Identification of Indian sub-continent as hotspot for HCV genotype 3a origin by Bayesian evolutionary reconstruction

Manish Chandra Choudhary, Vidhya Natarajan, Priyanka Pandey, Ekta Gupta, Shvetank Sharma, Rachana Tripathi, M. Shesheer Kumar, Syed N. Kazim, Shiv K. Sarin

Abstract

Background: Recent emphasis in Hepatitis C virus (HCV) evolutionary biology has focused on analysis using Core, E1/E2 and/or NS5b regions, with limited appreciation of full length genome. While HCV genotypes have been described as endemic in the Indian subcontinent, there has been no confirmation at the molecular evolutionary level of these genotypes. We have attempted here to determine the status of Indian HCV genotype 3a sequences in relation to similar genotypes from other parts of the world.

Methods: Cloning, sequencing and molecular characterization was performed on 9 Indian sequences and comparative analyses were performed with 46 sequences from other countries. Evolutionary-rate and molecular-clock hypothesis testing was addressed by Bayesian MCMC.

Results: Genetic analysis of full length genome revealed two hypervariable regions (HVR) in E2 region – HVR496 and HVR576, with a variable 5–8 amino-acid insertion sequence and a putative N-glycosylation site. Phylogenetic analysis revealed a divergence resulting in 2 distinct clades: clade-1 represented by HCV 3a subtype and clade-2 represented by other 3 subtype genomes. Clade-2 shows earlier divergence than clade-1. Analysis revealed that genotype 3a genomes from India roots out first (~99 years ago) in clade1. Bayesian skyline plot analysis revealed an increase in effective number of infections from 1940s to 1990s, followed by a gradual decrease after 2000.

Conclusions: Genotype 3a sequences appear to have originated in India and later dispersed to United Kingdom around mid 1940s, most likely around the time of Indian independence and World War II.
HCV genotype 4 is prevalent in North as well as central Africa and the Middle East (Iles et al., 2013). Genotype 5a is commonly found in the northern parts of South Africa (Murphy et al., 1996), central France (Henguell et al., 2004) and Belgium and is also sporadically found in other regions of Europe and Latin America (Chamberlain et al., 1997; Davidson et al., 1995; Jover et al., 2001; Levi et al., 2002), whereas genotype 6 so far is known to be confined to Hong Kong and few other South-Asian countries (Mondelli and Silini, 1999; Centers for Disease Control and Prevention, 1998). Genotype 7, the latest entrant in the family is said to have its origin in Central Africa (Smith et al., 2014).

Since HCV genotypes and subtypes respond differentially to antiviral therapy, it is important to have prior information regarding the same while determining the type and duration of therapy (Pang et al., 2009). Recently, Pang et al. determined the correlation between response to therapy to that of evolutionary branching order of major HCV genotypes (Pang et al., 2009). Although, HCV genotype 3a responds well to standard treatment regimen of Peg-interferon alfa-2a, it is also associated with significant hepatic steatosis and fibrosis (Singh et al., 2010; Hisar et al., 2009, 2006). Recent trials on HCV genotype 3a with Telaprevir and Boceprevir have reported negligible to modest decrease in RNA level (Foster et al., 2011; Narahari et al., 2009). Recently multiple investigations across the globe have attempted to trace the evolutionary history (Iles et al., 2013). Genotype 5a is highly prevalent among intravenous drug users and is showing increased incidence in various parts of the world (Samimi-Rad et al., 2012; Paintsil et al., 2009; Aitken et al., 2004).

Recent efforts to analyze HCV infection rates based on genotype/phenotype suggest that genotype 3a infections comprise almost 60–70% of HCV related infections in India (Sievert et al., 2011; Naranbar et al., 2009). Recently multiple investigations across the globe have attempted to trace the evolutionary history of various HCV genotypes (Verbeck et al., 2006; Pybus et al., 2007; Gededzha et al., 2013). Nearly all of these studies employed sub-genome regions (Core, E1E2, NS5a/5b) analysis to trace the evolutionary history. The evolutionary trace of HCV in context of its whole genome is not well-investigated. In light of this information, we performed molecular characterization of evolutionary history of HCV genotype 3a using smaller genomic fragments, i.e., E2p7NS2 and NS5b and the results were compared with full genome dataset.

2. Materials and methods

2.1. Patient selection

The study protocol was approved by the Institutional Review Board at the Institute of Liver and Biliary Sciences (ILBS), New Delhi, India and patients were enrolled after their informed consent. Only treatment naïve patients with HCV genotype 3a (n = 9) were included and patients with co-infection with either hepatitis B or HIV were excluded. Liver biopsy was performed and staging was done according to the Ishak’s scoring system.

2.2. Viral RNA extraction, sequencing data analysis and N-glycosylation site prediction

HCV RNA was extracted from 140 μL serum using the QIAamp viral RNA Mini Kit (QIAGEN GmbH, Germany) and cDNA synthesized as per manufacturer’s protocol. 3–5 μL of cDNA was used for PCR amplification of the full genome HCV using Phusion Hot Start II DNA polymerase (Thermo Fisher Scientific Inc., Waltham, MA). The PCR products were further cloned using CloneJet PCR cloning kit (Thermo Fisher Scientific Inc., Waltham, MA), and 3–5 colonies were picked for bidirectional sequencing using vector sequencing primers and additional internal primers (Table 2).

Sequence alignments were visualized, compared, and edited using ClustalW version in BioEdit program v7.2.0. The HCV genotype and subtype were confirmed by BLAST (NCBI, USA). N-glycosylation site prediction was performed by N-glycosylation tool (http://hcv.lanl.gov).

2.3. Compilation of HCV dataset

For phylogenetic analysis dataset was prepared consisting of 53 sequences; HCV genotype 3 full-length or near to full-length genomes retrieved from NCBI (n = 46) along with full genomes reported in this paper (n = 7). Sequences with known year and country of isolation were only selected for analysis. Recombinant sequences or sequences isolated from a single outbreak or same cohort were excluded. E2p7NS2 and NS5b sequences from the above datasets were additionally compared.

2.4. Determination of nucleotide substitution model and likelihood analysis

The most appropriate nucleotide substitution model for the phylogenetic analyses was determined by jMODELTEST v2.1.4 and likelihood scores were determined using hierarchical likelihood ratio tests (hLRT) and the Akaike Information Criterion (AIC) Sanchez et al. (2011). Likelihood-mapping analysis was performed on quartets using TreePuzzle to obtain an overall impression of the phylogenetic signal present in the HCV genotype 3 full genome sequences (Schmidt et al., 2002). For each analysis all 10,000 possible quartets for each of the full genome sequences were evaluated.

2.5. Time-based phylogeny and evolutionary reconstruction by Bayesian MCMC

Time-measured phylogeny and evolutionary rates were inferred using a Bayesian Markov Chain Monte Carlo (MCMC) method implemented in the BEAST package version 1.7.5 (http://beast.bio.ed.ac.uk/) (Drummond and Rambaut, 2007). MCMC analysis was run for a chain length of 50 million for 4 independent cycles and sampling was done every 1000 steps. Adequate chain convergence was assessed using Tracer software version 1.5 (beast.bio.ed.ac.uk). The effective sample size (ESS) was calculated for each parameter, and all ESS values > 200 were considered sufficient for sampling and chain mixing. The maximum clade credibility tree was selected from the posterior tree distribution after considering 10% burn-in using Tree Annotator v1.7.5. The final tree was analyzed and modified in FigTree version 1.4.0 (tree.bio.ed.ac.uk) for display purposes.

2.6. Bayesian coalescent analysis

Bayesian skyline plot (BSP) analysis was carried out using all possible combinations of the molecular clock models; relaxed uncorrelated lognormal, relaxed uncorrelated exponential and strict models using the coalescent analysis (non-parametric piecewise-constant model) with an external rate substitution rate of 1.3 × 10^(-3) per site per year (Gray et al., 2011). E2p7NS2 dataset was also analyzed using the same external rate while NS5b dataset was analyzed using rate obtained from E2p7NS2 analysis.

2.7. GenBank accession numbers

The GenBank accession numbers of the HCV genotype 3a sequences isolated from Indian patients reported in this paper...
are HQ738645, JN714194 and JQ717254–JQ717260. Details of all other sequences used in this study are available upon request.

3. Results

3.1. Patient characteristics

Detailed characteristics of the nine patients included in the study are summarized in Table 1. Baseline plasma samples were collected from HCV genotype 3a infected patients before the start of the treatment. All patients underwent liver biopsy, with mild liver disease defined as Ishak’s fibrosis stage 0–1, severe liver disease considered as Ishak’s fibrosis stages 5–6. AST and ALT levels were raised while AFP level was normal in all the 9 patients studied.

3.2. Sequence diversity in HCV GT3a vis-à-vis other genotype and subtypes

The sequences reported here have been submitted in the Genbank database with accession numbers HQ738645, JN714194 and JQ717254–JQ717260. The average nucleotide composition of nine HCV genotype 3a sequences reported here calculated by MEGA 5.2.1 was 21.1% A, 23.1% T, 27.6% G, 28.3% C.

Full genome amino-acid sequence alignment of HCV genotype 3a vis-à-vis other genotype and subtypes by ClustalX version 2.0 revealed two additional regions of hypervariability named as HVR496 and HVR576 respectively based on its amino-acid position in the full genome. HVR496 spans amino acids 496 to 502 and is seven amino acids long while HVR576 represents amino acids 576 to 584 and is nine amino acids long with a unique 5–6 amino-acid insertion site (Fig. 1). Interestingly, it was found that the number of amino-acids which were inserted at this site were variable as revealed by multiple alignment of available HCV genotype 3a sequences present in the database.

Next, N-linked glycosylation site prediction was performed using N-glycosylation tool represented by Nx[ST] patterns, where x can be any amino acid. It was found that within HVR576 lies a putative N-glycosylation site which was found to be highly conserved in HCV genotype 3a, as is evident from the alignment (Fig. 1).

3.3. Best-fit nucleotide substitution model and phylogenetic signal analysis

Best-fit nucleotide substitution model was determined by jMODELTEST by comparing the likelihood scores from full-length genome and using the hLRTs and the AIC. The General Time-Reversible (GTR) substitution model along with incorporated invariant sites (I) and assumed rate heterogeneity (C) across sites returned as the best-fit model and was subsequently employed.

The likelihood mapping analysis of the HCV genotype 3 data set revealed that the percentage of dots falling in the central region accounted for 0.1% of total samples, samples falling on the axes accounted for 0.4% of total samples (0.1% + 0.2% + 0.1%) while samples falling in the corners of the triangle accounted for 99.5% (32.9% + 32.8% + 33.8%) of the total samples, thus indicating a fully resolved phylogenetic tree (Fig. 2).

3.4. Evolutionary rate estimate and phylogeographic tree analysis of HCV 3a

The mean evolutionary rate of the whole genome sequences of HCV genotype 3 was estimated based on the alignment of dated

Table 1

<table>
<thead>
<tr>
<th>Genbank accession No.</th>
<th>Age (years)</th>
<th>Sex</th>
<th>HCV genotype</th>
<th>HCV RNA (IU/mL)</th>
<th>Ishak’s fibrosis score</th>
<th>AST (IU/L)</th>
<th>ALT (IU/L)</th>
<th>AFP (ng/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HQ738645</td>
<td>52</td>
<td>M</td>
<td>3a</td>
<td>1.69 × 10⁶</td>
<td>0.1</td>
<td>117</td>
<td>74</td>
<td>8.62</td>
</tr>
<tr>
<td>JN714194</td>
<td>43</td>
<td>F</td>
<td>3a</td>
<td>7.53 × 10⁶</td>
<td>2</td>
<td>219</td>
<td>73</td>
<td>5.1</td>
</tr>
<tr>
<td>JQ717254</td>
<td>35</td>
<td>F</td>
<td>3a</td>
<td>8.27 × 10⁶</td>
<td>2</td>
<td>133</td>
<td>133</td>
<td>4.7</td>
</tr>
<tr>
<td>JQ717255</td>
<td>39</td>
<td>M</td>
<td>3a</td>
<td>4.19 × 10⁷</td>
<td>0.1</td>
<td>97</td>
<td>204</td>
<td>4.1</td>
</tr>
<tr>
<td>JQ717256</td>
<td>46</td>
<td>M</td>
<td>3a</td>
<td>6.67 × 10⁷</td>
<td>5.6</td>
<td>69</td>
<td>79</td>
<td>7.2</td>
</tr>
<tr>
<td>JQ717257</td>
<td>50</td>
<td>M</td>
<td>3a</td>
<td>9.01 × 10⁴</td>
<td>1.2</td>
<td>45</td>
<td>87</td>
<td>6.9</td>
</tr>
<tr>
<td>JQ717258</td>
<td>41</td>
<td>F</td>
<td>3a</td>
<td>5.87 × 10⁷</td>
<td>5.6</td>
<td>214</td>
<td>77</td>
<td>4.2</td>
</tr>
<tr>
<td>JQ717259</td>
<td>38</td>
<td>M</td>
<td>3a</td>
<td>1.32 × 10⁷</td>
<td>5.6</td>
<td>139</td>
<td>73</td>
<td>3.8</td>
</tr>
<tr>
<td>JQ717260</td>
<td>47</td>
<td>F</td>
<td>3a</td>
<td>5.93 × 10⁷</td>
<td>0.1</td>
<td>151</td>
<td>81</td>
<td>6.1</td>
</tr>
</tbody>
</table>


All patients were infected with HCV genotype 3a, were having high viral load with presence of liver disease (indicated by fibrosis score) having raised AST and ALT with normal AFP values. All values are indicated in the units as mentioned in the table.

Table 2

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Product No.</th>
<th>Primer name</th>
<th>Position in whole genome</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>UF</td>
<td>1–25</td>
<td>ACCTGCCTCTTACGAGCGCAGACTC</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>UR1</td>
<td>339–315</td>
<td>GTTGCCGCGTCTACGAGACTCC</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>CF2</td>
<td>339–363</td>
<td>ATGACGACAACACCTCTCAAAAAACAAAA</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>NS1R2</td>
<td>2566–2545</td>
<td>GCCAAGAGCAGAGCAGAGCCAG</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>NS2F3</td>
<td>2545–2566</td>
<td>GCCCTGTCGCGCTTCGCTTGT</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>NS3R3</td>
<td>5328–5307</td>
<td>AGACCTCCTGGTGCTGCTTCGG</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>NS4aF4</td>
<td>5329–5348</td>
<td>ACAAAGCTGGTGTGCTGCTT</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>NS4bR4</td>
<td>6274–6250</td>
<td>ACAAAGCTGGTGTGCTTCTT</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>NS5aF5</td>
<td>6239–6264</td>
<td>ACCAGGGTGGAGCAGAAGCTCACCA</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>NS5aR5</td>
<td>7629–7608</td>
<td>GCAGCAGAGACACGCTGTGCTC</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>NS5b1F5</td>
<td>7600–8385</td>
<td>GACAGGGCAGACAGACGGCTGCT</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>NS5b1R6</td>
<td>8385–8361</td>
<td>TCACCTCGGCAGAGCAGAAAGCTC</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>NS5b2F6</td>
<td>8388–8413</td>
<td>TCCATCTCCAGGGACAGGGCTTTA</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>NS5b2</td>
<td>9425–8098</td>
<td>TGGAGTGTTCATTCGACAGTCCAC</td>
</tr>
</tbody>
</table>

M.C. Choudhary et al. / Infection, Genetics and Evolution 28 (2014) 87–94
sequences using an external substitution rate of $1.3 \times 10^{-3}$ per site per year using a relaxed uncorrelated lognormal which was the best molecular clock obtained when compared with other molecular clocks (data not shown). Under this condition, the estimate was $1.07 \times 10^{-3}$ (95% HPD 6.79 $\times 10^{-4}$–1.48 $\times 10^{-3}$) substitutions/site/year.

Tree branched with a high posterior probability of 0.99–1.0, indicative of fully resolved branched tree with a greater degree of confidence. The approximated age of the entire tree calculated based on genotype 3 full length genomes was estimated to be about 366 years (95% HPD; 226.64–528.05) (Fig. 3). Phylogenetic analysis comprising of Indian HCV genotype 3a and reference HCV genotype 3 full-genome by Bayesian MCMC algorithm revealed a bifurcation in divergence with 2 distinct clades: clade-1 represented by genotype 3a genomes, and clade-2 represented by other HCV genotype 3 sequences. While clade-2 appears to have diverged earlier around 366 years ago; among clade-1, HCV genotype 3a genomes from India rooted out earlier. Phylogenetic analysis using relaxed molecular clock approach revealed that most probable origin of HCV genotype 3a was in India about 99 years ago (95% HPD; 63.23–137.05) among clade-2, sequences from United Kingdom branched very closely together with Indian sequences (Fig. 3) with sequences branching out as early as 85 years ago. Phylogenetic reconstruction using E2p7NS2 yielded similar results (Fig. 4a and b) with Indian sequences rooting out earlier about 105 years ago (95% HPD; 66.73–153.50) as observed in phylogenetic tree with full genome dataset with an estimated substitution rate of $1.17 \times 10^{-3}$ (95% HPD; 6.88 $\times 10^{-4}$–1.47 $\times 10^{-3}$) substitutions/site/year, while in case of NS5b dataset, earliest sequence of HCV genotype 3a branched about 126 years ago (95% HPD, 63.23–137.05) after divergence from Indian sequences.
62.24–214.68) with a rate of \(6.23 \times 10^{-4}\) (95% HPD; \(1.56 \times 10^{-4}–1.08 \times 10^{-3}\)) substitutions/site/year.

3.5. Evolutionary demography of entire HCV-3 dataset

The Bayesian skyline-plot analysis using entire HCV genotype 3 dataset revealed that the number of infections remained relatively constant until the early 1940s, after which there was one log increase in effective number of infections till early 1990s (pre-anti-viral introduction period).

4. Discussion

Full length genome molecular characterization revealed that apart from commonly reported HVR1 located at 5’ end of E2 protein, HCV genotype 3a also contained 2 additional regions of hyper-variability termed as HVR496 and HVR576, also reported by Humphreys et al. (2009). HCV envelope proteins have a unique property of undergoing glycosylation (Helle et al., 2011, 2010). The E2 protein of HCV genotype 3a contain eight potential N-glycosylation sites (Helle et al., 2011). Sequence comparison analysis by ClustalX method revealed that HVR576 contained an additional putative glycosylation site. Various studies have indicated the role of N-linked glycosylation in virion assembly and infectivity (Goffard et al., 2007). Further, it has also been reported that mutation in these glycosylation sites leads to envelope protein instability and virion assembly defects (Helle et al., 2010). These results indicate that N-glycosylation sites are critically important for E1E2 folding and heterodimerization and thus for virion assembly.

Analysis of more datasets may be required to ascertain the functional relevance of putative N-glycosylation site found in HVR576 in HCV genotype 3a viral sequences.

HCV epidemiology at present is explained by isolated studies and data from blood bank (Chowdhury et al., 2003; Kumar et al., 2007). Among different HCV genotypes, genotype 3a is most common genotype circulating in Indian subcontinent including northern part of India (Sood et al., 2012; Singh et al., 2004; Sievert et al., 2011; Thakur et al., 2000) as well in Pakistan (Idrees and Riazuddin, 2008), Nepal (Tokita et al., 1994) and in United Kingdom. Narahari et al. (2009) performed HCV genotyping in 2118 patients from different geographic regions of India and reported that HCV3 (3a/3b) was prevalent in 62% and HCV1 (1a/1b) in 31% of the patients, with predominance of HCV3 in northern (\(p = 0.01\)) and eastern (\(p = 0.008\)) regions of India (Narahari et al., 2009). A recent study by Chakravarti et al. (2013) has highlighted the changing trends of prevalence and genotypic distribution of hepatitis C virus among high risk groups in North India. The authors report that genotype 3 is the predominant genotype, though the subtype distribution within genotype 3 may be changing and is well supported by previous studies (Tokita et al., 1994; Valliammai et al., 1995). The presence of high degree of genetic variability (as reported in this manuscript and discussed below) with the coexistence of different subtypes of 3 in the northern part of the India suggests that this is the most probable site of origin of HCV genotype 3, also discussed by Zehender et al. (2013).

Analyses from various countries have demonstrated the origin of HCV to be in and around Indian-subcontinent (Pakistan). In an attempt to investigate this fact the present study was carried out considering the fact that India and Pakistan share a fairly large
human ancestry. As the results suggest, the origin of HCV genotype 3a is in the Indian subcontinent, even when analyzed at the whole genome level as well as by E2p7NS2 region (Fig. 4a). The diversity of HCV in the Indian sub-continent has also not been thoroughly examined in this regard. Sequences from India tend to cluster together forming a distinct India-specific cluster genotype 3a.
The tMRCA of HCV genotype 3a was calculated at about 99 years ago. It is interesting to note that sequences from India rooted out at the same time and later diverged in the subcontinent. The analysis further showed that other HCV genotype 3a sequences from the Indian sub-continent (including Pakistan) rooted closely followed by sequences of United Kingdom, leading us to speculate that the viral strain would have disseminated to United Kingdom and later to other European countries. The Bayesian skyline-plot analysis revealed that the number of infections remained relatively constant until the early 1940s, after which there was one log increase in effective number of infections till early 1990s (pre-antiviral introduction period), a period also coincided with rapid industrialization and blood transfusions together with unsafe use of blood products (Fig. 5). The period also coincided with the time of Indian independence in 1940s and 1950s and end of British colonization as well as World war, which involved massive migration of human population and possibly have led to spread of HCV to other parts of the world. The estimation of evolutionary history of HCV genotype 3a infection in India and Indian sub-continent will not only will be beneficial for epidemiological and clinical analyses, but it will also help in a long way in devising strategy for effective disease management and surveillance program to facilitate better public health service.

References


