Genomic structure of the human prototype strain H of hepatitis C virus: Comparison with American and Japanese isolates

(RNA virus/polypeptide/Flaviviridae/Flavivirus/Poliovirus)

Genaivee Inchaustegi+, Suzanne Zebdedee‡, Dong-Hun Lee§, Masahiko Sugitani*, Marc Nasoff‡, and Alfred M. Prince*

*Lindsey F. Kimball Research Institute of The New York Blood Center, Laboratory of Virology and Parasitology; ‡Laboratory of Microchemistry, 310 East 67th Street, New York, NY 10021; and §Pharmacia Genetic Engineering, Inc., 10280 North Torrey Pines Road, Suite 485, La Jolla, CA 92037

Communicated by Rudi Schmid, August 12, 1991

ABSTRACT Genomic RNA from the human prototype strain H of the hepatitis C virus (HCV-H) has been molecularly cloned and sequenced. The HCV-H sequence reported consists of 9416 nucleotides including the 5' and 3' untranslated regions. HCV-H shows 96% amino acid identity with the American isolate HCV-1 but only 84.9% with the Japanese isolates HCV-J and HCV-BK. In addition to the hypervariable region (region V) previously identified in the putative E2 domain, three other variable domains were identified: region V₁ (putative E1), region V₂ (putative E2), and region V₃ (putative NS5). These regions appear to be conserved (86–100%) among the American isolates (HCV-1 and HC-J1) or among various Japanese isolates (HCV-J, HCV-BK, HCV-JH, and HC-J4) but show striking heterogeneity when the two subgroups are compared (42–87.5% amino acid difference). A structural similarity between the 5'-terminal hairpin structure of HCV and of poliovirus was observed. This study further suggests the existence of at least two genomic subtypes of HCV and confirms a distant relationship between HCV and pestiviruses.

Following the introduction of diagnostic methods to detect hepatitis B virus and hepatitis A virus, it was demonstrated that a third class of infectious agents referred to as non-A, non-B (NANB) hepatitis virus was responsible for 60–90% of posttransfusion NANB hepatitis (1–3). A predominant agent of NANB hepatitis has recently been identified and termed hepatitis C virus (HCV) (4). Expression of a HCV cDNA clone in yeast led to the development of assays that detect HCV antibodies in a high proportion of serum samples from patients with chronic NANB hepatitis (5–7). These data suggest that HCV causes most if not all blood-borne NANB hepatitis.

Isolation of HCV cDNA clones has been reported by a number of researchers in the United States and in Japan (4, 8–11). In addition, the cloning and complete genomic sequence of three distinct HCV isolates, one American and two Japanese (HCV-1, HCV-J, and HCV-BK), has been reported (12–14). These studies indicate that HCV is a positive-strand RNA virus with a genome size of ~10 kilobases (9401–9416 nucleotides (nt)) encoding a single continuous translationally open reading frame (3010–3011 amino acid (aa) residues). Analysis of the three HCV genomic sequences reveals homologies at the nucleotide sequence, amino acid sequence, and hydrophaticity-plot levels with the nonstructural (NS) proteins of the flaviviruses, as well as with animal pestiviruses and some plant viruses (15). Preliminary sequence comparisons using the existing data base indicate that the HCV family is a rather heterologous family of viruses (85–96% homology), exhibiting domains of great variability, in particular in the putative envelope region (16).

The American HCV prototype strain H (HCV-H) originally obtained from a patient with chronic posttransfusion NANB hepatitis (17) was isolated by inoculation in chimpanzees in 1979 (18) and has been used to carry out numerous experiments in the chimpanzee model. We decided to clone the genome of HCV-H in order to provide basic information that would help to better interpret data accumulated using this isolate. In this article, we report the complete nucleotide sequence of the genomic RNA of HCV-H and compare this with genomic sequences of the other reported HCV isolates.

MATERIALS AND METHODS

Isolation of HCV RNA. Two types of HCV-infected materials obtained from the same animal inoculated with HCV strain H [HCV isolate H77 (19)] were used: chimpanzee liver taken by biopsy during the acute phase of infection (4 weeks postinoculation) and chimpanzee plasma taken between 4 and 13 weeks postinoculation. The plasma had titers of chimpanzee 50% infectious doses 10^3.5–10^5 (CID₅₀) per ml. Nucleic acids were extracted from liver and plasma as described (20–22).

Cloning of HCV cDNA. Procedures for cDNA synthesis have been described in detail (21, 22). Specific oligonucleotide primers derived from published HCV sequences (10, 12, 23, 24) were used to prime the reaction. Selected target sequences were amplified by a PCR-based approach using a variety of oligonucleotide primers as mentioned above. Conditions for amplification were as described (21). Amplified sequences were subsequently isolated, blunt-ended, and inserted into pUC18 or pBluescript (Stratagene) cloning vectors by standard procedures (20).

Sequence Analysis of Cloned cDNA. Clones were sequenced by the dideoxy chain-termination method (25) using a DuPont Genesis 2000 automated sequencer or with the Pharmacia automated laser fluorescent DNA sequencer. To minimize sequencing errors due to PCR artifacts (misreading by Taq polymerase), for each target sequence, at least three independent clones were isolated and sequenced in both directions and their sequences were compared to derive the final consensus sequence representative of the HCV-H genome.

RESULTS

Characterization of cDNA Clones and Primary Structure of HCV-H. Forty cDNA clones were isolated, ranging in size

Abbreviations: HCV, hepatitis C virus; NANB, non-A, non-B; NS, nonstructural; aa, amino acid(s); nt, nucleotide(s).

†To whom reprint requests should be addressed.
‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. M67463).
between 124 and 615 nt. The putative map location of these clones is shown in Fig. 1. The sequence of the HCV genome was deduced, representing 9416 nt, which is similar in length to that of previously isolated HCV genomes. The sequence has a high G+C content (58.8%) and contains one large open reading frame (nt 1–9033) corresponding to a polyprotein of 3011 aa.

Sequences comprising segments form the 5' and 3' ends of the viral genome (nt 1–359 and 9120–9416) are shown in Fig. 2 A and B, respectively. We were able to identify HCV-H sequences from the 5' and 3' noncoding domains, comprising 341 and 42 nt, respectively. The first 12 nt and the last 20 nt (Fig. 2 A and B, shaded boxes) correspond to the nucleotide primers used in the amplification process and thus are not confirmed as HCV-H sequences. We could not obtain viral sequences when we used an oligo(dT) primer for cDNA synthesis followed by PCR amplification using different combinations of primers, suggesting the absence of internal A-rich tracts at the 3'-terminal end or of a 3'-terminal poly(A) sequence. Similarly, no sequences could be amplified when A-rich primers complementary to the 3'-end (U-rich) nucleotide sequence of the two reported Japanese isolates HCV-J and HCV-BK were used in the cDNA priming reaction, suggesting the absence of a U-rich terminal sequence in the genome of HCV-H. A small hairpin structure could be assigned at the 3' terminus (bottom of Fig. 2 B) with a calculated $\Delta G$ value of $-8\, \text{kcal}$. Independent approaches for the cloning of the 3' or 5' ends of the genome (e.g., inverted PCR) only resulted in the isolation of sequences already comprised in the previously reported HCV sequences.

The large open reading frame is preceded by five AUG codons (nt −329, −310, −256, −245, and −109) followed by stop codons, suggesting the existence of hypothetical small open reading frames in the 5' noncoding region of HCV genomes. Several repeated sequences (Fig. 2, R₁–R₃) were identified in the 5' and the 3' noncoding regions, and in the C-terminal part of the putative NS5 domain.

We identified, as described in previous reports for other HCV isolates, that the HCV-H genome or polyprotein shares only limited similarity with other known viral sequences, except for three domains: (i) a few stretches of nucleotides in the 5' noncoding region identical to those reported for the American prototype HCV-1 (13, 23) that correspond to conserved elements in pestiviruses (Fig. 2 A, dots); (ii) blocks of amino acids found in the putative NS3 domain (nt 3351–4856) corresponding to putative NTP-binding helicase and trypsin-like serine protease also identified in flaviviruses and pestiviruses (refs. 26 and 27; see ref. 13 for position); (iii) the Gly-Asp-Asp consensus sequence conserved among all virus-encoded RNA-dependent RNA polymerases (aa 2737–2739; ref. 28). In addition, 19 putative N-glycosylation sites were located (aa 204, 209, 234, 306, 325, 423, 428, 448, 532,

**Fig. 1.** Schematic representation of the HCV genome and location of HCV-H cDNA clones 0–39. Alignment with the polyprotein encoded by flaviviruses is shown as well as the putative domains in the HCV-encoded polyprotein. Regions of amino acid homology with the dengue type 2 NS3 and the carnation mottle virus (CARMV) replicase are indicated (hatched and open boxes) in the HCV genome. C, capsid; M, matrix; E, envelope; GDD, Gly-Asp-Asp; bp, base pairs.

**Fig. 2.** Nucleotide sequence of the HCV-H genome and deduced amino acid sequence of the 5' (nt 1–359) (A) and 3' (nt 9120–9416) ends. Nucleotides are numbered from the 5' end and amino acids from the first methionine of the polyprotein. Boxes (at the 5' and 3' ends) represent sequences of primers used for the amplification of clones 0 and 39. Black bars indicate the position of identical direct repeats of 6–8 nt (R₁–R₃) found in the 5' and 3' ends. Nucleotides conserved among pestiviruses and HCV-H are marked by dots (A). The putative small hairpin structure at the 3' end is represented below the sequence.
Table 1. Homology of nucleotide and deduced amino acid sequence between HCV-H and heterologous isolates

<table>
<thead>
<tr>
<th>Isolate*</th>
<th>5' NC (−326 to −1) nt</th>
<th>C (1–570) nt</th>
<th>E1 (571–1140) nt</th>
<th>E2/NS1 (1141–2186) nt</th>
<th>NS2 (2187–3017) nt</th>
<th>NS3 (3018–4844) nt</th>
<th>NS4a (4845–5585) nt</th>
<th>NS4b (5586–6038) nt</th>
<th>NS5 (6039–9029) nt</th>
<th>3' NC (9030–9084) nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCV-1</td>
<td>99.7</td>
<td>98.4</td>
<td>98.9</td>
<td>93.5</td>
<td>93.1</td>
<td>96.9</td>
<td>92.9</td>
<td>95.1</td>
<td>95.4</td>
<td>97.2</td>
</tr>
<tr>
<td>HC-J1</td>
<td>99.1</td>
<td>98.9</td>
<td>98.8</td>
<td>93.3</td>
<td>92.6</td>
<td>88.2</td>
<td>95.5</td>
<td>94.8</td>
<td>94.6</td>
<td>96.7</td>
</tr>
<tr>
<td>HC-J4</td>
<td>99.1</td>
<td>98.7</td>
<td>98.7</td>
<td>92.7</td>
<td>94.1</td>
<td>97.6</td>
<td>97.4</td>
<td>97.3</td>
<td>97.5</td>
<td>97.2</td>
</tr>
<tr>
<td>HCV-JH</td>
<td>98.9</td>
<td>98.3</td>
<td>98.4</td>
<td>92.1</td>
<td>88.2</td>
<td>85.7</td>
<td>92.3</td>
<td>96.1</td>
<td>96.2</td>
<td>95.7</td>
</tr>
<tr>
<td>HCV-J</td>
<td>98.0</td>
<td>92.1</td>
<td>98.9</td>
<td>73.9</td>
<td>78.4</td>
<td>97.2</td>
<td>79.3</td>
<td>97.2</td>
<td>92.7</td>
<td>98.0</td>
</tr>
<tr>
<td>HCV-BK</td>
<td>98.8</td>
<td>98.3</td>
<td>98.4</td>
<td>73.8</td>
<td>77.9</td>
<td>98.4</td>
<td>80.4</td>
<td>98.4</td>
<td>97.3</td>
<td>98.0</td>
</tr>
</tbody>
</table>

Nucleotide positions for the various domains are given in parentheses. Positions for C and E1 were deduced from ref. 16, and those for E2 and NS2–NS5 from ref. 14. NC, noncoding; —, not available for comparison.

*References: HCV-1 (13, 23); HC-J1 and HC-J4 (10); HCV-JH (11); HCV-J (12); HCV-BK (14).

540, 556, 576, 623, 645, 1213, 1255, 2041, 2364, and 2789.

These were mostly clustered between aa 196 and 647 and appeared similar in distribution and number to the N-glycosylation sites observed in the case of envelope proteins of pestiviruses.

Comparison of Nucleotide and Amino Acid Sequences of HCV-H and Heterologous HCV Isolates. The comparison between different genomic domains of HCV-H and previously reported HCV isolates is summarized in Table 1. The HCV genome encodes possibly eight proteins: three structural proteins [the nucleocapsid (C) and two envelope (E1 and E2) proteins] and five NS proteins (NS2, NS3, NS4a, NS4b, and NS5), thus delimiting eight putative coding domains. Sequence comparison was only partial with HC-J1, HC-J4, and HC-JH, as the complete sequence of the genome of these isolates has not been reported.

Our data indicated a very high degree of sequence homology in two genomic domains: (i) the 5' noncoding domain (98.2–99.7% nucleotide identity) and (ii) in the C domain (90.0–98.9% nucleotide identity and 97.9–98.9% amino acid identity). Interestingly, two sets of repeated sequences found in the 5' noncoding domain, R1 and R2 (Fig. 2A) [R2 has been reported by Takamizawa et al. (14)] are conserved among all reported isolates. Two copies of the repeated sequence R1 are also conserved between the two American isolates HCV-H and HCV-1, but only one copy is found in both Japanese isolates, HCV-J and HCV-BK. It is possible that the 5' noncoding sequences of the two Japanese genomes simply do not extend far enough to encompass the second copy. The nucleotide sequence reported for the other HCV isolates was too short to allow for comparison.

Regions of moderate identity were found throughout the NS domains, where a clear separation between American and Japanese isolates could be seen. Whereas 93.8–95.9% nucleotide identity was observed when HCV-H was compared with the American isolates, only 72.7–80.4% nucleotide identity was found with Japanese isolates (95.1–97.2% and 78.2–92.6% amino acid identity, respectively). One region, found in the putative NS5 domain (aa 2356–2379) that we have termed region V3, reflected even more striking divergence between the two HCV subgroups. This region was 100% identical at the amino acid level between the two American isolates (data not shown) but only 12.5% identical when compared with Japanese isolates (Fig. 3D). Most of the changes were conservative and thus may not result in functional modification of the protein.

Regions of greater divergence were found in the putative envelope region E1 (nt 571–1140) and E2 (nt 1141–2197), where 77.9–94.1% and 65.6–92.9% amino acid identity, respectively, was observed between HCV-H and the other isolates. In addition to the hypervariable region, or region V, identified by Weiner et al. (16) in E2, in which protein heterogeneity between HCV-H and other HCV isolates ranged from 51.7% to 72.4% (Fig. 3B), we identified two regions of high variability. Both regions, which we have termed regions V1 and V2 (aa 246–275 and 456–482, respectively) appeared very conserved among American or Japanese HCV (86–96% identity) but showed much greater heterogeneity when the two groups were compared (55–58% identity; Fig. 3A and C).

**DISCUSSION**

The human prototype strain H of HCV (HCV-H) was one of the originally identified isolates of the virus (18). We report here the cDNA sequence of HCV-H (9416 nt). To date, this is the second nucleotide sequence of a HCV genome determined for a single prototype strain, as the two other sequenced Japanese strains, HCV-J and HCV-BK (12, 14), were derived from clones isolated from pooled plasma and may therefore represent genomic sequences from diverse isolates.

The genome of HCV-H shows an overall amino acid identity of 96% with the American prototype HCV-1 and 84.9% with both HCV-J and HCV-BK isolates. Three new regions of high variability were identified within E1, E2, and NS5 (regions V1, V2, and V3). In these regions, sequence heterogeneity appears to be subgroup-specific (i.e., American vs. Japanese isolates), in particular for region V3, where up to 87.3% divergence was found between the two subgroups. Sequence heterogeneity has been observed in the envelope/NS1 regions of flaviviruses but not to the extent reported here for regions V1 and V2. That three of four variable regions within the HCV genome are located in the putative envelope domain suggests that these domains might be under immunological pressure. It remains to be determined whether the putative E1 and E2/NS1 regions of HCV are immunogenic and whether they can elicit protective immunity as seen with envelopes of pestiviruses (29, 30) or flaviviruses (31, 32). Possibly, the heterogeneity found at the structural level in these domains as well as within the NS5 domain in region V3 may also reflect antigenic heterogeneity among different HCV isolates.

Ogata et al. (19) recently published partial sequence of the original HCV-H strain directly isolated from patient H in 1977 and prior to transmission to chimpanzee. Comparison with the HCV-H sequence obtained in our study indicates 98.5% identity in the structural domain (aa 1–778), 98.8% identity in the NS3 domain (aa 1209–1636), and 98.8% identity in the NS5 domain (aa 2530–2869). The greatest rate of divergence between the two isolates was found in the hypervariable region V, with 2.5% nucleotide and 6.8% amino acid difference respectively, whereas 100% identity was observed in both regions V1 and V2 (region V3 was not available for
Despite geographical separation, different HCV isolates displayed nearly complete identity in the 5′ noncoding region. A similar observation has been made in flaviviruses that are members of the same serologically related subgroup (34), whereas members of different antigenic subgroups share only low levels of homology in that region. This is in contrast to what has been reported for polioviruses, for which sequence conservation is very high in the 5′ noncoding region of different serotypes (35). Antigenic grouping of HCV isolates remains to be determined.

No poly(A) sequences were identified in this study, in contrast to the report by Han et al. (23). Attempts to identify subgenomic poly(A) + RNA species for HCV-H also failed. We cannot exclude the possibility that our starting material (liver tissue or plasma) may not have contained all forms of viral RNA, although this seems unlikely. None of the HCV genomes so far reported contain the AAUAAA motif in the 3′-terminal noncoding sequence that has been implicated as polyadenylation signal for cellular mRNA (36), although poly(A) tails could be genetically coded. The HCV-H 3′-terminal sequence seems to be more closely related to the structure of Japanese HCV isolates, but because of the complex secondary structure of the HCV 3′ terminus these results may not be definitive.

The 5′ hairpin structure originally described by Han et al. (23) for HCV-1 has been also identified for HCV-H (nt 346–287). No hairpin loops have been reported at the 5′ end of pestivirus genomes, possibly because the authentic 5′-terminal sequence of these viruses has not been cloned. By contrast, highly conserved hairpin structures have been located at the 5′ end of viral RNA from enteroviruses, rhinoviruses, and picornaviruses. These structures have been implicated in translational control and efficient viral replication. Recently, Simoes and Sarnow (37) demonstrated that the 5′-proximal RNA hairpin may play a major role in the life cycle of poliovirus, since infectious cDNA clones mutated in that region were replication-deficient. We found that the HCV 5′ hairpin structure displays an interesting degree of homology (in terms of nucleotide sequence and configuration) with the structure described by those authors (Fig. 4). The homology observed here suggests that this sequence may represent the authentic 5′ end of the HCV genome. Based on the original observation by Kato et al. (12) [i.e., the short length of the 3′ noncoding region of HCV (54 nt) and the presence of a poly(U) stretch at the 3′ end] and our observation concerning the 5′ hairpin structure, HCV appears to show interesting homologies with both terminal ends of polioviruses. In addition, for HCV, a high number of unused AUG codons (up to eight) preceding the initial start

**Biochemistry: Inchauspe et al.**

**FIG. 3.** Alignment of the deduced amino acid sequences of variable and hypervariable regions from heterologous HCV isolates. Sequences from the HCV-H genome (top line) are compared with counterpart sequences from other HCV isolates (see Table 1 for references). Identical amino acid residues are denoted by dots whereas amino acid changes are indicated. Nonconservative changes are denoted by boxes. (A) Region V1 (in E2, aa 246–275). (B) Region V (in E2, aa 456–482). (C) Region V2 (in NS5, aa 2356–2379).

**FIG. 4.** Comparison of predicted RNA hairpin structures for poliovirus (37) and HCV genomes. Conserved nucleotides between the two structures are delimited by a dotted line.
codon have also been identified in the 5' noncoding region of polioviruses (38, 39). Thus HCVs may share some structural and possibly functional similarities with polioviruses with respect to cis-acting elements involved in the control of viral replication. Clearly, there is a need for an in vitro tissue culture system and/or the development of HCV infectious cDNA clones to allow further exploration of this hypothesis.

In conclusion, our data confirm the existence of at least two distinct genomic subtypes of HCV (i.e., American- and Japanese-type isolates) and suggest a similarity between structural and functional elements implicated in the regulation of viral replication for HCV, pestiviruses, and polioviruses.

We thank Dr. L. Andrus and Dr. S. Lustgman for helpful discussion during the course of this study. We thank Dr. R. Fields and Dr. K. Knowlbauch for computer assistance. This work was supported in part by a grant from Pharmacia Genetic Engineering, Inc.