Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East

Richard W. Chamberlain,¹ Neill Adams,² Awad A. Saeed,³ Peter Simmonds² and Richard M. Elliott¹

¹ Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, UK
² Department of Medical Microbiology, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK
³ Riyadh Armed Forces Hospital, PO Box 7897, Riyadh 11159, Saudi Arabia

Hepatitis C virus (HCV) type 4 is the predominant genotype found throughout the Middle East and parts of Africa, often in association with high population prevalence as in Egypt. To investigate more fully its evolutionary relationship with other genotypes of HCV, and to study its overall genome organization, we have determined the entire sequence encompassing the coding region of the genotype 4a isolate ED43, obtained from an HCV-infected individual from Egypt. The sequence of ED43 contained a single open reading frame encoding a polyprotein of 3008 amino acids (aa), smaller than that reported for other HCV genotypes which vary from 3010 aa to 3037 aa. The nucleotide and amino acid sequences were compared with the full-length sequences already reported for genotypes 1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b and those of isolates JK049 and JK046 described as types 10a and 11a. The differences in length of the polyprotein originated in variable regions in the E2 and NS5A genes. The complete sequence of ED43 confirmed the classification of type 4 as a separate major genotype.

Introduction

Hepatitis C virus (HCV) is the aetiological agent for the majority of cases of blood-borne non-A, non-B hepatitis. The HCV genome is a linear, single-stranded RNA of positive polarity, approximately 9500 nucleotides long, which contains a single open reading frame (ORF) encoding a large polyprotein of about 3000 amino acids (aa). The polyprotein is processed to give the mature polypeptides by both host and virus encoded proteases (reviewed by Houghton et al., 1994). This processing yields three structural proteins [core (C), envelope 1 (E1) glycoprotein and envelope 2 (E2) glycoprotein] and six nonstructural proteins (NS2, NS3, NS4A, NS4B, NS5A and NS5B). The ORF is flanked by a 5′ non-coding region (NCR) which serves as an internal ribosome entry site (IRES) and a 3′NCR which consists of a stretch of nucleotides (nt) of variable length (27–66 nt) and sequence, followed by a poly(U) tract and then a highly conserved 98 base element suggested to be required for replication and/or packaging (Tanaka et al., 1995, 1996; Kolykhalov et al., 1996).

HCV isolates have been classified based on comparisons of partial nucleotide sequences, for example the 5′NCR (Bukh et al., 1992; Simmonds et al., 1993a), or by phylogenetic analysis of sequences from selected regions of the genome, such as the core region (Bukh et al., 1994; Simmonds et al., 1994a), E1 (Bukh et al., 1993; Simmonds et al., 1994a; Stuyver et al., 1994) and NS5B (Simmonds et al., 1993b, 1994a; Stuyver et al., 1994). The majority of HCV isolates so far studied can be classified into six major groups (Simmonds et al., 1994b) designated genotypes 1–6, with subdivisions in each (subtypes a, b, c, etc.). A further three genotypes (7, 8 and 9) have been proposed based on partial sequences at the 5′ and 3′ ends of the genomes of isolates from Vietnam and Thailand (Tokita et al., 1994, 1995), and recently another two genotypes (10a and 11a) have been suggested based on comparisons of complete sequences of isolates from Indonesia with those of types 1, 2 and 3 (Tokita et al., 1996). HCV genotypes 1 and 2 have a broad distribution in the Far East, Europe, USA and parts of Africa. Genotype 3 shows a similar distribution, being found in Europe, USA, Thailand and India although rarely in Japan. Genotype 4 has been identified as the principal genotype amongst infected individuals from the Middle East, North Africa (particularly Egypt where it is associated with a high

Author for correspondence: Richard W. Chamberlain.
Fax +44 141 337 2236, e-mail cham01r@vir.gla.ac.uk

The nucleotide sequence of HCV genotype 4 (HCV-ED43) has been deposited in the EMBL database under accession number Y11604.
population prevalence (Saeed et al., 1991; Kamel et al., 1992) and Central Africa, although there is a low frequency of detection in populations outside these areas. Genotypes 5 and 6, however, show a highly restricted geographical distribution: genotype 5 has so far only been commonly found in South Africa and genotype 6 has only been found in Hong Kong, Macau and neighbouring regions in South-East Asia such as Vietnam (for review see Simmonds (1995) and references within). Within the genotypes, diversity between isolates can be considerable. This is clearly seen, for example, in genotypes 3 and 4 where up to nine subtypes have so far been identified in each genotype (Bukh, 1995; Mellor et al., 1995; Simmonds, 1995; Oni & Harrison, 1996).

Complete nucleotide sequences have been determined for at least 26 HCV genomes representing genotypes 1a, 1b, 1c, 2a, 2b, 2c, 3a and 3b and the putative genotypes 10a and 11a. However, genotype 4 has to date, only been defined by analysis of sequences representing the 5'NCR, core, E1, NS4 and NS5 regions (Simmonds et al., 1993a, b, 1994a; Bhattacherjee et al., 1995), and additional sequence at the 3'NCR beyond the poly(U) tract has been determined from the particular sample used in the present study (Kolykhalov et al., 1996). To enable a more complete comparison of type 4 with other variants of HCV we have determined the sequence for the ORF and 3'NCR, to the poly(U) tract, of HCV isolate ED43, genotype 4, from an Egyptian residing in Saudi Arabia, which represents the first complete sequence for this genotype.

### Methods

*Hepatitis C virus isolate ED43.* Isolate ED43 corresponds to the HCV genotype 4 described previously as type 4 or type 4a (Simmonds et al., 1993a, b). Fifty ml of plasma was donated by a 33-year-old Egyptian who was found to be anti-HCV positive (confirmed by RIBA-2) during a routine check-up. In 1982 he had received multiple blood transfusions following a car accident in Cairo. There was no history of overt hepatitis and he remained asymptomatic. However, his liver enzymes were elevated during the 1 year follow-up period prior to collection of this sample.

**RNA extraction and PCR.** HCV RNA was extracted from 100 µl of plasma and amplified by PCR using nested sets of primers following the method of Chan et al. (1992) with the following modifications; 5 µl of RNA was used to generate cDNA in a 20 µl reaction at 42 °C for 1 h with 5 units of avian myeloblastosis virus reverse transcriptase. Five µl of the RT reaction were used for the PCR in a 50 µl reaction. The primary amplification was carried out for 2 min at 80 °C, followed by 30 cycles of 18 s at 94 °C, 21 s at 45 °C and 1:5 min at 72 °C, and a final 6 min at 72 °C. One µl of the primary amplification was used for the secondary reaction, in a final volume of 20 µl, using identical cycling conditions.

Initial fragments (a–d in Fig. 1) were generated using primers already described in previous studies (Simmonds et al., 1993a, b, 1994a; Bhattacherjee et al., 1995). The PCR products were subjected to electrophoresis in a 1% low melting point agarose gel, visualized by ethidium bromide staining and UV illumination and purified by glass milk extraction (GeneClean; Bio101). The fragments were then blunt-end cloned into Smal-digested pUC119 DNA and sequenced. The nucleotide sequence for fragment c was used to design a specific set of primers used...
in conjunction with oligo(dT) to generate fragment f. Fragments g–r were generated using ED43-specific primers for one end of the fragment based on pre-determined sequence and primers made to conserved sequences for the other end, determined by aligning previously published sequences (see Table 1 for accession numbers). Finally, fragment s was produced using ED43-specific primers determined from sequences of fragments l and g, and fragment t generated also by using ED43-specific primers based on sequences from fragments r and i (all primer sequences available from the authors by request). At least two independent clones of each amplified DNA fragment were used for sequence determination; ten nucleotide differences were observed between different clones in the approx. 9.3 kb sequence obtained.

**Sequence assembly and analysis.** Sequences were joined using the Sequence Assembly Program (SAP) from the Staden package (Staden, 1987). Phylogenetic analysis of the completed sequence was carried out using the DNADIST and NEIGHBOR programs from the PHYLIP package (Felsenstein, 1993).

### Results and Discussion

A series of overlapping clones was generated that provided a continuous stretch of 9364 nt from the 5′NCR to the 3′NCR (Fig. 1). Analysis of the sequence showed a single ORF encoding a polyprotein of 3008 aa which is smaller than for other genotypes. Isolates belonging to genotypes 1a and 1c have a polyprotein of 3011 aa while the polyproteins of 1b isolates are 3010 aa (with the exception of isolate HCV-N which is 3014 aa due to a 4 aa insertion in NS5B). The polyproteins of genotype 2a and 2b isolates are 3033 aa, that of 2c is 3037 aa, compared with 3021 aa for genotype 3a, 3023 aa for 3b, 3021 aa for 3b, 3022 aa for JK049 and 3016 aa for JK046. The variation in size of the polyprotein is attributable to differences in the length of the E2 and NS5A genes (Fig. 2) with the exception of isolates HCV-Tr (genotype 3b) and HCV-JK046 (genotype 11a) which have an insertion and deletion, respectively, in E1. Isolate HCV-ED43 has an E2 protein of identical length to that predicted for types 1a, 1b and 1c, while isolates 11a (HCV-JK046, acc. no. D63821) and 11b (HCV-JK046, acc. no. D63822) have a polyprotein of 3011 aa while the polyproteins of 1b isolates are 3010 aa (with the exception of isolate HCV-N which is 3014 aa due to a 4 aa insertion in NS5B). The polyproteins of genotype 2a and 2b isolates are 3033 aa, that of 2c is 3037 aa, compared with 3021 aa for genotype 3a, 3023 aa for 3b, 3021 aa for 3b, 3022 aa for JK049 and 3016 aa for JK046. The variation in size of the polyprotein is attributable to differences in the length of the E2 and NS5A genes (Fig. 2) with the exception of isolates HCV-Tr (genotype 3b) and HCV-JK046 (genotype 11a) which have an insertion and deletion, respectively, in E1. Isolate HCV-ED43 has an E2 protein of identical length to that predicted for types 1a, 1b and 1c, while all the other genotypes have larger E2 proteins. The amino terminus of E2 forms the hypervariable region (Weiner et al., 1991) (amino acid residues 385–416; based on HCV-1 numbering, Fig. 2a). There is very little conservation of sequence in this region and isolate JK046 (genotype 11a, Tokita et al., 1996) even has an amino acid deletion. Conservation of the amino acid sequence is not seen until the N-linked glycosylation site at residues 417–419. The NS5A protein in isolate ED43 is the smallest of all the HCV isolates with apparent deletions at several places in the protein (Fig. 2b), including a four-amino-acid deletion within the predicted

<table>
<thead>
<tr>
<th>Genotype</th>
<th>n</th>
<th>1a</th>
<th>1b</th>
<th>1c</th>
<th>2a</th>
<th>2b</th>
<th>2c</th>
<th>3a</th>
<th>3b</th>
<th>jk049</th>
<th>jk046</th>
<th>ED43</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>4</td>
<td>15.2</td>
<td>14.5</td>
<td></td>
<td>28.1</td>
<td>28.4</td>
<td>27.5</td>
<td></td>
<td>25.7</td>
<td>26.1</td>
<td>25.6</td>
<td>24.7</td>
</tr>
<tr>
<td>1b</td>
<td>13</td>
<td>21.6</td>
<td>15.8</td>
<td></td>
<td>27.6</td>
<td>28.6</td>
<td>27.9</td>
<td></td>
<td>24.9</td>
<td>26.0</td>
<td>25.0</td>
<td>24.1</td>
</tr>
<tr>
<td>1c</td>
<td>1</td>
<td>20.1</td>
<td>22.1</td>
<td></td>
<td>27.6</td>
<td>28.0</td>
<td>27.6</td>
<td></td>
<td>25.2</td>
<td>25.5</td>
<td>24.3</td>
<td>24.3</td>
</tr>
<tr>
<td>2a</td>
<td>1</td>
<td>31.4</td>
<td>31.0</td>
<td>31.1</td>
<td>15.9</td>
<td>13.2</td>
<td></td>
<td>29.2</td>
<td>30.3</td>
<td>28.4</td>
<td>28.9</td>
<td>29.3</td>
</tr>
<tr>
<td>2b</td>
<td>1</td>
<td>32.2</td>
<td>32.2</td>
<td>32.0</td>
<td>22.9</td>
<td>16.2</td>
<td></td>
<td>29.4</td>
<td>30.7</td>
<td>29.0</td>
<td>29.2</td>
<td>29.4</td>
</tr>
<tr>
<td>2c</td>
<td>1</td>
<td>30.9</td>
<td>31.5</td>
<td>31.4</td>
<td>19.8</td>
<td>22.4</td>
<td></td>
<td>29.9</td>
<td>31.0</td>
<td>29.7</td>
<td>29.0</td>
<td>29.4</td>
</tr>
<tr>
<td>3a</td>
<td>2</td>
<td>31.1</td>
<td>30.9</td>
<td>30.8</td>
<td>32.4</td>
<td>33.4</td>
<td>33.6</td>
<td></td>
<td>14.7</td>
<td>17.8</td>
<td>26.3</td>
<td>26.7</td>
</tr>
<tr>
<td>3b</td>
<td>1</td>
<td>31.3</td>
<td>31.2</td>
<td>31.0</td>
<td>33.7</td>
<td>34.0</td>
<td>33.7</td>
<td></td>
<td>21.8</td>
<td>18.7</td>
<td>27.0</td>
<td>27.8</td>
</tr>
<tr>
<td>jk049 (10a)</td>
<td>1</td>
<td>31.1</td>
<td>31.0</td>
<td>30.8</td>
<td>32.9</td>
<td>32.9</td>
<td>33.1</td>
<td></td>
<td>24.8</td>
<td>25.0</td>
<td>25.9</td>
<td>26.1</td>
</tr>
<tr>
<td>jk046 (11a)</td>
<td>1</td>
<td>29.6</td>
<td>29.6</td>
<td>29.8</td>
<td>32.2</td>
<td>32.8</td>
<td>31.8</td>
<td></td>
<td>31.0</td>
<td>32.1</td>
<td>31.3</td>
<td>25.9</td>
</tr>
<tr>
<td>ED43 (4)</td>
<td>1</td>
<td>28.4</td>
<td>28.9</td>
<td>28.7</td>
<td>31.9</td>
<td>32.1</td>
<td>32.5</td>
<td></td>
<td>31.5</td>
<td>31.2</td>
<td>31.5</td>
<td>30.7</td>
</tr>
</tbody>
</table>

**Table 1. Pairwise comparisons using the complete genomic from different genotypes of HCV**

interferon-sensitivity determining region (amino acid residues 2214–2250, based on HCV-1 numbering; Enomoto et al., 1995). A similar deletion is also seen in the type 2a and 2b isolates although not in 2c. However, all the type 2 isolates did show the characteristic insertion of 16 amino acids at position 2402 either side of highly conserved regions. The type 1 isolates had a characteristic deletion at residue 2356 which was also the site of a similar deletion in ED43.

Certain important functional regions were maintained for all isolates: for example the catalytic triad of His-X$_{53}$-Asp-X$_{57}$-Ser for the serine protease (NS3) (Grakoui et al., 1993) was conserved in isolate ED43, as in all other genotypes. The serine protease-dependant cleavage sites (NS3/NS4A, NS4A/NS4B, NS4B/NS5A and NS5A/NS5B) were also very well conserved between isolates. The consensus sequence of an acidic residue (Asp or Glu) at the P6 position and cysteine or threonine at P1 with a serine or alanine residue at P1′ (Grakoui et al., 1993) were all identified in isolate ED43. The NS2/NS3 cleavage site (Grakoui et al., 1993) was also conserved in this isolate as with all the other genotypes. The cleavage sites for the other proteases (C/E1, E1/E2 and E2/NS2) were less well conserved. The core protein carboxy terminus of isolate ED43 consisted of Pro-Ala-Ser-Ala, which is also seen in all the genotype 1 isolates with the exception of HCV-JK1. The E2 protein amino terminus forms part of the hypervariable region and so is not conserved between isolates (Fig. 2) whereas the carboxy terminus of E2 (Ala-Tyr-Ala) was. There was also a strong conservation of potential N-linked glycosylation sites in the envelope proteins. In the E1 protein of HCV-ED43 four potential sites were identified, three of which were conserved between all isolates examined and in the E2 protein there were ten potential sites, of which seven were conserved between all isolates.

Phylogenetic analysis of the complete genomic sequence of ED43 revealed a similar relationship to other genotypes as found previously from sequence comparisons of this variant in subgenomic regions such as NS5 and NS4 (Bhattacherjee et al., 1995; Simmonds et al., 1993b) (Fig. 3). The branch leading to type ED43 originates close to the origin of the tree, and is of similar overall length to other genotypes. A slightly closer relationship between type ED43 and type 1 variants than to other genotypes is implied by the short branch leading to their common ancestor on the phylogenetic tree that has a bootstrap value of 86%, and is reflected in a slightly greater degree of nucleotide sequence similarity than found between other genotypes (Table 1). For example, type ED43 shows a divergence of 28.4–28.7% over the length of the genome compared with the range of 30.7–32.5% between other genotypes. However, type ED43 and type 1 sequences were more divergent from each other than the divergence seen between the most distantly related subtypes (22-9% for 2a compared with 2b), and more divergent than type 3 sequences are from JK049, which was originally proposed to be a new major genotype (24.8–25%) (type 10a; Tokita et al., 1996, and discussed in Simmonds et al., 1996).

Overall, the sequence of ED43 confirms its previously proposed classification (Simmonds et al., 1993a, b) and, by implication, of other similar isolates from Central Africa (Stuyver et al., 1994; Bukh et al., 1993) as examples of genotype 4. The problem that remains is the nomenclature of the numerous subtypes. The first published description of type 4 was based upon sequence comparisons of HCV variants similar to ED43 obtained from Egyptian blood donors (Simmonds et al., 1993a). However, variants termed type 4a, 4b, 4c and 4d on the basis of sequence comparisons in E1 were subsequently described amongst HCV infected individuals from Zaire (Bukh et al., 1994), even though their further analysis has indicated that type 4a from Egypt is a different subtype from the type 4a from Central Africa (Stuyver et al., 1994; Bukh et al., 1995). However, irrespective of nomenclature
there are a number of reasons to regard the type 4 variant from Egypt as the prototype sequence for type 4. Firstly, it was the first type 4 variant described, based upon its identification from comparisons in the core region (Simmonds et al., 1993a). Secondly, this paper is the first to describe the complete genome sequence of a type 4 variant. Thirdly, and more importantly, this variant is by far the most widely distributed of type 4 sequences, being the principal genotype in the Middle East, and accounts for a major proportion of cirrhosis and hepatocellular carcinoma in these populations. Recently, the ED43 isolate was selected for use as a candidate vaccine strain to provide titered challenge stocks with which to investigate cross-immunity between different genotypes (S. Feinstone, unpublished data). The determination of its complete sequence will be of value in exploring the antigenic differences of structural and non-structural proteins between different genotypes, and provides the resource to enable subunit vaccines based upon type 4 to be developed in the future.

We wish to express our gratitude to the donor of the plasma used in this study. This work was supported by a project grant from the Medical Research Council to R.M.E. and P.S.

References


Received 14 November 1996; Accepted 4 February 1997