Origin of hepatitis C virus genotype 3 in Africa as estimated through an evolutionary analysis of the full-length genomes of nine subtypes, including the newly sequenced 3d and 3e

Chunhua Li, Ling Lu, Donald G. Murphy, Francesco Negro and Hiroaki Okamoto

We characterized the full-length genomes of nine hepatitis C virus genotype 3 (HCV-3) isolates: QC7, QC8, QC9, QC10, QC34, QC88, NE145, NE274 and 811. To the best of our knowledge, NE274 and NE145 were the first full-length genomes for confirming the provisionally assigned subtypes 3d and 3e, respectively, whereas 811 represented the first HCV-3 isolate that had its extreme 3′ UTR terminus sequenced. Based on these full-length genomes, together with 42 references representing eight assigned subtypes and an unclassified variant of HCV-3, and 10 sequences of six other genotypes, a timescaled phylogenetic tree was reconstructed after an evolutionary analysis using a coalescent Bayesian procedure. The results indicated that subtypes 3a, 3d and 3e formed a subset with a common ancestor dated to ~202.89 [95 % highest posterior density (HPD): 160.11, 264.6] years ago. The analysis of all of the HCV-3 sequences as a single lineage resulted in the dating of the divergence time to ~457.81 (95 % HPD: 350.62, 587.53) years ago, whereas the common ancestor of all of the seven HCV genotypes dated to ~780.86 (95 % HPD: 592.15, 1021.34) years ago. As subtype 3h and the unclassified variant were relatives, and represented the oldest HCV-3 lineages with origins in Africa and the Middle East, these findings may indicate the ancestral origin of HCV-3 in Africa. We speculate that the ancestral HCV-3 strains may have been brought to South Asia from Africa by land and/or across the sea to result in its indigenous circulation in that region. The spread was estimated to have occurred in the era after Vasco da Gama had completed his expeditions by sailing along the eastern coast of Africa to India. However, before this era, Arabians had practised slave trading from Africa to the Middle East and South Asia for centuries, which may have mediated the earliest spread of HCV-3.

INTRODUCTION

Hepatitis C virus (HCV) is a blood-borne pathogen that infects an estimated >185 million people worldwide, which corresponds to 2.8 % of the global population (Alter, 1999).

The GenBank/EMBL/DDBJ accession numbers for the nucleotide sequences of the HCV-3 isolates are JF735126 and KJ470612-KJ470619. One supplementary table is available with the online version of this paper.

The infection is characterized by the establishment of chronic hepatitis in 70–85 % of the infected individuals, of whom >20 % subsequently develop cirrhosis, liver failure and hepatocellular carcinoma, (Hoofnagle, 2002; Zoulim et al., 2003). HCV is a cause of substantial morbidity and mortality, the rates of which are expected to increase over the next decades (Williams, 1999).

HCV possesses a positive-sense ssRNA genome of ~9600 nt in length and is classified into the Hepacivirus genus of the Flaviviridae family (Thiel et al., 2005). Based on a recent
paper on its expanded classification, HCV has now been divided into seven genotypes and 67 subtypes, of which 10 subtypes and an unclassified novel variant have been assigned within genotype 3 (Smith et al., 2014). Different genotypes exhibit different geographical distribution patterns and are associated with different treatment outcomes. In general, genotypes 1–3 are distributed worldwide, whereas genotypes 4–7 are restricted to certain regions (Simmonds et al., 2005). Clinically, those infected with genotypes 2 and 3 present a higher rate of sustained virological responses to therapy with IFN-α and ribavirin than those infected with genotypes 1 and 4 (Feld & Hoofnagle, 2005; Manns et al., 2006). With respect to complications, the isolates of genotype 3 are linked to the development of more pronounced hepatocellular steatosis (Abid et al., 2005; Hourioux et al., 2007; Jackel-Cram et al., 2007; Negro, 2006; Rubbia-Brandt et al., 2000), an accelerated fibrosis progression rate (Bochud et al., 2009), an increased risk of developing hepatocellular carcinoma (Nkontchou et al., 2011) and an increased all-cause mortality (van der Meer et al., 2012). Although the mechanisms involved in HCV-3-associated pathogenesis are not fully understood, they may be due to the genetic variation of the virus.

With the completion of one of our recent studies, six subtypes (3a, 3b, 3g, 3h, 3i and 3k) in genotype 3 have been confirmed and their full-length genomes have been characterized (Lu et al., 2013). However, due to the availability of only partial sequences, four other subtypes (3c, 3d, 3e and 3f) have not yet been sequenced completely. In the present study, we determined the full-length genomes of subtypes 3d and 3e to partially fill this information gap.

Among the subtypes in genotype 3, both 3a and 3b are distributed worldwide and are often detected in developing countries—a phenomenon that is ascribed to their propensity of being transmitted via intravenous drug use (IDU) (Bourliere et al., 2002; McCaw et al., 1997; Pawlotsky et al., 1995; Silini et al., 1995). In contrast, other subtypes have been detected mainly in South Asia and its neighbouring regions, and thus show a higher genetic complexity that indicates their possible long-term indigenous circulation and origin in the regions (Hotta et al., 1994; Lole et al., 2003; Lu et al., 2013; Murphy et al., 2007; Okamoto et al., 1994; Panigrahi et al., 1996; Simmonds et al., 1996, 2005; Tokita et al., 1996; Valliammai et al., 1995). Although a limited number of subtype 3h isolates have been detected in countries that are geographically well separated, these were obtained exclusively from immigrants who originated in Africa and showed the longest genetic distances compared with all of the other assigned subtypes (Abid et al., 2000; Bernier et al., 1996; Colson et al., 2011; Corbet et al., 2003; Murphy et al., 2007). In addition to some other variants originating in the Middle East, two isolates of the provisionally proposed subtype 3l were identified recently in Iran (Amini et al., 2006). These are two lines of evidence that may challenge the commonly accepted belief that HCV-3 may have its origin in South Asia (Simmonds et al., 2005). Instead, the ancestral origin may have been in Africa and this genotype may have been brought to areas outside of Africa via the Middle East or across the sea to South Asia. To test this hypothesis, we determined the full-length genomes of four subtype 3a, two subtype 3b and one subtype 3h isolates to not only calculate a molecular clock for genotype 3, but also perform evolutionary analysis. Using the sequence data retrieved from the Los Alamos HCV database (http://www.hcv.lanl.gov), and determined in this and previous studies, a timescaled phylogenetic tree was reconstructed. Based on this tree, the ancestral origin of HCV-3 was estimated and a new hypothesis was suggested.

RESULTS

Genome sequence and organization

The full-length genomes of nine HCV-3 isolates were characterized: NE145, NE274, 811, QC7, QC8, QC9, QC10, QC34 and QC88, each using 16–23 overlapping fragments. Of these isolates, 811 had a genome size of 9633 nt, beginning from the 5’ terminus and ending at the extreme 3’ end, with the latter obtained using 3’ RACE reverse transcription (RT)-PCR. Each of the other eight genomes spanned 9458–9465 nt from the 5’ UTR to the poly(U) track of the 3’ UTR (Table 1). These nine genomes each contained a single ORF of 9060–9069 nt. The 5’ UTRs were 339–340 nt in length, whereas the 3’ UTRs varied from 52 to 234 nt. The sizes of the 10 protein-coding regions were the same as those of the H77 genome with the exception of those of the E2 and NS5A regions, which were variable in length (Table 1).

Fig. 1(a) shows an alignment of the 3’ UTR sequences from six isolates, which were selected to represent different genotypes of HCV. The 811 isolate shows a 3’ UTR that is 234 nt in length, consisting of a type-specific region immediately following the stop codon at the end of the NS5B region, a poly(U) stretch, a C(U)_n-repeat and a highly conserved 98 nt region at the extreme 3’ end (Kolykhalov et al., 1996; Yamada et al., 1996). Using an online computer program (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi), we predicted the secondary structure for this 234 nt RNA sequence, which showed five typical hairpins, as reported previously (Fig. 1b). To obtain a better understanding of the structural similarity, we pairwise compared the 811 sequence in this region against that of the five other isolates, i.e. WS-#16, HCVgenotype4a-KM, JFH1, Th580 and H77, and revealed p distances of 0.225, 0.275, 0.298, 0.326 and 0.331, respectively. This finding indicated that the 811 sequence in this region was more closely related to WS-#16, which belongs to subtype 3a (Kolykhalov et al., 1996). The results also indicated that the 98 nt in the extreme 3’-terminus were highly conserved among the genotypes. Although a few substitutions were observed in the very last 46 nt, these did not appear to disturb the stability of the predicted secondary structure (Fig. 1c) (Blight & Rice, 1997).

Genotype 3 classification

Fig. 2 depicts a maximum-likelihood (ML) tree that was reconstructed with the 45 full-length HCV genome
Evolutionary analysis of HCV-3 full-length sequences

Table 1. Patient information for the nine genotype 3 isolates and the variable number of nucleotides/amino acids in certain genomic regions

<table>
<thead>
<tr>
<th>ID</th>
<th>Subtype</th>
<th>Origin</th>
<th>Full Model</th>
<th>ORF</th>
<th>5' UTR</th>
<th>E2</th>
<th>NS5A</th>
<th>3' UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>H77</td>
<td>1a</td>
<td>USA</td>
<td>9646</td>
<td>9036/3011</td>
<td>341</td>
<td>1089/363</td>
<td>1344/448</td>
<td>269</td>
</tr>
<tr>
<td>NE274</td>
<td>3d</td>
<td>Nepal</td>
<td>9458</td>
<td>9066/3021</td>
<td>339</td>
<td>1104/368</td>
<td>1359/453</td>
<td>53</td>
</tr>
<tr>
<td>NE145</td>
<td>3e</td>
<td>Nepal</td>
<td>9458</td>
<td>9066/3021</td>
<td>340</td>
<td>1107/369</td>
<td>1356/452</td>
<td>52</td>
</tr>
<tr>
<td>811</td>
<td>3h</td>
<td>Somalia</td>
<td>9633</td>
<td>9060/3019</td>
<td>339</td>
<td>1104/368</td>
<td>1353/451</td>
<td>234</td>
</tr>
<tr>
<td>QC07</td>
<td>3a</td>
<td>Canada</td>
<td>9462</td>
<td>9066/3021</td>
<td>340</td>
<td>1107/369</td>
<td>1356/452</td>
<td>57</td>
</tr>
<tr>
<td>QC08</td>
<td>3a</td>
<td>Canada</td>
<td>9458</td>
<td>9066/3021</td>
<td>339</td>
<td>1107/369</td>
<td>1356/452</td>
<td>53</td>
</tr>
<tr>
<td>QC09</td>
<td>3a</td>
<td>Canada</td>
<td>9465</td>
<td>9066/3021</td>
<td>339</td>
<td>1107/369</td>
<td>1356/452</td>
<td>53</td>
</tr>
<tr>
<td>QC10</td>
<td>3a</td>
<td>Canada</td>
<td>9464</td>
<td>9069/3022</td>
<td>339</td>
<td>1110/370</td>
<td>1356/452</td>
<td>56</td>
</tr>
<tr>
<td>QC34</td>
<td>3b</td>
<td>Canada</td>
<td>9458</td>
<td>9060/3021</td>
<td>339</td>
<td>1107/369</td>
<td>1356/452</td>
<td>53</td>
</tr>
<tr>
<td>QC88</td>
<td>3b</td>
<td>Canada</td>
<td>9463</td>
<td>9069/3022</td>
<td>339</td>
<td>1110/370</td>
<td>1356/452</td>
<td>55</td>
</tr>
</tbody>
</table>

The genotype 1 sequence H77 (GenBank accession number AF009606) is included for comparison.

sequences assembled in this study. The tree shows that the nine genomes determined in this study were all classified into genotype 3, but represented different subtypes. Based on the other sequences that were co-analysed, QC7, QC8, QC9 and QC10 can be classified into subtype 3a, QC34 and QC88 can be classified into subtype 3b, and 811 can be classified into subtype 3h. In contrast, NE145 and NE274 each led to a single branch, and both formed a distantly linked twin. In the tree, a total of eight subtypes and a novel subtype equivalent represented by the QC115 variant were classified under genotype 3. These formed two subsets and three independent subtypes. From the top of the tree, subset A contained 3a, 3d and 3e, and subset B contained 3i, 3b and 3g. Below and adjacent to these two subsets was the relatively distant subtype 3k, which used to be classified as genotype 10 because of its greater genetic differences from the members of the two subsets (Tokita et al., 1996). The base of the genotype 3 clade exhibited the divergent subtype 3h and the QC115 isolate, both of which showed the largest genetic differences from all other HCV-3 members. In general, when more than one sequence was grouped to form a subtype cluster, a bootstrap support of 100% was shown. Based on this tree, pairwise $p$ distances were calculated by comparing subtype 3h to the other seven assigned subtypes. The results showed distances ranging from 0.261 to 0.278; the largest was from the HCV-Tr/3b, NE145/3e and QC260/3g isolates, and the smallest was from the QC268/3i isolate. Apparently, these inter-subtype genome-wide distances were more closely related to an isolate of 3a, i.e. ZS633 (Xu et al., 2013), than to NE274 in the 5' UTR tree, 14980 may not represent a subtype 3d isolate. Six isolates corresponding to subtype 3e were archived. Two (15201 and 15301) showed 5' UTR sequences identical to that of QC08/3a, whereas four (NICED-B2, NICED-B3, NICED-B4 and NE145) formed a solid 3e cluster based on their core region sequences (Chowdhury et al., 2003). With the exception of NE145, which was completely sequenced in this study, NICED-B2, NICED-B3 and NICED-B4 were all sampled in India, whereas 15201 and 15301 were of unknown origin. The full-length genome sequence was available for one subtype 3h isolate while partial E1 sequences were available for seven others; six of these were obtained from African immigrants (SOM1, SOM2, SOM3, QC29 and 52735 were obtained from Somalia, and QC200 was obtained from Congo) and two (25186-R and 50556-R) were detected in the UK, but the patients’ information is unknown. In addition, 8332488, which was obtained from an immigrant originating in Djibouti, a country that neighbours Somalia, was also classified into subtype 3h based on its cloned NS3 sequences (Colson et al., 2011). Two novel HCV-3 variants were detected recently in Iran and have been assigned provisionally as subtype 3i (Amini et al., 2006). The core sequences of both of these and the partial NS5B sequence of one of these were available in the Los Alamos HCV database, and were also analysed in this study. In the tree based on the core region sequences, both isolates were led by a single internal branch parallel to that of subtype 3h, whereas in the tree based on the NS5B sequences, the single sequence was more closely related to that of the QC115 isolate and those of subtype 3h than to any other sequences. Regardless, we observed that the 5' UTR sequences poorly differentiated the subtypes within genotype 3.

**Evolutionary analysis**

Based on the 14 full-length subtype 3a sequences, which exhibited sampling dates covering 22 years from 1991 to
Fig. 1. (a) An alignment of the 3′ UTR sequences from the H77 (AF009606), JFH1 (AB047639), 811 (JF735126), WS-#16 (AF009075), HCVgenotype4a-KM (AB795432) and Th580 (D84262) isolates representing subtypes 1a, 2a, 3h, 3a, 4a and 6c, respectively. (GenBank accession numbers are given in parentheses.) The rectangle at the beginning of the alignment highlights the stop codons preceding each 3′ UTR. A vertical line in the middle marks the start of the highly conserved 98 bases at the extreme 3′ end. The ruler indicates the numbering of the H77 genome. The dots inside the alignment show the bases identical to those of the H77 isolate, whereas the short bars represent the absent bases. (b) The optimal RNA secondary structure of the 269 nt 3′ UTR of the H77 isolate was predicted, using an online computer program (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi), with a minimum free energy of −61.90 kcal mol⁻¹. (c) The optimal RNA secondary structure of the 234 nt 3′ UTR of the 811 isolate was predicted with a minimum free energy of −55.10 kcal mol⁻¹. In the highly conserved 98 nt of the extreme 3′ UTR (the 5′ starting nucleotide is boxed), the different nucleotides shown by 811 in comparison with H77 are indicated by arrows.
2012, an evolutionary rate of $1.289 \times 10^{-3} \pm 1.47 \times 10^{-5}$ substitutions per site per year (mean $\pm$ 1.96 SE) was estimated (see insert in Fig. 4). By increasing the SE 14.14-fold and applying this rate in a normal distribution as the prior rate to a subsequent Markov chain Monte Carlo (MCMC) procedure under the uncorrelated log-normal clock model, we analysed the 45 full-length HCV genomes assembled in this study to yield the timescaled phylogenetic tree presented in Fig. 4. This tree showed a topology highly similar to that of the tree shown in Fig. 2, with the exception that all of the tips were aligned to the right end, and all of the nodes and branches were measured by grids corresponding to the timescale shown at the tree base. This scale indicates the year at which the different branches diverged. It was estimated that subtype 3a has a common ancestor dated to $\sim 78.26$ [95% highest posterior density (HPD): 63.04, 97.05] years ago (point A), which is after the era in which the ancestral 3a strains diverged from the ancestor of both subtypes 3d and 3e $\sim 202.89$ (95% HPD: 160.11, 264.6) years ago (point B). Considering both subsets A and B as a whole, the most recent common ancestor was dated to $\sim 253.78$ (95% HPD: 199.69, 325.51) years ago (point C), whereas the analysis of genotype 3 as a single lineage dated the divergence time to $\sim 457.81$ (95% HPD: 350.62, 587.53) years ago (point D). The inclusion of all of the analysed 45 HCV genomes representing genotypes 1–7 as a whole in the analysis resulted in the dating of the common ancestor to $\sim 780.86$ (95% HPD: 592.15, 1021.34) years ago (point E). It was shown that both subtypes 3d and 3e were closer relatives to subtype 3a than to any other subtypes, and that these three may have represented a single epidemiological lineage during a certain period in the past. Similarly, the past co-origin of both subsets A and B was also indicated – both appeared to have diverged from each other $\sim 100$ years ago.

**Fig. 2.** ML tree estimated using 45 full-length genome sequences. In addition to the nine isolates determined in this study (●), references from subtypes 1a, 1b, 2a, 2b, 3a, 3b, 3d, 3e, 3g, 3k, 3i, 3h, 4a, 4d, 5a, 6a, 6i, 7a and a novel HCV-3 variant QC115 were also included. Each branch is led by an isolate and denoted by the ID of the isolate (from this and our previous studies) or its GenBank accession number. The bootstrap supports are shown in italics. Bar, 0.1 nt substitutions per site. For descriptive purposes, the genotypes and subtypes are indicated above the related branches, and the two subsets A and B are highlighted by two shaded rectangles.
after the era when their ancestor diverged from subtype 3k (point F). Among all of the HCV-3 lineages, however, subtype 3h and the novel variant QC115 were the oldest, and both were separated from each other ~320.42 (95% HPD: 232.18, 422.18) years ago (point G). In the tree, all of these major divergences showed full posterior values, indicating robust classification and reliable MCMC processing.

Geographical distribution of HCV-3 isolates

By December 2011, the Los Alamos HCV database had archived a total of 7840 genotype 3 sequences. After removing those that did not include information regarding their sampling countries, those obtained from experimental animals, those representing various genomic regions of the same isolates and multiple clones from
single subjects, a total of 5846 individual isolates were differentiated. These were obtained from different patients and their sampling countries were known. The information on their geographical distribution is summarized in Tables 2 and S1 (available in the online Supplementary Material). Table 2 shows that subtype 3a isolates had been identified worldwide and accounted for 88.6% of the total 5846 isolates. In contrast, the numbers of other subtypes were remarkably decreased. With the exception of subtype 3b, which was found in 18 countries/regions and accounted for 8.96% of the total 5846 isolates, the other subtypes each accounted for no more than 1.09% of the isolates and were limited to certain regions. The results showed that the highest diversity of the lineages of HCV-3 was found in Asia (nine of the 10 subtypes plus one recombinant) followed by North America (six subtypes and an unclassified variant). Based on the number of isolates, however, the majority of the lineages were reported in Europe (2662/5846, 45.5%) followed by Asia, America, Australia and then the other continents.

**DISCUSSION**

In this study, the full-length genome sequences of nine genotype 3 isolates of HCV, denoted QC7, QC8, QC9, QC10, QC34, QC88, NE145, NE274 and 811, were characterized. These isolates represent subtypes 3a, 3b, 3d, 3e and 3h. In previous studies, partial sequences had been determined for these isolates, but their full-length
genomes had not been obtained (Abid et al., 2000; Murphy et al., 2007; Tokita et al., 1994). By applying our modified approaches (Lu et al., 2005), these nine isolates were entirely sequenced in the present study not only for the confirmation of their subtype assignments within genotype 3, but also for the reconstruction of a timescaled phylogenetic tree to estimate the origin and divergence of this HCV genotype.

To the best of our knowledge, NE274 and NE145 represent the first full-length genomes of subtypes 3d and 3e, respectively. In a recent paper on the expanded HCV classification, both 3d and 3e were categorized into the provisionally assigned subtypes (Smith et al., 2014). As their full-length genomes were determined in this study, these two subtypes are now confirmed, which allowed us to fill the information gap. Although 811 represents the second full-length genome of subtype 3h, it is, to the best of our knowledge, the first HCV-3 full-length genome to have its extreme 3' UTR sequence characterized. Based on this information, we were able to predict its secondary structure to show its similarity to other HCV genotypes. Together with the full-length genomes of subtypes 3g, 3h, 3i and 3k obtained in one of our recent studies (Lu et al., 2013), and those of four subtype 3a and two subtype 3b isolates, which were completely sequenced in this study, an expanded panel of HCV-3 full-length genomes is provided. These genomes represent all of the assigned subtypes within HCV-3 with the exception of 3c and 3f, which remain our future goals.

However, our reanalysis revealed that the latter two isolates can be reclassified into subtype 3a. Consequently, NE048 represents the only subtype 3c isolate that has been identified to date. Regarding subtype 3f, two isolates have been recorded: NE125 from Nepal (Tokita et al., 1994) and PK64 from Pakistan (Stuyver et al., 1996). Although we attempted to sequence the NE048/3c and NE125/3f isolates in the present study, we failed to provide their full-length genomes due to the limited number of samples with low viral titres. Therefore, our future studies will rely on more surveys to find additional 3c and 3f isolates in the regions in which these subtypes are indigenous.

Our recent studies have shown a very high degree of genetic diversity among the HCV-6 isolates found in South-East Asia, which therefore suggests South-East Asia as the region of origin for this oldest HCV lineage (Wang et al., 2013). However, to a lesser extent, the present study also revealed the high complexity of HCV-3 isolates in the Indian subcontinent and in countries around the Indian Ocean, where 10 of the 11 assigned HCV-3 subtypes have been detected (Table 2). Based on these data, we may conclude that the Indian subcontinent is the region of origin for HCV-3. In fact, this conclusion reflects a commonly accepted belief. Regardless, the number of HCV-3 isolates reported in the Los Alamos HCV database is highest in Europe (2662/5846 = 45.5%) compared with those reported in other continents. This difference is observed because almost all of the European HCV-3 isolates belong to subtype 3a, with the exception of five that can be classified into other subtypes (Castillo et al., 2006; Colson et al., 2011; Corbet et al., 2003; Ross et al., 2000; van Doorn et al., 1994). In

### Table 2. Geographical distribution of HCV-3 subtypes

<table>
<thead>
<tr>
<th>Subtype</th>
<th>No. of countries* in which subtype is detected</th>
<th>No. of isolates (n=5846)</th>
<th>Percentage of isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of countries* in which subtype is detected</td>
<td>Into subtypes</td>
<td>Not into subtypes</td>
</tr>
<tr>
<td>3a</td>
<td>55 (worldwide)</td>
<td>4632</td>
<td>545</td>
</tr>
<tr>
<td>3b</td>
<td>18 (AU, BD, BR, CA, CN, DE, GB, HK, ID, IN, JP, KH, MM, MY, NP, PK, TH, VN)</td>
<td>482</td>
<td>42</td>
</tr>
<tr>
<td>3a/3b</td>
<td>1 (CN)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td>1 (NP)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3d</td>
<td>1 (NP)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3e</td>
<td>1 (NP)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3f</td>
<td>2 (NP, PK)</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3g</td>
<td>2 (CA, IN)</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>3h</td>
<td>4(CA, DK, FR, SO)</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>3i</td>
<td>2 (CA, IN)</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>3k</td>
<td>6 (CA, ES, ID, IN, NL, PK)</td>
<td>49</td>
<td>16</td>
</tr>
<tr>
<td>3l</td>
<td>1 (IR)</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6 (BR, CA, ID, IN, TH, TR)</td>
<td>5211</td>
<td>635</td>
</tr>
</tbody>
</table>

*Country codes officially assigned in ISO 3166-1: AU, Australia; BD, Bangladesh; BR, Brazil; CA, Canada; CN, China; DE, Germany; DK, Denmark; ES, Spain; FR, France; GB, Great Britain; HK, Hong Kong; ID, Indonesia; IN, India; IR, Iran; JP, Japan; KH, Cambodia; MM, Myanmar; MY, Malaysia; NL, Netherlands; NP, Nepal; PK, Pakistan; SO, Somalia; TH, Thailand; TR, Turkey; VN, Viet Nam (http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2).
Evolutionary analysis of HCV-3 full-length sequences

addition, subtype 3a strains have been found to be distributed worldwide because of their long-distance transmission via the IDU network. Furthermore, molecular epidemiological studies of HCV have been more frequently performed in Europe than in other continents, which has resulted in the determination of a large number of 3a sequences.

With the exception of 3a and 3b that are distributed worldwide, only a small number of the isolates of the other subtypes in genotype 3 have been detected in geographical regions outside South and South-East Asia. These include three 3g (QC260 and QC299 in Canada, BID-G1243 in Europe), 10 3i (QC100, QC102, QC125, QC238, QC268, QC270 and QC310 in Canada, BID-G1244, BID-G1245 and 17872-R in Europe), five 3k (DQ641982-85, HE974762 and NL96 in Europe, QC105 in Canada, KC118332 in Iran), nine 3h (QC29 and QC200 in Canada, SOM1, SOM2, SOM3, 52735, 50556-R and 8332488 in Europe) and two 3l isolates (DQ202323 and DQ065830 in Iran) (Abid et al., 2000; Amini et al., 2006; Bernier et al., 1996; Castillo et al., 2006; Colson et al., 2011; Corbet et al., 2003; Murphy et al., 2007; Newman et al., 2013; Samimi-Rad et al., 2013; van Doorn et al., 1994). Of these, five 3i isolates (QC100, QC102, QC125, QC238 and QC310) have been detected among Caucasians in Canada and these five isolates shared a unique insertion of 12 nt, which may indicate a single source of infection, likely via IDU (Murphy et al., 2007). With the exception of these five 3i isolates, the majority of the above-described non-3a/3b isolates were from immigrants originating in South Asia or Africa, although a few were detected in the Middle East. These findings appear to agree with the above-described and commonly accepted belief that HCV-3 may have its origin in South Asia and subsequently spread to other regions. However, the ancestral origin of HCV-3 could have been in Africa, as suggested by the timescaled phylogenetic tree reconstructed in the present study. The tree shows that subtype 3h and the branch led by the QC115 isolate represent the oldest HCV-3 lineages. As all of the 3h isolates had their origins in Africa, the tree may suggest that the ancestral origin of HCV-3 is in Africa. From Africa, the ancient HCV-3 strains may have spread to South Asia either across the sea or by land via the Middle East. In support of the latter hypothesis, the novel QC115 isolate was detected in an immigrant from the Middle East (Lu et al., 2013) and isolates of the newly proposed subtype 3l, which are more genetically related to subtype 3h and the Q115 isolate than to any other genotype 3 members, have been detected in Iran (Amini et al., 2006). These findings suggest that there could be additional genetic lineage 3 lineages of HCV that have not been detected in the Middle East. Based on the timescaled phylogenetic tree, the period during which the spread of the ancestral HCV-3 strains from Africa to the Middle East and South Asia could have occurred was estimated. This period was dated to ~457 (95% HPD: 351, 588) years ago, i.e. in the middle of the sixteenth century (1553, 95% HPD: 1422, 1660). Coincidently, this time frame corresponds to the era during which Vasco da Gama completed his expeditions by sailing along the eastern coast of Africa to India (http://en.wikipedia.org/wiki/Vasco_da_Gama). However, before da Gama’s voyages, slave trading from Eastern Africa to the Middle East and South Asia had taken place for centuries (Gordon, 1989; http://www.arabslavetrade.com/). Since the seventh century, Arabian traders had brought Africans across the Indian Ocean from the coasts of the Red Sea in Eritrea and Ethiopia, from the Swahili Coast of present-day Kenya, Mozambique and Tanzania, and from elsewhere in South-East Africa to present-day Iraq, Iran, Kuwait, Turkey and other parts of the Middle East, and to Pakistan and India in South Asia. Such slave-trading activities may have mediated the spread of the earliest HCV-3 strains from Eastern Africa to South Asia and the islands of Indonesia. However, additional studies on the molecular epidemiology of HCV not only in Africa but, more importantly, in the Middle East are required to obtain more solid evidence in support of this hypothesis.

In this study, we analysed the geographical distribution patterns of HCV-3 only based on the sequences archived in the Los Alamos HCV database. Their percentages represented the frequencies in the database, but not in the population. These frequencies may not really reflect the epidemiological situation although they can be a good approximation. Therefore, these percentages should be extrapolated with caution.

Errors may exist in the evolutionary analysis. For example, in this study only a limited number of HCV-3 sequences, sampled during a very short period of time from 1991 to 2011, were used for the time calibration. Among these sequences there may exist some sampling or stochastic errors, whilst such a relatively small error may be magnified in the subsequent estimation of the past histories of HCV that have lasted for centuries or millennia. Other potential errors may involve the possible saturation of nucleotide substitution and different rates along branches or within genomic regions. All evolutionary analyses of the highly variable RNA viruses such as HCV are subjected to these errors. Theoretically, the resulting data should be adjusted; unfortunately, however, we still lack the knowledge for these adjustments.

METHODS

Samples. Three serum samples, NE274, NE145 and 811, were obtained from patients with chronic liver disease. Of these, both NE274 and NE145 were sampled in Nepal (Tokita et al., 1994). These were selected because partial sequences that had been determined previously in these two samples represented the provisionally assigned subtypes 3d and 3e. The 811 sample was collected from a Somali refugee in Switzerland, in which one of the first partial 3h sequences was isolated (Abid et al., 2000). To yield the dated full-length genome sequences for the reconstruction of the ancestral relationship among genotype 3 isolates, six additional samples, denoted QC7, QC8, QC9, QC10, QC34 and QC88, were also included. These were obtained from patients living in the Canadian province of Quebec (Murphy et al., 2007). Among these, QC7, QC8, QC9 and QC10 were sampled...
in 1991 and represented subtype 3a isolates. QC34 and QC88 were sampled in 1995 and 2002, respectively, and represented subtype 3b isolates.

**PCR amplification and sequencing.** From each 100 μl serum sample, the viral RNA was extracted using the QIAamp Viral RNA Mini kit (Qiagen) following the manufacturer’s protocol. The cDNA was synthesized using the RevertAid First-Strand cDNA Synthesis kit (Fermentas Life Science) under slightly modified conditions: the RNA pellet was dissolved in a random primer solution and incubated at 95 °C for 5 min before the other reagents were added. The resulting cDNA was used as a template to amplify the target sequences through conventional nested PCR. With the exception of the above-described protocols, all of the other procedures, such as the degenerate primers, the overlapping PCR strategies, the BigDye DNA sequencing method and the full-length genome sequence assembling, were the same as those reported previously (Li et al., 2012; Lu et al., 2013).

**Amplification of the extreme 5’ end.** The extreme 5’ UTR termini of the HCV-3 isolates were highly conserved. To amplify such terminal sequences through conventional PCR, a set of 5’-end degenerate primers was designed based on the known 3a and 3k sequences (forward, 5′-ACCTGCTCCTTTTCCAGGACACT-3′; inner reverse, 5′-TGGTGGCAAGTCTAGGAGCCT-3′; outer reverse, 5′-CTTTTGAGGTGCTTACGTC-3′).

**Amplification of the 3′ end.** For the QC7, QC8, QC9, QC10, QC34, QC88, NE145 and NE274 isolates, the 3′ termini were extended to the poly(U) tracks using conventional PCR (Li et al., 2012). However, for the 811 isolate, the extreme 3′ terminus was determined based on 3′ RACE RT-PCR, in which the RNA was polyadenylated using the A-Plus Poly(A) Polymerase Tailing kit (Epicentor Biotechnologies) and then extracted using TriPure (Lu et al., 2005). After the RNA pellet was resuspended, RT was performed and the cDNA was synthesized based on the poly(T) primer CACT35, which contained the 35mer (T). The resulting cDNA was then used to amplify the poly(T) ends using the CACT35 as the downstream primer and an 811-specific primer 5′-GCTATTTCCTCTGGGCTACTCC-3′ (corresponding to nt 9332–9355 in the 811 genome) as the upstream primer. The PCR was performed using FastStart Taq DNA Polymerase (Roche) with the following strategy: a first cycle of denaturation at 95 °C for 3 min, 35 cycles of 95 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s, and a final cycle of extension at 72 °C for 7 min. After amplification, the products were cloned into a TA vector and 10 clones were selected for DNA sequencing (Lu et al., 2008).

**Inspection of genetic variations and phylogenetic analyses.** The nine full-length HCV genomes obtained were denoted according to the nucleotide numbering in the H77 genome. To reconstruct the phylogenetic trees, the BI software (Drummond & Rambaut, 2007) was used, under the most likely substitution model and the TreeAnnotator program to reconstruct a tree from the resulting set of credible trees. The best trees were heuristically searched under the most probable substitution model GTR + I + F6. To assess the statistical robustness of the phylogenetic groupings, bootstrap analyses were conducted with 100 replicates.

**Evolutionary analysis.** Based on the 45 full-length HCV genome sequences assembled, an evolutionary timescale tree was further estimated using the MCMC algorithm implemented in the BEAST package (version 1.6.1) (Drummond & Rambaut, 2007). We recently analysed the full-length subtype 1a and 1b sequences, and identified that the relaxed log-normal clock model is better than the exponential and strict models (Yuan et al., 2013). Therefore, we used this model in combination with the GTR + I + F substitution model and the Bayesian skyline model in this study for the estimation of the timescale tree in the MCMC analysis, but we inputted an evolutionary rate of 1.289 × 10^{-2} ± 1.47 × 10^{-3} substitutions per site per year as the prior rate, which was estimated directly using the 14 dated 3a sequences assembled in this study using a previously described method (Yuan et al., 2013). After these parameters were set, we ran the MCMC procedure for 300 million states and obtained a tree every 30,000 states. After discarding the first 10% burn-in, the output was examined for convergence through a visual inspection of the chain lengths and by comparing the statistics of the effective sample size (ESS) using the Tracer program (http://tree.bio.ed.ac.uk). In this study, sufficient sampling was considered to have been achieved when all of the ESS numbers were > 200. We used the TreeAnnotator program to reconstruct a tree from the resulting set of credible trees and the final tree was denoted the MCC (maximum clade credibility) tree. As a molecular clock was incorporated, the branch lengths and the tree node heights were marked in units of years. The phylogenetic structure was then displayed using the FigTree program, and the clades, lineages and internal node heights were indicated as necessary.

To further comprehend the genetic complexity of subtypes 3e, 3d, 3h and 3l, the related partial sequences available in the Los Alamos HCV database were retrieved to represent these four subtypes and analysed phylogenetically. Four datasets, denoted 5′ UTR, core, E1 and NS5B, were therefore assembled.

To exclude the recent viral recombination events (Lee et al., 2010; Legrand-Abravanel et al., 2007), RNA virus software (Martin et al., 2010) was run using the modified settings that we described previously (Lu et al., 2007). This analysis was only performed for the full-length genome sequences.

**Geographical distribution of genotype 3 sequences.** A total of 7840 HCV isolates from the Los Alamos HCV database accessed on December, 2011 were classified as genotype 3. Of these, 5846 had known sampling countries and represented individual isolates. Based on these 5846 isolates, the correlation between subtypes and sampling countries was investigated.

**ACKNOWLEDGEMENTS**

The study described was supported by a grant from the National Institute of Allergy and Infectious Diseases (5 R01 AI080734-03A).

**REFERENCES**


Nkontchou, G., Ziol, M., Aout, M., Lhabadie, M., Baazia, Y., Mahmoud, A., Roulot, D., Ganne-Carrié, N., Grande-Lemaire, V. & other authors (2011). HCV genotype 3 is associated with a higher


Yuan, M., Lu, T., Li, C. & Lu, L. (2013). The evolutionary rates of HCV estimated with subtype 1a and 1b sequences over the ORF length and in different genomic regions. *PLoS ONE* 8, e6498.