Sequence analysis of hepatitis C virus genotypes 1 to 5 reveals multiple novel subtypes in the Benelux countries

L. J. van Doorn,1* G. E. M. Kleter,2 L. Stuyver,3 G. Maertens,3 J. T. Brouwer,4 S. W. Schalm,4 R. A. Heijtink2 and W. G. V. Quint1

1Department of Molecular Biology, Diagnostic Center SSDZ, R. de Graafweg 7, PO Box 5010, 2600 GA Delft, Netherlands, Luxembourg) were genotyped by means of reverse hybridization Inno-LiPA (line probe assay). Genotypes 1a, 1b, 2a, 2b, 3a, 4a and 5a were detected. From the cohort, isolates representing all types and those showing an aberrant LiPA pattern were further analysed by sequencing parts of the 5' UTR, core (nt 1 to 326; aa residues 1 to 108) and core/E1 (nt 477 to 924; aa residues 159 to 308) regions. Molecular evolutionary analysis of the core and core/E1 regions allowed discrimination between known and additional subtypes, especially within types 2 and 4. The core region is not suitable for classification of new subtypes because of the relatively high level of conservation. The core/E1 region displays a higher level of sequence variation and allows much more distinct discrimination between subtypes. Genotypes 2 and 4 are particularly heterogeneous, with at least 7 and 10 subtypes, respectively. In contrast to previous reports from Europe, HCV isolates from the cohort constituted a highly heterogeneous population of virus variants, especially within genotypes 2 and 4.

Hepatitis C virus (HCV) isolates from a cohort of 315 patients from the Benelux countries (Belgium, The Netherlands, Luxembourg) were genotyped by means of reverse hybridization Inno-LiPA (line probe assay). Genotypes 1a, 1b, 2a, 2b, 3a, 4a and 5a were detected. From the cohort, isolates representing all types and those showing an aberrant LiPA pattern were further analysed by sequencing parts of the 5' UTR, core (nt 1 to 326; aa residues 1 to 108) and core/E1 (nt 477 to 924; aa residues 159 to 308) regions. Molecular evolutionary analysis of the core and core/E1 regions allowed discrimination between known and additional subtypes, especially within types 2 and 4. The core region is not suitable for classification of new subtypes because of the relatively high level of conservation. The core/E1 region displays a higher level of sequence variation and allows much more distinct discrimination between subtypes. Genotypes 2 and 4 are particularly heterogeneous, with at least 7 and 10 subtypes, respectively. In contrast to previous reports from Europe, HCV isolates from the cohort constituted a highly heterogeneous population of virus variants, especially within genotypes 2 and 4.

Hepatitis C virus (HCV) is the major aetiological agent of parenterally transmitted non-A, non-B hepatitis (Choo et al., 1989; Kuo et al., 1989). The virus contains a positive-sense, ssRNA genome of approximately 9400 nt. Sequence comparisons revealed the existence of multiple HCV strains or genotypes. Recently, a classification system has been proposed (Chan et al., 1992; Simmonds et al., 1994a; Stuyver et al., 1993), differentiating between types, subtypes and isolates. HCV genotyping may have clinical relevance, e.g. the efficacy of interferon therapy (Takada et al., 1992; Yoshioka et al., 1992; Tsubota et al., 1994).

Genotyping is accomplished by analysis of parts of the genome, such as the 5' UTR (Stuyver et al., 1993; Simmonds et al., 1993) and the core region (Okamoto et al., 1992b). Phylogenetic studies of the core (Bukh et al., 1994; Simmonds et al., 1994b), the E1 (Bukh et al., 1993; Simmonds et al., 1994b) and the NS5 region (Chayama et al., 1993; Simmonds et al., 1994b) suggested that any region of the genome could be used for classification of HCV genotypes (Chan et al., 1992), although the level of sequence heterogeneity differs considerably between different parts. More detailed studies indicated that the NS5B region is probably most suited for classification (Stuyver et al., 1994).

A reverse hybridization line probe assay (LiPA) has been developed for genotyping of HCV isolates by 5' UTR analysis (Stuyver et al., 1993; van Doorn et al., 1994). The 5' UTR is highly conserved but shows significant sequence variation (Kleter et al., 1994; van Doorn et al., 1994), which allows the use of this region for HCV genotyping (Stuyver et al., 1994).

HCV isolates were obtained from participants in a trial of treatment with interferon-α, coordinated by the Benelux Study Group on the treatment of chronic hepatitis C. All patients resided in either Belgium, The Netherlands or Luxembourg (Benelux), although their ethnic origin was diverse. All samples used in this study were obtained before onset of therapy.

HCV RNA was isolated by a modification of the guanidinium isothiocyanate method (Chomczynski & Sacchi, 1987; Kleter et al., 1993). Detection of HCV RNA was performed by RT-PCR with primers HCV35 (sense, positions −318 to −296; 5' TTGGCGGCGCC GCACTCCACCATGAATCACTCCCC 3'; underlined sequence is non-HCV specific) and HCV19 (antisense, 0001-3054 © 1995 SGM
positions −1 to −20; 5′ GTGCACCGTGCTACGAGACC-T 3′). Genotyping was performed by a prototype Inno-LiPA HCV genotyping assay (Stuyver et al., 1993). PCR products from the 5′ UTR were reamplified by nested PCR with primers NCR3 (sense, positions −314 to −288; 5′ GGGCGGGCGGCCAACCATA-RRATCACCTCCCTGAGGG 3′; R indicates A or G; underlined sequences are non-HCV specific) and LD58B (antisense, positions −35 to −64; 5′ GGCGGGCGGCCGCGCCAAGCGCATACTCCCTC-3′; biotinylated at the 5′ end).

RT–PCR was also performed with primers HCV983 (antisense, positions 963 to 983; 5′ GGGGCGGCCAACCATA-CATTCAGTACCAAGGC 3′; biotinylated at the 5′ end). Synthesis of cDNA was primed with HCV983. First round PCR products were reamplified with LD58C and 186c (anti-biotinylated LD58C (sense, positions −57 to −34; 5′ GGTACTGCGATAGGGTGCTG 3′) in a semi-nested PCR for analysis of the core region. To analyse core/E1 sequences, first-round PCR products were reamplified by nested PCR with primers NCR3 (sense, positions −35 to −64; 5′ ATiTACCCCATGAGGGCGCCG 3′) in a semi-nested PCR for analysis of the core region. To analyse core/E1 sequences, first-round PCR products were reamplified by nested PCR with primers NCR3 (sense, positions −35 to −64; 5′ ATiTACCCCATGAGGGCGCCG 3′) in a semi-nested PCR for analysis of the core region. To analyse core/E1 sequences, first-round PCR products were reamplified by nested PCR with primers NCR3 (sense, positions −35 to −64; 5′ ATiTACCCCATGAGGGCGCCG 3′) in a semi-nested PCR for analysis of the core region.

All biotinylated nested PCR fragments were subjected to direct sequence analysis as described earlier (Kleiter et al., 1994; van Doorn et al., 1994). Phylogenetic trees were constructed using the Phylogeny Inference Package (PHYLIP; version 3.5c; Felsenstein, 1993).

In this study, 315 HCV isolates, obtained from patients living in the Benelux area of Western Europe were all genotyped by LiPA. Genotypes 1, 1b, 2a, 2b, 3a, 4a and 5a were detected and the prevalence of each genotype is shown in Table 1. Only four mixed infections were detected. One isolate (NE-92) could be identified as type 2, but subtyping was not possible (designated type 2?). One isolate (NL96) resulted in an aberrant LiPA pattern and could not be typed. Within type 1, the prototype LiPA allowed positive identification of subtype 1b only, typing the remaining isolates as type 1. The genotypic distribution (Table 1) differs from earlier reports (Dusheiko et al., 1994; McOmish et al., 1994) and illustrates that the major HCV genotypes are all present in Western Europe, including genotypes 4 and 5 which were regarded as African genotypes. The version of LiPA used in this study did not permit the recognition of type 6, which seems to be exclusively present in the Far East.

Randomly selected isolates as well as isolates showing an aberrant LiPA pattern were subjected to further analysis in order to validate the 5′ UTR classification in coding regions and to assess the genomic variability of HCV isolates from the Benelux region.

First, a total of 37 isolates (nine of type 1, four of 1b, three of 2a, two of 2b, one of type 2?, four of 3a, eight of 4a, five of 5a and one untypeable) were further analysed by sequencing of the 5′ UTR. Sequence results were in complete agreement with LiPA patterns for each isolate.

Secondly, part of the core region sequence (nucleotides 1–326) was amplified from 36 isolates (12 of type 1, four of 1b, three of 2a, two of 2b, one of 2?, four of 3a, four of 4a, five of 5a and one untypeable, according to LiPA) and subjected to direct sequence analysis. Phylogenetic distances among these and several reference sequences were calculated. The phylogenetic distances did not segregate into three non-overlapping distance ranges (data not shown) for types, subtypes and isolates. Further analysis revealed that isolates originally typed as genotypes 2a or 4a were much more heterogeneous than other types, suggesting the existence of multiple subtypes within these groups. Therefore, 36 isolates (seven of type 1, four of 1b, five of 2a, one of 2b, one of 2?, four of 3a, eight of 4a, five of 5a and one untypeable, according to LiPA) were subjected to sequencing of the more variable C-terminal core/N-terminal E1 region (nucleotides 477 to 924). Phylogenetic distances were calculated and the resulting phylogenetic tree is presented in Fig. 1. The frequency distribution of pairwise phylogenetic distances in the core/E1 region showed some overlap (Fig. 2), but the segregation into types, subtypes and isolates was much better than for the N-terminal core region. Remarkably, the average pairwise phylogenetic distance of all type 2 isolates compared to other types was 0.68 ± 0.05, whereas for all other intertype distances this was only 0.57 ± 0.08. The frequency peak of 0.24 is mainly related to pairwise distances within type 4 (0.23 ± 0.04) compared to 0.36 ± 0.05 between subtypes of all other types. Although overlap regions were small, it was impossible to calculate exact border values for types, subtypes and isolates. Therefore, the pairwise distances in this core/E1 region were also compared with border values obtained by analysis of a larger core/E1 region from isolates of which NS5B sequences were also known.

Table 1. Prevalence of HCV genotypes in a cohort of 315 patients from the Benelux countries as determined by LiPA

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Number of patients</th>
<th>Prevalence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>10-1</td>
</tr>
<tr>
<td>1b</td>
<td>187</td>
<td>59-4</td>
</tr>
<tr>
<td>2a</td>
<td>20</td>
<td>6-3</td>
</tr>
<tr>
<td>2b</td>
<td>3</td>
<td>0-9</td>
</tr>
<tr>
<td>2?</td>
<td>1</td>
<td>0-3</td>
</tr>
<tr>
<td>3a</td>
<td>45</td>
<td>14-3</td>
</tr>
<tr>
<td>4a</td>
<td>15</td>
<td>4-8</td>
</tr>
<tr>
<td>5a</td>
<td>7</td>
<td>2-2</td>
</tr>
<tr>
<td>Multiple</td>
<td>4</td>
<td>1-3</td>
</tr>
<tr>
<td>Untypeable</td>
<td>1</td>
<td>0-3</td>
</tr>
</tbody>
</table>

Table 1 illustrates that the major HCV genotypes are all present in Western Europe, including genotypes 4 and 5 which were regarded as African genotypes. The version of LiPA used in this study did not permit the recognition of type 6, which seems to be exclusively present in the Far East.
Fig. 1. Phylogenetic tree drawn from the phylogenetic analysis of nucleotide sequences by the neighbour-joining method (Saitou & Nei, 1987). Reference sequences are: type 1a, HCV1 (Choo et al., 1991); type 1b, HCV-J (Kato et al., 1990) and HCV-BK (Takamizawa et al., 1991); type 1c, HC-G9 (Okamoto et al., 1994); type 2a, HC-J6 (Okamoto et al., 1991); type 2b, HC-J8 (Okamoto et al., 1992a); type 2d, NE-92 (Stuyver et al., 1994); type 2c, S83 (Bukh et al., 1993); type 3a, NZL1 (Sakamoto et al., 1994); type 3b, NE137 (Tokita et al., 1994); type 4d, DK13 (Bukh et al., 1993); type 4c, GB358 (Stuyver et al., 1994); type 5a, BE95 (Stuyver et al., 1994); type 6a, HK2 (Bukh et al., 1993).

Isolates NL29 and NL35 were obtained from patients who probably contracted their HCV infection in Morocco. These two isolates could be classified into a
separate subtype of genotype 1, distinct from the proposed genotype 1c isolates (HCG9, Okamoto et al., 1994; Td-6, Td-34/92, Hotta et al., 1994a, b) from Indonesia.

Initial phylogenetic analyses have suggested the importance of the G at position -99 in the 5' UTR for identification of subtype 1b isolates. Recently, some 1b isolates containing an A at -99 have been reported (Bukh et al., 1992, 1993), suggesting that this position cannot be consistently used for recognition of subtype 1b. In our limited number of isolates we observed only two discrepancies (isolate NL43 and NL69), which were classified as 1a by LiPA, but as 1b by sequence analysis. Low discrepancy levels have also been reported by others (Mahaney et al., 1994; Gianinni et al., 1994).

Based on core/E1 sequences, three type 2 isolates (NL49 and NL33 from Surinam and NL50 from The Netherlands) could each be classified into a novel subtype, apart from subtype 2a (HCJ-6), 2b (HCJ-8), 2c (S83) and 2d (NE92). Isolate NL50 contained a mutation at position −127, whereas the other two subtypes were indistinguishable from subtype 2a in the 5' UTR. Subtypes 2b and 2d could both be identified by covariance in the 5' UTR sequence motifs, resulting in specific LiPA patterns. The single type 2 isolate that could not be subtyped by LiPA was also analysed in the NS5B region; it has been described earlier as NE-92, classified as subtype 2d (Stuyver et al., 1994). Thus, genotype 2 isolates show remarkable heterogeneity in the core and core/E1 regions. From phylogenetic analysis of the core/E1 region six subtypes can be detected in this group of patients. All type 2 isolates from our randomly selected group were different from the Japanese type 2a.

This has also been confirmed for most of the remaining type 2 isolates of the entire cohort of 315 patients, by subtype-specific PCR (Okamoto et al., 1992b; data not shown). Similar to observations by Bukh et al. (1993) our study also revealed that type 2 is the most distant group among HCV genotypes 1 to 5 and confirmed the high degree of heterogeneity within this genotype.

The four selected type 3 isolates all belonged to genotype 3a. However, the single untypeable isolate, NL96, could be classified as a seventh subtype of type 3, different from genotypes 3a to 3f (Tokita et al., 1994). This isolate has also been analysed in the E1/E2 region and the NS5 region (data not shown), confirming classification as a separate subtype within genotype 3. The characteristic type 3 'TCA motif' at position −95 to −92 is not present in this isolate. A new sequence motif is present between positions 167 and 159. This isolate was obtained from a patient who probably contracted the HCV infection in Indonesia, which may explain the rarity of such isolates in Western Europe. Recently, this isolate was found to be highly homologous to Indonesian isolate Td-3/93 (Hotta et al., 1994b) in the NS5b region (data not shown).

Like type 2, type 4 isolates are a highly heterogeneous group. Since the frequency distribution (Fig. 2) showed some overlap between isolates and subtypes, several of our type 4 isolates could not be formally classified as different subtypes after core/E1 analysis. However, among eight genotype 4 isolates, at least four subtypes could be distinguished. B203 clustered with subtype 4c isolates and NL52 with subtype 4d isolate DK13. Comparison with published type 4 sequence data (Stuyver et al., 1994; Bukh et al., 1993, 1994) indicated that the remaining six isolates (from Egypt, Zaire and The Netherlands) clustered into two novel subtypes. Five of these isolates belong to the same newly identified subtype, which seems to be the predominant type 4 subtype in the Benelux countries. The existence of at least eight subtypes within type 4 has been reported earlier (Stuyver et al., 1994). Therefore the total number of subtypes within type 4 increases to at least 10. However, formal classification of all type 4 subtypes requires additional sequence analyses, preferably in the NS5B region (Stuyver et al., 1994).

The two sequence motifs in the 5' UTR of type 4 isolates were completely conserved in our group of patients. Additional point mutations, located outside the motifs, were present and some of the subtypes could be identified by the presence of specific 5' UTR sequences.

The five isolates of genotype 5 were highly homogeneous and were classified into a single subtype (5a).

Classification of isolates into the major genotypes based on either 5' UTR, core or E1 was completely consistent. It is remarkable that analyses of 5' UTR and the core/E1 regions result in similar classifications. The 5' UTR is a non-coding region containing a highly conserved, presumably functional, element with stringent...
constraints on the secondary structure of the RNA (Yoo et al., 1992; Wang et al., 1993). In contrast, the core and E1 regions encode the putative nucleocapsid and envelope proteins, and would therefore be subjected to completely different selective pressures from the 5′ UTR.

Comparison of core sequences described in this study and subtype-specific primers as described by Okamoto et al. (1992b) revealed that specific primer target sequences contained several mismatches, especially in type 2 isolates. It seems likely, therefore, that subtype-specific amplification with multiple different type-specific primers will not be an efficient genotyping method in heterogeneous HCV populations.

It appears that the majority of HCV genotypes known today can be differentiated in the 5′ UTR. However, sequence variation in this region is not sufficient to recognize every single subtype. Whereas the 5′ UTR and core region are very conserved, the E1 and NS5 regions reveal additional variation. At present, it is unclear whether detailed subtyping of HCV isolates has any clinical significance. This is the subject of further study.

The classification system of HCV into types, subtypes and isolates and the nomenclature of the major genotypes appears now to be generally accepted. However, classification and nomenclature of novel isolates is still a problem that requires consensus among the scientific community. Based on this and other studies (Bukh et al., 1993; Stuyver et al., 1994) the E1 and NS5B regions might be the most appropriate regions to use for detailed phylogenetic analysis and classification.

In conclusion, genotyping of HCV isolates from a large patient cohort revealed the presence of a highly heterogeneous population of HCV types and subtypes in the Benelux countries of Western Europe. However, it should be noted that the Benelux region has a multiracial population, which might explain detection of HCV variants that had only been reported in distant regions of the world. It can be speculated that most HCV genotypes have a worldwide distribution, although the relative regional prevalence of each type may vary considerably. Therefore, detection of minor subtypes may require analysis of larger patient cohorts.

The authors wish to thank the Benelux Study Group for providing plasma samples from the patient cohort.

References


(Received 5 December 1994; Accepted 10 February 1995)