Complete genomes of hepatitis C virus (HCV) subtypes 6c, 6l, 6o, 6p and 6q: completion of a full panel of genomes for HCV genotype 6

Ling Lu, Chunhua Li, Yongshui Fu, Fengxiang Gao, Oliver G. Pybus, Kenji Abe, Hiroaki Okamoto, Curt H. Hagedorn and Donald Murphy

Division of Gastroenterology/Hepatology, Department of Medicine, University of Kansas Medical Center, 4035 Delp, MS 1023, Kansas City, KS 66160, USA
The First People’s Hospital of Yunnan Province, Kunming, China
Gunagzhou Blood Center, Guangzhou City, Guangdong Province, China
Public Health Laboratories, New Hampshire Department of Health and Human Services, Concord, NH, USA
Department of Zoology, Oxford University, Oxford, UK
Department of Pathology, National Institute of Infectious Diseases, Tokyo, Japan
Division of Virology, Department of Infection and Immunity, Jichi Medical University School of Medicine, Tochigi-Ken, Japan
Institut National de Sante Publique du Quebec, Laboratoire de Sante Publique du Quebec, Sainte-Anne-de-Bellevue, Quebec, Canada

Received 24 December 2006
Accepted 7 February 2007

Hepatitis C virus (HCV) is classified in the genus Hepacivirus of the family Flaviviridae. It has a single-stranded, positive-sense RNA genome of approximately 9600 nt. The genome contains a single open reading frame (ORF) that spans nearly the complete length of the genome. The ORF encodes three structural (core, E1 and E2) and seven non-structural (P7, NS2, NS3, NS4A, NS4B, NS5A and NS5B) proteins (Major & Feinstone, 1997). Phylogenetic analysis has classified HCV into six genotypes and, within each genotype, closely related variants are grouped into a number of subtypes (Robertson et al., 1998). In a recent HCV nomenclature proposal, up to 71 subtypes have been classified; within genotype 6, 17 subtypes in total (6a–6q) have been assigned (Simmonds et al., 2005). Complete genome sequences represent the ‘gold standard’ for HCV classification and molecular epidemiology; they are vital for the identification of putative recombinant strains and can be compared and analysed with any set of subgenomic sequences.

Geographically, genotype 6 variants have been identified exclusively in south-eastern Asia or immigrants from this region (Apichartpiyakul et al., 1994; Bernier et al., 1996; Doi et al., 1996; Mellor et al., 1995, 1996; Noppornpanth et al., 2006a; Shinji et al., 2004; Stryver et al., 1995; Thaikruay et al., 2004; Theamboonlers et al., 2002). In previous studies, complete genomic sequences representing 12 subtypes within genotype 6 have been isolated from

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this study are EF424625–EF424629.

A table showing pairwise nucleotide similarities of the five HCV complete genomic sequences with other genotype 6 subtypes and a diagram outlining the strategy used to amplify the five complete HCV genomic sequences are available as supplementary material in JGV Online.
samples obtained from this region (Tokita et al., 1998; Simmonds et al., 2005; Li et al., 2006; Lu et al., 2006a, 2007). However, five other officially declared subtypes (6c, 6l, 6o, 6p and 6q) have not yet had their genomes sequenced fully. Here, we report and analyse the complete genome sequences of these five subtypes.

Five samples were used in this study. Th846 was from a commercial blood donor in Thailand and represents a subtype 6c infection (Tokita et al., 1995). Sample 537796 was from an Asian immigrant in the USA and represents subtype 6l. This isolate was identified in the US Third National Health and Nutrition Examination Survey (NHANES III) (Nainan et al., 2006). Three samples (QC227, QC216 and QC99) represent subtypes 6o, 6p and 6q, respectively. QC227 was obtained from a Caucasian individual who was infected following a blood transfusion in Canada in the mid-1980s, whilst QC216 and QC99 were respectively from a Vietnamese and a Cambodian immigrant in Quebec, Canada.

From 100 μl of each of these samples, complete HCV genomes were amplified by using strategies listed in Supplementary Fig. S1 (available in JGV Online). The amplicons were sequenced directly and sequence information was analysed as described elsewhere (Lu et al., 2007). The lengths and genomic organizations of the resultant five genomes are described in Table 1. Pairwise comparison of the five sequences with 14 genotype 6 reference sequences revealed that they have nucleotide similarities of 72.6–81.5% (mean, 75.3%) over the entire genome. The highest sequence similarities were observed between isolates 537796 (6l), VN405 (6k) and km41 (6f). The lowest similarities were found when the five sequences were compared with those of subtypes 6a and 6b. Similarities were the highest in the core (mean, 87.4%) and NS5B (mean, 78.2%) genes and lowest in the P7 (mean, 65.8%) and NS2 (mean, 68.2%) regions (see Supplementary Table S1, available in JGV Online). According to the criterion that HCV subtypes are defined by nucleotide similarities of <75–80% over the entire genome (Simmonds et al., 1994), the similarities calculated here indicate that these five sequences represent five distinct HCV subtypes.

Phylogenetic analysis demonstrated that the five isolates were all classified into genotype 6. These strains therefore complete a panel of 17 subtypes with determined complete genomes, designated 6a–6q within genotype 6 (Simmonds et al., 2005). Novel isolates km41 and gz52557, for which the entire genome sequences are also available, may represent two novel subtypes (Lu et al., 2006a). In the phylogenetic tree, each isolate is represented by a separate branch, forming a highly diverse genotype 6 clade (Fig. 1a). Within this clade, four subclades can be determined, with each subclade having a bootstrap value of 100%. The first subclade contains subtypes 6a and 6b; the second consists of 6c, 6d, 6e, 6f, 6o, 6p and 6q; the third includes 6g and the unassigned isolate gz52557; and the fourth contains 6h, 6i, 6j, 6l, 6m, 6n and the unassigned km41. The second subclade can be further split into three pairs (6c/6d, 6e/6q and 6o/6p) and a single branch, 6f. The fourth subclade can be divided into three groups (6h/6i/6j, 6k/km41/6l and 6m/6n), with each group having a bootstrap value of 100%.

A second phylogeny was reconstructed by using the deduced amino acid sequences of the isolates. The phylogeny had an identical genotype 6 topology, indicating the robustness of this branching pattern (Fig. 1b). Nucleotide sequences from ten protein-encoding regions were analysed separately, which also showed that the 19 isolates were consistently distinct (data not shown). To identify potentially recombinant isolates, pairwise similarity scores (i.e. diversity plots) were calculated along the HCV genome between the five isolates and 14 reference genotype 6 sequences. Each genome showed similar patterns of genetic similarity to other isolates along the genome and showed no evidence of recent recombination events (data not shown).

To relate subtypes 6c, 6l, 6o, 6p and 6q to HCV variants characterized by partial sequences, three short, subgenomic regions of the five new isolates were analysed alongside other sequences retrieved from GenBank. The subgenomic segments were (i) 350 nt of the core gene, (ii) 424 nt of the E1 gene, and (iii) 329 nt of the NS5B gene, corresponding to positions 340–689, 867–1290 and 8295–8623 of the H77 genome (GenBank accession no. NC_004102), respectively. Phylogenetic trees estimated from these three segments

---

**Table 1. Genomic organization of the five HCV isolates**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Amplicons (n*)</th>
<th>Genomic ORF</th>
<th>5’ NCR</th>
<th>Core</th>
<th>E1 Length (nt)</th>
<th>E2</th>
<th>P7</th>
<th>NS2</th>
<th>NS3</th>
<th>NS4A</th>
<th>NS4B</th>
<th>NS5A</th>
<th>NS5B</th>
<th>3’ NCR</th>
<th>Poly(U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Th846/6c</td>
<td>25</td>
<td>9459</td>
<td>9048</td>
<td>333</td>
<td>573 576</td>
<td>1089</td>
<td>189</td>
<td>651</td>
<td>1893</td>
<td>162</td>
<td>783</td>
<td>1356</td>
<td>1776</td>
<td>72</td>
<td>24</td>
</tr>
<tr>
<td>537796/6l</td>
<td>22</td>
<td>9453</td>
<td>9051</td>
<td>338</td>
<td>573 576</td>
<td>1095</td>
<td>189</td>
<td>651</td>
<td>1893</td>
<td>162</td>
<td>783</td>
<td>1353</td>
<td>1776</td>
<td>64</td>
<td>28</td>
</tr>
<tr>
<td>QC227/6o</td>
<td>34</td>
<td>9450</td>
<td>9051</td>
<td>338</td>
<td>573 576</td>
<td>1092</td>
<td>189</td>
<td>651</td>
<td>1893</td>
<td>162</td>
<td>783</td>
<td>1356</td>
<td>1776</td>
<td>61</td>
<td>28</td>
</tr>
<tr>
<td>QC216/6p</td>
<td>31</td>
<td>9453</td>
<td>9051</td>
<td>339</td>
<td>573 576</td>
<td>1092</td>
<td>189</td>
<td>651</td>
<td>1893</td>
<td>162</td>
<td>783</td>
<td>1356</td>
<td>1776</td>
<td>63</td>
<td>31</td>
</tr>
<tr>
<td>QC99/6q</td>
<td>19</td>
<td>9463</td>
<td>9051</td>
<td>339</td>
<td>573 576</td>
<td>1092</td>
<td>189</td>
<td>651</td>
<td>1893</td>
<td>162</td>
<td>783</td>
<td>1353</td>
<td>1776</td>
<td>76</td>
<td>28</td>
</tr>
</tbody>
</table>

*Number of overlapping fragments covering the complete genomic sequence.
Fig. 1. Phylogenetic tree based on (a) complete nucleotide sequences and (b) deduced amino acid sequences. The six HCV genotypes are indicated by numbers 1–6; subtypes are designated 1a–6q and 6? (subtype unassigned). Reference HCV sequences are indicated by an isolate name followed by a GenBank accession number in parentheses. The five HCV variants sequenced completely in this study are shown in bold. Bootstrap analysis values are shown in italics. Bars, 0.10 nucleotide substitutions per site.
demonstrated that Th846 represents the sole published representative of subtype 6c, whereas 537796, QC227, QC216 and QC99 clustered with six, four, four and five other sequences, respectively (Fig. 2). Among these other sequences, six 6l (VN507, VN530, VN531, VN606, VN655 and VT681), two 6o (VN4 and VT316) and two 6p (D3 and VN12) sequences had origins in Vietnam (Noppornpanth et al., 2006b; Stuyver et al., 1995; Tokita et al., 1994). Two 6o (QC30 and QC33), two 6p (QC123 and QC190) and four 6q (QC57, QC1564, QC176 and QC189) sequences were isolated in Vietnamese or Cambodian immigrants in Quebec, Canada (Bernier et al., 1996; Murphy et al., 2003). Another 6q sequence (IG93336) was recorded from Laos, but detailed information on this isolate has not been published. Bootstrap values for these clusters varied from 98 to 100%.

Several studies have suggested that unique and highly divergent variants of HCV genotype 6 exist in southeastern Asia. In one study, two partial HCV sequences (1582 and 1142 nt long; GenBank accession numbers D37843 and D37857, respectively) were determined for strain Th846, which was isolated from a commercial blood donor in Thailand. Analysis of these two fragments initially led to the classification of Th846 as a novel genotype 7d virus (Tokita et al., 1995). After reclassification of

Fig. 2. Phylogenetic trees based on (a) partial sequences from the core region corresponding to nt 340–689, (b) partial E1 region sequences corresponding to nt 867–1290, (c) partial NS5B region sequences corresponding to nt 8295–8623 (numbering according to the H77 genome; GenBank accession no. NC_004102). Designations are the same as those described in the legend to Fig. 1.
genotypes 7, 8 and 9 as subtypes within genotype 6, this isolate (Th846) was assigned to subtype 6c (Simmonds et al., 1996). The latter designation has now been used in the consensus proposal for HCV nomenclature (Simmonds et al., 2005). Although several studies of HCV infection among blood donors in Thailand and other south-eastern Asian countries have been performed, no other 6c isolate has been reported to date (Thaikruea et al., 2004), suggesting that Th846 represents a rare HCV variant. In this regard, Th846 is comparable to g252557, which is also not related closely to other genotype 6 variants (Lu et al., 2006a). As the prevalence and epidemiological features of these strains are not well understood, sequencing their entire genomes may provide important new insights regarding the evolution and spread of genotype 6 variants. In addition, characterization of the full range of global HCV diversity is essential for the development and implementation of automated virus-genotyping tools (de Oliveira et al., 2005).

In the US NHANES III study, partial sequences of the core, E1 and NS5B regions were determined for HCV isolate 537796, obtained from an Asian immigrant (Nainan et al., 2006). Initial phylogenetic analysis classified this isolate as subtype 6l. Here, we obtained the entire genome sequence for this isolate. Although isolate 537796 is distinct from subtype 6k and the unassigned km41, it is related to these strains (genome similarities of 81.0 and 81.5 %, respectively). These similarities are comparable to those between subtypes 6i and 6j, and between subtypes 6m and 6n (genome similarities of 81.3–82.7 %). Furthermore, subtypes 6k, 6l and the unassigned km41 form a clade of closely related subtypes (Fig. 1), suggesting that this genotype 6 lineage has been present and circulating in Vietnam for a considerable period of time. In support of this, a search of the Los Alamos HCV database identified two subtype 6k (VN405 and D33) and six subtype 6l (VN507, VN530, VN531, VN606, VN655 and VT681) isolates. All 6k and 6l isolates were sampled exclusively from Ho Chi Minh City, Vietnam (Tokita et al., 1994), except for VN405, which was from Hanoi. Although isolates km41 and km45 were isolated in China, they were from the Kunming region, bordering Vietnam. The isolation of strain 537796 from an Asian immigrant suggests that it might also originate from Vietnam, but this information was not reported in the NHANES III study (Nainan et al., 2006).

Previously, partial sequences have been determined for a number of HCV variants from patients in Quebec, Canada (Murphy et al., 2003). Among these, variants QC227, QC216 and QC99 were assigned to recently recognized subtypes 6o, 6p and 6q, respectively (Simmonds et al., 2005), and were selected for complete sequencing here. VN4 was the first isolate of subtype 6o (obtained in 1995 from a Vietnamese patient), although it was initially designated subtype ‘7c’ (Stuyver et al., 1995). QC30 was the second isolate of subtype 6o detected from a Vietnamese immigrant in Quebec, Canada (Bernier et al., 1996). QC33 and QC227 were later isolated from another Vietnamese immigrant and a Caucasian in Quebec, respectively, and both were classified as subtype 6o (Murphy et al., 2003). Interestingly, the Caucasian had acquired HCV infection in Quebec after a blood transfusion in the mid-1980s. This would be the first report of post-transfusional hepatitis C caused by a genotype 6 isolate in North America. VT316 is another subtype 6o strain isolated from Vietnam (http://hcv.lanl.gov/content/hcv-db/index).

Five subtype 6p isolates have been characterized so far, all with origins from south-eastern Asia. VN12 was the first isolate of subtype 6p discovered in a Vietnamese patient (Stuyver et al., 1995), followed by QC123, QC190 (identified among Cambodian immigrants in Quebec) and QC216 (from a Vietnamese immigrant; Murphy et al., 2003). Recently, isolate D3 was obtained from a blood donor in Ho Chi Minh City. It represents a recombinant between subtypes 2i and 6p (Noppornpanth et al., 2006a). A recombination event could reflect a relatively high degree of prevalence or co-infection.

In addition to QC99, which was sequenced fully in this study, five subtype 6q sequences are available from GenBank: QC57, QC164, QC176, QC189 and IG93336, all isolated from Vietnamese or Cambodian immigrants in Quebec (Murphy et al., 2003) or recorded to be from Cambodia (IG93336; http://hcv.lanl.gov/content/hcv-db/index). Isolate IG93335 from Laos is recorded as subtype 6q (http://hcv.lanl.gov/content/hcv-db/index) but, in our phylogenetic analysis, it appears to be a distinct, novel subtype (IG93335 has similarities of 82.4–85.3 % with the six subtype 6q sequences). This range of nucleotide similarity is comparable to those observed for 6i versus 6j (82.9 %), 6c versus 6d (84 %) and 6k versus km41 (85.1 %) when the same NS5B regions were compared pairwise.

Almost all subtypes of HCV genotype 6 (except for subtype 6a) are characteristic of endemic transmission. In other words, isolates of these subtypes appear to be relatively rare, sampled from a restricted geographical region (or from immigrants from that area) and highly genetically diverged from each other. The low frequency of detection of these strains could be due to biased or incomplete sampling or low prevalence. The latter could be a result of either inefficient, non-parenteral transmission or the presence of background immunity among local populations because of a long-term endemic infection. In contrast, only a single case of person-to-person transmission of these strains was confirmed outside south-eastern Asia (the Caucasian QC227, infected in Canada). It suggests a long-term co-circulation and evolution of genotype 6 variants in south-eastern Asia, with little exchange of virus lineages with outside populations. Collectively, these features are consistent with a zoontic origin for HCV genotype 6, which in some ways resembles the pattern observed for genotypes III and IV of hepatitis E virus (Lu et al., 2006b).

The accelerating rate of globalization and economic development in south-eastern Asia will lead to increased
urbanization, commercialization, transportation and tourism. More immigrants and visitors are expected to travel between south-eastern Asia and other parts of the world. This could enhance the spread of genotype 6 variants to new regions. Modern, efficient transmission routes, such as blood transfusion, injection drug use and unsafe medical practice, are thought to have been crucial for the worldwide spread of epidemic HCV strains during the past 50–70 years (Pybus et al., 2001). These transmission routes, especially injection drug use, may have increased the likelihood of HCV co-infection and recombination, such as that observed recently between subtypes 1b and 2k in St Petersburg, Russia (Kalinina et al., 2002, 2004). Whilst a similar recombination event was found between subtypes 6p and 2i in Vietnam, its prevalence and possible impact remain to be addressed (Noppornpanth et al., 2006a).

To date, a high prevalence has not been recorded for any HCV genotype 6 subtype, except for subtype 6a. Although lesser symptoms and better responses to interferon therapy have been reported for genotype 6 infections (Dev et al., 2002; Doi et al., 1996), this does not necessarily indicate that such strains are less pathogenic, as they might have caused less recognizable, atypical hepatitis in many cases. In addition, current HCV detection systems were established to detect the highly prevalent worldwide epidemic strains of HCV, and may be failing to detect genetically diverse genotype 6 strains correctly. Furthermore, these variants display 5′ non-coding region (NCR) sequences that are similar to those of genotype 1, whilst showing greater genetic variability in other regions (Noppornpanth et al., 2006a). Thus, these variants would be identified as being of genotype 1 and consequently missed by 5′ NCR-based genotyping assays. The completion of full genomic sequences for a panel of 17 assigned genotype 6 subtypes reported here will hopefully provide useful sequence information to enable more effective diagnostic testing to be developed.

This study will form the basis for a more precise estimation of the origin of HCV and the history of HCV evolution. Previously, the origin of HCV was estimated to be 500–2000 years ago by using a constant-rate molecular-clock model and based on analysis of short segments of the E1 and NS5B regions (Pybus et al. 2001; Simmonds, 2004; Smith et al., 1997). The estimation of the origin of HCV should be improved in the future by combining the complete HCV genomic sequences now available with a newly developed and accurate ‘relaxed-clock’ model of evolution (Drummond et al., 2006).

Acknowledgements

This work was supported by NIH grants 5 P20 RR016443-04 and CA063640.

References


1524 Journal of General Virology 88


