Hepatitis C virus (HCV) is an enveloped positive-stranded RNA virus that belongs to the family Flaviviridae and it is the leading cause of acute and chronic hepatitis in about 3% of the human population (Shepard et al., 2005). The entire viral genome contains about 9600 nt encoding structural (core, E1 and E2) and non-structural (P7 and NS2–NS5B) proteins flanked by untranslated regions (Choo et al., 1991). The virus shows extreme genetic diversity and currently six major groups (genotypes 1–6) and a large number of subtypes have been identified (Simmonds, 2004). Viral genomes of different genotypes may vary by 30–35%, while nucleotide sequences of subtypes may differ by 15–20% (Simmonds et al., 2005).

HCV genotype 6 viruses mainly circulate in South East Asia (SEA), including Vietnam (Tokita et al., 1994), Indonesia (Tokita et al., 1996), Thailand (Chinchai et al., 2003; Thaikruia et al., 2004), Cambodia (Caruana et al., 2005) and Myanmar (Shinji et al., 2004). Currently, genotype 6 viruses are classified into 17 subtypes (6a–6q) based on complete genome analysis (Lu et al., 2006, 2007a; Tokita et al., 1996, 1998). Recently, the existence of subtype 6r and 6s viruses was proposed based on sequence analysis of core/E1 and NS5B regions from unclassified genotype 6 isolates from Asian immigrants in Canada revealed that two other viruses (D49 and D83) could be classified as novel candidates of HCV subtypes 6t and 6u.

High nucleotide sequence heterogeneity hampers complete genome characterization of new subtypes due to inefficient primer annealing. We developed techniques to obtain large RT-PCR fragments amplified from HCV-positive sera using degenerate primer sets and obtained entire HCV genome sequences from 100 μl HCV-positive sera (Noppornpanth et al., 2006a). This led to the complete genome characterization of a Vietnamese HCV recombinant genotype 2i/6p strain (Noppornpanth et al., 2006a). In this report, we selected several sera from HCV-positive blood donors from Vietnam and Thailand based on a preliminary screen to include rare or unclassified HCV subtypes 6 (Noppornpanth et al., 2006a, b) and performed full-genome sequencing to characterize further HCV subtypes 6.

Total RNA was isolated from 100 μl serum by proteinase K/phenol/chloroform extraction as described previously (Rispens et al., 1997). The entire HCV genome was amplified from three different cDNA reverse transcription reactions from positions 5460, 8625 and 9325 (numbering system according to prototype strain H77; GenBank accession no. AF009606). Three overlapping large PCR fragments were first-round amplified by degenerate primers, followed by nested PCR as described previously (Noppornpanth et al., 2006a). Amplified products were gel-purified with the QIAquick gel extraction kit (Qiagen) and ligated directly into plasmid pCR2.1 (TA cloning kit; Invitrogen). For each amplicon, five positive clones were selected and sequenced. Multiple sequence alignment was performed with the BioEdit program (version 7.0.1) according to amino acid alignment. Phylogenetic trees

The GenBank/EMBL/DDBJ sequence accession numbers for the sequences reported in this study are EU246930–EU246940.
were constructed from tree puzzle 5.2 (Schmidt et al., 2002) and MrBayes version 3.0b4 (Ronquist & Huelsenbeck, 2003) programs using genetic distance calculated by maximum-likelihood with rate heterogeneity and Bayesian methods, respectively. Transition/transversion ratio, proportion of invariant sites and gamma distribution factors were estimated based on the real sequence datasets. Nucleotide and amino acid substitutions were estimated by using HKY and WAG models, respectively. Bootstrapping was calculated from 10,000 tree-puzzle steps or 5000 trees generated from the MrBayes program. Pairwise nucleotide comparison was estimated based upon p distances with MEGA3 software (Kumar et al., 2004). Phylogenetic analysis and sequence similarity test were performed using reference sequences of HCV subtypes 6a–6q, published in the Los Alamos HCV database (http://hcv.lanl.gov/content/hcv-index). Genetic recombination events were examined by similarity plot and bootscanning analysis, using the SimPlot program (version 3.5.1 for Windows98/NT/2000).

We characterized 11 complete HCV genotype 6 genomes, including seven isolates from Vietnam (D9, D33, D42, D49, D83, D85 and D88) and four isolates from Thailand (TH22, TH24, TH31 and TH52). These genomes varied in length from 9373 to 9423 nt, encoding 3016 to 3023 aa (Supplementary Table S1 available in JGV Online).

The whole genome of TH52 shared the main branch with samples C-0044 and C-0046, representative of HCV subtype 6f (Fig. 1). There is a high sequence similarity (94.7 %) compared to the prototype strain and results from the short fragment analysis confirmed that TH52 is a member of HCV subtype 6f (Table 1 and Supplementary Fig. S2 available in JGV Online). Interestingly, 52/53 samples of subtype 6f classified in the Los Alamos database were obtained from Thailand (Kuiken et al., 2005). Full-length genomes of HCV subtype 6i (Th602 and C-0159), a

![Phylogenetic tree](image-url)

Fig. 1. Phylogenetic tree based on complete nucleotide sequences. The six genotypes are indicated in bold, subtypes are designated 1a–6q and 6 (unassigned subtype). The HCV prototype sequences are indicated by isolate name followed by the GenBank accession number in parentheses. Confidence values (>70 %) calculated from 10,000 tree-puzzle steps are indicated at the major branching points. Branch lengths are drawn to scale. The 11 HCV strains determined in this study are shown in bold with asterisks. Unassigned samples are specified with circular marks. Nucleotide sequences presented in this study were submitted to GenBank; accession numbers are EU246930–EU246940.
subtype found only in Thailand, have been sequenced recently (Lu et al., 2007b). TH24 complete genome and core, E1 and NS5B phylogenetic trees showed 99–100% bootstrap support related to the C-0159 sequence (Fig. 1). We identified TH22 and TH31 strains as HCV subtype 6n based on phylogenetic analysis of full-length nucleotide and amino acid sequences (Fig. 1 and Supplementary Fig. S1), while whole genome sequence similarity was 93.9 and 95.8%, respectively, when compared with subtype 6n reference strain D86/93 isolated from Thailand (Lu et al., 2007b). Both strains clustered with KM42 isolated from China (Lu et al., 2005) and D86/93 with high bootstrap support. HCV subtype 6n was first isolated from Thailand (Apichartpiyakul et al., 1994), and then later reported to be isolated from India, China and Myanmar (Shinji et al., 2004).

D9 showed 92.6% sequence similarity with subtype 6a (strain 6a33, GenBank accession no. AY859526) and 100% bootstrapping phylogeny support using complete, short fragment nucleotide and full-length amino acid sequences (Table 1, Fig. 1 and Supplementary Figs S1 and S2). HCV subtype 6a was first discovered in 1993 in Hong Kong (Simmonds et al., 1993) and has been distributed throughout Asia, including China (Lu et al., 2005), Vietnam (Noppornpanth et al., 2006b), Taiwan and Singapore (Los Alamos HCV database).

Phylogenetic analysis of short regions (core, E1 and NS5B) indicated that D33 clustered to the subtype 6l prototype strain 537796 (Supplementary Fig. S2). Complete genome phylogeny and pairwise similarity (95.1%) confirmed classification of D33 as subtype 6l (Fig. 1 and Table 1). Remarkably, subtypes 6l and 6k were isolated from Vietnamese individuals (Lu et al., 2006, 2007a). D83 on the other hand had 92.5% sequence similarity to subtype 6o prototype strain QC227 isolated from an Asian immigrant in Canada (Murphy et al., 2007), confirmed by phylogeny of whole genome and short fragment analysis (Fig. 1 and Supplementary Fig. S2). Subtype 6o is a rare subtype in SEA and we detected only one subtype 6o virus out of 95 HCV-positive samples collected from Vietnam (data not shown).

Based on full-length nucleotide and amino acid sequences D42, D88 and D83 are closely related to subtype 6e strain GX004, recently reported from China (Li et al., 2006) (Fig. 1 and Supplementary Fig. S1). Sequence similarities compared with GX004 were 87.0, 90.8 and 83.6% for D42, D88 and D83, respectively (Table 1). However, phylogenetic analysis of complete genome, core and NS5B sequences revealed that D83 branched separately from other subtypes (Fig. 1 and Supplementary Fig. S2). D83 was about 18% different from D42 and D88 at nucleotide level and about 10% different at amino acid level (data not shown).

Based on nucleotide and amino acid sequences of the whole genome, D49 branched closely to subtype 6q (Fig. 1 and Supplementary Fig. S1). However, short genome fragment analysis using core and NS5B revealed branching...
with subtypes 6q and 6d, respectively, while E1 analysis showed a separate branch linked to subtype 6p (Supplementary Fig. S2). Sequence similarities compared with subtypes 6a–6q prototype sequences were 72.6–79.2% (Table 1).

Although complete genome sequences of HCV subtypes 6a–6q were used as reference strains, some genomes such as GZ52557 and KM41 isolated from China (Lu et al., 2005), and at least 23 strains (provided in the Los Alamos HCV database) based on short fragment sequences of core/E1 and NS5B (Supplementary Table S2), could not be classified. Because we could not subtype D49 and D83, we reanalysed the phylogenetic tree including all subtypes (6a–6s) and unassigned subtype samples reported thus far. Now the core/E1 region of D49 showed close relationship with subtype 6p and clustered with QC240, QC131 and QC145, while NS5B clustered with the same viruses but separated more clearly from other subtypes (Fig. 2a and b). Unexpectedly, both core/E1 and NS5B from D83 grouped with QC191 and QC323 viruses and clearly clustered separately from subtype 6e (Fig. 2a and b). Thus, viruses D49, QC240, QC131 and QC145 were assigned to the novel HCV subtype 6t, while D83, QC191 and QC323 are candidate subtype 6u viruses.

Using full-length genome sequences is considered the gold standard of HCV genotype/subtype identification. Due to the heterogeneity, whole genome amplification using specific primers is troublesome and 25–30 amplicons are required (Lu et al., 2006). We used degenerate primers to amplify a large PCR fragment and generated three to five amplicons, resulting in cost and time efficient characterization of the HCV genomes. Based on the pairwise sequence similarity and phylogenetic analysis of full-length and short fragment sequences, we identified complete genomes of HCV subtype 6a (D9), 6e (D42 and D88), 6f (TH52), 6i (TH24), 6l (D33), 6n (TH22 and TH31) and 6o (D85). Moreover, two new subtypes 6t (D49) and 6u (D83) were assigned. Recombination between the two newly

![Fig. 2. Rooted neighbour-joining trees depicting the phylogenetic relationship among HCV genotype 6 unassigned sequences and prototype strains. Trees were constructed for (a) a 424 nt segment of the core/E1 genes (positions 869–1292) and (b) a 324 nt segment of the NS5B gene (positions 8282–8605). Bootstrap values (5000 replicates) are shown along each main branch. Branch lengths are drawn to scale. The prototype HCV genotype 6 strains, indicated by isolate name, are designated 6a–6s and 6 (unassigned subtype). GenBank accession numbers of subtype 6s and 6r are indicated in parentheses after the isolate name. HCV genotype 1a (H77, GenBank accession no. AF009606) was used as an outgroup sequence. The 11 HCV strains obtained in this study are shown in bold with asterisks. The new assigned subtypes 6t and 6u are indicated.](http://vir.sgmjournals.org)
described genomes and other HCV sequences was not observed (data not shown).

Although phylogenetic analysis based on the full-length genome is the most accurate method for HCV classification, most HCVs have been classified based on short sequence analysis of the core, E1 and NS5B regions (Robertson et al., 1998; Sandres-Saune et al., 2003; Simmonds et al., 2005). Historically, HCV classification based on short sequence phylogenetic analysis using a simple nucleotide substitution model has led to incorrect branching between closely related subtypes. HCV genotypes 7, 8 and 9 from SEA have been identified using the unweighted pair-group method with arithmetic mean (UPGMA) for phylogenetic tree construction (Tokita et al., 1994). However, combining the complex nucleotide substitution model with neighbour-joining results in grouping of these sequences to HCV genotype 6 subtypes (Mizokami et al., 1996; Simmonds et al., 1994). In our study, we performed the classification based on the HKY nucleotide substitution model and maximum-likelihood method to increase the accuracy of nucleotide phylogenetic analysis. Moreover, we generated protein phylogenetic trees based on amino acid sequences to confirm our results.

Even though a range of HCV subtype 6a–6q whole genomes has been characterized, a number of viruses cannot be classified (Murphy et al., 2007). These unassigned samples may represent novel candidate HCV subtypes. Assignment of subtypes 6t and 6u in our study is in line with the HCV genotyping consensus proposal (Simmonds et al., 2005). NS5B and core/E1 sequences of at least three examples are required for a provisional designation of HCV new subtype and at least 15% difference at the nucleotide level from other HCV subtypes should be present (Simmonds et al., 2005). We have defined two full-length genome sequences that represent new subtypes of HCV genotype 6. Subtype 6t comprises D49, QC240, QC131 and QC145, all isolated from Vietnamese individuals, and subtype 6u includes D83, which grouped to QC191 and QC323 and shows a separate branch from subtype 6e. Sample QC191 and QC323 originated from Asian immigrants in Canada but remained unassigned with respect to subtype classification (Murphy et al., 2007). Grouping of D83 sequence obtained from our study with these strains suggests that subtype 6u circulates in SEA. High sequence similarity of HCV subtypes 6e and 6u indicates that both may share the same origin but have shown genetic drift as a result of long-term circulation.

Molecular evolution of HCV genotype 6 based on phylogenetic relationship separates all subtypes into three groups. Subtypes 6a and 6b are related to each other, subtypes 6c, 6d, 6e, 6f, 6g, 6p, 6q and 6o share the same phylogenetic branch and the last group contains subtypes 6i, 6j, 6k, 6l, 6m, 6n and 6h. Epidemiological studies of HCV genotype 6 have revealed that certain subtypes, such as 6d, 6h and 6i, are found in Vietnam while 6c, 6f, 6j and 6l are found in Thailand, whereas subtype 6m and 6n were isolated mainly from Thailand and Myanmar. Although there may be some sampling bias and information about subtype distribution is lacking from some countries such as Laos, Cambodia and Myanmar, these results suggest that spread of HCV occurs in closed populations. These data are consistent with the idea that HCV genome diversity is related to geographical distribution and routes of viral transmission in certain areas. Whereas genotypes 1, 2 and 3 have spread worldwide mainly through contaminated blood and intravenous drug users (Alter, 1999), the genotype 4, 5 and 6 viruses are more restricted to African and Asian countries (Simmonds et al., 2005) and vary genetically as a result of long endemicity in certain areas (Pybus et al., 2001).

Acknowledgements

This work was supported by the Commission of the European Community (HECSA Project ICA4-CT-1999-10009). We would like to thank Dr Vu Thuy Yen, Jan Brouwer, Paula van Luijt, Apiradee Theamboonlers and Kavita Bedi for their technical assistance.

References


Complete genome analysis of HCV subtypes 6t and 6u


