Full-length sequence of the genome of hepatitis C virus type 3a: comparative study with different genotypes

Noriko Yamada,1 Kiyomi Tanihara,1 Masashi Mizokami,2 Kenichi Ohba,2 Akira Takada,3 Mikihiro Tsutsumi3 and Takayasu Date1*

1 Department of Biochemistry, Kanazawa Medical University, Uchinada, Ishikawa 920-02, 2 Second Department of Internal Medicine, Nagoya City University, Kawasumi Mizuho, Nagoya 467 and 3 Department of Gastroenterology, Department of Internal Medicine, Kanazawa Medical University, Uchinada, Ishikawa 920-02, Japan

Hepatitis C virus (HCV) type K3a (type 3a), which represents a minor genotype in Europe, the U.S.A. and Asia, appears to be significantly distributed throughout Australia and Brazil. We amplified the HCV-K3a/650 genome by reverse transcription polymerase chain reaction in ten overlapping fragments and determined the nucleotide sequences. The total sequence was 9454 bases in length and contained an open reading frame of 3021 amino acids, which is 10 or 11 amino acids longer than in HCV type 1 and 12 amino acids shorter than the sequence of type 2. These differences were due to the different lengths of both the putative envelope protein E2 and the NS5A regions, whose nucleotide lengths differ between types 1 and 2 also. Phylogenetic analysis of the putative core region and a portion of NS5B encoding the Gly-Asp-Asp motif indicated that HCV-K3a closely matched the corresponding type 3a group. The deletion and addition of amino acids in both E2 and NS5A may be associated with their pathobiological features.

Hepatitis C virus (HCV) is the major causative agent of non-A, non-B hepatitis (Choo et al., 1989). A number of HCV strains from different areas of the world have been isolated and subjected to sequence analysis. The viral genome is a positive strand RNA molecule, approximately 9400 nucleotides in length, containing an open reading frame (ORF) which encodes 3010 to 3033 amino acid residues (Kato et al., 1990; Choo et al., 1991; Takamizawa et al., 1991; Okamoto et al., 1991, 1992b). By analogy with flaviviruses, this polypeptide can be divided into a 5′ structural region, consisting of the putative core and envelope proteins, and a 3′ region corresponding to non-structural (NS1 to NS5) proteins. HCV shows substantial nucleotide sequence variation throughout the viral genome and has been classified into several genotypes to date. Since we discovered that HCV isolates could be assigned to two genotypes and four subtypes based on analysis of a portion of the NS5 domain (Enomoto et al., 1990), many variants have been isolated and recent phylogenetic analysis of the NS5 region has led us to classify HCV into six major groups and 11 subtypes (Simmonds et al., 1993a). HCV has also been grouped into several distinct genotypes on the basis of sequence variation at the 5′ non-coding region (5′ NCR; Bukh et al., 1992; Chan et al., 1992; Simmonds et al., 1993b), core gene (Simmonds et al., 1993b), E1 gene (Bukh et al., 1993) and other regions (Okamoto et al., 1991, 1992; Chan et al., 1992; Simmonds et al., 1992). Previously, we described a scheme of HCV types – PT (prototype), K1, K2a, K2b and K3a (Enomoto et al., 1990; Nakao et al., 1991; Takada et al., 1993b) – which correspond to types 1a, 1b, 2a, 2b and 3a, respectively (Chan et al., 1992). Other groups use different schemes. In this paper we use the terminology as described by Chan et al. (1992) to distinguish between different genetic variants of HCV.

The major HCV types distributed widely around the world are types 1a and 1b. Type 2 has also been found in almost all countries, but it preferentially distributes in the Far East (Takada et al., 1992a, 1993a; Bukh et al., 1993). Types 4, 5 and 6 are found mainly in Africa (Cha et al., 1992; Bukh et al., 1992, 1993; Simmonds et al., 1993a). Type 3a, which was first found in isolates from Thailand (Mori et al., 1992), and was then reported from Europe (Chan et al., 1992) and the U.S.A. (Lee et al., 1992), appears to be a common genotype worldwide with the exception of Japan (Takada et al., 1993b). Here we report the full sequence of type 3a (HCV-K3a/650) and compare it with other isolates.

We previously reported the world distribution profile of HCV types obtained from a limited number of collections and demonstrated that their distribution

The sequence data from this article have been deposited with the GSDB, DDBJ, EMBL, NCBI DNA databases under accession number D28917.
Table 1. Distribution of HCV type 3a in different countries*

<table>
<thead>
<tr>
<th>Country</th>
<th>Occurrence</th>
<th>No. of samples</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>13</td>
<td>57</td>
<td>22.8</td>
</tr>
<tr>
<td>Brazil</td>
<td>5</td>
<td>58</td>
<td>8.6</td>
</tr>
<tr>
<td>Spain</td>
<td>2</td>
<td>26</td>
<td>7.7</td>
</tr>
<tr>
<td>U.S.A.</td>
<td>2</td>
<td>63</td>
<td>3.2</td>
</tr>
<tr>
<td>Japan</td>
<td>3</td>
<td>376</td>
<td>98</td>
</tr>
<tr>
<td>China</td>
<td>0</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>Thailand</td>
<td>0</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>

* In other countries such as Germany (13 samples), France (31) and Egypt (25), no type 3a was detected in our samples.

appears to differ from country to country. However, some isolates could not be classified (Takada et al., 1992a, b). Cloning and sequence analysis in the region of NS5B of three unclassified samples revealed that they were all part of a third group named K3a (Takada et al., 1993b) which corresponds to type 3a (Chan et al., 1992). A specific probe was synthesized from the cloned fragment and used for slot-blot analysis of PCR products. Table 1 shows the world distribution of HCV type 3a. Type 3a seems to be a minor genotype in most countries but it is significantly distributed in Australia and Brazil. In contrast to the finding by Mori et al. (1992), no type 3a was detected in our collection from Thailand. Because our samples were not representative, the distribution of type 3a appears to be localized in specific areas even within one country. Recently, we found three Japanese patients with HCV type 3a.

In order to clone the type 3a HCV genome, cDNAs were obtained by reverse transcription polymerase chain reaction (RT–PCR) using samples (HCV-K3a/650) (10^7 HCV/ml) from the U.S.A. Amplified fragments were obtained from both the NS5B (Enomoto et al., 1990) and 5' NCR (Okamoto et al., 1990) regions, since their sequences had been determined previously. The primer used for cDNA synthesis was a random hexamer, except for synthesis of the 3'-end fragment, for which oligo-(dA)14 was used. A major difficulty with PCR cloning is the possibility that mismatches between the primers and the variant sequence will prevent amplification. To overcome this problem, the sequences of PCR primers for the side whose sequence had not been defined were selected from among relatively conserved regions of different genotypes. Most of the PCR primers were also designed to contain a restriction enzyme site near their 5' end to assist in subsequent cloning. We amplified ten overlapping fragments: they were between nucleotides 1 to 362, 41 to 1597, 1449 to 2584, 2456 to 3664, 3607 to 4795, 4685 to 5443, 5304 to 6418, 6414 to 8286, 7635 to 8725 and 8447 to 9454. The 5' terminal sequence of the 5' NCR was obtained by using a 5'-

Amplifier race kit (Clontech) followed by PCR according to the manufacturer's protocol with minor modifications. The nucleotide sequences in these fragments were determined using an automated DNA sequencer (ABI 373A-70).

The resulting HCV-K3a genome was 9454 bases in length and contained an ORF encoding a polypeptide of 3021 amino acids (aa). The genome was organized from 5' to 3' as genes C (191 aa), E1 (192 aa), E2–NS1 (433 aa), NS2 (170 aa), NS3 (677 aa), NS4A (54 aa), NS4B (261 aa), NS5A (452 aa) and NS5B (591 aa), based on cleavage sites for the processing of the precursor polyprotein obtained from types 1a and 1b (Hijikata et al., 1991 a, 1993; Grakoui et al., 1993 a, b). The ORF of HCV-K3a is 10 and 11 amino acids longer than that of types 1a and 1b, respectively (Kato et al., 1990; Choo et al., 1991), and 12 amino acids shorter than that of types 2a and 2b (Okamoto et al., 1991, 1992). These differences are due mainly to deletion or addition to the E2–NS1 and NS5A regions.

Comparative analysis of nucleotide sequences was carried out to clarify the relationships between our isolate and other isolates. Phylogenetic trees for the core protein region and a portion of the NS5B region in types 1 to 3 were constructed using evolutionary analysis software (Ina; ODEN ver 1.1, National Institute of Genetics Mishima, Japan, 1990). In both cases variants associated into three major groups – types 1, 2 and 3 – and each group split into two major clusters (Fig. 1). Based on analysis of the NS5 region, we had previously classified HCV isolates corresponding to the phylogenetic groups PT/K1 and K2, the first being subdivided into PT and K1 and K2 comprising K2a and K2b (Enomoto et al., 1990). Phylogenetic analysis supports this classification and has also been confirmed by other research groups. These analyses indicated that HCV-K3a/650 is most closely related to E-groups such as E-b1, E-b2, E-b3 and E-b7 and next is closely related to HCV-TR, which includes type 3a and type 3b, respectively (Simmonds et al., 1993a, b). The NS5B sequence of HCV-K3a/650 analysed in this study was more than 93% identical to all other reported sequences in group 3a and about 80% identical to that in group 2b.

The phylogenetic trees constructed from the E1 regions were similar and essentially the same. Classification using partial sequences of the genome is controversial, but the determination of full genomic sequences of HCV is impractical for most laboratories. The full genomic sequence of our subtype 3a may be useful when new sequences are determined and analysed.

The nucleotide and deduced amino acid sequences of the genes in our isolate were compared within ten different regions with those of other HCV genotypes (1a, 1b, 2a and 2b). As shown in Table 2, nucleotide sequences
Fig. 1. Phylogenetic tree for the HCV core (a) and NS5B (b) regions of HCV-K3a/650 and previously reported isolates. The DDBJ/GenBank/EMBL accession numbers of the clones are as follows: HCV-1, M62321; HCV-J1, D00831; HCVH, M67463; HCVJ, D00574; HCV-J, D90208; HCV-BK, M58335; HCVT, M84754; HCV-JT, D11168; HCVJK1, X61596; HC-J4, D00832; HC-J5, D10075; HC-J6, D00944; HC-J7, D10077; HC-J8, D10988; E-b1, D10123; HCV-TR, D11443. The lower scale refers to the number of the nucleotide substitutions per site.

Table 2. Comparison of the nucleotide and deduced amino acid sequences of ten different regions of the HCV-K3a genome with the corresponding regions of seven other HCV isolates *

<table>
<thead>
<tr>
<th>Region (aa)†</th>
<th>HCV-1 (1a)</th>
<th>HCV-J (1b)</th>
<th>HCV-JT (1b)</th>
<th>HACV-BK (1b)</th>
<th>HC-CHINA (1b)</th>
<th>HC-J6 (2a)</th>
<th>HC-J8 (2b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' NCR</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>91</td>
<td>92</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td>C (191)</td>
<td>82 (89)</td>
<td>80 (89)</td>
<td>82 (89)</td>
<td>82 (89)</td>
<td>81 (89)</td>
<td>80 (85)</td>
<td>78 (84)</td>
</tr>
<tr>
<td>E1 (192)</td>
<td>64 (66)</td>
<td>65 (67)</td>
<td>64 (66)</td>
<td>65 (67)</td>
<td>64 (65)</td>
<td>59 (54)</td>
<td>56 (54)</td>
</tr>
<tr>
<td>E2-NS1 (433)</td>
<td>66 (72)</td>
<td>65 (71)</td>
<td>66 (72)</td>
<td>66 (70)</td>
<td>65 (70)</td>
<td>63 (66)</td>
<td>63 (68)</td>
</tr>
<tr>
<td>NS2 (170)*</td>
<td>54 (52)</td>
<td>59 (59)</td>
<td>57 (61)</td>
<td>57 (57)</td>
<td>56 (56)</td>
<td>54 (50)</td>
<td>53 (48)</td>
</tr>
<tr>
<td>NS3 (677)</td>
<td>69 (80)</td>
<td>71 (82)</td>
<td>70 (82)</td>
<td>69 (82)</td>
<td>70 (81)</td>
<td>68 (81)</td>
<td>68 (80)</td>
</tr>
<tr>
<td>NS4A (54)</td>
<td>69 (76)</td>
<td>67 (74)</td>
<td>65 (74)</td>
<td>66 (72)</td>
<td>68 (76)</td>
<td>65 (65)</td>
<td>61 (61)</td>
</tr>
<tr>
<td>NS4B (261)</td>
<td>69 (79)</td>
<td>69 (77)</td>
<td>69 (76)</td>
<td>67 (76)</td>
<td>69 (76)</td>
<td>65 (71)</td>
<td>62 (70)</td>
</tr>
<tr>
<td>NS5A (452)</td>
<td>63 (59)</td>
<td>64 (72)</td>
<td>64 (71)</td>
<td>64 (71)</td>
<td>64 (72)</td>
<td>59 (62)</td>
<td>61 (62)</td>
</tr>
<tr>
<td>NS5B (591)</td>
<td>71 (75)</td>
<td>71 (76)</td>
<td>71 (76)</td>
<td>71 (76)</td>
<td>70 (75)</td>
<td>69 (75)</td>
<td>69 (75)</td>
</tr>
</tbody>
</table>

* HCV-1 (Choo et al., 1991), HCV-J (Kato et al., 1990), HCV-JT (DDBJ/GenBank/EMBL no. D11168), HCV-BK (Takamizawa et al., 1991), HC-CHINA (DDBJ/GenBank/EMBL no. L02836), HC-J6 (Okamoto et al., 1991) and HC-J8 (Okamoto et al., 1992) were compared for nucleotide sequence identities throughout the entire genome except for the 3' NCR. The relevant genotype designation is indicated in parentheses below each isolate.
† The number of (deduced) amino acids encoded in HCV-K3a/650.
‡ The NS2 region is processed into at least two proteins (Grakoui et al., 1993a).

in the 5' NCR of HCV-K3a/650 show 91 to 92% identity to those of types 1a and 1b and 89% and 88% identity to those of type 2a and type 2b, respectively. Most of the sequence heterogeneity in the 5' NCR is found in the vicinity of the 5' end between positions 1 and 18. Other regions in the 5' NCR show 93% and 91% identity to the nucleotide sequences of type 1 and type 2, respectively.

When HCV-K3a/650 was compared at the amino acid level with other isolates, the gene encoding the putative core protein was 89% and 85% identical to type 1 and type 2, respectively. However, putative envelope proteins
E1 and E2–NS1, which contain a hypervariable region at the N terminus (Hijikata et al., 1991b), show fewer similarities within the two groups. The NS3 gene products, which have protease and probably RNA helicase activities, are more conserved than other regions such as the core protein. All residues required for protease activity reported thus far are conserved in both the NS2 and NS3 regions. The NS4A and NS4B gene products, p27 and p70, respectively (Grakoui et al., 1993a, b), show about 61 to 77% identity in types 1 and 2. The 50 amino acid residues at the N-terminus of NS4B varied considerably, showing 50% identity to the corresponding region in other genotypes. The NS5A gene product of HCV-K3a/650 shows less similarity to those encoded by other genotypes because of varied sequences between the middle region and the C terminus (Fig. 2b). The nucleotide sequence of NS5B (a putative RNA polymerase gene) of HCV-K3a/650 is the same length as in other genotypes and shows about 75% identity to types 1 and 2.

The deduced amino acid sequences of the E2- and NS5A-encoded proteins are of interest since these regions show type-specific deletions or additions in several positions. As shown in Fig. 2(a), the optimal alignment indicated that several gaps were present in the middle of E2–NS1. The total number of amino acid residues in E2 for types 1, 2 and 3a was estimated to be 427, 431 and 433, respectively. The other type 1a groups, HC-J1 and HCVH, showed the same gaps as found in HCV1, while other type 1b groups, HCV-BK, HCV-JT and HCV-CHINA, showed the same additions or deletions as found in HCV-J1. Since the cDNA of HCV-K3a/865, another isolate of type 3a from Australia, had an almost identical sequence in the E2 region, these deletions and additions could be type-specific.

Vaccinia virus transient-expression assays revealed that the E2 gene encodes 70K (gp70) and 88K (gp88) proteins containing N-linked glycans (Grakoui et al., 1993b). The latter glycoprotein sequence extends into the NS2 region. The cleavage site of the N-terminus of gp70
is between amino acids 383 and 384, while that of E2-NS2 is predicted to be near the middle of the E2 gene where type-specific additions or deletions in the amino acid sequence are found. Alternative proteolytic processing or other post-translational modifications are also thought to occur in the E2-NS2 region, leading to the production of multiple forms of E2, possibly with distinct biological functions in HCV replication (Garakoui et al., 1993b). Thus, it is possible that these addition or deletion sequences in the E2 region may affect the properties of the multiforms of E2 which in turn changes the immunorespons, proliferation and stability of the virus particle in blood.

The deduced amino acid sequences in the NS5A regions also show type-specific additions and deletions (Fig. 2b). Type 2 has the largest NS5A gene of the reported genotypes (a 20 amino acid insertion at its C terminus in NS5A) whereas type 1b has the smallest one. The regions between aa 260 and 263 and aa 384 and 388 are missing in types 2 and 1, respectively, whereas type 2 has 20 amino acids inserted between aa 433 and 434. Furthermore, the sequence between aa 375 and 412 shows hypervariation among different genotypes, although only a few amino acids are substituted within the same genotype. For example, in the HCV-K3/865 genome, Ala at aa 373 is substituted by Arg.

Two proteins are derived from the NS5 region through processing by the NS3-encoded protease; these are NS5A (58K) and the C-terminal NS5B (66K; Grakoui et al., 1993b; Manabe et al., 1994). NS5B is predicted to contain the RNA-dependent RNA polymerase activity on the basis of the presence of the characteristic Gly-Asp-Asp sequence (aa 2747 to 2749) and surrounding conserved motifs (Poch et al., 1989). The function of the NS5A product has not yet been elucidated. Recently, we demonstrated that in samples from patients with chronic hepatitis only a single 84K protein specifically reacted with antibodies to the NS5B protein (Tsutsumi et al., 1994). This 84K protein is 18K larger than the NS5B product obtained by vaccinia virus transient-expression assays. Therefore, it is possible that the NS5B product is post-translationally modified or that alternative processing occurs in human liver cells. In flaviviruses the NS5 region is not processed further but remains as a single polypeptide of about 100K (Chambers et al., 1990); it would thus be interesting to determine whether the C terminus of the NS5A region, which shows type-specific heterogeneity in length, is the component of native RNA polymerase isolated from human hepatocytes.

There are several lines of evidence to indicate that infection with different HCVs leads to different clinical pictures. HCV type 2 groups seem to be less able to replicate in human liver compared to type 1b, because of a significantly lower concentration of RNA in the blood (Yoshioka et al., 1992). It has been reported that type 2a and 2b groups are more likely to respond to interferon than type 1b (Takada et al., 1992b; Yoshioka et al., 1992; Kanai et al., 1992). Although there are no data to link regions in the viral genome with each clinical picture, we hope that the deletion and addition of amino acid sequences in both E2 and NS5A may be associated with different pathobiological features.

We thank Dr Lawrence Lumeng, Indiana University, U.S.A. and Dr Geoff Farrell, University of Sydney, Australia for gifts of serum samples from patients with HCV. This work was supported by grants from the Hokkoku Cancer Research Foundation, Ishikawa, Japan and the Foundation for Promotion of Cancer Research backed by the Japan shipbuilding Industry Foundation.

References


(Received 18 April 1994; Accepted 4 July 1994)