Evolutionary analysis of variants of hepatitis C virus found in South-East Asia: comparison with classifications based upon sequence similarity

P. Simmonds,1 J. Mellor,1 T. Sakuldamrongpanich,2 C. Nuchaprayoon,2 S. Tanprasert,2 E. C. Holmes3 and D. B. Smith1

1 Department of Medical Microbiology, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK
2 National Blood Centre, Thai Red Cross Society, Henri Durant Road, Bangkok 10330, Thailand
3 Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Variants of hepatitis C virus (HCV) have been classified by nucleotide sequence comparisons in different regions of the genome. Many investigators have defined the ranges of sequence similarity values or evolutionary distances corresponding to divisions of HCV into types, subtypes and isolates. Using these criteria, novel variants of HCV from Vietnam, Thailand and Indonesia have been classified as types 7, 8, 9, 10 and 11, many of which can be further subdivided into between two to four subtypes. In this study, this distance-based method of virus classification was compared with phylogenetic analysis and statistical measures to establish the confidence of the groupings. Using bootstrap resampling of phylogenetic trees in several subgenomic regions (core, E1, NS5) and with complete genomic sequences, we found that one set of novel HCV variants ('types 7, 8, 9 and 11') consistently grouped together into a single clade that also contained type 6a, while 'type 10a' grouped with type 3. In contrast, no robust higher-order groupings were observed between any of the other five previously described HCV genotypes (types 1–5). In each subgenomic region, the distribution of pairwise distances between members of the type 6 clade were consistently bi-modal and therefore provided no justification for classification of these variants into the three proposed categories (type, subtype, isolate). Based on these results, we propose that a more useful classification would regard all these variants as subtypes of type 6 or type 3, even though the level of sequence diversity within the clade was greater than observed for other genotypes. Classification by phylogenetic relatedness rules out simple sequence similarity measurements as a method for assigning HCV genotypes, but provides a more appropriate description of the evolutionary and epidemiological history of a virus.

Introduction

Hepatitis C virus (HCV) shows a degree of genetic heterogeneity similar to that of other RNA viruses with different isolates of HCV showing as much as 30% nucleotide sequence divergence over the entire genome. It is suspected that such divergence may lead to serological and biological differences between variants of HCV. For example, the diversity of the amino acid sequence of the HCV E1 and E2 glycoproteins of HCV (around 34–40% for E1, and 26–29% for E2) is similar to that found in the E gene of different serotypes of dengue fever virus (26–40%) which are known not to cross-neutralize. In the absence of biological or true neutralization assays for HCV, classification of HCV is currently based upon nucleotide sequence comparisons. Sequence relationships appear equivalent in different parts of the genome. For example, the six main groups of variants found in the NS5 gene (Simmonds et al., 1993) are paralleled by equivalent groupings in E1 (Bukh et al., 1993), core (Bukh et al., 1994) and NS4 (Bhattacherjee et al., 1995), and by comparison of available complete genomic sequences (types 1a, 1b, 1c, 2a, 2b, 3a, 3b) (Sakamoto et al., 1994). So far, there is no evidence...
Fig. 1. Phylogenetic analysis of nucleotide sequences of HCV variants from Hong Kong, Vietnam, Thailand, Burma and Indonesia. (a) A 222 base fragment (positions 7975–8196); (b) a 366 base fragment (positions 8464–8829) of
NS5. The sequence of type 1a (Choo et al., 1991) was used as an outgroup. Previously described genotype designations (column 1) or provisional clade identifiers (column 2) are shown on the right.
for recombinants of HCV that have sequences of one genotype in one part of the genome and of a different one elsewhere (Smith et al., 1995; Stuyver et al., 1994; Simmonds et al., 1994b).

How long and from where in the genome a region should be to enable reliable classification by this method has been a matter for debate. Although it is generally true that longer sequences are more informative for classification, it is usually possible to identify genotypes by sequence comparison of relatively short subgenomic regions of HCV (Mellor et al., 1995; Tokita et al., 1994; Bukh, 1995). However, the identification of novel genotypes of HCV is generally considered to require sequences from at least two parts of the genome (Simmonds et al., 1994a). The 5' non-coding region (5'NCR) is too conserved to be useful for this purpose, although some subtypes and almost all types of HCV show distinct sequences in this region (Mellor et al., 1994a). The classification based upon this method provides the most suitable method for the classification of HCV.

Another more fundamental discussion concerns the methods by which sequences should be compared. To date, it has been possible to obtain equivalent classifications of HCV by either phylogenetic analysis of sequences or by measures of pairwise sequence similarity. In the former method, genotypes have been identified by the branching order of a phylogenetic tree. For example, sequences of subgenomic fragments of types 1–6 always produce trees containing six main branches (clades) which split further into groupings corresponding to the subtypes of HCV.

The basis for the second method is that the sequences of the different types are all similarly divergent from each other, as are virus subtypes, and so it is possible to define ranges of sequence similarities that correspond to the type and subtype categories. For example, pairwise comparisons of 76 sequences from a 222 base fragment of NS5 (positions 7975–8196) produce ranges of similarity values from 56.2–72.1% between types of HCV, 74.8–86.0% between subtypes and greater than 87.8% between variants of the same subtype (Simmonds et al., 1993). Different regions of the genome produce different ranges depending on how conserved or variable the region is, but in each case the ranges are non-overlapping. For E1 and NS4, the maximum and minimum pairwise distances between type and subtype are greater than for NS5 (Bhattacherjee et al., 1995; Bukh et al., 1993), while they are smaller for core (Bukh et al., 1994). For many investigators the sequence similarity of a novel sequence to those of established genotypes has become the primary means to identify and classify new genotypes.

Recently, a series of novel variants of HCV has been found in Vietnam, Thailand, Burma and Indonesia (Doi et al., 1996; Mellor et al., 1995, 1996; Tokita et al., 1994, 1995, 1996; Sugiyama et al., 1995; Apichartpiyakul et al., 1994). On the basis of previously established ranges for type and subtypes in NS5 and core/E1, it was proposed that they should be classified as types 6a, 6b, 7a, 7b, 7c, 7d, 8a, 8b, 9a, 9b, 9c, 10a and 11a (Tokita et al., 1994, 1995, 1996). Sugiyama et al. (1995) described similar variants from Thailand as types VII and VIII.

In the current study, we have carried out phylogenetic analysis of several subgenomic regions of HCV, followed by bootstrap resampling to determine the robustness of the observed groupings. The classification based upon this method was compared with that derived from pairwise sequence measurements in order to critically evaluate which method provides the most suitable method for the classification of HCV.

**Methods**

**Samples.** Nucleotide sequences in the core and NS5 regions were amplified by PCR from plasma from blood donors from Bangkok, Thailand and Burma as previously described (Mellor et al., 1995). Amplified DNA was directly sequenced between base positions 15–322 (core) and 7975–8196 (NS5; sequences numbered as in Choo et al., 1991) as previously described. The analysis also included previously published sequences from core, E1 and different regions of NS5 (Doi et al., 1996; Tokita et al., 1994, 1995, 1996; Mellor et al., 1995, 1996; Sugiyama et al., 1995; Apichartpiyakul et al., 1994).

In order to compare the ability of fragments of different sizes to differentiate between genotypes, as well as to include as much of the published sequence information as possible, we carried out analysis on a total of six datasets. Sequences corresponding to 308 bases at the 5' end of the core region have been obtained from 56 Thailand, Burmese and Vietnamese samples (Mellor et al., 1995, 1996; Tokita et al., 1994, 1995, 1996); more sequences (n = 66) were available from a shorter region (positions 27–277; length 251 bases) (Sugiyama et al., 1995). A total of 26 sequences were obtained from Vietnamese, Thailand and Indonesia samples in the E1 region (positions 574–1149) (Tokita et al., 1994, 1995, 1996). Three sets of sequences were compared in the NS5 region. The first consisted of the 3' terminal 1093 bases of NS5B, for which 25 sequences were compared (Tokita et al., 1994, 1995, 1996). Thirty-six sequences were available from a 366 nucleotide fragment of NS5B between positions 8464–8829 (Tokita et al., 1994, 1995, 1996; Sugiyama et al., 1995). The largest dataset for NS5 was obtained between positions 7975–8196 (n = 65) (Doi et al., 1996; Tokita et al., 1994, 1995, 1996; Mellor et al., 1995, 1996; Apichartpiyakul et al., 1994). Because of the very large number of sequences available for core, E1 and 222 base NS5 regions for genotypes 1–5, it was necessary to use a subset in the phylogenetic analyses. Consequently, phylogenetic trees were reconstructed using two representatives, where available, of every other previously described type and subtype (listed in Mellor et al., 1995).

**Sequence analysis.** Evolutionary distances (i.e. those corrected for multiple substitution) between pairs of sequences were estimated using the DNADIST program in the PHYLIP package (Felsenstein, 1993). Frequency analysis of pairwise distances was assisted by the use of standard statistical software (SYSTAT). Phylogenetic analysis was carried out using the programs DNADIST, PROTDIST, NEIGHBOR and DRAWTREE in the PHYLIP package (Felsenstein, 1993). Trees (rooted and unrooted) are shown with branch lengths drawn to scale. The robustness of the groupings was assessed using bootstrap resampling of 100 replicate neighbour-joining trees (programs SEQBOOT, DNADIST, NEIGHBOR and CONSENSE in the PHYLIP package; Felsenstein, 1993).
Results

Identification of HCV genotypes in South-East Asia

Eight previous studies describe the detection of novel genotypes of HCV from samples in Thailand, Burma, Vietnam and Indonesia (Doi et al., 1996; Tokita et al., 1994, 1995, 1996; Mellor et al., 1995, 1996; Sugiyama et al., 1995; Apichartpiyakul et al., 1994). These independent studies have sometimes analysed different regions of the HCV genome which complicates the analysis of their inter-relationships. However, phylogenetic trees of sequences in the 222 base and 366 base NS5 regions allowed almost all of variants to be compared (Fig. 1).

Sequences described as NG(I) (Mellor et al., 1995) were the same genotype as those described as 7c (Tokita et al., 1995), type VII (Sugiyama et al., 1995) and as BB7 and LD47/93 (Doi et al., 1996), and represent one of the most frequently detected variants in the four different studies. Some variants found in northern Thailand (Apichartpiyakul et al., 1994) were equivalent to NG(I) (Mellor et al., 1995), while others (B492, PC) were similar to a variant found in Burma, EUBUR1 (Mellor et al., 1996). Neither of these variants was found in Vietnam (Tokita et al., 1994). Similar analysis of a 251 base fragment of the core region allowed comparison of sequences from all nine previous studies, and of the 13 new core sequences obtained in this study. Each of the groupings observed in the two regions of NS5 were closely reproduced in the core region (data not shown).

Combining the results from NS5 and core, a total of 14 distinct genotypes could be identified in four South-East Asian countries. Clusters a, d, e, k, l and h (described as types 6a, 7a, 7b, 8a, 8b and 9a) were found in Vietnam, while clusters b, c, f, i, j, m and n were found in Thailand [described as 6b, 7d, 7c/NG(I)/VII, 9b, 9c, B492/PC and NG(I)] and in Burma (EUBUR1). Cluster g (type 11a) was found in Indonesia. So far, none of these variants show an overlapping geographical distribution with any other variant.

Classification by phylogenetic analysis

Evolutionary relationships between the South-East Asian variants and genotypes 1–5 and with type 6a were analysed by phylogenetic analysis of nucleotide sequences in different subgenomic regions. In order to compare all of the published sequences, five datasets were used corresponding to core (308 bases), E1 and three different regions in NS5. The unrooted phylogenetic tree for the 222 base region of NS5 was the most complete (Fig. 2) but otherwise showed a topology typical of each of the regions analysed, i.e. six major branches corresponding to genotypes 1–5 and a sixth branch that contains type 6a as well as all of the variants from Vietnam and Thailand and the variant from Indonesia described as ‘type 11a’ (Tokita et al., 1996) (Fig. 2). Another Indonesian variant described as ‘type 10a’ grouped with type 3 sequences and corresponded to TD3, a variant previously found in Indonesia (Hotta et al., 1994a).

Bootstrap resampling of phylogenetic trees was carried out to test the robustness of the observed major clades (Table 1). Although there has been discussion about the statistical meaning of bootstrap values (Li et al., 1994; Felsenstein &
Table 1. Bootstrap resampling of phylogenetic groupings in different regions of the HCV genome

<table>
<thead>
<tr>
<th>Region</th>
<th>Nucleotide positions</th>
<th>Clade*</th>
<th>Next most frequent grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3t 4 5 6</td>
<td>Branching order</td>
</tr>
<tr>
<td>Core</td>
<td>15–3220</td>
<td>78 100 84 100 100 90</td>
<td>2(3(5(4,1)))</td>
</tr>
<tr>
<td>E1</td>
<td>574–1149</td>
<td>100 100 100 100 100 100</td>
<td>4((1,5)(2,3))</td>
</tr>
<tr>
<td>NS5</td>
<td>7975–8196</td>
<td>98 100 90 96 100 90</td>
<td>3(4(2(1,5)))</td>
</tr>
<tr>
<td>NS5</td>
<td>8454–8829</td>
<td>100 100 100 NA NA NA</td>
<td>2,(1,3)</td>
</tr>
<tr>
<td>NS5</td>
<td>7938–9030</td>
<td>100 100 100 NA NA NA</td>
<td>1(2,3)</td>
</tr>
<tr>
<td>Complete</td>
<td>–341–9033</td>
<td>100 100 100 NA NA NA</td>
<td>1(2,3)</td>
</tr>
</tbody>
</table>

* Number of 100 bootstrap resamplings that group variants into six clades corresponding to HCV genotypes 1–5 and a sixth group containing type 6a and variants from South-East Asia.
† Indonesian variants TD3 and ‘type 10a’ included in type 3 clade.
‡ Next most frequent grouping of genotypes.
§ Bootstrap value of next most frequent grouping.
NA. Not available.

Table 2. Distribution of evolutionary distances between type, subtype and isolates of HCV genotypes 1–6 in different regions

<table>
<thead>
<tr>
<th>Region</th>
<th>Core</th>
<th>E1</th>
<th>NS5 222 bases</th>
<th>NS5 366 bases</th>
<th>NS5 1093 bases</th>
<th>Complete genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start*…</td>
<td>15</td>
<td>574</td>
<td>7975</td>
<td>8464</td>
<td>7938</td>
<td>–341</td>
</tr>
<tr>
<td>End*…</td>
<td>322</td>
<td>1149</td>
<td>8196</td>
<td>8829</td>
<td>9030</td>
<td>9030</td>
</tr>
<tr>
<td>Length (bases)…</td>
<td>308</td>
<td>576</td>
<td>222</td>
<td>366</td>
<td>1093</td>
<td>9371</td>
</tr>
<tr>
<td>No. of sequences†…</td>
<td>93</td>
<td>87</td>
<td>173</td>
<td>37</td>
<td>33</td>
<td>21‡</td>
</tr>
</tbody>
</table>

| Isolate    | Minimum value | 0.000 | 0.000 | 0.000 | 0.011 | 0.008 | 0.0095 |
| Subtype    | Maximum value | 0.082 | 0.185 | 0.151 | 0.075 | 0.074 | 0.099  |
|            | Median        | 0.033 | 0.080 | 0.052 | 0.057 | 0.060 |        |
| Type       | Minimum value | 0.057 | 0.260 | 0.149 | 0.107 | 0.141 | 0.200  |
|            | Maximum value | 0.153 | 0.450 | 0.349 | 0.246 | 0.227 | 0.231  |
|            | Median        | 0.099 | 0.329 | 0.235 | 0.192 | 0.197 |        |
|            | Minimum value | 0.084 | 0.434 | 0.317 | 0.291 | 0.340 | 0.312  |
|            | Maximum value | 0.295 | 0.873 | 0.743 | 0.573 | 0.487 | 0.343  |
|            | Median        | 0.192 | 0.592 | 0.532 | 0.377 | 0.410 |        |

* Sequences numbered as in Choo et al. (1991).
† Number of sequences of established genotype used to define ranges.
‡ Genotypes 1a (n = 3), 2b (n = 13), 1c(C) (n = 1), 3a (n = 1), 3b (n = 2) and 3b (n = 1).
§ Uncorrected pairwise distances used to compare complete genomic sequences.

Kishino, 1993; Hillis & Bull, 1993), a cut-off value of around 75% is generally considered to establish confidence in the phylogenetic groupings. Bootstrap values obtained for the main clades were almost always much greater than this. Whichever subgenomic region was analysed, variants of HCV corresponding to genotypes 1–5 consistently formed separate groups with bootstrap values usually of 100% (the lowest being for the type 1 clade in the core region with a bootstrap value of 78 out of 100 replications).

Supporting the classification of South-East Asian variants (apart from type 10a) as type 6 variants, bootstrap resampling showed consistent grouping of these variant within the type 6 clade with values ranging from 90 to 100 per 100 replications. Similarly, TD3 and ‘type 10a’ consistently grouped with other...
type 3 variants (bootstrap values 84 to 100%). In contrast, no robust higher-order grouping of other genotypes was found. For example, the grouping of type 3 with type 4 sequences (Fig. 2) was not supported by bootstrap analysis (Table 1). In this region, the most commonly observed grouping was between type 1 and 5 sequences (66% of trees, a value insufficient to confer significance). Bootstrap resampling of sequences in other subgenomic regions also failed to reveal associations between genotypes 1–5, and the groupings that were observed varied between core, E1 and NS5. For example, the most frequent pairwise grouping in E1 was between genotypes 2 and 3 (76%), compared with 5 and 1 + 4 combined for core (46%) and 1 and 5 for NS5 (66%). Even using complete genomic sequences, no grouping other than into genotypes 1, 2 and 3 and a separate branch for ‘type 11a’ was observed.

A separate bootstrap analysis of trees produced by the program PROTDIST using the inferred amino acid sequences from each NS5 region and E1 produced similar results to those in Table 1 for nucleotide sequences. Each region supported the grouping of South-East Asian variants with type 6a into a single clade, and none provided consistent evidence for significant grouping of other genotypes. The amino acid sequences in the core region were too invariant to produce reproducible groupings.

**Classification by sequence similarity**

Using collections of sequences of established genotype (types 1–5 and 6a), the distributions (minimum and maximum values, medians) of pairwise evolutionary distances between type, subtype and isolate for different subgenomic regions (core, 308 bases; E1, 576 bases; NS5, 222 bases; NS5, 366 bases; NS5, 1093 bases) and of complete genomic sequences were calculated (Table 2).

The distribution of evolutionary distances within the different clades observed in the subgenomic regions has been used as an alternative method to determine how variants may
be classified. For example, the distribution of sequence distances between the currently classified genotypes (1–5, 6a) is tri-modal. The first peak corresponds to the set of pairwise distances within types, the second to distances between subtypes, while the third corresponds to distances within subtypes. This analysis was used previously to substantiate the proposed type, subtype and isolate categories used for HCV classification (Simmonds et al., 1994a).

This analysis was extended to the five subgenomic regions described above to investigate the relationships between sequences falling within the type 6 clade (Fig. 3). The finding of three separate peaks of sequences distances would support the previous proposal to classify them into a novel series of types, subtypes, and isolates (types 6a, 6b, 7a, 7b, 7c, 7d, 8a, 8b, 9a, 9b, 9c and 11a), while two distributions would support their classification as subtypes of type 6.

In practice, each of the subgenomic regions produced a distribution with two distinct peaks (Fig. 3a–e). For example, in the core region the first peak of evolutionary distances had a median value of 0.021, similar to that observed between isolates within established HCV subtypes (0.033; Fig. 3a; Table 2). The second peak had a median value of 0.149, intermediate between the values previously calculated between subtypes and between genotypes of currently classified variants of HCV (0.099 and 0.192 respectively; Fig. 3a; Table 2). Similar bimodal distributions of sequence distances were observed in each of the other four subgenomic regions amongst the set of sequences obtained from South-East Asia. For example, the median of the second peak for the 222 base fragment of NS5 was 0.414, almost halfway between the median values for subtype (0.235) and type (0.532; Table 2). Even using the longest fragment of NS5 (1093 nucleotides),

Fig. 4. Phylogenetic analysis of (a) nucleotide or (b) inferred amino acid sequences of complete genomes of types 1, 2, 3, 'type 10a' and 'type 11a' (Tokita et al., 1996), shown as unrooted trees. Bootstrap values for the type 1, 2 and 3 (including JK049) clades are shown in Table 1.
Fig. 5. Ranges of pairwise distances between type, subtype and isolate of 22 complete genomic sequences of types 1a (n = 3), 1b (n = 13), 1c (n = 1), 2a (n = 1), 2b (n = 1), 3a (n = 2) and 3b (n = 1), and ‘type 10a’. In (a), ranges for complete genomic sequences and for subgenomic fragments corresponding to each of the HCV genes, core through NS5B, were calculated based upon the classification of ‘type 10a’ as a separate major genotype. The ranges obtained by its alternative classification as a subtype of type 3 allow a much clearer discrimination of types and subtypes (b). Vertical lines on scale bars represent increments of 10% sequence divergence.

the median value of distances between South-East Asian genotypes was intermediate (0.291 compared with medians of 0.197 and 0.410 for subtype and type respectively). In this region, none of the pairwise distances between variants within the type 6 group overlapped with distances between genotypes 1-3 (maximum value for type 6 sequences 0.336, minimum distance between previously defined genotypes 0.340; Table 2), showing that the intermediate status of the South-East Asian variants did not arise from simply comparing inadequate lengths of sequence.

Analysis of complete genomic sequences

Tokita et al. (1996) have suggested that the current uncertainty and difficulty in classifying type 6-group variants may be resolved by obtaining complete genomic sequences. Although the complete sequence of the variant from Indonesia (‘type 11a’) was described, this remains of little use in the absence of any other variant in the type 6 group with which to compare it. However, an analogous comparison can be made between the other Indonesian variant (‘type 10a’) with the complete genomic sequences of type 3a and 3b, as the sequences show similar ambiguous inter-relationships to those found within the type 6 clade.

By phylogenetic analysis in any of the subgenomic regions, ‘type 10a’ consistently grouped within the type 3 clade (Table 1). Similarly, analysis of complete nucleotide (Fig. 4a) sequences showed the close evolutionary relationship of ‘type 10a’ with type 3a and 3b. The branching point for ‘type 10a’ with type 3 was only slightly displaced from the branch of type 3a and 3b. In contrast, no grouping is found between any other pair of genotypes (bootstrap values in Table 1), and each
clade appears to radiate from a single point. Similar relationships exist between amino acid sequences translated from the complete genomes (Fig. 4b). Using evolutionary or phenotypic analysis, there is no justification for regarding ‘type 10a’ as a separate genotype from type 3.

However, Tokita et al. proposed that ‘type 10a’ belonged to a separate major genotype because of its greater divergence from all known complete genome sequences (25%) than was previously found between subtypes (range 20.0–23.1%; Table 2). Similarly, pairwise distances outside the normal subtype range were also found in each of the subgenomic regions (data not shown). However, the pairwise distance between type 3b and ‘type 10a’ was also considerably smaller than between other genotypes (25% compared with 31.2%). Indeed, it might be more parsimonious to regard ‘type 10a’ as a subtype of type 3 (and extend the subtype range by 1.9%) rather than to regard ‘type 10a’ as a separate genotype [and therefore reduce the range of distances between types by 6.2% (from 31.2% to 25%)].

The effects of these two alternative classifications on the ranges between type and subtype in subgenomic regions is shown in Fig. 5. Classifying ‘type 10a’ as a separate type reduces the lowest pairwise distance between types to a point where it becomes close to or even overlaps the upper limit of the subtype range (Fig. 5a) in each subgenomic region. In contrast, classification of ‘type 10a’ as a subtype of 3 slightly increases the subtype range in each subgenomic region, but maximizes the separation between categories (Fig. 5b). Clear differentiation of type and subtype can therefore be observed in each of the larger and more variable subgenomic regions (such as NS5B, NS5A, NS4B, NS3 and E1).

Discussion

Implications for HCV classification

The main finding of this analysis of HCV variants from South-East Asia was the consistent grouping of most of these variants with type 6a upon phylogenetic analysis of several subgenomic regions. This grouping was found in each subgenomic region analysed and each appeared to be equally informative in the phylogenetic analysis. In contrast, no robust grouping was observed between other genotypes (types 1–5; Table 1). The second finding was that the set of pairwise distances between different genotypes amongst the South-East Asian variants formed a single distribution in each subgenomic region. Minimum and maximum values ranged continuously from those observed between subtypes of other genotypes through to distances found between different major genotypes (Table 2).

Clearly, variability within this group is structured differently from that between other HCV variants, making classification into types and subtypes according to previous criteria difficult to apply (Simmonds et al., 1994a). For example, genotypes from Vietnam were classified as types 7a, 7b, 8a, 8b and 9a by comparison of their pairwise distances with ranges of distances previously found between other genotypes and subtypes. Variants with pairwise distances below the previously established ‘cut-off’ value of 20% (Tokita et al., 1994) or evolutionary distances of 0.23; Mellor et al., 1995) for the 1093 base fragment of NS5, would be subtypes (7a, 7b and 8a, 8b) while those with values greater would be different types. This simple approach is justified for classification of sequences when it is applied to discrete distributions of pairwise distances (Simmonds et al., 1993). However, the same ‘cut-off’ value cannot be used to classify sequences with a unimodal distribution of pairwise distances that has a median distance close to this cut-off value, as is the case for the type 6 clade. Indeed, we have previously found it impossible to consistently classify as types or subtypes variants from Thailand previously described as NG(I) and NG(II). For example, the pairwise distance between NG(II) and type 7a for the 222 nucleotide region of NS5 was 0.31 (in the subtype range), but 0.35 with type 7b (type range). Similar difficulties will undoubtedly be encountered as more sequences from the type 6 clade become available, and perhaps for other clades such as type 3.

To resolve this difficulty, we have previously suggested that this group might more usefully be regarded as a single major genotype that contains members that are more divergent from each other than are subtypes of other genotypes (Mellor et al., 1995). This proposal fits better with the observation of a single distribution of sequence distances, and their consistent grouping by phylogenetic analysis. As an evolutionary classification it would also be more appropriate for the description of a continuous range of variants found in a relatively restricted geographical area (Vietnam, Thailand, Burma, Indonesia; Mellor et al., 1996). However, this proposal would effectively rule out the straightforward use of sequence similarity as a means of virus classification.

The nomenclature for South-East Asian variants is influenced by the method of sequence comparison. Whereas analysis of these variants on the basis of sequence similarity has led to the suggestion of four new major genotypes (types 7, 8, 9 and 11), each with three or four subtypes, as well as a new subtype of type 6, 6b (Tokita et al., 1994, 1995), phylogenetic analysis implies a single genotype (type 6) comprising 14 relatively diverse subtypes (Figs 1 and 2).

Although most of the analysis described in this study has been confined to variants of HCV grouping within the type 6a clade, there is evidence that the phenomenon of highly divergent subtypes may also occur amongst other HCV genotypes. For example, the variants ‘type 10a’ and TD3 obtained from Indonesia (Tokita et al., 1996; Hotta et al., 1994b) consistently grouped with type 3 variants upon phylogenetic analysis (Fig. 2; Table 1), although they showed sequence similarity values with type 3 variants generally in the type range (evolutionary distances in the NS5 222 base fragment of 0.323–0.412). It is possible that more extensive analysis of HCV variants in Indonesia and in neighbouring
countries might reveal a similarly diverse range of genotypes phylogenetically related to type 3, and pose the same problems for classification as do the type 6 variants.

**Aims of HCV classification**

The classification of virus variants must not only be accurate in describing real and consistent grouping of variants as discussed above, but should also be of some practical benefit. So far, biological differences have been consistently observed between individuals infected with HCV type 1 and types 2 and 3 in terms of the response to interferon treatment (Miyakawa et al., 1995; Davis, 1994; Bukh et al., 1994), but there are no convincing data for clinical or virological differences between virus subtypes (Lau et al., 1996; Mahaney et al., 1994). Similarly, serological differences have been demonstrated between virus types 1–6 (Bhattacherjee et al., 1995; Tsukiyama Kohara et al., 1993; McOmish et al., 1993) but not yet between virus subtypes.

As yet, clinical or virological information about type 6 clade viruses is sparse, and it is unclear whether members of the clade behave as diverse subtypes with common biological properties, or as genotypes with distinct characteristics. However, we have recently observed that samples from individuals infected with type 6 clade variants (other than 6a) generally show a specific serological reactivity to type 6a peptides in a serotyping ELISA (Bhattacherjee et al., 1995) (and unpublished results). This suggests that the phylogenetic grouping of variants within the type 6 clade may also reflect serological properties of the variants. In the future it will be possible to investigate whether other biological and clinical similarities are also found amongst members of the type 6 clade.

**Molecular epidemiology of HCV genotypes**

Although a number of issues remain concerning the nomenclature of HCV variants based upon different methods of sequence comparison, the analysis presented in this paper provides a new insight into the process of HCV sequence diversification. We have previously suggested that the observed diversity of HCV variants in different geographical regions may reflect the history of transmission within a community. For example, the numerous subtypes of type 4 in central Africa and of type 3 in the Indian subcontinent may reflect a long-standing endemic pattern of virus transmission, while the restricted subtype distribution elsewhere (such as 4a in Egypt, and type 3a in drug-users in Europe) is evidence for their recent spread into new risk groups (Mellor et al., 1995).

In the case of the variants found in South-East Asia, even highly limited sampling of the population provides an unusually diverse range of sequences. The existence of numerous co-existing lineages suggests a long-term stability in virus transmission within these populations. At the same time, the repeated detection of certain genotypes such as type 7c/VII/NG(II) in Thailand indicates the more recent and more rapid spread of certain variants possibly in association with new risk groups for HCV infection (Holmes et al., 1995).

Over much longer periods, it is possible that the process of diversification within the type 6 clade will proceed further to produce HCV genotypes with no detectable residual commonality, as currently appears to be the case for genotypes 1–5. However, this process is relatively slow. Estimates of the rate of sequence change of HCV from a large cross-sectional study of individuals infected with HCV 17 years previously from a common source ranged from 0.11% per site per year in NS5 to 0.16% in E1 at silent sites. Using these rates, the current diversity between different isolates within a subtype such as type 1b predicts a common ancestor approximately 70–80 years ago (D. B. Smith and others, unpublished data). Although constraints on sequence change make it more difficult to extrapolate divergence times between different subtypes, such as type 1a and 1b, a minimum estimate would be 300 years, correcting for differences in the transitions and transversions and silent and non-silent substitutions.

The current diversity observed in Vietnam and Thailand therefore represents a period of transmission within the local population considerably longer than any of these estimates. The existence of ancient lineages heightens the difficulties in understanding the persistence of a virus infection within human societies in which parenteral exposure has been documented to be the only efficient route of transmission.

**References**


Received 17 June 1996; Accepted 6 August 1996