individuals in Zaire and Gabon (Bukh et al., 1993, 1994; Stuyver et al., 1994). The discovery of three novel subtypes of type 1 in three different Nigerian blood donors is preliminary evidence for the same phenomenon elsewhere in Africa (Table 2). In geographical regions where HCV sequence diversity corresponds to pattern B, the value of naming every novel subtype might be questioned, especially in extreme cases where each individual sampled is infected with a hitherto novel HCV genotype.

Origins of HCV

If the rate of nucleotide sequence change of HCV was known it would be possible to calculate the period of endemicity required to produce the observed ranges of sequence diversity associated with pattern B. Although rates of sequence change over relatively short intervals of time have been calculated (Ogata et al., 1991; Abe et al., 1992; Okamoto et al., 1992a; Cuypers et al., 1991), such methods may lead to significant underestimates of the time of divergence of less closely related variants, such as the HCV subtypes (Simmonds, 1995). The existence of numerous subtypes in a single geographical area suggests the long-term presence of HCV in the population.

Phylogenetic analysis and calculation of mean pairwise distances between subtypes reveals differences in the degree of diversity of sequences within a major genotype, as well as the number of component subtypes. For example, type 3 sequences are, in general, much more divergent from each other than are subtypes of type 1 or type 2 (Fig. 5). The variant from Thailand (TD3) described as a separate major genotype (Apichartpiyakul et al., 1994) could be equally well interpreted as a particularly divergent type 3 variant, even though it has diverged beyond the upper limit of the subtype range. This variation may reflect the relative duration of HCV infection with that genotype within particular populations. This model suggests an explanation for the pattern of sequence diversity elsewhere in South-East Asia. The closer than expected relatedness of variants described as type 7a(O), 7(b), 8a(O), 8b(O), 9a(O), H1, H2, NG(I) and NG(II) with each other and with type 6a, the existence of a common ancestor for them that is distinct from that of the other HCV genotypes (Fig. 5), and their presence within a continuous and relatively restricted geographical region strongly suggests that
these variants were originally subtypes of a single major genotype, but the process of diversification has proceeded further than in other genotypes.

Endemic and epidemic spread of HCV

The high level of diversity of particular major genotypes within restricted geographical regions (pattern B) may be interpreted as long-term, endemic infection within a community. Conversely, epidemic spread of HCV, such as its entry into new risk groups, might be represented by pattern A. For example, the widespread occurrence of recent and rapid transmission of this particular variant throughout Europe, North America, parts of Africa and the Far East. The degree of diversity of type 1b sequences is around 9% over the complete genome, which dates the start of its dissemination to relatively recent times, perhaps over the last 40–50 years. How this genotype was transmitted remains unclear, although examples are known of type 1b transmission to large sections of particular populations through the use of contaminated blood products (Hohne et al., 1994; Power et al., 1994). Similarly, the close relatedness of type 3a isolates suggests more recent spread to a range of Western countries, in this case possibly through its association with intravenous drug abuse. Other examples of possible epidemic patterns of HCV transmission include the extensive distribution of type 4a(E) in the Middle East, particularly in Egypt, where a particularly high rate of adult infection is found (Saeed et al., 1991; Abdelaal et al., 1994; Kamel et al., 1994).

The existence of discrete subtypes in countries showing pattern A may reflect the past epidemiology of HCV transmission, where the founder effect from a single introduction of HCV may be apparent for several decades. Consequently, classification of HCV into subtypes provides convenient epidemiological labels for variants within countries where HCV has been relatively recently spread, but may turn out to be of different relevance in areas where HCV infection shows a more endemic pattern of variability. There are clear parallels between this model for the epidemiology of HCV infection with current knowledge of the spread of human immunodeficiency virus type 1 (HIV-1). The discovery of an extremely wide range of subtypes of HIV-1 in Africa (Louwagie et al., 1993; Bruce et al., 1994) supports the hypothesis for its origin from that continent. In contrast, the restricted variability of HIV-1 in other countries, for example, of subtype B in USA and Europe, and subtype C in Thailand drug users, is evidence for its relatively recent spread in particular risk groups elsewhere.

References


(Received 21 February 1995; Accepted 26 May 1995)