Hepatitis C virus complete genome sequences identified from China representing subtypes 6k and 6n and a novel, as yet unassigned subtype within genotype 6

Ling Lu,1† Tatsunori Nakano,2† Chunhua Li,3,4† Yongshui Fu,5 Steve Miller,1 Carla Kuiken,6 Betty H. Robertson7† and Curt H. Hagedorn1

1Division of Gastroenterology/Hepatology, Department of Medicine, University of Kansas Medical Center, 4035 Delp, MS 1023, Kansas City, KS 66160, USA
2Department of Internal Medicine, Ichinomiya Nishi Hospital, Ichinomiya, Aichi, Japan
3Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China
4Graduate School of the Chinese Academy of Sciences, Beijing, China
5Guangzhou Blood Center, Guangzhou City, Guangdong Province, China
6Los Alamos National Laboratory, Los Alamos, NM, USA
7National Center for Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

Received 2 August 2005
Accepted 17 November 2005

Here, the complete genome sequences for three hepatitis C virus (HCV) variants identified from China and belonging to genotype 6 are reported: km41, km42 and gz52557. Their entire genome lengths were 9430, 9441 and 9448 nt, respectively; the 5’ untranslated regions (UTRs) contained 341, 342 and 339 nt, followed by single open reading frames of 9045, 9045 and 9057 nt, respectively; the 3’ UTRs, up to the poly(U) tracts, were 41, 51 and 52 nt, respectively. Phylogenetic analyses showed that km41 is classified into subtype 6k and km42 into subtype 6n. Although gz52557 clustered distantly with subtype 6g, it appeared to belong to a distinct subtype. Analysis with 53 and 105 partial core and NS5B region sequences, respectively, representing 17 subtypes from 6a to 6q and three unassigned isolates of genotype 6 in co-analyses demonstrated that gz52557 was equidistant from all of these isolates, indicating that it belongs to a novel subtype. However, based on a recent consensus that three or more examples are required for a new HCV subtype designation, it is suggested that gz52557 remains unassigned to any subtype.

Hepatitis C virus (HCV) is divided into six genotypes and, within each genotype, closely related variants are grouped into subtypes (Robertson et al., 1998). Aside from the six genotypes, additional HCV variants were historically proposed as genotypes 7, 8, 9, 10 and 11 (Tokita et al., 1998). Phylogenetically, excluding genotype 10, these variants are all related closely to genotype 6 and have been proposed to represent multiple subtypes of genotype 6 (Simmonds et al., 1996). In the Los Alamos HCV database (http://hcv.lanl.gov/content/hcv-db/index), as many as 17 subtypes (6a–6q) are assigned to genotype 6. However, except for six subtypes (6a, 6b, 6d, 6g, 6h and 6k), all other subtypes still lack complete genomic sequences. Geographically, these variants were isolated exclusively from South-East Asia or from immigrants from this geographical region (Bernier et al., 1996; Mellor et al., 1996; Shinji et al., 2004; Stuyver et al., 1995; Thaiakrua et al., 2004; Theamboonlers et al., 2002). In one of our recent studies, four related variants of genotype 6 were also discovered in southern and south-western China (Lu et al., 2005a). Initial analysis of partial sequences using the Kimura two-parameter method (Kimura, 1980) showed that isolates km41 and km45 grouped ambiguously between subtypes 6k and 6l, and km42 was classified as subtype 6n. The fourth variant, gz52557, was tentatively assigned to a

†Deceased 21 November 2005; this paper has been dedicated to her memory by her former research fellows Ling Lu and Tatsunori Nakano.
‡These authors contributed equally to this work.
The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are DQ278892–DQ278894.

Supplementary material is available in JGV Online.
HCV genomic sequences were amplified with the primers listed in Supplementary Table S1 (available in JGV Online) by using previously described methods (Lu et al., 2005a, b) or a SMART RACE PCR kit (BD Clontech). The amplicons were sequenced directly and sequence analysis was performed by using GCG (version 10.0), PHYLML (http://atgc.lirmm.fr/phylml) and MEGA3 software (Kumar et al., 2004). Phylogenetic trees were reconstructed by using genetic distances calculated with the maximum-likelihood method (Guindon & Gascuel, 2003), of which the transition/transversion ratios and the proportion of invariant sites were estimated based on the real sequence datasets. Nucleotide and amino acid substitutions were estimated by using the HKY and Dayhoff models. Bootstrap analysis was performed in 500 replicates. Pairwise nucleotide similarities were estimated based upon p distances with MEGA3 software.

The entire genomes of km41 and km42 were amplified in 11 and nine fragments, respectively. The longest fragment was 4327 nt long and was amplified with strain-specific primers by using long RT-PCR. The two shortest fragments were 145 and 155 nt long and were amplified with degenerate and adapter primers by using rapid amplification of cDNA ends (RACE) PCRs (see Supplementary Table S1 in JGV Online). These fragments overlapped the entire genomes of km41 and km42, which were 9430 and 9441 nt, respectively. The 5’ untranslated regions (UTRs) were 341 and 342 nt, respectively, followed by single open reading frames (ORFs) of 9048 nt. The 3’ UTRs were 41 and 51 nt, respectively, containing 16 and 26 nt poly(U) tracts. The nucleotide compositions were 21-24 % A, 28-52 % C, 27-10 % G, 22-98 % T and 0-16 % mixed nucleotides for km41, and 21-34 % A, 28-73 % C, 26-54 % G, 23-23 % T and 0-16 % mixed nucleotides for km42. Both genomes shared common sizes with VN004 (GenBank accession no. D84265) and km45 (GenBank accession no. AY878650) in their complete ORFs.

Table 1. Nucleotide similarities (%) of km41, km42 and gz52557 with seven genotype 6 reference sequences

<table>
<thead>
<tr>
<th>Variant</th>
<th>Compared*</th>
<th>Complete†</th>
<th>5’ UTR</th>
<th>Core</th>
<th>E1</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’ UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>km41</td>
<td>EUHK2/6a</td>
<td>72.3</td>
<td>95.7</td>
<td>87.3</td>
<td>65.2</td>
<td>69.2</td>
<td>57.7</td>
<td>68.3</td>
<td>74.1</td>
<td>76.4</td>
<td>72.5</td>
<td>64.2</td>
<td>74.7</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>TH580/6b</td>
<td>72.9</td>
<td>96.7</td>
<td>86.6</td>
<td>63.4</td>
<td>68.4</td>
<td>56.6</td>
<td>68.6</td>
<td>74.5</td>
<td>77.5</td>
<td>73.6</td>
<td>67.4</td>
<td>75.6</td>
<td>58.5</td>
</tr>
<tr>
<td></td>
<td>VN235/6d</td>
<td>74.0</td>
<td>98.8</td>
<td>83.4</td>
<td>69.4</td>
<td>68.8</td>
<td>66.1</td>
<td>67.8</td>
<td>75.3</td>
<td>70.9</td>
<td>74.4</td>
<td>67.7</td>
<td>77.5</td>
<td>70.7</td>
</tr>
<tr>
<td></td>
<td>JK046/6g</td>
<td>73.3</td>
<td>98.2</td>
<td>84.5</td>
<td>66.5</td>
<td>68.7</td>
<td>55.6</td>
<td>66.6</td>
<td>73.8</td>
<td>78.6</td>
<td>74.6</td>
<td>68.5</td>
<td>76.4</td>
<td>61.0</td>
</tr>
<tr>
<td></td>
<td>VN004/6h</td>
<td>77.0</td>
<td>98.8</td>
<td>87.3</td>
<td>66.5</td>
<td>73.8</td>
<td>64.6</td>
<td>70.6</td>
<td>78.6</td>
<td>79.1</td>
<td>78.1</td>
<td>72.6</td>
<td>79.2</td>
<td>75.6</td>
</tr>
<tr>
<td></td>
<td>VN405/6k</td>
<td>83.4</td>
<td>97.0</td>
<td>92.2</td>
<td>82.7</td>
<td>78.6</td>
<td>75.1</td>
<td>78.5</td>
<td>83.1</td>
<td>83.0</td>
<td>81.3</td>
<td>85.1</td>
<td>86.6</td>
<td>92.7</td>
</tr>
<tr>
<td></td>
<td>km45/6k</td>
<td>93.7</td>
<td>99.4</td>
<td>96.5</td>
<td>92.0</td>
<td>92.6</td>
<td>88.4</td>
<td>92.2</td>
<td>93.7</td>
<td>93.1</td>
<td>95.9</td>
<td>92.7</td>
<td>94.8</td>
<td>92.7</td>
</tr>
<tr>
<td>km42</td>
<td>EUHK2/6a</td>
<td>72.3</td>
<td>95.7</td>
<td>86.6</td>
<td>66.5</td>
<td>68.3</td>
<td>54.0</td>
<td>67.4</td>
<td>75.2</td>
<td>77.0</td>
<td>70.4</td>
<td>63.9</td>
<td>76.1</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>TH580/6b</td>
<td>73.2</td>
<td>95.9</td>
<td>85.8</td>
<td>67.0</td>
<td>67.8</td>
<td>51.9</td>
<td>69.6</td>
<td>75.4</td>
<td>76.0</td>
<td>73.3</td>
<td>67.6</td>
<td>75.7</td>
<td>63.8</td>
</tr>
<tr>
<td></td>
<td>VN235/6d</td>
<td>74.6</td>
<td>97.0</td>
<td>84.6</td>
<td>69.6</td>
<td>71.3</td>
<td>61.4</td>
<td>68.4</td>
<td>76.5</td>
<td>72.1</td>
<td>75.0</td>
<td>67.3</td>
<td>78.0</td>
<td>76.1</td>
</tr>
<tr>
<td></td>
<td>JK046/6g</td>
<td>73.5</td>
<td>97.0</td>
<td>85.1</td>
<td>68.3</td>
<td>69.4</td>
<td>59.3</td>
<td>64.4</td>
<td>74.4</td>
<td>76.5</td>
<td>74.1</td>
<td>68.1</td>
<td>77.1</td>
<td>68.6</td>
</tr>
<tr>
<td></td>
<td>VN004/6h</td>
<td>77.8</td>
<td>96.4</td>
<td>89.6</td>
<td>70.4</td>
<td>73.7</td>
<td>69.3</td>
<td>71.6</td>
<td>80.5</td>
<td>78.7</td>
<td>78.2</td>
<td>72.7</td>
<td>79.0</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>VN405/6k</td>
<td>78.7</td>
<td>95.4</td>
<td>87.2</td>
<td>75.7</td>
<td>73.7</td>
<td>69.3</td>
<td>72.4</td>
<td>80.5</td>
<td>82.0</td>
<td>78.5</td>
<td>74.6</td>
<td>81.2</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>km41/6k</td>
<td>78.7</td>
<td>97.4</td>
<td>88.8</td>
<td>76.2</td>
<td>72.8</td>
<td>68.3</td>
<td>73.1</td>
<td>79.7</td>
<td>79.1</td>
<td>79.4</td>
<td>73.3</td>
<td>82.5</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td>km45/6k</td>
<td>78.5</td>
<td>96.8</td>
<td>88.8</td>
<td>73.7</td>
<td>72.6</td>
<td>67.7</td>
<td>71.7</td>
<td>79.5</td>
<td>82.0</td>
<td>80.6</td>
<td>73.4</td>
<td>81.6</td>
<td>88.2</td>
</tr>
<tr>
<td>gz52557</td>
<td>EUHK2/6a</td>
<td>71.7</td>
<td>96.8</td>
<td>83.1</td>
<td>66.2</td>
<td>68.8</td>
<td>57.7</td>
<td>63.9</td>
<td>74.0</td>
<td>75.4</td>
<td>72.1</td>
<td>65.0</td>
<td>74.3</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>TH580/6b</td>
<td>72.8</td>
<td>97.0</td>
<td>84.1</td>
<td>66.7</td>
<td>69.4</td>
<td>60.8</td>
<td>63.1</td>
<td>74.8</td>
<td>76.5</td>
<td>72.3</td>
<td>67.9</td>
<td>74.8</td>
<td>83.0</td>
</tr>
<tr>
<td></td>
<td>VN235/6d</td>
<td>73.8</td>
<td>99.4</td>
<td>83.1</td>
<td>69.4</td>
<td>68.1</td>
<td>64.6</td>
<td>68.4</td>
<td>74.2</td>
<td>71.6</td>
<td>74.6</td>
<td>67.0</td>
<td>77.6</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td>JK046/6g</td>
<td>75.6</td>
<td>98.8</td>
<td>86.9</td>
<td>72.0</td>
<td>70.2</td>
<td>62.4</td>
<td>69.3</td>
<td>75.1</td>
<td>78.7</td>
<td>76.3</td>
<td>71.0</td>
<td>79.7</td>
<td>92.3</td>
</tr>
<tr>
<td></td>
<td>VN004/6h</td>
<td>72.9</td>
<td>99.4</td>
<td>83.4</td>
<td>65.8</td>
<td>68.8</td>
<td>63.0</td>
<td>67.4</td>
<td>75.5</td>
<td>71.6</td>
<td>70.9</td>
<td>67.0</td>
<td>73.9</td>
<td>73.3</td>
</tr>
<tr>
<td></td>
<td>VN405/6k</td>
<td>73.3</td>
<td>97.9</td>
<td>81.7</td>
<td>67.5</td>
<td>68.2</td>
<td>66.1</td>
<td>66.2</td>
<td>75.0</td>
<td>69.4</td>
<td>72.5</td>
<td>68.0</td>
<td>76.6</td>
<td>76.9</td>
</tr>
<tr>
<td></td>
<td>km41/6k</td>
<td>73.7</td>
<td>98.5</td>
<td>82.0</td>
<td>66.8</td>
<td>68.7</td>
<td>64.6</td>
<td>67.4</td>
<td>75.6</td>
<td>74.2</td>
<td>75.1</td>
<td>68.2</td>
<td>76.3</td>
<td>70.7</td>
</tr>
<tr>
<td></td>
<td>km45/6k</td>
<td>73.8</td>
<td>98.5</td>
<td>82.7</td>
<td>67.5</td>
<td>68.6</td>
<td>64.6</td>
<td>68.4</td>
<td>75.8</td>
<td>73.3</td>
<td>74.2</td>
<td>68.2</td>
<td>76.4</td>
<td>82.7</td>
</tr>
<tr>
<td>km42/6n</td>
<td>73.4</td>
<td>96.7</td>
<td>83.7</td>
<td>67.1</td>
<td>69.0</td>
<td>54.5</td>
<td>65.9</td>
<td>75.1</td>
<td>77.0</td>
<td>72.5</td>
<td>67.2</td>
<td>77.3</td>
<td>76.5</td>
<td>–</td>
</tr>
</tbody>
</table>

*Isolate compared/subtype.
†Sequences over the complete genome length.
(9045 nt or 3015 aa) and in the ten protein-encoding regions. These include the core (573 nt or 191 aa), E1 (576 nt or 192 aa), E2 (1092 nt or 364 aa), P7 (189 nt or 63 aa), NS2 (651 nt or 217 aa), NS3 (1893 nt or 631 aa), NS4A (162 nt or 54 aa), NS4B (783 nt or 261 aa), NS5A (1353 nt or 451 aa) and NS5B (1776 nt or 591 aa) regions.

Pairwise comparisons of nucleotide sequences over the entire genome and in the ten protein-encoding regions showed that km41 was related more closely to VN405 and km45 than to any other genotype 6 isolate (Table 1). This was confirmed by phylogenetic analyses with complete nucleotide and deduced amino acid sequences from 25 reference isolates representing various HCV genotypes and subtypes (Fig. 1). The phylogenetic trees showed that km41 clustered closely with VN405 and km45. VN405 represents the prototype subtype 6k isolate in the nomenclature where HCV is classified into six genotypes (Simmonds et al., 1994). By this nomenclature, HCV genotypes differ by 31–33 % nucleotides and subtypes by 20–25 % over the entire genome range (Simmonds et al., 1994). Based on the phylogenetic analyses using the maximum-likelihood and neighbour-joining methods, recommended in a recent consensus (Simmonds et al., 2005), and the 83–4 % nucleotide similarity by which both km41 and km45 resembled VN405 over the entire genomes (Table 1), we can classify km41 and km45 as new members of subtype 6k.

Similar pairwise comparisons of nucleotide sequences showed that km42 was related slightly more closely to VN004 and VN405. However, the nucleotide similarities were lower than those observed between km41 and VN405. Over the entire genome, the similarities of km42 to VN004 and VN405 were 77-8 and 78-7 %, respectively (Table 1). These fall into the typical range of different HCV subtypes within a genotype (Simmonds et al., 1994, 1996). Phylogenetic analysis using 25 referenced complete genome sequences showed that km42 resulted in a branch between subtypes 6k and 6l. This suggests that km42 represents a different subtype that still lacks a full genomic sequence, as all available complete genome sequences for genotype 6 isolates were included in the trees (Fig. 1). We previously classified km42 tentatively as subtype 6n by using partial sequence data (Lu et al., 2005a). In order to relate km42 to additional isolates characterized by using partial sequences, we further analysed km42 in three different genomic regions. These included a 321 nt partial core region, a 468 nt partial E1 region and a 329 nt partial NS5B region. Phylogenies using the core and E1 regions demonstrated that km42 clustered with five and 11 sequences, respectively, all of which were isolated from Thailand (with bootstrap values of 58 and 100 %; see Supplementary Fig. S1 in JGV Online) (Doi et al., 1996; Mellor et al., 1995; Thaikruea et al., 2004; Theamboonlers et al., 2002). However, the tree based on the NS5B region showed that km42 grouped with a cluster of sequences that were found exclusively in Myanmar (with a bootstrap value of 100 %, identified by a dashed circle in Fig. 2a) (Shinji et al., 2004). These findings suggest a common source of infection for these subtype 6n isolates.

Fig. 1. Table-like phylogenetic trees based on (a) the complete HCV genome sequences and (b) the complete amino acid sequences deduced. The six HCV genotypes are indicated by numbers 1–6, subtypes are designated 1a–6n and gz52557 is identified by 6? (subtype unassigned). Reference HCV sequences are indicated by an isolate name followed by GenBank accession number in parentheses, except for NMR2A75-5, which was not submitted to GenBank (Inoue et al., 2000). The four distinct Chinese HCV variants determined in this and our previous study are shown in bold. Bootstrap-analysis values are shown in italics. Bars, 0–10 nucleotide substitutions per site.
Because of a relatively low virus titre (10^4 genome copies ml^{-1}), gz52557 could not be amplified by using long RT-PCR. However, a series of 14 overlapping PCR fragments of 156–1876 nt was amplified by using primers listed in Supplementary Table S1 (available in JGV Online). Among them, there was a fragment of 402 nt from the core region, a second fragment of 377 nt from the NS5B region and a third fragment of 423 nt from the NS3 region. They were amplified with nested or semi-nested PCR by using degenerate primers. Their sequence information was subsequently used as ‘islands’ to guide further amplifications with a PCR strategy described previously (Lu et al., 2005b). This resulted in the acquisition of nine additional fragments for sequencing. The three longest fragments were 1876, 1647 and 1245 nt and were amplified with a combination of degenerate primers at the 5' ends and strain-specific primers at the 3' ends. In addition, two fragments were amplified by using RACE PCR to gain the 5' and 3' UTR ends. Collectively, the 14 fragments overlapped the entire 9448 nt gz52557 genome. The 5' UTR was 339 nt, followed by a single ORF of 9057 nt (capable of encoding 3018 aa) and 52 nt of 3' UTR, including a poly(U) stretch of 22 nt. The nucleotide compositions were 20–99% A, 29–64% C, 27–15% G, 22–18% T and 0–04% mixed nucleotides. Except for the E2 (1104 nt or 368 aa) and NS5A (1350 nt or 450 aa) regions, gz52557 shared common sizes with km41 and km42 in eight protein-encoding regions. Pairwise comparison of nucleotide sequences over the entire genome indicated that gz52557 was related most closely to JK046 (GenBank accession no. D63822), but was equidistant from other isolates of genotype 6. It had a nucleotide similarity of 75·6% to JK046 over the entire genome, and in seven protein-encoding regions the similarities were 69·3–79·7%. The 5' UTR and core region were most conserved, whilst in the P7 region, gz52557 resembled VN405 more closely, and in the NS3 region, gz52557 was related more closely to km41 and km45 (Table 1). Phylogenetic analyses using 25 reference nucleotide and deduced amino acid sequences over the entire genome (Fig. 1) and in the eight protein-encoding regions, excluding the P7 and NS3 regions (trees not shown), demonstrated that gz52557 clustered distantly with JK046, but appeared to represent a different subtype. The inclusion of 59 and 104 partial core and NS5B region sequences representing subtypes 6a–6q and three unclassified genotype 6 isolates (IG57272, QC66 and MYAN-3E-3) in co-analyses further demonstrated that gz52557 formed a single branch

**Fig. 2.** Star-like phylogenetic trees based on (a) 105 partial NS5B region sequences (corresponds to nt 8294–8622 of the JK046 genome) and (b) 53 partial core region sequences (corresponds to nt 339–737 of the JK046 genome) (for names of sequences and their GenBank accession numbers, see Supplementary Table S3 in JGV Online). For both trees, subtypes are identified at branch ends or outside the dashed curves that define clusters. The ED43 (GenBank accession no. Y11604) sequence of genotype 4a is used as an outgroup. The symbol 6? in bold indicates the isolates with their subtypes unassigned. The four sequences determined in this and our previous study are shown in bold. The dashed circle in (a) identifies the group of subtype 6n sequences that were isolated exclusively from Myanmar and with which the km42 variant clustered with a bootstrap value of 100%. Bars, 0·05 nucleotide substitutions per site (a); 0·02 nucleotide substitutions per site (b).
that should represent a new subtype (Fig. 2). Pairwise comparisons of gz52557 with representative partial sequences from these 17 subtypes (6a–6q) and three unclassified isolates in the NS5B region showed that they were nearly equidistant from gz52557, with 68–1–76·3% nucleotide similarities (see Supplementary Table S2 in JGV Online). Collectively, these results support the proposal that gz52557 represents a novel subtype of HCV genotype 6.

Because km42 was positioned between subtypes 6h and 6k and gz52557 between subtypes 6d and 6g in phylogenetic trees (Fig. 1), it remains possible that these two isolates represent recombinants of the related subtypes. However, when the similarities of km42 and gz52557 were plotted against any of the following reference sequences over the entire genome (JK046/6g, TH580/6b, VN235/6d, VN405/6k, VN004/6h, EUHK2/6a), almost-identical similarity-distribution curves were observed (data not shown). This provided evidence that km42 and gz52557 are not recombinants of these six subtypes.

In one of our recent studies, variant km41 was found to cluster ambiguously between subtypes 6k and 6l based on the analysis of partial genomic sequences by using the Kimura two-parameter method (Lu et al., 2005a). After determining its entire genome sequence followed by pairwise nucleotide comparison and phylogenetic analysis using a more precise maximum-likelihood method, we determined that this variant belonged to subtype 6k. A single isolate of subtype 6k, VN405, was previously discovered from one of four commercial blood donors from Hanoi, Vietnam (Tokita et al., 1994). The km41 isolate represented the third sequence for subtype 6k and the second isolate of this subtype from China. Another subtype 6k isolate, km45, was also reported by us and was characterized from a sample in the same collection from Kunming in south-west China (Lu et al., 2005b). The lack of additional subtype 6k isolates appearing between the time points at which VN405 (1994), km41 (2002) and km45 (2002) were identified makes it difficult to link the variants by a chain of transmission.

A third HCV variant, km42, was sequenced in its entirety from a sample that was also collected in Kunming, China. This variant represents the first isolate of subtype 6n from China. Collectively, eight previous studies identified a total of 42 subtype 6n isolates. Among them, 11 were from Yangon, Myanmar, and the remainder were from Chiang Mai, Thailand (Apichartpiyakul et al., 1994; Doi et al., 1996; Mellor et al., 1995, 1996; Shinji et al., 2004; Simmonds et al., 1996; Theamboonlers et al., 2002; Thaikruenae et al., 2004). A major epidemiological feature of these variants was that they were identified most frequently in blood donors or intravenous drug users without obvious manifestations of chronic hepatitis (Doi et al., 1996). Of the 42 subtype 6n isolates, the first three to be described were BB9, D86/93 and D97/93 (Apichartpiyakul et al., 1994). EUTH19, EUTH49 and EUTH86 were initially designated NG(I) to denote a new genotype (Mellor et al., 1995, 1996). Together, they were later reclassified as subtype 6n (Simmonds et al., 1996).

In addition, nine isolates were mistyped as subtype 6a (Theamboonlers et al., 2002), 11 isolates from Yangon were designated the M6-2 group (Shinji et al., 2004) and three of the remaining 11 sequences from Chiang Mai were not yet assigned to any subtype (Thaikruenae et al., 2004). Phylogenetically, km42 clustered closely with these isolates and Kunming, Chiang Mai and Yangon are in geographically adjacent regions (Mellor et al., 1996). Therefore, a common source of infection might be suggested for these subtype 6n variants.

Some HCV variants have been misgenotyped or designated unclassified, due to a lack of complete genomic sequence information. One isolate, 99CHNGX004 (GenBank accession no. AY732110), provides an example of this. It was recently characterized from Guangxi Province, China, from a human immunodeficiency virus-positive blood sample and was thought to be an unclassified genotype 6 variant, based on a partial E2 region sequence (http://hcv.lanl.gov/content/hcv-db/index) (Zhang et al., 2004). However, after sequencing the E1 and NS5B regions, we found that it belonged to subtype 6e (C. Li, Y. Fu, L. Lu, C. H. Hagedorn, J. Yu & L. Zhang, unpublished data). This example emphasizes why a complete genomic sequence should be required for each HCV subtype and how this information is critical for genotyping HCV isolates precisely. Based on this consideration, we determined the entire genomic sequence for km42. Although many isolates have been classified into subtype 6n, km42 now represents the first complete genomic sequence for this subtype. Out of 17 subtypes within genotype 6, complete genomic sequences have been determined for only seven subtypes [6a, 6b, 6d, 6g, 6h, 6k and 6n (from this study)]. In order for a comprehensive phylogenetic analysis to be performed for any new genotype 6 isolate, we recommend that at least one entire genomic sequence be determined for each of these subtypes.

Using the serum sample of a patient from Guangzhou, China, a complete genomic sequence for the gz52557 variant was determined. Phylogenetic analysis and pairwise comparison for nucleotide similarities demonstrated that it was nearly equidistant from all other genotype 6 subtypes. These data suggest that gz52557 may represent a novel subtype. However, a recent consensus stated that sequence from both the core/E1 and NS5B regions of three or more examples of infection are required to designate a new subtype (Simmonds et al., 2005). Based on this, it would be premature to assign gz52557 to a new subtype, and we therefore suggest that gz52557 remains an unclassified isolate. The origin of gz52557 is obscure. Because it was related slightly more closely to subtype 6g than to any other genotype 6 subtypes, gz52557 may have originated from another South-East Asian country, such as Indonesia, where subtype 6g is prevalent (Inoue et al., 2000; Tokita et al., 1998). Alternatively, it may be indigenous to southern China. The failure of previous studies to find this may have been due to technical limitations (Mellor et al., 1996), a low prevalence of infection or undersampling. To answer these questions
and define the subtype designation for gz52557 would require a more extensive molecular-epidemiological survey to be carried out in southern China. Such a survey should include the analysis of additional sequences from related isolates and may provide a chance for identifying new HCV variants.

Acknowledgements

This work was supported by NIH grants CA-063064 and 5 P20 RR016443-04.

References


Tokita, H., Okamoto, H., Tsuda, F. & 7 other authors (1994). Hepatitis C virus variants from Vietnam are classifiable into the seventh, eighth, and ninth major genetic groups. Proc Natl Acad Sci USA 91, 11022–11026.
