Mutations in the Hepatitis C Virus core Gene Are Associated with Advanced Liver Disease and Hepatocellular Carcinoma

Sarah L. Fishman,1 Stephanie H. Factor,1 Cinzia Balestrieri,1 Xiaofeng Fan,2 Adrian M. DiBisceglie,2 Suresh M. Desai,3 Gary Benson,4 and Andrea D. Branch1

Abstract

Purpose: Hepatitis C virus (HCV) infection can promote the development of hepatocellular carcinoma (HCC). Published data implicate the HCV core gene in oncogenesis. We tested the hypothesis that core gene sequences from HCC patients differ from those of patients without cirrhosis/HCC.

Experimental Design: Full-length HCV sequences from HCC patients and controls were obtained from the investigators and GenBank and compared with each other. A logistic regression model was developed to predict the HCC risk of individual point mutations and other sequence features. Mutations in partial sequences (bases 36–288) from HCC patients and controls were also analyzed. The first base of the AUG start codon was designated position 1.

Results: A logistic regression model developed through analysis of full-length core gene sequences identified seven polymorphisms significantly associated with increased HCC risk (36G/C, 209A, 271U/C, 309A/C, 435A/C, 481A, and 546A/C) and an interaction term (for 209A–271U/C) that had an odds ratio < 1.0. Three of these polymorphisms could be analyzed in the partial sequences. Two of them, 36G/C and 209A, were again associated with increased HCC risk, but 271U/C was not. The odds ratio of 209A–271U/C was not significant.

Conclusions: HCV core genes from patients with and without HCC differ at several positions. Of interest, 209A has been associated with IFN resistance and HCC in previous studies. Our findings suggest that HCV core gene sequence data might provide useful information about HCC risk. Prospective investigation is needed to establish the temporal relationship between appearance of the viral mutations and development of HCC.

Viruses are among the most important human carcinogens. Several viruses, including the human papillomavirus, encode oncogenic proteins that promote cellular transformation (1). Human papillomavirus strains differ in their oncogenic potential (2), showing that specific mutations can modulate viral transforming activities. Chronic infection with the hepatitis C virus (HCV) increases the risk of hepatocellular carcinoma (HCC) in patients who progress to liver cirrhosis. The role of HCV proteins in the development of HCC is unclear. Based on the human papillomavirus model, if HCV has direct oncogenic effects, it is likely that the sequences of HCV in patients with HCC are different from the viral sequences in patients with early-stage liver disease.

Of the several HCV proteins reported to alter cellular growth (3–6), the core protein is the one most strongly associated with cellular transformation. Expression of the core protein enhances cell proliferation, DNA synthesis, cell cycle progression, cellular transformation, and liver cancer in experimental systems (7–15). Transgenic mice containing the HCV core gene develop steatosis and liver cancer (16–18).

The HCV core gene is a complex genetic region of the viral genome. It contains two open reading frames (19, 20), three confirmed RNA structures (20, 21), and several additional putative RNA signals/structures (21). Mutations in the core gene therefore have the potential to alter the proteins encoded by the two open reading frames, the RNA secondary structures, and/or the RNA signals.

HCV sequences differ in patients with early-stage versus late-stage liver disease (22–27). It is not known whether the sequence differences are present because viruses with certain polymorphisms have enhanced potential to cause liver damage; consequently, viruses that carry these mutations are more likely to be found in patients with liver cancer. Alternatively, the sequence differences arise because liver disease progression changes the cellular environment in which HCV replicates and selects for variants carrying polymorphisms that enhance survival in the damaged liver. HCV core genes differ in tumor versus nontumor samples of the same liver (27–32). Interestingly, analyses of full-length core gene sequences from serum

Authors' Affiliations: 1Department of Medicine, Mount Sinai School of Medicine, New York, New York; 2Department of Internal Medicine, Saint Louis University, Saint Louis, Missouri; 3Abbott Laboratories, Abbott Park, Illinois; and 4Departments of Biology and Computer Science, Boston University, Boston Massachusetts

Received 9/18/08; revised 1/1/09; accepted 1/28/09; published online First 4/21/09.

Grant support: NIH grants DA016156 and DK066939 (A.D. Branch).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Requests for reprints: Andrea D. Branch, 1 Gustave L. Levy Plaza, Box 11-20, New York, NY 10029. Phone: 212-659-8371; Fax: 212-348-3571; E-mail: andrea.branch@mssm.edu.

©2009 American Association for Cancer Research. DOI:10.1158/1078-0432.CCR-08-2418
Our study shows that sequences of the hepatitis C virus (HCV) in patients with hepatocellular carcinoma differ from those of patients with early-stage liver disease. We map sequence differences in the viral core gene, the HCV gene most strongly implicated in cellular transformation and the development of liver cancer. One polymorphism was particularly strongly associated with liver cancer. Specifically, 209A was present in 56% of the full-length sequences from patients with liver cancer but only 35% of patients with early-stage disease. Our findings contribute to information about the evolution of advanced liver disease and liver cancer in patients with chronic HCV infection. Future studies are planned to determine the prognostic significance of the sequence features we detected and to determine their effect on HCV gene expression and core protein function.

Translational Relevance

Our study shows that sequences of the hepatitis C virus (HCV) in patients with hepatocellular carcinoma differ from those of patients with early-stage liver disease. We map sequence differences in the viral core gene, the HCV gene most strongly implicated in cellular transformation and the development of liver cancer. One polymorphism was particularly strongly associated with liver cancer. Specifically, 209A was present in 56% of the full-length sequences from patients with liver cancer but only 35% of patients with early-stage disease. Our findings contribute to information about the evolution of advanced liver disease and liver cancer in patients with chronic HCV infection. Future studies are planned to determine the prognostic significance of the sequence features we detected and to determine their effect on HCV gene expression and core protein function.

Materials and Methods

Assembly of sequences for analysis. Full-length core gene sequences (bases 1-573) were downloaded from GenBank using the HCV database interface in November 2007. Genotype 1b sequences were included in further analysis if they were isolated from human serum and represented a unique patient who was treatment-naive; had not undergone liver transplantation; was not actively infected with either HIV or hepatitis B virus; whose liver disease could be identified as either noncirrhotic, cirrhotic without HCC, or HCC; did not have additional noted comorbidities (e.g., hemophilia); and contained the complete HCV core gene without deletions or insertions. Publications linked to the sequences or the GenBank entries were reviewed to identify sequences that satisfied the inclusion criteria. Sequences were excluded if insufficient information was available. Sequences meeting the inclusion criteria were separated into two groups based on the disease state of the patients: (a) acute and chronic HCV infection without cirrhosis and (b) chronic infection with HCC, recognizing that most, if not all, HCV patients with HCC also have cirrhosis. To provide a second set of sequences to use in limited confirmatory testing, partial core gene sequences were downloaded from GenBank using the HCV database interface in November 2007. Partial sequences were retained in the study if they met the inclusion criteria applied to the full-length sequences, contained bases 36 to 288, but did not contain the complete core gene sequence.

Identifying polymorphisms associated with the country of origin. We first analyzed the full-length and partial sets of control sequences to identify any polymorphisms associated with either Japanese or Spanish origin. This was done because many of the full-length HCC sequences came from Japan and all of the partial HCC sequences came from Spain. To seek region-associated polymorphisms, we first constructed a consensus sequences derived from the multiple sequence alignment of the full-length non-HCC (control) HCV sequences. The majority nucleotide at each position was designated the consensus nucleotide. The first nucleotide of the AUG start codon was numbered nucleotide 1. At each position of the full-length sequences, the \( \chi^2 \) test or Fisher's exact test (R package), as appropriate, was used to compare the proportion of the non-HCC (control) sequences from Japan that had the consensus nucleotide to the proportion of control sequences from other locations that had the consensus nucleotide. Polymorphisms with \( P \leq 0.05 \) were identified and analyzed further using forward, backward, and stepwise multivariable logistic regression. Polymorphisms that remained significantly associated with Japanese origin were identified \( (P \leq 0.05) \) and all subsequent analyses of full-length sequences were adjusted for these Japan-associated polymorphisms and for Japanese origin. A similar process was used to seek Spain-associated polymorphisms among the partial sequences, but none were found and no adjustment was made for Spanish origin in analyses of the partial sequences.

Building a logistic regression model of the HCC risk associated with polymorphisms in the HCV core gene. We used a multistep process to identify mutations that were more prevalent among the HCV sequences from HCC patients than among the HCV sequences from patients without HCC. The proportion of the consensus nucleotide at each position in the full-length HCC HCV sequences was first compared with the proportion of the consensus nucleotide at each position in the full length non-HCC HCV sequences using a multivariable logistic regression model, which included the following covariates: the Japan-associated mutations and Japanese origin, and the position under analysis. Polymorphisms that were more common in the HCC HCV set \( (P \leq 0.05) \) were investigated further using multivariable logistic regression to identify mutations that remained significantly associated with HCC \( (P \leq 0.05) \). We next created interaction covariates using all possible two-way combinations of the point mutations found to be associated with HCC in the previous analysis and analyzed a series of multivariable logistic regression models to find the best fit. All tested models included the covariates for Japan-associated mutations and Japanese origin. We used forward, backward, and stepwise logistic regression to identify the polymorphisms and interactions significantly associated with HCC \( (P \leq 0.05) \). The best fit was determined by the Wald statistic.

Analysis of partial core gene sequences. To look for additional evidence that the polymorphisms identified by multivariable logistic regression were associated with HCC, the prevalence of these mutations in partial core gene sequences (bases 36-288) of HCC and control patients was evaluated using \( \chi^2 \) or Fisher's exact test, as appropriate, to the extent permitted by the length constraints. Sequence features with \( P \) values \( < 0.05 \) were considered significant. Calculations were done using SAS version 9.0 and Epi Info version 3.4.3.

\footnote{http://hcv.lanl.gov/components/sequence/HCV/combinedsearch/search.html}
Rat 3. Analysis of potential RNA structures. RNA secondary structural models were generated using M-fold\(^{(33, 34)}\) with default parameters. The inputs were portions of a consensus sequence of all the full-length genotype 1b core sequences analyzed in this study. Structures of interest were evaluated using R-fold analyzer, which is freely available at http://tandem.bu.edu/foldsupport/foldsupport.html and by request, in PERL format.

Results

Sequence collection, selection, and characterization

A total of 1,439 full-length sequences were downloaded from GenBank. Eight hundred ninety-eight were eliminated for the following reasons: 532 lacked sufficient information for inclusion, 208 were isolated from liver tissue, 104 were from later time points of patients whose earlier time point was already included, 19 were from chimpanzees, 10 were from pooled serum samples, 8 were cell culture-adapted, 4 were from patients with HIV and/or hepatitis B virus coinfection, 4 were nongenotype 1b recombinants, 3 were post-IFN treatment, 3 were post-liver transplantation, 2 contained insertions or deletions, and 1 was isolated from ascitic fluid. In total, 541 satisfied the inclusion criteria. Two hundred sixty sequences were products of direct sequencing. The remaining 281 were quasi-species clones from 18 patients. When only 2 clones were available, 1 was chosen at random; when >2 clones were available, the consensus sequence of the clones was constructed and used.

Table 1. Gene sequences of patients with HCV infection and HCC compared with gene sequences of subjects with HCV infection and without HCC

<table>
<thead>
<tr>
<th>Gene position</th>
<th>HCC sequence set ((n = 65), %)</th>
<th>Control sequence set ((n = 214), %)</th>
<th>Individual analysis, OR ((95% \text{ confidence interval}))</th>
<th>(P)</th>
<th>Multivariable analysis, OR ((95% \text{ confidence interval}))</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>36C/G 36A</td>
<td>3 (5)</td>
<td>3 (2)</td>
<td>8.38 (1.28-54.62)</td>
<td>0.026</td>
<td>14.79 (2.01-108.35)</td>
<td>0.008</td>
</tr>
<tr>
<td>78U 78C</td>
<td>2 (3)</td>
<td>43 (20)</td>
<td>0.16 (0.03-0.71)</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>209A 209G</td>
<td>37 (56)</td>
<td>74 (35)</td>
<td>2.93 (1.57-5.42)</td>
<td>&lt;0.001</td>
<td>13.50 (4.48-40.68)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>271U/C 271A</td>
<td>42 (65)</td>
<td>61 (29)</td>
<td>3.84 (1.91-7.70)</td>
<td>&lt;0.001</td>
<td>10.57 (3.42-32.66)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>309C/A 309U</td>
<td>29 (45)</td>
<td>55 (26)</td>
<td>2.04 (1.05-3.98)</td>
<td>0.034</td>
<td>2.66 (1.17-6.05)</td>
<td>0.018</td>
</tr>
<tr>
<td>384U 384C</td>
<td>3 (5)</td>
<td>1 (1)</td>
<td>12.70 (1.07-149.80)</td>
<td>0.043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>408U 408C</td>
<td>4 (6)</td>
<td>2 (1)</td>
<td>9.98 (1.38-72.12)</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>435A/C 435G</td>
<td>23 (35)</td>
<td>27 (13)</td>
<td>2.12 (1.05-4.26)</td>
<td>0.035</td>
<td>4.09 (1.73-9.68)</td>
<td>0.001</td>
</tr>
<tr>
<td>465U 465C</td>
<td>5 (8)</td>
<td>4 (2)</td>
<td>4.93 (1.07-22.64)</td>
<td>0.040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>481A 481G</td>
<td>13 (20)</td>
<td>8 (4)</td>
<td>3.47 (1.31-9.13)</td>
<td>0.012</td>
<td>3.825 (1.2-11.89)</td>
<td>0.02</td>
</tr>
<tr>
<td>546A/C 546G</td>
<td>8 (12)</td>
<td>9 (4)</td>
<td>3.85 (1.16-12.80)</td>
<td>0.027</td>
<td>10.04 (2.17-46.37)</td>
<td>0.003</td>
</tr>
<tr>
<td>209A/271U/C</td>
<td></td>
<td></td>
<td>0.16 (0.03-0.72)</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: ORs and \(P\) values were calculated using logistic regression while controlling for Japanese origin.

Analysis of potential RNA structures. RNA secondary structural models were generated using M-fold\(^{(33, 34)}\) with default parameters. The inputs were portions of a consensus sequence of all the full-length genotype 1b core sequences analyzed in this study. Structures of interest were evaluated using R-fold analyzer, which is freely available at http://tandem.bu.edu/foldsupport/foldsupport.html and by request, in PERL format.

Analysis of sequence diversity and complexity. Genetic distances at the nucleotide level were calculated using the Kimura two-parameter metric and these distances were used to construct phylogenetic trees using the neighbor joining algorithm (Phylip). Consensus trees were constructed from 1,000 bootstrap repetitions. Protein distances were calculated using the Dayhoff PAM model in Protidist (Phylip). The number of nucleotide and amino acid substitutions in the sequences of the HCC set and the control set were tallied relative to Con 1 (GenBank accession no. AF238799), which served as a genotype 1b reference sequence, using custom Perl scripts and compared using the Student’s \(t\) test.

The genetic complexity of the sequences from patients with HCC was compared with that of sequences from the control group by calculating the Shannon entropy at each position of the gene using the Los Alamos National Lab Entropy-Two tool\(^7\). Global Shannon entropies were compared by the Wilcoxon rank-sum test (R package version 2.6.0).

---

\(6\) http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi

\(7\) http://hcv.lanl.gov/content/sequence/ENTROPY/entropy.html
In total, 278 of the full-length sequences in this study were obtained from GenBank: 214 from patients with either acute or chronic HCV and without cirrhosis or HCC (the control group), 58 from patients with HCC, and 6 from cirrhotic patients without HCC. Seven additional sequences from patients with HCC were produced by the investigators, yielding a total of 65 sequences from patients with HCC (the HCC group; Supplementary Fig. S1A; Supplementary Table S1A). The sequences from the control group included 97 (45%) from Japan and the sequences from the HCC group included 58 (89%) from Japan. We assume that most of the patients with HCC were cirrhotic, as HCC in HCV patients almost always arises in the background of cirrhosis. The number of available sequences from patients with cirrhosis but without HCC was too small to allow detailed analysis.

A total of 1,824 partial core sequences were downloaded from GenBank. We verified that all these sequences were from patients not represented in the full-length set. Of these, 1,132 were eliminated for the following reasons: 744 lacked sufficient information for inclusion, 153 were from patients with HIV

### Table 2. Effects of significant substitutions on potential secondary structures and coding function in the core gene

<table>
<thead>
<tr>
<th>Position</th>
<th>Control type</th>
<th>HCC type</th>
<th>RNA effect</th>
<th>Core effect</th>
<th>ARFP effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Control</td>
<td>HCC</td>
</tr>
<tr>
<td>36</td>
<td>A</td>
<td>G/C</td>
<td>None</td>
<td>Lys</td>
<td>Silent/Asn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Weakens SLV</td>
<td>Silent (Gly)</td>
<td>Ala</td>
</tr>
<tr>
<td>78</td>
<td>U</td>
<td>C</td>
<td>GC→GU</td>
<td>Gin</td>
<td>Arg</td>
</tr>
<tr>
<td>209</td>
<td>G</td>
<td>A</td>
<td>None</td>
<td>Strengthens SL_248</td>
<td>Met</td>
</tr>
<tr>
<td>271</td>
<td>A</td>
<td>U/C</td>
<td>Introduces GU or GC pair</td>
<td>Weakens SL_248</td>
<td></td>
</tr>
<tr>
<td>309</td>
<td>U</td>
<td>C/A</td>
<td>Disrupts AU base pair</td>
<td>Silent (Ser)</td>
<td>Leu</td>
</tr>
<tr>
<td>384</td>
<td>C</td>
<td>U</td>
<td>None</td>
<td>Silent (Cys)</td>
<td>Ala</td>
</tr>
<tr>
<td>408</td>
<td>C</td>
<td>U</td>
<td>None</td>
<td>Silent (Tyr)</td>
<td>Thr</td>
</tr>
<tr>
<td>435</td>
<td>G</td>
<td>A/C</td>
<td>None</td>
<td>Silent (Gly)</td>
<td></td>
</tr>
<tr>
<td>465</td>
<td>C</td>
<td>U</td>
<td>Weakens TSLE (GC→GU)</td>
<td>Silent (Val)</td>
<td></td>
</tr>
<tr>
<td>481</td>
<td>G</td>
<td>A</td>
<td>Strengthens TSLE, UG→UA</td>
<td>Gly</td>
<td>Ser</td>
</tr>
<tr>
<td>546</td>
<td>G</td>
<td>A/C</td>
<td>None</td>
<td>Silent (Leu)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Consensus nucleotide is shown in bold.

Fig. 1. Characteristic substitutions in the core gene may affect putative RNA secondary structures and alter core and ARFP function. White boxes, position of the nucleotide mutations on the RNA secondary structures of the core gene region; gray boxes, ARFP mutations; black boxes, core mutations. Domains of the core protein are shown below the secondary structure map of the core region.
and/or hepatitis B virus infection, 68 were isolated from liver tissue, 59 were from pooled serum samples, 39 were isolated from nonserum extrahepatic compartments, 38 were post-liver transplantation, 18 were from later time points of patients whose earlier time point was already included, 11 were nongenotype 1b recombinants, 1 contained deletions, and 1 was isolated from a patient with hemophilia. In total, 692 sequences satisfied the inclusion criteria. Of these, 71 were quasi-species clones from a total of 7 patients. A consensus of the clones from each subject was determined and used. The remaining 621 sequences were either directly sequenced or represented one of two available clones. In total, 628 sequences were obtained: 543 from patients with acute or chronic HCV without cirrhosis or HCC (the control group), 55 from patients with HCC (the HCC group), and 30 from cirrhotic patients without HCC (Supplementary Fig. S1B; Supplementary Table S1B). Because the region spanning nucleotides 36 to 288 had the highest representation in the partial sequences, the analysis was carried out on this region. Sequences from 49 from patients with HCC and 309 from control patients were analyzed.

### Table 3. Risk associated with HCC in the partial sequences

<table>
<thead>
<tr>
<th>Gene position</th>
<th>HCC HCV (n = 49), %</th>
<th>Control HCV (n = 309), %</th>
<th>OR (95% confidence interval)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>36C/G 36A</td>
<td>5 (10)</td>
<td>4 (1)</td>
<td>8.64 (1.92-40.14)</td>
<td>0.003</td>
</tr>
<tr>
<td>209A 209G</td>
<td>38 (77) 11 (23)</td>
<td>116 (37) 192 (63)</td>
<td>5.72 (2.69-12.40) &lt;0.001</td>
<td></td>
</tr>
<tr>
<td>271U/C 271A</td>
<td>11 (23) 38 (77)</td>
<td>87 (28) 221 (72)</td>
<td>0.74 (0.34-1.57) 0.39</td>
<td></td>
</tr>
<tr>
<td>209A/271UC</td>
<td>8 (16) 26 (8)</td>
<td>2.12 (0.82-5.34) 0.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Analysis of two potential stem loop structures in the distal region of the core gene. A, original model of the TSLE. B, modified model of the TSLE to include position 435. Sections highlighted in black are common between the two structures. Phylogenetic analysis of the original TSLE (A) and modified TSLE (B) as assessed by R-fold Analyzer are shown surrounding the helices of interest. The percentage of sequences in an alignment of HCC sequences supporting base-pair formation is given above each base pair in the helix. Below the helix, alternative base-pair combinations found in an alignment of HCC sequences (black), an alignment of HCC genotype consensus sequences (gray), and an alignment of control sequences (unshaded) are shown. Base pairs are presented with the 3' base on the left and the 5' base on the right. Bold, combinations of nucleotides that support Watson-Crick or wobble base pairing. The number following the base pair indicates the number of sequences observed in the alignment with the given combination of nucleotides.

Identification of HCV core gene polymorphisms associated with HCC in full-length sequences and construction of a logistic regression model

Univariable and then multivariable analysis found five Japan-associated polymorphisms: 9A, 264C, 273A, 330G, and 534C. All subsequent analyses controlled for both Japanese origin and the effects of the Japan-associated polymorphisms (see Materials and Methods).

The multivariable logistic regression models built to evaluate individual positions identified 10 HCV core gene polymorphisms significantly associated with increased HCC risk, 36G/C, 209A, 271C/U, 309A/C, 384U, 408U, 435A/C, 465U, 481A, and 546A/C, and 1 significantly associated with decreased HCC risk, 78U (Table 1). Each of these polymorphisms was examined to determine its effect on the amino acid sequence of the core protein and/or ARFP. Of the 11, 4 alter the amino acid sequence of the core protein: 36G/C (K12Silent/N), 209A (Q70R), 271U/C (M91L), and 481A (G161S). Six are predicted to change the amino acid sequence of the ARFP: 36G/C (K12Silent/N), 209A (Q70R), 271U/C (M91L), and 481A (G161S). Six are predicted to
modify RNA structural elements. Specifically, 36G/C alters the sequence of the portion of the internal ribosome entry site that overlaps with the beginning of the core protein coding sequence, 78U alters SLV (a required translation element), 271U/C and 309A/C alter SLV, and 465U and 481A alter the terminal stem loops element (TSLE). In addition, 271U/C changes the codon at a site of internal protein synthesis initiation (ref. 35; Table 2; Fig. 1).

Multivariable logistic regression analysis of the 11 polymorphisms identified in the individual analysis showed that 7 of these mutations were significantly associated with increased risk: 36G/C, 209A, 271C/U, 309A/C, 435A/C, 481A, and 546A/C. These 7 mutations and their 21 pairwise combinations were evaluated in a final logistic regression analysis. All 7 polymorphisms and the 209A-271C/U pair were significantly associated with HCC in the best-fit model ($P \leq 0.05$). The odds ratio (OR) of the 209A-271C/U interaction term was 0.16, indicating that the presence of both mutations in the same sequence conferred less risk than the product of their individual ORs.

Analysis of a second set of partial sequences

The OR of four of the sequence features in the best-fit model could be examined in a second set of partial sequences, which spanned bases 36 to 288. The ORs for 36G/C, 209A, 271C/U, and the 209A-271C/U pair are presented in Table 3. The significant increased risk of HCC associated with 36G/C and 209A was also found in this analysis ($P = 0.003$ and $P < 0.001$, respectively) both at the usual level of significance ($P \leq 0.05$) and when a Bonferroni correction is applied ($P < 0.0125$). Contrary to the first analysis, 271U/C was not significantly associated with HCC as an individual term or as part of the 209A-271C/U pair.

Further analysis of the full-length sequences

Analysis of positions in and around the TSLE. The TSLE is a putative RNA secondary structure composed of bases 438 to 516 (36). Mutations in the TSLE region of HCV RNA are associated with HCC (27). Multivariable logistic regression showed that 435A/C and 481A individually increased HCC risk, and individual analysis showed that 465U was associated with increased HCC risk. We explored the possibility that the TSLE might have a 5' extension involving base 435. M-fold identified a 10 bp helix in which bases 427 and 435 are predicted to pair; this pair cannot form in sequences with the HCC 435A/C polymorphism (Fig. 2). R-fold Analyzer (see Materials and Methods) and a custom bioinformatics tool that reports the percentage of the sequences in an alignment capable of forming predicted interactions, showed that the new helix has phylogenetic support in all three sets of full-length sequences that we analyzed: HCV HCC ($n = 65$); HCV non-HCC ($n = 214$); and consensus sequences of genotypes 1a, 1b, and 2 to 6 ($n = 8$). The effects of the 435A/C, 465U, and 481A HCC polymorphisms on helical elements in the TSLE are depicted in Fig. 2.

Analysis of sequence diversity and complexity. We compared full-length HCV HCC and control HCV sequences to identify regions where one group of sequences was more variable than the other. The core gene, core protein, and ARFP were individually investigated. In the first analysis, we identified nucleotides in the RNA and amino acids in the core protein and ARFP that differed from an external reference sequence, Con1 (GenBank accession no. AJ238799), and then compared the number of substitutions per sequence between groups. The average number of nucleotide substitutions in full-length core gene sequences of patients with HCC was 22.5 $\pm$ 4.8 (mean $\pm$ SD) and the average number of nucleotide substitutions in full-length core gene sequences of patients without HCC was 20.6 $\pm$ 5.4 ($P = 0.01$, Student's $t$ test). The average number of amino acid substitutions in ARFP of HCC patients was 18.3 $\pm$ 3.9 versus 16.8 $\pm$ 4.3 in ARFP of control patients ($P = 0.01$). The average number of amino acid substitutions in the core protein did not differ between HCC and non-HCC patients. Within the RNA structural element SLV, the sequences from patients with HCC had fewer substitutions than the control sequences ($0.44 \pm 0.66$ versus $0.72 \pm 0.78$; $P = 0.010$), suggesting that the sequences from patients with HCC were under a greater selection pressure to preserve the wild-type function of this translation control element than the sequences from control patients, although the high degree of conservation of SLV makes it difficult to draw firm conclusions. The sequences from patients with HCC did...
not differ from sequences from control patients in the density of mutations in SLV and SL_248 but had more mutations in the TSLE (3.50 ± 2.2 versus 2.76 ± 1.7; P = 0.004). The number of sequences containing at least one mutation in the TSLE was 34 of 65 (52%) in the sequences from patients with HCC versus 35 of 214 (16%) of the sequences from the control patients (P < 0.001).

We next examined the Shannon entropy and mean genetic distance. The total entropy of the HCC set was slightly higher (42.6 versus 41.0; P < 0.001, Wilcoxon rank-sum test). The Shannon entropy at each nucleotide position of the core gene was compared between the two groups. A Monte Carlo randomization strategy was used to identify positions with a significant difference in entropy. In the HCC set compared with the control set, 12 positions, 28, 270, 271, 378, 384, 408, 435, 465, 472, 481, 517, and 546, were more variable, and one position, 78, was less variable (Fig. 3). In the sequences from patients with HCC, the mean pairwise genetic distance for the core protein was 17% greater than in the sequences from control patients (0.027 versus 0.023; P < 0.001), and for the ARFP, it was 11% greater (0.063 versus 0.056; P < 0.001). Taken together, these analyses show that there is a small but statistically significant increase in the variability and diversity of the sequences from patients with HCC. Countering this overall trend, SLV is more conserved and less diverse in the HCV HCC set. Of note, 3 of the 7 HCV polymorphisms associated with HCC, 36G/C, 209A, and 309A/C do not involve positions that have greater entropy in the HCC set than in the control set. Thus, it is unlikely that they reflect random sequence change or pressure to escape from anti-core antibodies, as both of these processes typically increase entropy.

Lack of clustering by disease status. A phylogenetic tree was generated from an alignment of all full-length HCC and control sequences. The sequences from patients with HCC did not form a separate cluster (data not shown).

Discussion

In this study, we examined HCV core gene sequences to identify features associated with HCC. All available genotype 1b core sequences were examined and those with sufficient information about the clinical status of the patient were included. The most striking result of this study is the finding that although the HCC core genes do not represent a separate clade or strain they do have characteristic mutations and other distinctive sequence features. HCC core gene sequences differed from controls at seven positions that remained associated with increased risk of HCC after controlling for country of origin. Although sequences from patients with HCC were significantly more diverse than sequences from the control patients, the actual difference in the number of mutations relative to an external consensus was small, ~8%, suggesting that the mutations in the sequences from patients with HCC do not reflect nonspecific sequence drift but rather are the result of positive selection pressure. Longitudinal studies are needed to determine whether the HCC-associated polymorphisms are associated with the risk of “developing” HCC or with the presence of HCC.

The HCV core gene is a complex genetic region with overlapping functions. Many of the nucleotides have additional functions above coding for the amino acids of core protein including encoding the ARFP and maintaining required RNA signals and structures. Other research will be needed to determine how the polymorphisms identified in this study contribute to the oncogenesis observed.

Seven polymorphisms were identified as having significant associations with HCC: 36G/C, 209A, 271U/C, 309A/C, 435A/C, 481A, and 546A/C. Further investigation is needed to determine whether the four polymorphisms that did not remain significantly associated with HCC in the final multivariable logistic regression analysis (78U, 384U, 408U, and 465U) are in fact associated with HCC. Positions of particular interest and their potential effect on HCV gene expression are discussed below.

36G/C may be involved in translation of the HCV polyprotein. The 36C/G polymorphism had an OR of 14.79 (95% confidence interval, 2.01-108.35) and a P = 0.008 in multivariable logistic regression analysis of full-length sequences and an OR of 8.64 (95% confidence interval, 1.92-40.14) and P = 0.003 in the analysis of the partial sequences. Because 36C/G was associated with HCC in both sets of sequences, the association between mutations at position 36 and HCC is likely to be biologically significant. The prevalence of the A36G/C mutation was low in both full-length (5%) and partial (10%) sequences from patients with HCC, in keeping with the high level of conservation, >95% across HCV genotypes, at this position in particular, and in the region from codons 4 to 15. Although this polymorphism may not be epidemiologically significant, the multiple roles of this position, in coding for both the core protein, (K12), ARFP (N11) and as part of the IRES (37, 38) are noteworthy, as mutations at this position may have a global effect on HCV protein production and replication.

G209A, a mutation previously linked to IFN resistance and HCC. The G209A substitution replaces the basic amino acid, arginine, with the neutral amino acid glutamine. The 209A polymorphism was present in 56% of the full-length sequences from patients with HCC sequences and in only 35% of the control sequences (P < 0.001); 209A was present in 74% of the partial sequences from patients with HCC and in only 40% of the control partial sequences (P < 0.001). The 209A mutation remained significantly associated with HCC in multivariate analysis of the full-length core sequences. The increased risk associated with this polymorphism was also observed in the set of partial sequences. In addition, this mutation was previously reported in 4 of 5 liver sequences from patients with HCC (39). These results indicate that there is a strong association between the G209A substitution and the presence of HCC.

In studies of Akuta et al., the G209A polymorphism has been linked to both IFN treatment failure and HCC (40). Genotype 1b sequences obtained from the study of Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (41), a prospective study of pretreatment sequences from serum designed to assess rates of response to peg-IFN and ribavirin therapy, were also analyzed at position 209. The 209A polymorphism was present in 4 of 16 (25%) of the subjects with a marked response to IFN treatment and in 12 of 16 of subjects with a poor response (75%; P = 0.01). Thus, the significance of 209A is heightened by its association with both IFN resistance and HCC risk. The antiproliferative effects of IFN pathway help protect cells from neoplastic transformation. It is important to determine if the
209A polymorphism confers IFN resistance and the pathway of this activity. This critical pathway may not only lead to treatment failure but also override cellular control mechanisms and contribute to cellular transformation and the development of HCC.

**Significance of position 271.** Further investigations are needed to determine the significance of mutations at position 271. Base 271 is part of codon 91, which encodes either leucine (271U/C) or methionine (271A) in the genotype 1b core protein. Analysis of position 271 in >300 sequences from Japanese patients with high viral load genotype 1b HCV showed associations between methionine 91 (271A) and IFN resistance and HCC (40). In contrast, smaller direct sequencing studies from Japan involving 28 HCC and 15 control sequences showed an association between leucine 91 (271U/C) and HCC (24, 26). In our investigation, 271U/C was associated with increased HCC risk in the full-length sequences but not in the partial sequences. The relationship between 271U/C and HCC risk may be clearer when the molecular changes that accompany mutations at position 271 are understood. Recent data indicate that the 271U/C polymorphism decreases the strength of a signal for the internal initiation of HCV protein synthesis (42). This modulation of a translation signal may be the key functional change that occurs as a result of mutations at position 271. The associated amino acid change (methionine to leucine) is highly conservative and not predicted to cause significant changes in the structure of the core protein (43).

**TSLE and positions 435, 465, and 481.** The distal position of the core gene contains a putative RNA structure called the TSLE originally proposed to comprise nucleotides 438 to 516. Three of the 11 positions identified in our individual analysis are in this region. Positions 465 and 481 alter the proposed structure of the TSLE. Our analysis of the region upstream of the TSLE suggested that the structural model of the TSLE is improved by the addition of a third helix. In this new helix, base 435 is predicted to be base-paired in 187 of 214 (87%) control sequences but is only paired in 43 of 65 (64%) of sequences from patients with HCC due to the 435A/C polymorphism. In our set of full-length sequences, mutations that altered the TSLE were present in 34 (52%) of the sequences from patients with HCC versus 35 (16%) of the control sequences (P < 0.001). In addition to the individually significant point mutations, we found that sequences from patients with HCC have a significantly, but only slightly higher, diversity in this region. This characteristic was also observed by Ogata et al., who reported a high density of mutations in this region in HCV sequences from patients with HCC (27).

**Strengths and limitations of this study.** In this study, we characterized all available sequences in GenBank. We examined every position in the core gene and combined multiple approaches to uncover several distinguishing features of sequences from patients with HCC. Because the majority of sequences in the full-length set were from Japan, we adjusted for Japanese origin in our multivariable analyses. Although the high percentage of Japanese sequences may have biased our findings, two of the three mutations associated with HCC in the full-length set of sequences were also associated with HCC in the partial set of sequences, which were from Spain. This result suggests that our results have global relevance. Finally, the positions identified in this study provide an important foundation for future experiments to understand how the mutations change the biochemical properties of the core gene and confer a selective advantage. In the future, it will be important to learn whether these mutations enhance cellular proliferation and IFN resistance.

Limitations of our study include its cross-sectional design. It is possible and likely that several control sequences were from patients who went on to develop HCC or who had undiagnosed HCC at the time of specimen collection. This may have lead to an underestimate of the risk caused by some sequence features and/or an underestimate of the mutations associated with HCC. Now that polymorphisms associated with HCC have been identified, it is important to determine when during the course of disease progression these mutations arise. Finally, our data set did not contain information about certain possible confounders such as age, race, sex, duration of infection, and long-range interactions between nucleotides in the core gene and in other portions of the HCV genome.

**Conclusions.** We identified seven polymorphisms in the HCV core gene associated with increased HCC risk in a multivariable logistic regression model. Two of them, 209A and 36G/C, were also associated with increased HCC risk in an independent set of partial sequences. Many of the seven polymorphisms alter predicted RNA structures, which may indicate that these structures are under selection pressure and that they regulate processes that contribute to oncogenesis. Although many host and viral factors contribute to the development of HCC, this study identified polymorphisms in the HCV core gene that may be indicative of the presence or forthcoming development of HCC. It remains unclear whether these mutations are inherently oncogenic or if they are benign adaptations to the altered environment that exists in damaged livers. In either case, they are associated with HCC and thus may serve as clinical markers of HCC. Patients harboring HCV strains with these mutations may benefit from closer surveillance. Further studies are warranted to assess the diagnostic value of HCV sequencing in cirrhosis/HCC identification and to identify the molecular pathways underlying the association between certain HCV mutations and advanced liver disease and HCC.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.

**References**


Mutations in the Hepatitis C Virus core Gene Are Associated with Advanced Liver Disease and Hepatocellular Carcinoma

Sarah L. Fishman, Stephanie H. Factor, Cinzia Balestrieri, et al.


Updated version
Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/15/9/3205

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2009/05/20/1078-0432.CCR-08-2418.DC1

Cited articles
This article cites 42 articles, 13 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/15/9/3205.full.html#ref-list-1

Citing articles
This article has been cited by 3 HighWire-hosted articles. Access the articles at:
/content/15/9/3205.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.