Complete genomic sequences for hepatitis C virus subtypes 4b, 4c, 4d, 4g, 4k, 4l, 4m, 4n, 4o, 4p, 4q, 4r and 4t

Chunhua Li, Ling Lu, Xianghong Wu, Chuanxi Wang, Phil Bennett, Teng Lu and Donald Murphy

Correspondence
Ling Lu
ling.lu@hsc.utah.edu
Donald Murphy
donald.murphy@inspq.qc.ca

1 Division of Gastroenterology-Hepatology and Nutrition, Department of Medicine, University of Utah, 30N 1900E, Salt Lake City, UT 84132, USA
2 Department of Biology, Kunming Teacher's College, Kunming, Yunnan, PR China
3 Guangzhou Blood Center, Guangzhou, Guangdong, PR China
4 Micropathology Ltd, University of Warwick Science Park, Coventry, UK
5 University of Georgia, Athens, GA, USA
6 Institut national de santé publique du Québec, Laboratoire de santé publique du Québec, Sainte-Anne-de-Bellevue, QC, Canada

In this study, we characterized the full-length genomic sequences of 13 distinct hepatitis C virus (HCV) genotype 4 isolates/subtypes: QC264/4b, QC381/4c, QC382/4d, QC193/4g, QC383/4k, QC274/4l, QC249/4m, QC97/4n, QC93/4o, QC139/4p, QC262/4q, QC384/4r and QC155/4t. These were amplified, using RT-PCR, from the sera of patients now residing in Canada, 11 of which were African immigrants. The resulting genomes varied between 9421 and 9475 nt in length and each contains a single ORF of 9018–9069 nt. The sequences showed nucleotide similarities of 77.3–84.3 % in comparison with subtypes 4a (GenBank accession no. Y11604) and 4f (EF589160) and 70.6–72.8 % in comparison with genotype 1 (M62321/1a, M58335/1b, D14853/1c, and 1?/AJ851228) reference sequences. These similarities were often higher than those currently defined by HCV classification criteria for subtype (75.0–80.0 %) and genotype (67.0–70.0 %) division, respectively. Further analyses of the complete and partial E1 and partial NS5B sequences confirmed these 13 ‘provisionally assigned subtypes’.

Although highly prevalent in the Middle East and Central/West Africa with increasing emergence in Europe and North America, many subtypes of hepatitis C virus (HCV) genotype 4 lack full-length genomic sequences (Kamal & Nasser, 2008). According to the updated HCV database at the Viral Bioinformatics Resource Center (http://www.hcvdb.org/index.asp), this genotype now has 19 assigned subtypes (a–h and k–u). However, only subtypes 4a, 4d, 4f and 4k currently have full-length or near full-length reference sequences (Hmaied et al., 2007; Timm et al., 2007; Kuntzen et al., 2008). The aim of this study was to completely sequence 13 HCV isolates each representing a distinct genotype 4 subtype: 4b–d, 4g, 4k, 4l, 4m–r and 4t.

All serum samples were obtained from patients (five male and eight female) living in Quebec, Canada, between September 2002 and March 2007. Their ages ranged between 31 and 68 years old. Except for isolates QC382/4d and QC274/4l, which were collected from Caucasian individuals, the remaining 11 specimens were collected from immigrants of African origin (Table 1). HCV infection was confirmed in these individuals by routine virological assays for clinical diagnosis (Murphy et al., 2007). Starting with 100 μl samples of serum, HCV genomic fragments were amplified using previously described methods (Li et al., 2006). Briefly, RNA was extracted using TriPure (Roche) and cDNA was generated using random hexamers (Promega) and avian myeloblastosis virus reverse transcriptase (Roche). After a 1:3 dilution with RNase/DNase free water, a 1–2 μl aliquot of the diluted cDNA was used as PCR template. Overlapping fragments were amplified by using conventional PCR (Roche) with primers listed in Supplementary Table S1, available in JGV Online. The fragments covering the extreme 5’ UTR termini were amplified using semi-nested

The GenBank/EMBL/DDBJ accession number for sequences reported are FJ462431 (QC139/4p), FJ462432 (QC193/4g), FJ462433 (QC249/4m), FJ462434 (QC262/4q), FJ462435 (QC264/4b), FJ462436 (QC381/4c), FJ462437 (QC382/4d), FJ462438 (QC383/4k), FJ462439 (QC384/4r), FJ462440 (QC393/4o), FJ462441 (QC97/4n), FJ839869 (QC155/4l) and FJ839870 (QC274/4l).

Supplementary tables are available with the online version of this paper.
Full-length sequences of 13 HCV-4 subtypes

Because the nucleotides in the core, E1, E2, P7, NS2, NS3, NS4A, NS4B, and NS5B regions are identical among all HCV isolates, these regions are not shown.

Isolate | Genotype | Status | Country/continent | Gender | Age (years) | Isolation date | Fragments amplified | Full-length ORF genome (nt) | 5’UTR genome (nt) | 3’UTR genome (nt) | Poly(U) (nt) | UT R poly(T) tracts | UT R poly(A) tracts |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>QC264</td>
<td>4b</td>
<td>Immigrant</td>
<td>DR Congo</td>
<td>Female</td>
<td>45</td>
<td>9 May 2005</td>
<td>23</td>
<td>9410</td>
<td>9036</td>
<td>340</td>
<td>1089</td>
<td>1344</td>
<td>34</td>
</tr>
<tr>
<td>QC381</td>
<td>4c</td>
<td>Immigrant</td>
<td>DR Congo</td>
<td>Male</td>
<td>59</td>
<td>6 March 2007</td>
<td>20</td>
<td>9468</td>
<td>9069</td>
<td>340</td>
<td>1086</td>
<td>1332</td>
<td>28</td>
</tr>
<tr>
<td>QC382</td>
<td>4d</td>
<td>Non-immigrant</td>
<td>Canada</td>
<td>Male</td>
<td>31</td>
<td>11 February 2005</td>
<td>18</td>
<td>9421</td>
<td>9021</td>
<td>340</td>
<td>1089</td>
<td>1332</td>
<td>30</td>
</tr>
<tr>
<td>QC193</td>
<td>4g</td>
<td>Immigrant</td>
<td>Africa</td>
<td>Female</td>
<td>47</td>
<td>13 July 2004</td>
<td>10</td>
<td>9433</td>
<td>9036</td>
<td>341</td>
<td>1089</td>
<td>1344</td>
<td>31</td>
</tr>
<tr>
<td>QC262</td>
<td>4q</td>
<td>Immigrant</td>
<td>Rwanda</td>
<td>Female</td>
<td>43</td>
<td>8 February 2005</td>
<td>16</td>
<td>9433</td>
<td>9036</td>
<td>341</td>
<td>1089</td>
<td>1344</td>
<td>34</td>
</tr>
<tr>
<td>QC384</td>
<td>4r</td>
<td>Immigrant</td>
<td>Burundi</td>
<td>Female</td>
<td>47</td>
<td>15 March 2007</td>
<td>18</td>
<td>9440</td>
<td>9024</td>
<td>341</td>
<td>1089</td>
<td>1332</td>
<td>38</td>
</tr>
<tr>
<td>QC155</td>
<td>4t</td>
<td>Immigrant</td>
<td>Africa</td>
<td>Female</td>
<td>34</td>
<td>23 December 2003</td>
<td>20</td>
<td>9488</td>
<td>9026</td>
<td>341</td>
<td>1089</td>
<td>1332</td>
<td>34</td>
</tr>
</tbody>
</table>

*RACE PCR amplification of the 3’UTR was not successful for this isolate.

PCR with the ‘5’end’ primer. This primer’s design was based upon the conserved region of already published HCV-4 sequences. Fragments covering the 3’ UTR poly(T) tracts were amplified using a RACE PCR protocol. The first-round amplification used a forward HCV strain specific primer and a ‘poly(A)’ primer. The second-round amplification used an inner forward primer that is HCV strain specific and the ‘NUP’ primer (see Supplementary Table S1). Standard procedures were adopted to prevent nested RT-PCR false positives (Kwok & Higuchi, 1989). After PCR, the amplicons were purified with the QIAquick PCR purification kit (Qiagen) according to the manufacturer’s protocol. In order to generate consensus sequences that reflected the genetic heterogeneity expected at numerous positions through the HCV genome, the purified amplicons were sequenced directly. This was done bi-directionally using ABI Prism BigDye 3.0 terminators and an appropriate primer on an ABI Prism 3137 genetic analyser (PE Applied Biosystems).

All observable ambiguous nucleotides were recorded using standard IUPAC codes. Sequence information was analysed using GCG (Wisconsin Sequence Analysis Package; Genetic Computer Group, version 10.0), PHYML (Guindon & Gascuel, 2003), MEGA4 (Kumar et al., 2004), and RDP3 (Martin et al., 2005) as previously described (Lu et al., 2007). Alignments were created using the pileup program; further adjustments to the alignments were then made manually. Phylogenetic trees were estimated using the maximum-likelihood method under the HKY + I + Γ0 substitution model (gamma distribution approximated by using six rate categories) (Posada & Crandall, 1998). The transition/transversion ratio, proportion of invariable sites and gamma distribution shape parameter were estimated from the sequence data. Base frequencies were adjusted to maximize the likelihood. Bootstrap resampling was performed using 500 neighbour-joining replicates. To exclude recent virus recombination events (Kalina et al., 2002, 2004; Colina et al., 2004; Legrand-Abravanel et al., 2007; Noppornpanth et al., 2002, 2004; Colina et al., 2004), the RDP3 software was run with default settings modified as follows: (i) window size was 40 nt; (ii) the option of linear sequences was chosen; (iii) six different methods (RDP, GENECONV, MaxChi, Bootscan, Chimera and SiScan) were run simultaneously against the sequences listed below; and (iv) only the events detected by more than two of the above methods were considered positive. For pair-wise comparison and detection of HCV recombination, the following 39 complete HCV genomic sequences (subtypes) were co-analysed: GenBank accession no. M62321 (1a), M58335 (1b), D14853 (1c), AJ651228 (1d), D00944 (2a), D10988 (2b), D50409 (2c), DQ155561 (2i), AF031663 (2k), D17763 (3a), D49374 (3b), D63821 (3k), Y11604 (4a), EF589160 (4f), Y12083 (5a), D84262 (6b), EF424629 (6c), D84263 (6d), DQ314805 (6e), DQ35760 (6f), D63822 (6g), D84265 (6h), DQ35762 (6i), DQ35761 (6j), D84264 (6k), EF424628 (6l), DQ35763 (6m), DQ278894 (6n), EF424627 (6o), EF424626 (6p), EF424625 (6q), EU408328 (6r), EU408329 (6s), EF632071 (6t), EU408330 (6u), EU798760 (6v), DQ278893 (6w), DQ278892 (6x), DQ278891 (6y).
Complete genomic sequences were amplified from the 13 isolates each using 16–23 overlapping fragments. The genomes of these 13 isolates were 9388–9488 nt in length, starting from the extreme 5'UTR termini to the 3'UTR poly(U) tracts. Individually, the QC264/4b genome contained 9440 nt, QC381/4c 9468, QC382/4d 9421, QC193/4g 9435, QC383/4k 9438, QC274/4l 9388, QC249/4m 9426, QC97/4n 9435, QC93/4o 9422, QC139/4p 9475, QC262/4q 9433, QC384/4r 9440, and QC155/4t 9488 nt. Excluding isolate QC274/4l, for which amplification of the 3'UTR poly(U) tract failed, all the genomes had sequences starting from the extreme 5'UTR termini to the 3'UTR poly(U) tracts (Table 1). Of these genomes the single ORF ranged between 9018 and 9069 nucleotides, containing ten protein-coding regions of the following sizes: core (573 nt/191 aa), E1 (576 nt/192 aa), E2 (1080–1092 nt/360–364 aa), P7 (189 nt/63 aa), NS2 (651 nt/217 aa), NS3 (1893 nt/631 aa), NS4A (162 nt/54 aa), NS4B (783 nt/261 aa), NS5A (1332–1377 nt/444–459 aa), and NS5B (1776 nt/591 aa). Among them, only the E2 and NS5A regions displayed variation in nucleotide lengths (Table 1).

Pair-wise comparisons were performed between the 13 isolated sequences and two reference sequences: Y11604/4a and EF589160/4f. This resulted in genome-wide nucleotide similarities of 77.3–84.3 %. Although these observed similarities were higher than those suggested by the updated HCV nomenclature (75.0–80.0 %), for subtypes of a single genotype (Simmonds et al., 2005), these isolates represented 13 distinct genotype 4 subtypes. Among the ten protein-coding regions, core (86.7–94.8 %) and NS5B (80.2–88.2 %) showed the highest similarities, while P7 (65.1–81.4 %) and NS2 (67.2–79.0 %) showed the lowest (Supplementary Table S2, available in JGV Online). Reference sequences from genotypes 1, 2, 3, 5 and 6 were also subjected to pair-wise comparison (see above). This showed that all 13 isolated sequences were more similar to genotype 1 sequences (70.6–72.8 %) than to any other, apart from genotype 4 sequences. Again, this similarity was higher than that suggested by the updated HCV nomenclature (67–70 %) for the maximum similarity between different genotypes.

An un-rooted phylogenetic tree was estimated using the full-length genomic sequences, starting from the extreme 5'UTR termini to the 3'UTR poly(T) tracts representing various genotypes and subtypes. This showed a complex genotype 4 clade (Fig. 1). This clade contained two major clusters, each having a full bootstrap support of 100 %, and a single 4b branch. As a whole, this clade showed a full bootstrap support of 100 %. From the top of the tree, the upper cluster included 11 subtypes that were grouped into four subsets: one contained subtypes 4f, 4p and 4t; the second contained 4m and 4d; the third contained 4o, 4a and 4c; and the fourth contained 4q, 4l and 4n. The bootstrap supports for these four subsets were 85, 72, 27 and 83 %, respectively. Although further groupings were formed by the latter three subsets, the bootstrap supports ranged between 24 and 64 %. The lower cluster looked simpler; it contained only three subtypes: 4g, 4k and 4r and had a full bootstrap support of 100 %. At a deeper branching point, the 4k-4r grouping showed strong bootstrap support of 99 %. Statistically, a bootstrap support >70 % equated to a P-value <0.05 (Robertson et al., 1998). This implies that all groupings having bootstrap supports >70 % may be not reliable, and could change when a different combination of isolates or different lengths of sequences are analysed. This was exemplified by the 4o-4a-4c subset; when the full-length genomic sequences were estimated, the 4o branch joined with the 4a-4c grouping to form a subset, but showed a weak bootstrap support of only 27 % (Fig. 1). However, when partial E1 and NS5B sequences were analysed, the 4o branch was shown to grow distantly from the 4a-4c grouping (see Fig. 2). In contrast, the 4a-4c grouping is a reliable structure; it was maintained irrespective of whether full-length or partial genomic sequences were analysed.

![Fig. 1. Unrooted phylogenetic tree based on complete nucleotide sequences. HCV genotypes are indicated by numbers 1–7; subtypes are indicated 1a to 7a; ‘1?’ represents a genotype 1 subtype not yet assigned. HCV sequences are designated by a name followed by their GenBank accession number in parentheses. The 13 isolates completely sequenced in this study are shown in bold. For a simpler tree structure, the genotype 6 clade was compressed. Bootstrap analysis values are shown in italics. Bars, 0.10 nucleotide substitutions per site.](image-url)
Partial E1 and NS5B region sequences (corresponding to nt 869–1289 and 8276–8615 in the numbering of the H77 genome) were further analysed (Fig. 2a–b). In addition to the 13 isolates from this study, many other distinct isolates were also included to better illustrate how subtypes related. These included one isolate of subtype 4b, 11 of 4c, two each of 4d and 4g, 17 of 4k, 18 of 4l, 14 of 4m, 10 of 4n, 25 of 4o, nine of 4p, four of 4q, 33 of 4r and 17 of 4t (see Supplementary Table S3, available in JGV Online). Previously, these subtypes were provisionally assigned based only on partial sequences. Hence the current study confirmed these assignments and provided full-length reference sequences. Consistent across all the phylogenetic trees obtained, both subtypes 4a and 4c were closely grouped. When the partial NS5B sequences were analysed, only the QC264 isolate represented subtype 4b; it formed a subset with 4f, but showed a bootstrap support of 66% (Fig. 2b). Due to the lack of statistical significance, the 4b-4f grouping was not observed when the partial E1 region sequences were analysed (Fig. 2a). Regardless, the 13 related genotype 4 subtypes all showed significant bootstrap supports to confirm clusters when analysed with either region sequences. Originally, these genotype 4 isolates were identified in Central African countries.
Recombinants of different HCV genotypes and subtypes have been described in patients from Russia (2k/1b) (Kalinina et al., 2002, 2004), Peru (1a/1b) (Colina et al., 2004), France (2k/5) (Legrand-Abravanel et al., 2007) and Vietnam (2i/6p) (Nopcopornpath et al., 2006). In order to exclude such a possibility, pair-wise nucleotide similarity curves were plotted along HCV genomes. When comparing the 13 complete genomic sequences against each other and against the 39 reference sequences representing different genotypes and subtypes, no such evidence was revealed (data not shown).

In this study, we have characterized 13 complete genomic sequences representing 13 distinct genotype 4 subtypes: 4b, 4c, 4d, 4g, 4k, 4l, 4m, 4n, 4o, 4p, 4q, 4r and 4t. In the updated HCV nomenclature these 13 subtypes have all been categorized as 'provisionally assigned subtypes'. With their complete genomic sequences determined and their assignments proven to be consistent with existing designations, these 13 subtypes are now confirmed (Simmonds et al., 2005). Although several genotype 4 subtype sequences (4a, d, f and k) have recently been described that were 9269–9303 nt long, parts of their 5' and 3' UTRs were missing (Hmaied et al., 2007; Timm et al., 2007; Kuntzen et al., 2008). Accounting for these sequences, genotype 4 to date has complete or nearly complete genomic sequences for 15 subtypes. These sequences can be used not only for robust HCV classification, but also for clinical purposes. They will facilitate more accurate genotypeing approaches and further the study of HCV genotype effects in differential patient treatment options.

The recent consensus paper provided updated criteria for the nomenclature of HCV isolates (Simmonds et al., 2005). Major genotypes were defined as phylogenetic clusters that differ from each other by 31–33 % at the nucleotide level, while subtypes differ by 20–25 %. In this study, nucleotide differences <30 % (70.6–72.8 % similarities) were observed between genotypes 1 and 4, and differences as low as 15.7 % (84.3 % similarity) were observed between subtypes 4a and 4c. The former may reflect the fact that both genotypes 1 and 4 have origins from Central and West Africa (Kamal & Nasser, 2008) and historically may have belonged to a single ancient ‘genotype’ (Njouom et al., 2003). The latter may indicate a continual genetic variation of HCV with many variants still missing. Indeed, subtypes 4a and 4c may ultimately qualify as a single subtype if such missing variants are found (Fig. 2a). However, this would confirm a statement made in the consensus paper suggesting that subtype labels are of little or no value in the designation of HCV variants in high-diversity areas such as in sub-Saharan Africa and South-east Asia (Simmonds et al., 2005). As more HCV variants are characterized with full-length sequences, new statistics will be obtained and can be used to develop supplementary nomenclature criteria.

Studies have indicated that HCV genotype 4 strains originated and propagated in Central and West Africa before spreading to other regions (Genovese et al., 2005; Ndjomou et al., 2003; Ndong-Atome et al., 2008; Njouom et al., 2003, 2005, 2007; Pybus et al., 2003; Ray et al., 2000; Simmonds, 2004; Smith et al., 1997; Xu et al., 1994). Since the 1930s some strains have become prevalent in North Africa as a result of large-scale vaccination campaigns (Angelico et al., 1997; Tanaka et al., 2004). Recently, a variety of these strains have been found in Southern European countries on the Mediterranean Sea (Ansaldi et al., 2005; Fernández-Arcás et al., 2006; Franco et al., 2007; Katsoulidou et al., 2006; Nicot et al., 2005; Payan et al., 2005). Although infection due to genotype 4 is not common in the USA and Canada, most of the reported cases were clustered among IDUs or immigrants from countries where HCV-4 is known to be most prevalent or among individuals who acquired the infection in these countries (Lyra et al., 2004; Timm et al., 2007; Zein et al., 1996). In this study, 11 of the 13 complete genomic sequences determined were from immigrants living in Canada. Although the country origin was not known for all, they were all from Africa. Co-analysis with other previously reported sequences also provided similar results (Fig. 2a, b, Supplementary Table S3). These data are consistent with those obtained from Europe, where emerging HCV-4 infections have now become a serious public health problem (Cenci et al., 2007; Nicot et al., 2005). In order to prevent such a problem occurring in North America, emphasis should be placed on detecting these viral strains. For this purpose, an improved HCV genotyping approach utilizing such full-length sequence data as described in this study is considered useful.

With regards to potential limitations of the current study, the lack of any viral load information on the collected sera, combined with the relatively low starting volumes for RNA extraction and subsequent RT-PCR, means that the sequences obtained may not necessarily reflect the full extent of quasi-species variation present in the infected individual. Nevertheless, these sequences will serve as useful references not only for confirming the 13 previously provisionally assigned HCV-4 subtypes, but also for better HCV classification and molecular evolutionary studies.

Acknowledgements

We thanks Dr Curt Hagedorn for his suggestions for completing this work.
References


Morice, Y., Roulot, D., Grando, V., Stirnemann, J., Gault, E., Jeantil, V., Bentata, M., Jarrousse, B., Lortholary, O. & other authors (2001). Phylogenetic analyses confirm the high prevalence of hepatitis C virus (HCV) type 4 in the Seine-Saint-Denis district (France) and indicate seven different HCV-4 subtypes linked to two different epidemiological patterns. J Gen Virol 82, 1001–1012.


