The Core/E1 domain of hepatitis C virus genotype 4a in Egypt does not contain viral mutations or strains specific for hepatocellular carcinoma

Xiaoan Zhang\textsuperscript{a,c}, Soo Hyung Ryu\textsuperscript{a}, Yanjuan Xu\textsuperscript{a}, Tamerl Elbaz\textsuperscript{d}, Abdel-Rahman N. Zekri\textsuperscript{e}, Ashraf Omar Abdelaziz\textsuperscript{d}, Mohamed Abdel-Hamid\textsuperscript{f}, Valerie Thiers\textsuperscript{g,h,i}, Santiago F. Elena\textsuperscript{j,k}, Xiaofeng Fan\textsuperscript{a,b,\ast}, and Adrian M. Di Biseglie\textsuperscript{a,h}

\textsuperscript{a} Division of Gastroenterology and Hepatology, Department of Internal Medicine, Saint Louis University School of Medicine, St. Louis, MO 63104, USA
\textsuperscript{b} Saint Louis University Liver Center, Saint Louis University School of Medicine, St. Louis, MO 63104, USA
\textsuperscript{c} The Third Affiliated Hospital of Zhengzhou University, Zhengzhou 450001, Henan, China
\textsuperscript{d} Division of Gastroenterology and Hepatology, Department of Tropical Medicine, National Cancer Institute, Cairo University School of Medicine, Cairo, Egypt
\textsuperscript{e} Virology and Immunology Unit, Cancer Biology Department, National Cancer Institute, Cairo University School of Medicine, Cairo, Egypt
\textsuperscript{f} Viral Hepatitis Research Laboratory, National Hepatology and Tropical Medicine Research Institute, Cairo, Egypt
\textsuperscript{g} Virology Department, Institute Pasteur, Paris, France
\textsuperscript{h} INSERM, U 785, Villejuif, France
\textsuperscript{i} Université Paris SUD, Faculté de Médecine, Orsay, France
\textsuperscript{j} The Santa Fe Institute, Santa Fe, NM 87501, USA
\textsuperscript{k} Instituto de Biologia Molecular y Celular de Plantas, CSIC-UPV, 46022 Valencia, Spain

1. Background

The causal relationship between hepatitis C virus (HCV) infection and hepatocellular carcinoma (HCC) is well documented.\textsuperscript{1,2} About 1–3\% patients with chronic HCV infection will develop HCC in the United States.\textsuperscript{3} HCV-related tumorigenesis has been studied extensively and almost all HCV-encoded viral proteins, especially Core protein, can cause cellular transformation through multiple mechanisms.\textsuperscript{4} As a positive, single-strand RNA virus, a remarkable feature of HCV genome is the high genetic diversity with at least six major genotypes and more than 100 subtypes.\textsuperscript{5} Based on underlying mechanisms in HCV-related HCC, it is important to know whether such a high viral genetic diversity plays a differential role in the incidence of HCC. In other words, is the incidence of HCC preferentially associated with specific HCV genotype(s)/subtype(s)/strain(s) or particular mutations in the HCV genome? Due to the lack of appropriate small animal models supporting the HCV life cycle, these issues have been studied mostly in clinical settings.\textsuperscript{6–17} However, published reports have yielded conflicting data concerning these questions.\textsuperscript{6–17} The development
of HCC is a long-term, multi-step process affected by many factors from both the host and the virus. To assess the role of viral genetic diversity in the incidence of HCC, it is therefore necessary to have a well-designed experimental strategy that minimizes the interference from other factors contributing to carcinogenesis.

2. Objectives

In the present study, the potential association between HCV genetic mutations or strain variability and HCC incidence has been examined through a comparative genetic analysis merely focused on a single HCV subtype (genotype 4a) in a single country (Egypt).

3. Study design

3.1. HCV sequence data collection

Three HCV sequence datasets were included in this study. The first HCV dataset was derived from a nationwide epidemiological study designed to evaluate the prevalence of HCV in Egyptian blood donors. The dataset consists of 49 HCV genotype 4a E1/Core sequences with assigned GenBank accession numbers from AF271835 to AF271873, representing a subset of blood donors from 15 geographically diverse governorates in Egypt. The second dataset was generated in our laboratory in a study to investigate the role of HCV genotype in end-stage liver disease in Egypt. This dataset includes a total of 146 HCV E1/Core sequences corresponding to 97 patients with HCC, 43 patients with cirrhosis and 6 individuals without end-stage liver disease (GenBank accession numbers HQ615723–HQ615868). The final dataset includes 36 HCV E1/Core sequences from a study that investigated familial transmission of HCV in an Egyptian village. A summary for three datasets is presented in Table 1.

3.2. Control experiment to estimate data quality

The three datasets were generated from three different laboratories. It is thus possible that some nucleotide differences may simply result from differences in experimental protocols. In each laboratory, the sequences were obtained from serum samples through direct sequencing of reverse transcription-polymerase chain reaction (RT-PCR) product. Thus, the use of different primers in the RT-PCR protocols is a potential concern. We determined the HCV Core/E1 sequences for five randomly selected samples using primer sets from all three laboratories (Table 2). Sequences were compared for the estimation of potential influence by primer selection. In our experimental protocol, RT and PCR were respectively conducted with M-MLV reverse transcriptase (Promega) and AmpliTaq DNA polymerase (Applied Biosystems) as we described previously.

3.3. Genetic analysis

The genetic analysis was performed between datasets 1 and 2, while dataset 3 was used as a reference control. The target domain for comparative analysis was an overlapping region among these datasets, 387bp in length from nucleotide position 873 to 1259 (all position numbering in the study is based on HCV strain H77, GenBank accession number AF009606). A consensus sequence corresponding to this target domain was first generated from 41 unrelated HCV genotype 4a sequences deposited in the Los Alamos HCV database. Nucleotide (387 sites) and amino acid (129 sites) frequencies were calculated against the consensus sequence at each site, followed by Chi-square test. Next, we evaluated intra-group mutation patterns and selection pressure by both Tajima’s D test and the calculation of genetic diversity parameters, including genetic distance (d), the number of synonymous substitutions per synonymous site (dS), the number of non-synonymous substitutions per non-synonymous site (dN) and dN/dS values. Tajima’s D test (coding region) was done with program Dnasp and genetic parameters were analyzed with either maximum composite likelihood (d) or Nei–Gojobori method (dN, dS) implemented in the molecular evolutionary genetics analysis software package (MEGA, version 4.0).

3.4. Phylogenetic analysis

Phylogenetic analysis was performed for the combination of datasets 1 and 2. The best-fit nucleotide substitution model was first estimated through a hierarchical likelihood ratio test (hLRT) with ModelTest. Under the best-fit model, the unrooted maximum-likelihood (ML) tree was generated in program PHYML and used as the template to evaluate the extent of clock-like evolution between the datasets 1 and 2 through a regression analysis of root-to-tip distance against sampling dates in program Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen). Bayesian Markov chain Monte Carlo (MCMC) phylogenetic trees were simulated in BEAST package under the best-fit nucleotide substitution model as well as additional parameter settings, including a relaxed molecular clock (uncorrelated, lognormal), a Bayesian skyline coalescent prior, and a total run of 50 million generations to reach relevant parameter convergence as estimated by Tracer. The inferred MCMC trees then served as the input to estimate the strength of HCV strain clustering in terms of disease patterns or sample dates in program BaTS with 1000 replications and the removal of the first 10% trees as burn-in. Both the association index (AI) and the parsimony score (PS) were computed to see whether disease patterns or sampling dates are more strongly associated with the underlying phylogeny than expected by chance alone.

3.5. Statistical analysis

The significance of changes in either nucleotide or amino acid frequency was examined by Chi-square test. Other differences with regard to genetic parameters were assessed for statistical significance using two-tailed Student’s t-test.

---

Table 1
Summary of the three datasets used in this study.

<table>
<thead>
<tr>
<th>Data</th>
<th>Subjects</th>
<th>Number</th>
<th>Sampling date (year)</th>
<th>Sampling pattern</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blood donors</td>
<td>49</td>
<td>1993</td>
<td>Nationwide</td>
<td>Egypt</td>
</tr>
<tr>
<td>2</td>
<td>HCC</td>
<td>97</td>
<td>2003</td>
<td>Nationwide</td>
<td>Egypt</td>
</tr>
<tr>
<td>3</td>
<td>Cirrhosis</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CHC</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CHC</td>
<td>36</td>
<td>2002</td>
<td>A single village</td>
<td>Egypt</td>
</tr>
</tbody>
</table>

CHC, chronic hepatitis C infection.
Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Primers</th>
<th>Polarity</th>
<th>Sequence (5′ - 3′)</th>
<th>Position</th>
<th>Application</th>
<th>Product size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>987R_H77</td>
<td>Anti-sense</td>
<td>ctagggagacgtcatcatct</td>
<td>1305–1328</td>
<td>RT/1st round PCR</td>
<td>474 bp</td>
</tr>
<tr>
<td></td>
<td>4635_H77</td>
<td>Sense</td>
<td>gcacagggacacgctctgct</td>
<td>834–859</td>
<td>1st round PCR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5025_H77</td>
<td>Sense</td>
<td>aacctcctgctgtcttcctct</td>
<td>843–868</td>
<td>2nd round PCR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>975R_H77</td>
<td>Anti-sense</td>
<td>gtctctcatctcctactcatc</td>
<td>1293–1316</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CER1414</td>
<td>Anti-sense</td>
<td>ccctcagggcccccagt</td>
<td>1395–1414</td>
<td>RT/1st round PCR</td>
<td>596 bp</td>
</tr>
<tr>
<td></td>
<td>CEF696</td>
<td>Sense</td>
<td>tggtaaggtgcgtcatctac</td>
<td>697–716</td>
<td>1st round PCR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CEF723</td>
<td>Sense</td>
<td>ggccttcggcaacctcgct</td>
<td>724–743</td>
<td>2nd round PCR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CER1318</td>
<td>Anti-sense</td>
<td>cagctctcatctcactcctc</td>
<td>1299–1319</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E4R</td>
<td>Anti-sense</td>
<td>gttggtctctccatcactc</td>
<td>1307–1327</td>
<td>RT/1st round PCR</td>
<td>469 bp</td>
</tr>
<tr>
<td></td>
<td>Core823</td>
<td>Sense</td>
<td>ggtacatagtggcgacgg</td>
<td>850–869</td>
<td>2nd round PCR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Core850</td>
<td>Sense</td>
<td>cgctgcttcgctgcttctct</td>
<td>850–869</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E2R</td>
<td>Anti-sense</td>
<td>cggtctctctctctctctct</td>
<td>1299–1318</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RT, reverse transcription.

4. Results

4.1. Lack of amplification bias by different primer sets

The potential effect of different primer sets on the amplification of HCV Core/E1 domain was tested in five serum samples. Direct sequencing of amplicons with the primer sets from either dataset 2 or 3 showed the complete identity. Over 1935 bp amplification sequence (387 bp × 5), the primer set from dataset 1 generated one silent mutation (A → T), indicating a 99.95% match in comparison with the primer sets from datasets 2 and 3. Therefore, the use of different primer sets did not result in noticeable bias on the amplification of the targeted domain, allowing valid comparative analysis to be performed based on these three datasets.

4.2. Comparative analysis and Taq DNA polymerase-associated errors

In comparison with dataset 1, the HCC and cirrhosis groups in dataset 2 showed four distinct nucleotide substitutions at positions 891 (T → G, p < 0.0001), 1138 (T → G, p < 0.0001), 1161 (G → C, p < 0.0001) and 1187 (C → G, p < 0.0001). Due to the potential significance of such a nearly complete sweep-out in HCV strains associated with the end-stage liver disease, we repeated the experiment in five samples carrying HCV strains with all four substitutions. Surprisingly, none of these samples showed the initially observed substitutions. We then conducted additional experiments. First, an additional 25 samples were processed starting from the step of RNA extraction. Sequence alignment showed

![alignment](image-url)

Fig. 1. Alignment of full analytical domain (387 bp) of five representative sequences from each group. The HCV genotype 4a consensus sequence is shown on the top line. HCC, the HCC group from dataset 2; Taq and Tth indicates the same HCC samples repeated with either AmpliTag DNA polymerase and rTth DNA polymerase, Xl, respectively; CHC, chronic HCV infection without end-stage liver disease. The four nucleotide substitutions from either initial comparative analysis or the repeated experiments are marked as red and blue, respectively. Dot indicates nucleotide identity.
the same result, with the lack of nucleotide substitutions seen in the initial analysis. Instead, there were four alternative nucleotide substitutions at positions 923, 1084, 1131 and 1226 (Fig. 1). Second, these 30 samples were re-analyzed using a new RT-PCR protocol in which M-MLV reverse transcriptase and AmpliTaq DNA polymerase were replaced with SuperScript III reverse transcriptase (Invitrogen) and rTth DNA polymerase, XL (Applied Biosystems), which contain Deep Vent DNA polymerase with exonuclease activity. This experiment confirmed the result from the repeated experiment with AmpliTaq DNA polymerase (Fig. 1). Finally, with the use of AmpliTaq DNA polymerase or rTth DNA polymerase, XL, the PCR step, consisting of 70 cycles of two rounds, was used to amplify two independent HCV clones from our previous study. Direct amplicon sequencing indicated complete identity to the cloned HCV sequences (data not shown).

4.3. Genetic and phylogenetic analysis

All genetic and phylogenetic analyses showed similar results with either inclusion (the sequence being analyzed: 387 bp in length) or exclusion (the sequence being analyzed: 363 bp length) of the codons containing the eight potential PCR-associated errors as described above. For simplicity, only results generated under 363-bp analytical domain were presented.

There was no obvious difference between the HCC and cirrhosis groups from dataset 2 in terms of genetic diversity and Tajima’s D test (Fig. 2). In comparison with the dataset 1, both HCC and cirrhosis groups from the dataset 2 had higher genetic diversity, especially with significantly increased $d_N$ values ($p < 0.001$) (Fig. 2). Accordingly, the HCC and the cirrhosis group of dataset 2 had increased $d_0/d_2$ values, corresponding to Tajima’s D test that showed the stronger negative values in groups HCC ($−1.46$) and cirrhosis ($−1.35$) than the dataset 1 ($−1.25$).

The regression analysis of root-to-tip distance against sampling dates did not support a clock-like evolution in the ML tree constructed with the datasets 1 and 2 ($R^2 = 0.026$). In subsequent MCMC simulation, a relaxed molecular clock (uncorrelated, lognormal) was then applied. By giving each HCV strain a defined trait, either disease status or sampling time, BaTS analysis was run in two types of data combinations, HCC/cirrhosis and dataset 1/dataset 2. The former did not show obvious branch clustering in terms of disease status (HCC or cirrhosis) in MCMC trees ($AI = 8$, $p = 0.35$; $PS = 40$, $p = 0.025$). When including all HCV strains from the datasets 1 and 2, tree topologies were significantly associated with the distribution of qualitative traits, either disease status ($AI = 12.5$, $p < 0.001$; $PS = 73$, $p < 0.001$) or sampling dates ($AI = 5.68$, $p < 0.001$; $PS = 35$, $p = 0.001$) (Fig. 3).

5. Discussion

Identification of HCC-specific mutations is a challenging endeavor. HCV’s great diversity makes it difficult to perform a comparative analysis among different HCV genotypes or subtypes. The existence of ethnically or geographically specific mutations is also a concern. More importantly, even if putative HCC-associated mutations are observed, it is not known if these mutations are responsible for the HCC incidence or a simple result of evolutionary adaptation. The current study was designed to focus on a single HCV genotype (4a) in a single geographical region (Egypt). All three datasets have explicit sampling dates, patterns, and adequate numbers to provide a unique opportunity to explore the possibility of an epidemiological relationship between HCV mutations and HCC incidence.

Initial comparative analysis identified four statistically significant nucleotide substitutions in the HCC and cirrhosis groups. However, in repeated experiments, these four mutations were completely lost with the consistent appearance of alternative four mutations (Fig. 1). In sequence chromatograms, almost all eight mutations showed single peaks, suggesting that these mutations are not located in highly variable sites. Experimental contamination is not supported because all other sites from the same HCV isolates appear the same (Fig. 1). Under 70 PCR cycles, four putative false mutations over 387-bp domain give an error rate at
1.5 × 10^{-4} substitutions per base pair, which is well within the range of Taq DNA polymerases 1.1.10^{-4} to 2.0 × 10^{-5} errors per base pair.44–49 Thus the eight nucleotide substitutions observed are most likely not authentic. Under the same experimental procedure, the appearance on different positions in a non-random pattern from repeated experiments may be attributable to the batch to batch difference of AmpliTaq DNA polymerase. Another factor is the subtle alteration of template heterogeneity due to additional 1-year storage. The role of template heterogeneity contributing to the error rate of DNA polymerase has been ignored largely.40–42 Because of the complete sequence identity after 70-cycle PCR upon the use of plasmid DNA as the template, template heterogeneity may be a more possible factor to explain our observation. Finally, the four nucleotide mutations from the initial and repeated experiments are also present on the healthy volunteers from datasets 2 and 3, respectively (Fig. 1). Thus, even assuming a real nature, these mutations may just be a result of adaptive evolution without having any relationship with end-stage liver disease, either HCC or cirrhosis.

At the phylogenetic level, BaTS analysis revealed no apparent clustering in terms of their disease traits in HCC and cirrhosis. However, the inclusion of the dataset 1 (blood donors) resulted in strong association between disease traits (HCC/cirrhosis or blood donors) (Fig. 3). Since the dataset 1 were sampled in 1993, such an observation may be largely due to different sampling dates rather than disease traits. Because of a small number (n = 6) of HCV sequences from blood donors in the dataset 2, a univocal answer may require the analysis with the inclusion of more contemporaneously collected HCV sequences from patients without HCC/cirrhosis.

The HCC group, cirrhosis group and the dataset 1 all have significantly negative Tajima’s D values, indicating an excess of low-frequency mutations and therefore a positive selection pressure. Among datasets the HCC group has the strongest negative Tajima’s D value, corresponding with its highest d0/d2 ratio (Fig. 2). Actually, while having similar d2 values, the HCC and cirrhosis have significantly higher d0 values than the dataset 1 (p < 0.001) (Fig. 2). Taken together, these data suggest a strong evolutionary pressure of HCV in patients with end-stage liver diseases, which is consistent with previous reports in HCC patients infected with HCV genotype 1b.8,13 An important implication from this observation is a theoretically enhanced chance for the detection of putative HCC or cirrhosis-specific mutations, which requires caution in data interpretation since the mutations identified may simply be the consequence of adaptation.

It should be noted that our analysis was based on a short HCV domain, the 387-bp partial Core/E1 region. Comprehensive understanding of HCC-specific mutations and/or strains may require a full-length HCV genome scanning as well as the availability of adequate number of samples collected in both simultaneous and longitudinal patterns. In this setting, the current study represents a proof-of-concept investigation in terms of experimental approaches and phylogenetic techniques to address this elusive but clinically important issue.

Funding
This work was supported by NIH grants R01 DK80711 (Dr. Xiaofeng Fan), R21 AI076834 (Dr. Adrian M. Di Biscaglie) and USA and Egypt Science and Technology Joint Fund BIOG-002-004 (Dr. Adrian M. Di Biscaglie).

Competing interest
All authors with this manuscript have no conflict of interests.

Ethical approval
This study is purely a genetic analysis with sequence data from published studies and thus does not involve any ethical issues.

Acknowledgments
This work was supported by NIH grants R01 DK80711 (XF), R21 AI076834 (AMD) and USA and Egypt Science and Technology Joint Fund BIOG-002-004 (AMD).

References


41. Loewen PC, Switala J. Template secondary structure can increase the error frequency of the DNA polymerase from Thermus aquaticus. Gene 1995;164:59–63.