Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 9524–9528, December 1990
Biochemistry

Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis

(RNA virus/gene structure/nucleotide sequence)

NOBUYUKI KATO, MAKOTO HIJIKATA, YUKO OOTSUYAMA, MASANORI NAKAGAWA, SHOWGO OHKOSHI, TAKASHI SUGIMURA, AND KUNITADA SHIMOTOHNO

Virology Division, National Cancer Center Research Institute, 5-1-1, Tsukiji, Chuo-ku, Tokyo 104, Japan

Contributed by Takashi Sugimura, August 31, 1990

ABSTRACT **The nucleotide sequence of the Japanese type of hepatitis C virus (HCV-J) genome, consisting of 9413 nucleotides, was determined by analyses of cDNA clones from plasma specimens from Japanese patients with chronic hepatitis. HCV-J genome contains a long open reading frame that can encode a sequence of 3010 amino acid residues. Comparison of HCV-J with the American isolate of HCV showed 22.6% difference in nucleotide sequence and 15.1% difference in amino acid sequence. Thus HCV-J and the American isolate of HCV are probably different subtypes of HCV. The relationship of HCV-J with other animal RNA virus families and the putative organization of the HCV-J genome are discussed.**

Non-A, non-B hepatitis (NANBH), which occurs in some patients after blood transfusion, is caused by viral infection and frequently develops into chronic hepatitis and cirrhosis (1) and hepatocellular carcinoma (2). A cDNA clone derived from the most probable etiological agent, named hepatitis C virus (HCV), was obtained by immunoscreening a cDNA library derived from the plasma of a chimpanzee chronically infected with HCV (3). By using an anti-HCV ELISA with recombinant HCV antigen (C100) expressed in yeast, serum antibody against C100 protein was detected in 70–80% of the chronic NANBH patients examined in Japan, Italy, and the United States (4). Furthermore, a high incidence of the antibody in patients with hepatocellular carcinoma was also reported (5–7).

The HCV genome is considered to be a positive-stranded RNA molecule of about 10 kilobases and its partial nucleotide sequence (7310 nucleotides) has been determined (8). A search for sequence homology of the HCV genome indicated that this virus is slightly related to flaviviruses and pestiviruses (9). We have frequently detected the HCV genome in plasma and liver specimens from patients with NANBH by reverse transcription followed by the polymerase chain reaction (RT–PCR) (10, 11) and found that Japanese isolates (HCV-J) showed sequence variations from the original isolate in the United States (HCV-US) (10, 12). On RT–PCR sequence analyses of the region of the putative nonstructural protein 5 (NS5) of 19 HCV-J isolates, we found that 14–17% of the nucleotide sequence differed from that of HCV-US (12). The nucleotide sequences of the HCV-J isolates showed several characteristic common differences from that of HCV-US, although 2.5–11% sequence diversity was detected among the HCV-J isolates (12). These results suggested the existence of at least two subtypes of HCV in the world. To clarify this point, we molecularly cloned the almost whole HCV-J viral genome and compared its nucleotide sequence* with the published portions of genome sequence of HCV-US (8).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

METHODS

Preparation of Nucleic Acids. RNAs from samples of plasma from NANBH patients with high values of alanine aminotransferase were prepared as described (10).

RT–PCR. RT–PCR using RNA from plasma was carried out as described (12, 13). Oligonucleotide primers were designed from the nucleotide sequence of the HCV-J or HCV-US (8, 10, 12).

Construction of cDNA Libraries. RNAs from plasma samples of nine patients with NANBH, in which the HCV genome was detected by the RT–PCR, were used for making the cDNA libraries. A cDNA library of these RNAs was prepared in λgt11 using several primers, which were designed from the nucleotide sequence of HCV-J obtained by RT–PCR, and a cDNA synthesis kit from Bethesda Research Laboratory (14). For isolation of the 3′ end of the HCV-J genome, another cDNA library was prepared using oligo(dT) primer, after polyadenylylation of plasma RNAs with *Escherichia coli* poly(A) polymerase (15).

Cloning and Sequencing. RT–PCR products were molecularly cloned as described (10, 12). To avoid the possibility of misreading by *Thermus aquaticus* polymerase, we isolated three clones and determined their nucleotide sequences by the dideoxy-nucleotide chain-termination method (DNA sequencing kit, United States Biochemical). cDNA libraries were screened by the method of Benton and Davis (16) with cDNA fragments of the HCV-J genome obtained by RT–PCR as probes. The cDNA inserts of λgt11 obtained were subcloned into pTZ19R plasmid vector and used for sequence analysis.

RESULTS

Nucleotide Sequence and Deduced Amino Acid Sequence of the HCV-J Genome. The nucleotide sequences of clones 1–23, which were obtained by the RT–PCR and cloning of the cDNA library, were determined (Fig. 1). Adjacent sequences of these clones were overlapped to construct the sequences of the almost whole HCV-J genome of 9413 nucleotides. The base composition of this sequence is 20.0% adenine, 21.4% thymine, 30.0% cytosine, and 28.6% guanine. This sequence has a single open reading frame (ORF) of 9030 nucleotides beginning at the first ATG at nucleotide 330. This ORF spans virtually the entire length of the RNA, excluding 329 nucleotides of 5′ untranslated region and only 54 nucleotides of 3′ untranslated region. No other long ORF that is able to code for more than 134 amino acid residues was found in either the

Abbreviations: HCV, hepatitis C virus; NANBH, non-A, non-B hepatitis; HCV-J, Japanese type of HCV; HCV-US, American type of HCV; PCR, polymerase chain reaction; RT, reverse transcription; RT–PCR, RT followed by PCR; NS, nonstructural; ORF, open reading frame.

*The sequence reported in this paper has been deposited in the GenBank data base (accession no. D90208).

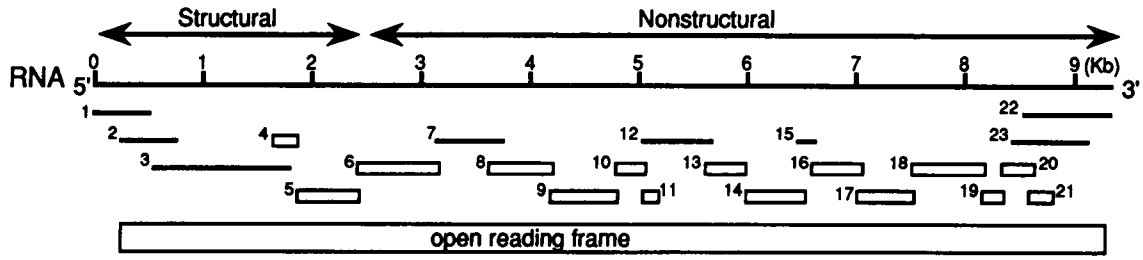

FIG. 1. Physical maps of HCV-J-derived cDNA clones chosen for nucleotide sequencing and their positions relative to the RNA genome. Thick bars indicate the cDNA clones obtained from cDNA libraries and open boxes indicate the cDNA clones obtained by RT–PCR. Assignment of the structural region is described in the *Discussion*.

sense or the complementary strand of the RNA genome. The complete nucleotide sequences of the HCV-J cDNA clones and the deduced amino acid sequence of the long ORF are shown in Fig. 2.

Comparison of Nucleotide and Deduced Amino Acid Sequences of the HCV-J and HCV-US Genomes. The nucleotide sequences of the HCV-US genome from residue 1282 to residue 1672 and from residue 1678 to residue 9178, which have been reported (8, 17), were compared with the corresponding sequences in HCV-J. The percentage homologies in blocks of 99 nucleotides or 33 amino acids were calculated, and the values are shown schematically in Fig. 3. No region shows more than 90% homology of nucleotide sequences, but at the amino acid sequence level several portions, including the central and C-terminal portions, of the putative protein show nearly 100% homology with those of the HCV-US genome. Nucleotide changes appear to be distributed throughout the genome, with divergences ranging from 10.1 to 39.4% in the translated region. Only 22.3% of the base changes result in amino acid substitutions. The difference in nucleotide sequences between HCV-J and HCV-US is 22.6% and the difference in amino acid sequences is 15.1%. Notably, one amino acid is missing between positions 2413 and 2414 of HCV-J, as judged by comparison with the sequence of HCV-US (Fig. 2). This amino acid was also missing in three other cDNA clones derived from three patients (data not shown), and so this deletion appears to be characteristic of the HCV-J genome.

DISCUSSION

In this report, HCV-J was molecularly cloned from the plasma of NANBH patients by the RT–PCR method and from cDNA libraries primed with specific primers, and the nucleotide sequences of the cDNAs obtained were determined. We found that the HCV-J genome has a long ORF that can encode a polyprotein of 3010 amino acids and shows sequence differences from HCV-US. Previous analyses of limited regions of HCV-J, which corresponds to the NS5 of flavivirus and is the most conservative domain, showed sequence differences in HCV-J and HCV-US (10, 12). In the present study, we found that sequence differences were located throughout the genomes and that approximately 23% of the nucleotide sequences of the two is different. Therefore, we propose that HCV-J and HCV-US are different subtypes of HCV.

A homology search using the partial sequence of the HCV-US genome indicated that HCV is related to flaviviruses and pestiviruses (9). As for HCV-US (9), the Ala-Gln-Arg-Arg-Gly-Arg-Xaa-Gly-Arg amino acid sequence (amino acid residues 1485–1493 in this work) is conserved in HCV-J, Dengue virus (18), Kunjin virus (19), Japanese encephalitis virus (15), yellow fever virus (20), West Nile virus (21), hog cholera virus (22), and bovine viral diarrhea virus (23), except that, in HCV-US, the first position of this conserved amino acid sequence is threonine instead of alanine. This conserved sequence is located in the central portion of the NS3 protein and is specific in flaviviruses and pestiviruses. HCV-J polyprotein and the NS5 proteins of flaviviruses and pestiviruses contain the third motif (Gly-Asp-Asp motif, amino acid residues 2736–2738 in HCV-J) of the four conserved motifs in RNA-dependent-polymerase-encoding elements (24). The distance between Ala-Gln-Arg-Arg-Gly-Arg-Xaa-Gly-Arg and the Gly-Asp-Asp motif in HCV-J polyprotein is 1243 amino acids. This distance is more like those in flaviviruses (amino acids 1213–1228) than those in pestiviruses (amino acids 1528–1529).

The polyprotein of HCV-J could be processed by cellular proteases (including signalase or Golgi proteases) and viral protease, as proposed for other flaviviruses (25). The first 120 amino acids of the N-terminal portion of the ORF are rich in basic amino acids (arginine and lysine, 23%) and proline (13%), like those in the core proteins of other flaviviruses (25). Furthermore, 15 putative N-glycosylation sites are clustered from amino acid 196 to amino acid 645, as in the envelope proteins of other pestiviruses (22, 23). The hydrophobicity profile of the N-terminal region is also partially similar to those of flaviviruses (25) and pestiviruses (22, 23). These features of the HCV-J RNA genome suggest that the N-terminal portion (700–800 amino acids) of this virus encodes the viral structural protein. In particular, the viral core protein may be created by cleavage at about amino acid 190, which is located after signal sequences that contain hydrophobic regions preceding the envelope protein.

The structure of the HCV-J genome determined in this work has mosaic features of the Japanese type, because it has been determined from a mixture of nine plasma samples that were positive for the HCV-J genome by RT–PCR. It should be noticed that there is 2.5–11% sequence diversity among the HCV-J isolates in the region of clone 20 (see Fig. 1), which is the most conserved region (12). In addition to 23 cDNA clones from which the sequence of the whole genome of HCV-J was determined, we obtained several other cDNA clones during screening of the cDNA library. The nucleotide sequences of these clones also show approximately 10% sequence diversity from the representative sequence of the HCV-J genome (data not shown).

The nucleotide sequences of two NS regions derived from HCV-J were reported (26, 27). These sequences of 583 (corresponding to residues 4676–5258 of HCV-J) and 269 (corresponding to residues 6856–7124 of HCV-J) nucleotides also showed 7% and 10% sequence diversity, respectively, from that of the HCV-J genome. Furthermore, two other groups reported sequences of the 5′ terminus of the HCV genome; Okamoto *et al.* (28) reported the sequences of two strains, HC-J1 and HC-J4 (corresponding to residues 6–1868 of HCV-J) and Takeuchi *et al.* (29) reported the sequence of one strain (corresponding to residues 240–1652 of HCV-J). All these nucleotide sequences except that of HC-J1 showed more than 90% homology with each other. The sequences of these strains can be included in the category of sequence variations of HCV-J isolates. However, the HC-J1 strain

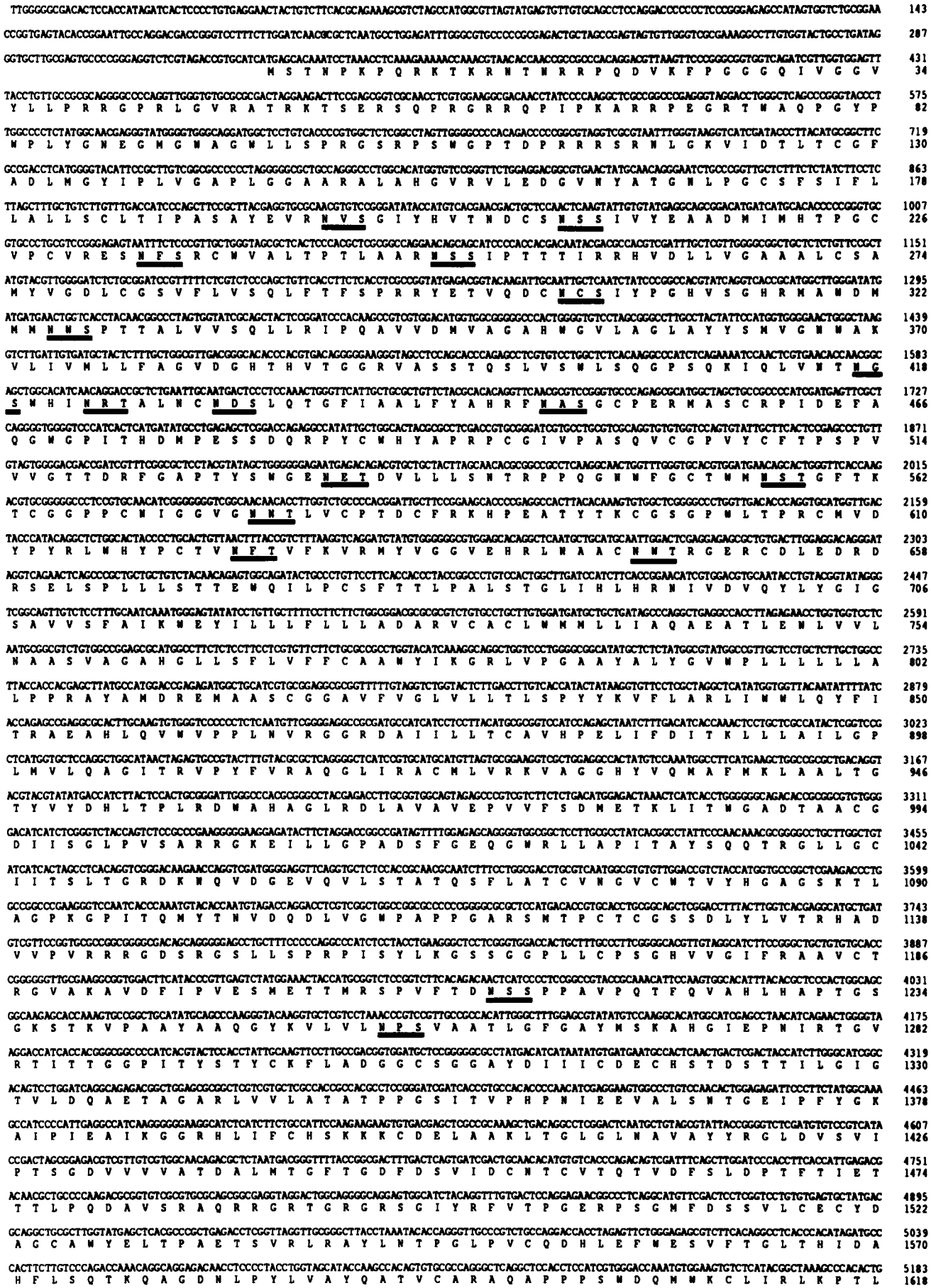

FIG. 2. (*Figure continues on opposite page.*)

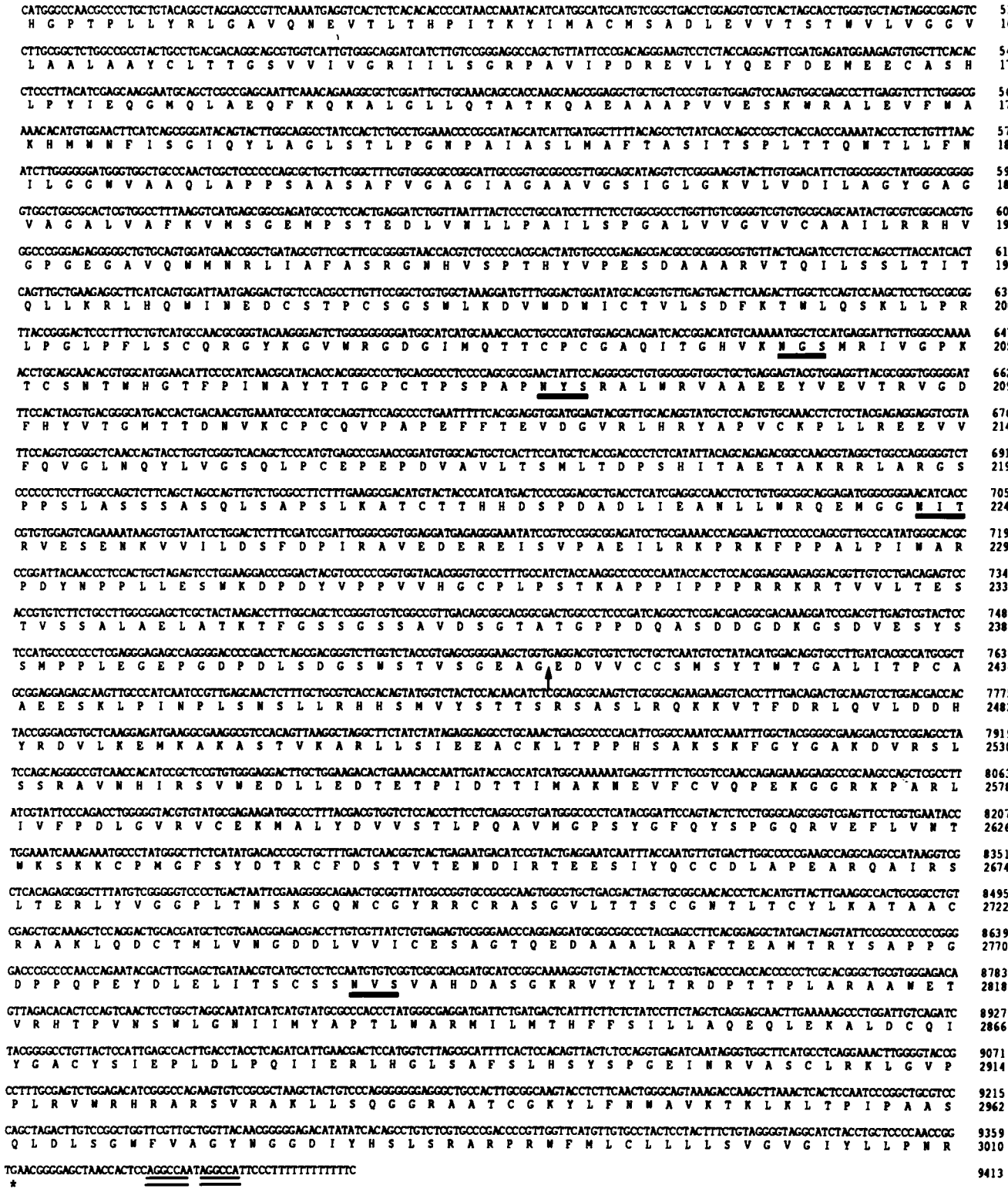

FIG. 2. Nucleotide sequence of HCV-J and deduced amino acid sequence of the long ORF. Nucleotide residues are numbered from the 5′ end of the genomic RNA, and amino acid residues are numbered from the first methionine of polyprotein. Underlines indicate potential sites of N-linked glycosylation. The arrow shows the position of the one amino acid missing compared with the sequence of the HCV-US genome. Repeated nucleotide sequences in the 3′ untranslated regions are double underlined. For nucleotide exchanges between overlapping cDNA clones, upstream cDNA clones are given priority for convenience.

shows only 80% homology with the other HCV-J strains and is highly homologous (more than 95%) with HCV-US. Since we found 2–7.6% sequence diversity among HCV-US isolates derived from plasma samples of Japanese hemophiliacs (M.H. and K.S., unpublished data), HC-J1 must belong to the HCV-US group.

Clone 1 in this work starts at nucleotide −5 of HC-J4, and so this HCV-J cDNA may cover the 5′ terminus of the viral genome. It is noteworthy that we found the thymine stretch in the 3′ terminal portion. The cDNA clone 22, with the sequence of the most 3′ terminal portion of the viral genome, contained TTTTTTTTTTTTC at its 3′ terminus. This se-

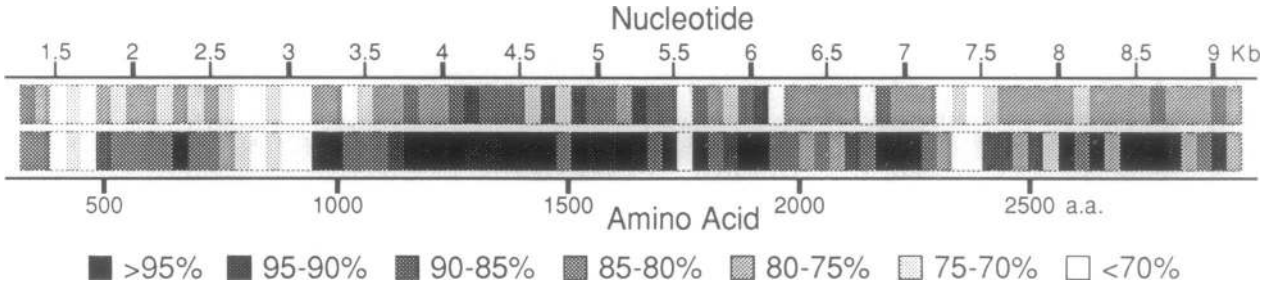

FIG. 3. Homology of nucleotide and deduced amino acid sequences of the HCV-J and HCV-US genomes. Successive sequences of 99 nucleotides or 33 amino acids are compared. Shades indicate degrees of homology. Numbers of nucleotide and amino acid sequences of the HCV-J genome are shown.

quence might be the actual 3′ terminus of the viral genome, although other possibilities are that degradation of the template RNA may have occurred before addition of adenine residues by poly(A) polymerase or that the 3′ terminus is not polymerized because of its strong secondary structure, as observed with other flaviviruses (25). This thymine stretch and the length of the 3′ noncoding region (54 nucleotides) are similar to those of poliovirus (30), but poliovirus has a poly(A) tail. Therefore, the 3′ terminus of the HCV-J genome may differ from those of other viruses including flaviviruses. The genetic information on the HCV-J genome obtained in this study will be useful in future studies, including those on infection, proliferation, the pathogenesis of HCV, and development of a vaccine.

This study was supported by Grants-in-Aid for Cancer Research and for a Comprehensive 10-Year Strategy of Cancer Control from the Ministry of Health and Welfare, Japan. M.H. is recipient of Research Resident Fellowships from the Foundation for Promotion of Cancer Research.

1. Dienstag, J. L. (1983) *Gastroenterology* **85,** 439–462.
2. Sakamoto, M., Hirohashi, S., Tsuda, H., Ino, Y., Shimosato, Y., Yamasaki, S., Makuuchi, M., Hasegawa, H., Terada, M. & Hosoda, Y. (1988) *Cancer Res.* **48,** 7294–7297.
3. Choo, Q.-L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W. & Houghton, M. (1989) *Science* **244,** 359–362.
4. Kuo, G., Choo, Q.-L., Alter, H. J., Gitnick, G. L., Redeker, A. G., Purcell, R. H., Miyamura, T., Dienstag, J. L., Alter, M. J., Syevens, C. E., Tegtmeier, G. E., Bonino, F., Colombo, M., Lee, W.-S., Kuo, C., Berger, K., Shuster, J. R., Overby, L. R., Bradley, D. W. & Houghton, M. (1989) *Science* **244,** 362–364.
5. Brouix, J., Barrera, J. M., Calvet, X., Ercilla, G., Costa, J., Sanchez-Tapias, J. M., Ventura, M., Vall, M., Bruguera, M., Bru, C., Castillo, R. & Rodes, J. (1989) *Lancet* **ii,** 1004–1006.
6. Colombo, M., Kuo, G., Choo, Q.-L., Danato, M. F., Minno, E. D., Tommasini, M. A., Dioguardi, N. & Houghton, M. (1989) *Lancet* **ii,** 1006–1008.
7. Ohkoshi, S., Kojima, H., Tawaraya, H., Miyajima, T., Kamimura, T., Asakura, H., Satoh, A., Hirose, S., Hijikata, M., Kato, N. & Shimotohno, K. (1990) *Jpn. J. Cancer Res.* **81,** 550–553.
8. Houghton, M., Choo, Q.-L. & Kuo, G. (1988) *Eur. Patent Appl.* 88,310,922.5 and Publ. 0,318,216.
9. Miller, R. H. & Purcell, R. H. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2057–2061.
10. Kato, N., Ohkoshi, S. & Shimotohno, K. (1989) *Proc. Jpn. Acad.* **65B,** 219–223.
11. Ohkoshi, S., Kato, N., Kinoshita, T., Hijikata, M., Ootsuyama, Y., Okazaki, N., Ohkura, H., Hirohashi, S., Honma, A., Ozaki, T., Yoshikawa, A., Kojima, H., Asakura, H. & Shimotohno, K. (1990) *Jpn. J. Cancer Res.* **81,** 862–865.
12. Kato, N., Hijikata, M., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S. & Shimotohno, K. (1990) *Mol. Biol. Med. in press.*
13. Kinoshita, T., Shimoyama, M., Tobinai, K., Ito, M., Ito, S., Ikeda, S., Tajima, K., Shimotohno, K. & Sugimura, T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5620–5624.
14. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25,** 263–269.
15. Sumiyoshi, H., Mori, C., Fuke, I., Morita, K., Kuhara, S., Kondou, J., Kikuchi, Y., Nagamatu, H. & Igarashi, A. (1987) *Virology* **161,** 497–510.
16. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
17. Takeuchi, K., Boonmar, S., Kubo, Y., Katayama, T., Harada, H., Ohbayashi, A., Choo, Q.-L., Kuo, G., Houghton, M., Saito, I. & Miyamura, T. (1990) *Gene* **91,** 287–291.
18. Mackow, E., Makino, Y., Zhao, B., Zhang, Y. M., Markoff, L., Buckler-White, A., Guiler, M., Chanock, R. & Lai, C. J. (1987) *Virology* **159,** 217–228.
19. Coia, G., Parker, M. D., Speight, G., Byrne, M. E. & Westaway, E. G. (1988) *J. Gen. Virol.* **69,** 1–21.
20. Rice, C. M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L. & Strauss, J. H. (1985) *Science* **229,** 726–733.
21. Castle, E., Leidner, U., Nowak, T., Wengler, G. & Wengler, G. (1986) *Virology* **149,** 10–26.
22. Meyers, G., Rumenapf, T. & Thiel, H.-J. (1989) *Virology* **171,** 555–567.
23. Collett, M. S., Larson, R., Gold, C., Strick, D., Anderson, D. K. & Purchio, A. F. (1988) *Virology* **165,** 191–199.
24. Poch, O., Sauvaget, I., Delarue, M. & Tordo, N. (1989) *EMBO J.* **8,** 3867–3874.
25. Rice, C. M., Strauss, E. G. & Strauss, J. H. (1986) in *The Togaviridae and Flaviviridae*, eds. Schlesinger, S. & Schlesinger, M. J. (Plenum, New York), pp. 279–326.
26. Kubo, Y., Takeuchi, K., Boonmar, S., Katayama, T., Choo, Q.-L., Kuo, G., Weiner, A. J., Bradley, D. W., Houghton, M., Saito, I. & Miyamura, T. (1989) *Nucleic Acids Res.* **17,** 10367–10372.
27. Maeno, M., Kaminaka, K., Sugimoto, H., Esumi, M., Hayashi, N., Komatsu, K., Abe, K., Sekiguchi, S., Yano, M., Mizuno, K. & Shikata, T. (1990) *Nucleic Acids Res.* **18,** 2685–2689.
28. Okamoto, H., Okada, S,. Sugiyama, Y., Yotsumoto, T., Tanaka, T., Yoshizawa, H., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1990) *Jpn. J. Exp. Med.* **60,** 167–177.
29. Takeuchi, K., Kubo, Y., Boonmer, S., Watanabe, Y., Katayama, T., Choo, Q.-L., Kuo, G., Houghton, M., Saito, I. & Miyamura, T. (1990) *Nucleic Acids Res.* **18,** 4626.
30. Nomoto, A., Omata, T., Toyoda, H., Kuge, S., Horie, H., Kataoka, Y., Genba, Y., Nakano, Y. & Imura, N. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5793–5797.