Assessment of Hepatitis C Virus Sequence Complexity by Electrophoretic Mobilities of Both Single- and Double-Stranded DNAs

YU-MING WANG,1 STUART C. RAY,1 OLIVER LAEYENDECKER,1 JOHN R. TICEHURST,2,3 AND DAVID L. THOMAS1*

Departments of Medicine1 and Pathology,2 Johns Hopkins University School of Medicine, Baltimore, and Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville,3 Maryland

Received 4 May 1998/Returned for modification 15 June 1998/Accepted 16 July 1998

To assess genetic variation in hepatitis C virus (HCV) sequences accurately, we optimized a method for identifying distinct viral clones without determining the nucleotide sequence of each clone. Twelve serum samples were obtained from seven individuals soon after they acquired HCV during a prospective study, and a 452-bp fragment from the E2 region was amplified by reverse transcriptase PCR and cloned. Thirty-three cloned cDNAs representing each specimen were assessed by a method that combined heteroduplex analysis (HDA) and a single-stranded conformational polymorphism (SSCP) method to determine the number of clonotypes (electrophoretically indistinguishable cloned cDNAs) as a measure of genetic complexity (this combined method is referred to herein as the HDA+SSCP method). We calculated Shannon entropy, incorporating the number and distribution of clonotypes into a single quantifier of complexity. These measures were evaluated for their correlation with nucleotide sequence diversity. Blinded analysis revealed that the sensitivity (ability to detect variants) and specificity (avoidance of false detection) of the HDA+SSCP method were very high. The genetic distance (mean ± standard deviation) between indistinguishable cloned cDNAs (intraclonotype diversity) was 0.6% ± 0.9%, and 98.7% of cDNAs differed by <2%, while the mean distance between cloned cDNAs with different patterns was 4.0% ± 3.2%. The sensitivity of the HDA+SSCP method compared favorably with either HDA or the SSCP method alone, which resulted in intraclonotype diversities of 1.6% ± 1.8% and 3.5% ± 3.4%, respectively. The number of clonotypes correlated strongly with genetic diversity (R2 = 0.93), but this correlation fell off sharply when fewer clones were assessed. This HDA+SSCP method accurately reflected nucleotide sequence diversity among a large number of viral cDNA clones, which should enhance analyses to determine the effects of viral diversity on HCV-associated disease. If sequence diversity becomes recognized as an important parameter for staging or monitoring of HCV infection, this method should be practical enough for use in laboratories that perform nucleic acid testing.

Genetic variation of certain RNA viruses may explain their capability to cause persistent infections and evade traditional treatment and prevention efforts. Hepatitis C virus (HCV) frequently establishes chronic infection and has considerable sequence variation, especially in putative envelope proteins E1 and E2, for which <60% amino acid identity has been described worldwide (3, 15, 17, 18, 22, 37, 46).

Genetic variation may also refer to differences among the swarm of viral variants within a person, often called a quasispecies (10, 11). The variants in an HCV quasispecies generally have 94 to 99% nucleotide identity (2, 25, 35). Within a single specimen, such variation can be characterized in terms of diversity or complexity. Diversity is the mean genetic distance calculated for all pairs of sequences (26), where genetic distance is directly proportional to the number of nucleotide differences between two variants. Complexity refers to the population distribution of variants and has been calculated from sequence data (26) but has also been estimated more practically on the basis of either the number of distinct gel bands resulting from single-stranded conformational polymorphism (SSCP) analysis (24) or the number of indistinguishable cDNA clones (clonotypes) recognized by gel shift analysis (32).

Nucleic acid sequencing of cloned cDNAs remains the “gold standard” for the assessment of viral variation but is too cumbersome to be applied to large, population-based studies. Electrophoretic analysis of SSCP has been more expedient, but its sensitivity (ability to identify distinct clones) is limited and it does not provide an estimate of genetic distance (5, 6, 23). Heteroduplex analysis (HDA) is also convenient and provides information on both genetic complexity and distance (4, 8, 9, 13, 14, 21, 27, 32, 47). However, HDA alone may not be sufficiently sensitive (6). We sought to develop a method combining HDA and SSCP analysis (referred to herein as the HDA+SSCP method) that could assess genetic complexity with high sensitivity (ability to discriminate between distinct sequences) and specificity (chance that clones detected as distinct truly represent distinct sequences) and that could provide accurate estimates of genetic diversity in a large prospective investigation by sampling a sufficiently large number of cloned cDNAs. In addition, we tested the value of estimating quasispecies complexity by calculating the Shannon entropy of the clonotype distribution, thus incorporating information about both the number of clonotypes and their respective proportions.

MATERIALS AND METHODS

Study subjects. As part of a prospective study of acute HCV infection, serial serum samples were obtained from individuals in the ALIVE cohort of injection drug users in Baltimore, Md. (45). These samples were tested for antibodies to HCV by using the second-generation HCV 2.0 enzyme immunoassay (Ortho...
Diagnostic Systems, Raritan, N.J.) as previously described (45). Individuals were identified as seroconverters when a sample tested positive following at least one negative result. Positive results were supplemented by a recombinant immuno- blot assay (Corona RIBA HCV 2.0 strip immunoblot assay; Chiron Corporation, Emeryville, Calif.) and confirmed by the detection of HCV RNA by a reverse transcriptase PCR (RT-PCR) assay (Ambion HCV Monitor; Roche Diagnostic Systems, Branchburg, N.J.) as previously described (44). Twelve samples from seven subjects were arbitrarily selected for this investigation, without knowledge of risk factors or disease state (Table 1). Genotyping by analysis of Core–E1 HCV sequence, according to the nomenclature of Simmonds et al. (36), revealed that all subjects were infected with HCV genotype 1a except subject 11469 (sample E), which was infected with genotype 1b (44).

Reverse transcription and nested-PCR amplification. Total RNA was extracted from 100 μl of plasma or serum by using 1 ml of Trizol LS Reagent (Life Technologies, Gaithersburg, Md.) at room temperature, followed by chloroform extraction and isopropanol precipitation in the presence of 20 μg of glycogen (Boehringer Mannheim, Indianapolis, Ind.). The RNA pellet was washed with 75% (vol/vol) ethanol and then air dried briefly and redissolved in 50 μl of diethyl pyrocarbonate-treated water with 10 mM dithiothreitol (Promega, Madison, Wis.) and 10 mM Tris–HCl, pH 8.3, 75% (vol/vol) ethanol, and then air dried briefly and redissolved in 50 μl of water.

Cloned cDNA was amplified from bacterial colonies as follows. Each colony was diluted in 100 μl of distilled water, mixed with 10 μl of MDE HDA (FMC Bioproducts, Rockland, Maine) at a 1:4 concentration according to the manufacturer’s protocol, and then immediately plunged into crushed dry ice. Five microliters of each reaction mixture plus 1 μl of Triple Dye loading buffer (FMC Bioproducts) was loaded on an MDE gel (19.0 by 10.0 cm) in a Micro-Well gel apparatus (Bio-Rad Laboratories, Hercules, Calif.), followed by electrophoresis at 140 V for 4,500 V·h (≈32 h). Gels were stained in a 1:10,000 dilution of SYBR Green II (FMC Bioproducts) for 30 min and documented with an Eagle Eye II Still Video System (Stratagene, Carlsbad, Calif.) with a SYBR Green Filter (Stratagene).

Selection of subject-specific driver. For each sample, a preliminary HDA+SSCP gel was performed with 20 cloned cDNAs, 1 of which was also randomly selected for use as a preliminary driver. Inspection of the preliminary gel always yielded a cloned cDNA with a relatively large gel shift as well as an SSCP pattern with four distinct bands that represented a minority of the SSCP patterns among the cloned cDNAs. This process resulted in the selection of a subject-specific driver that optimized resolution of all HDA and SSCP bands, in contrast to results obtained with a suboptimal driver (Fig. 1).

Complexity analysis. For each of the three electrophoretic methods (HDA, SSCP analysis, and the HDA+SSCP method), a clonotype was defined as a group of indistinguishable cloned cDNAs based on inspection of the gel for the number and mobility of bands. The number of clonotypes and the number of clones comprising each clonotype were recorded.

To quantify complexity, the Shannon entropy (H) calculation (34, 40) was applied. Shannon entropy incorporates both the number of clonotypes and the number of cloned cDNAs in each clonotype. It is defined as $H = -\sum_i P_i \ln(P_i)$, where $N$ is the total number of clonotypes and $P_i$ is the number of clones represented in clonotype $i$. Because each sample in our study could have up to 33 clonotypes, which would yield a maximum value of $H(33)$ for $H$, a normalized value of $H$, denoted $H'$, was defined as $H(33)$.

Nucleotide sequencing. Sequences were unidirectionally determined from the M15 primer binding site of plasmid clones by using a PRISM automated sequencer (version 2.1.1; Applied Biosystems, Inc., Foster City, Calif.). Except as noted, sequences were obtained from up to three representatives of each clonotype. Sequences were assembled by using the ESEE3s program (E. Cabot, Madison, Wis.), and primer sequences were removed.

Calculation of diversity. For a pair of sequences, distance was calculated as the Hamming distance, or number of nucleotide differences, per 100 bases. Intra-clonotype diversity (d) was defined as the mean of pairwise distance values for cloned cDNAs from clonotype $i$. Interclonotype diversity (d) was defined as the mean of distance values for clones being all cDNAs from clonotypes $i$ and $j$, where $i \neq j$. The weighted diversity ($D_w$) for a specimen was defined as the mean of distance values for all pairs of cloned cDNAs from that specimen, with each pair’s contribution weighted according to the proportion of cloned cDNAs represented by its clonotype. This calculation was simplified by summing the intraclonotype and interclonotype diversities as separate terms and can be represented as follows:

$$D_w = \frac{\sum_{i,j} w_{ij} d_{ij}}{\sum_{i,j} w_{ij}}$$

where $w_{ij}$ is the fraction of the total number of clones represented by clonotype $j$.

### Table 1. Characteristics of subjects and samples and results of the HDA+SSCP method and sequence analysis

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sample</th>
<th>Log_{10} [HCV RNA] $^a$</th>
<th>Clonotype distribution $^b$</th>
<th>No. of clonotypes $^b$</th>
<th>Entropy $^c$</th>
<th>$D_w$ $^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10388</td>
<td>A1</td>
<td>4.5</td>
<td>27,3,1,1,1</td>
<td>5</td>
<td>0.20</td>
<td>0.52</td>
</tr>
<tr>
<td>10718</td>
<td>B1</td>
<td>5.4</td>
<td>9,6,2,2,1,1,1,1,1,1,1,1</td>
<td>17</td>
<td>0.70</td>
<td>ND $^d$</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>5.4</td>
<td>4,3,3,2,2,1,1,1,1,1,1,1,1</td>
<td>22</td>
<td>0.85</td>
<td>ND $^d$</td>
</tr>
<tr>
<td>10960</td>
<td>C1</td>
<td>5.0</td>
<td>30,1,1,1,1,1</td>
<td>18</td>
<td>0.74</td>
<td>ND $^d$</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>5.1</td>
<td>25,2,2,1,1,1,1</td>
<td>7</td>
<td>0.28</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>C3</td>
<td>6.5/9/6</td>
<td>4,9,15,3,2,1,1,1,1,1,1</td>
<td>12</td>
<td>0.55</td>
<td>1.30</td>
</tr>
<tr>
<td>10976</td>
<td>D</td>
<td>3.0</td>
<td>26,1,1,1,1,1,1</td>
<td>8</td>
<td>0.27</td>
<td>1.23</td>
</tr>
<tr>
<td>11469</td>
<td>E</td>
<td>3.5/9/3</td>
<td>12,7,4,2,1,1,1,1,1,1</td>
<td>9</td>
<td>0.57</td>
<td>0.94</td>
</tr>
<tr>
<td>12951</td>
<td>F</td>
<td>2.7</td>
<td>6,2,2,2,2,1,1,1,1,1,1,1,1</td>
<td>23</td>
<td>0.85</td>
<td>6.17</td>
</tr>
<tr>
<td>40209</td>
<td>G</td>
<td>7.6</td>
<td>14,11,2,2,1,1,1,1,1,1,1,1</td>
<td>8</td>
<td>0.43</td>
<td>0.52</td>
</tr>
</tbody>
</table>

$^a$ [HCV RNA], HCV RNA copies per milliliter of serum.

$^b$ Number of electrophoretically indistinguishable cloned cDNAs.

$^c$ Normalized Shannon entropy, a calculated value that incorporates the number and distribution of clonotypes.

$^d$ Calculated value that incorporates the distribution of clonotypes and their respective genetic diversities.

$^e$ ND, not done. The normalized entropy was not calculated for these subjects.

$^f$ Insufficient seq data were obtained for this calculation.
DISCUSSION

Our data indicate that the HDA+SSCP method identified distinct cloned cDNAs with high sensitivity and specificity. In addition, by screening 33 cloned cDNAs per sample (versus 2 to 10 as is customary), more precise measurements of viral
complexity were obtained. Viral complexity measured by this method also correlated with diversity determined by nucleotide sequencing. When analyses were done with fewer cloned cDNAs, much less accurate estimates were obtained (Table 2), underscoring the merits of our approach (analysis of 33 cloned cDNAs by the HDA\textsuperscript{1}SSCP method).

In developing this method we sought a highly sensitive means for detecting distinct sequences in a mixture of cloned cDNAs, so that we can screen a large number of samples for sequence variation in future investigations. The sensitivity we report for the HDA\textsuperscript{1}SSCP method is comparable to that reported for more laborious and expensive methods which employ radiolabeling and purification of single-stranded drivers (14). Because we defined a clonotype according to any difference in electrophoretic migration, we expected that the HDA\textsuperscript{1}SSCP method would be at least as sensitive as either method alone. This is illustrated by Fig. 1A, in which lanes 1 and 2 appear identical by HDA but are clearly different by SSCP analysis and lanes 3 and 4 appear identical by SSCP analysis but have different HDA patterns. Importantly, and not guaranteed by this design, the HDA\textsuperscript{1}SSCP method maintained high specificity by not assigning identical cloned cDNAs to different clonotypes. Although methods combining HDA and SSCP for detection of genetic variation have been reported previously (1), either combined in one gel or performed separately, they used fewer samples and were not applied to sequences as variable as HCV-E2 (43).

We developed a subject-specific process for selecting a cloned cDNA driver because selection of an appropriate driver to which each variant is annealed is crucial for HDA. For other viruses, such as human immunodeficiency virus type 1, a single reference driver for cross-sectional or longitudinal assays allows comparison of gel shift both between individuals and over time (41). Like others, however, we were unable to form het-

![Figure 2](http://jcm.asm.org/)
eroduplexes when a driver representing this hypervariable region from another HCV quasispecies (i.e., another subject) was used (data not shown). Therefore, we could not use a single driver for all samples. We found that other approaches to driver selection, including random choice (14, 47) or selection of a majority-cloned cDNA (32), resulted in decreased sensitivity. Use of a driver which represented a majority clonotype for that specimen resulted in a large number of gel lanes with overlapping SSCP bands (rather than four distinct ones) and overlapping hetero- and homoduplex bands, both potentially obscuring small differences in gel shift (Fig. 1B). The approach described in Materials and Methods resulted in the selection of divergent minor variants for use as subject-specific drivers, maximizing clonotype identification while maintaining simplicity.

The combination of HDA and SSCP in one gel raises some unique methodological challenges. In order to obtain adequate resolution for both HDA and SSCP bands, which migrate with different rates, we found that at least a 15-cm migration distance was needed (data not shown). Prior studies of the HDA SSCP method utilized either denaturing conditions which interfered with heteroduplex formation, producing indistinct bands (1, 28, 39, 42), or slow cooling, which decreased the yield of SSCP by favoring formation of homo- and heteroduplexes (33). We found that rapid cooling under nondenaturing conditions resulted in the best yield and resolution (data not shown). We also found that SYBR Green II or conventional silver staining gave equivalent results, whereas SYBR Green I or ethidium bromide yielded unacceptably faint staining of SSCP bands. In order to preserve the SSCP bands it was

![Graphs showing frequency distributions of intraclonotype and interclonotype diversity for all samples.](http://jcm.asm.org/)

**FIG. 3.** Frequency distributions of intraclonotype and interclonotype diversity for all samples. All possible pairwise sequence comparisons were performed and classified as interclonotype or intraclonotype based on the results of HDA, the SSCP method, or the HDA+SSCP method. A histogram of the resulting percent diversity is displayed.
study the five samples from later time points did show in-
cloned cDNAs examined.
and G were plotted, and the normalized Shannon entropy values ($H^9$)
indicated in the inset.

duration of infection, with estimates of 1.44 
(16). The diversity of an HCV quasispecies increases with
from selecting samples collected soon after HCV infection
estimated. Rather, the relatively low diversity probably resulted
sequenced, it is unlikely that diversity was substantially underes-
32), relatively little viral diversity was detected in our 12 serum
necessary to perform electrophoresis in a 4°C chamber or at a
low voltage for 32 h as described in Materials and Methods.

Compared to earlier evaluations of HCV quasispecies (19,
32), relatively little viral diversity was detected in our 12 serum
samples. Because a large number of cloned cDNAs was se-
quenced, it is unlikely that diversity was substantially underes-
imated. Rather, the relatively low diversity probably resulted
from selecting samples collected soon after HCV infection
(16). The diversity of an HCV quasispecies increases with
duration of infection, with estimates of $1.44 \times 10^{-3}$ to $1.92 \times 10^{-3}$
nucleotide substitutions per site per year (29, 30). In this
study the five samples from later time points did show in-
creases in entropy, $D_w$, and the number of clonotypes, com-
pared with values from the earlier sample(s) from the same
subject (Table 1). Viral loads also generally increased at later
time points, but viral load did not correlate with diversity (e.g.,
Table 1, samples F and G).

A potential source of error in this type of investigation is
inherent in amplification and sequencing with currently avail-
able polymerases. Even as higher-fidelity thermostable en-
zymes become available, investigators need to remain aware of
polymerase error as a potential source of “unique clones”
which differ by one or two bases from other variants (38).
Gel-based methods are ideal for studying large cohorts be-
cause the same errors will affect both cases and controls, and
their effects will be further reduced by assessing large numbers
of samples. We propose that gel-based methods like ours can
be used to assign a weight to each sequence in a diversity

calculation, according to frequency (clonotype distribution),
be used to assign a weight to each sequence in a diversity

calculation, according to frequency (clonotype distribution),
providing a bias of giving equal weight to an infrequent vari-
cant, regardless of its origin (polymerase artifact or divergent
viral clone).

We calculated Shannon entropy (34) to evaluate its utility as
a measure of complexity that reflects both the number and
distribution of the clonotypes. In contrast, earlier reports have
estimated complexity only from the number of clonotypes. The
effect of neglecting clonotype distribution can be illustrated by
samples D and G (Fig. 4). Both samples yielded eight clon-
types, but the distributions are clearly different, with sample D
dominated by a major clonotype comprising nearly 80% of 33
cloned cDNAs while the major clonotype in sample G com-
prises less than a third of 33 clones. This difference, reflected
in the higher entropy value for sample G, is ignored when only
the number of clonotypes is used to estimate complexity. Un-
balanced distribution of clones is the rule rather than the
exception in HCV (Table 1), some of which may be due to
artifacts incorporated during amplification and sequencing.
Such errors give rise to “solitary clones” which have little effect
on the Shannon entropy calculation but are given equal weight
when the number of clonotypes is used as a measure of com-
plicity. Shannon entropy has been used to describe the com-
plexity of individual amino sequence positions in sequences
representing human (40) and human immunodeficiency virus
type 1 (20) genomes, as well as for automated gel analysis (7),
and has recently been applied to HCV clone distributions (31).

Despite the theoretical power of entropy to model complex-
ity, the number of clonotypes was more strongly correlated to
sequence diversity in this study. While both methods of esti-
mating complexity are very sensitive to the number of cloned
cDNAs assessed (Table 2), our assessment of 33 clones per
sample in this study favored the use of the number of clono-
types. By extrapolating from the data in Table 2, we predict
that Shannon entropy would be more useful for assessing more
than 50 cloned cDNAs per sample. These results are consistent
with the findings of Pawlotsky et al., who used SSCP of a
185-bp fragment to identify clonotypes, examined 30 HCV
clones amplified from each of 13 subjects, and sequenced up to
three clones per clonotype. They found linear regression of the
normalized Shannon entropy versus the (unweighted) diversity
to have a correlation coefficient ($R^2$) of 0.331 (31). Whether
this lower correlation was due to the region amplified, the size
of the amplicon, the gel-shift assay used to identify clonotypes,
or weighting of the diversity calculation remains to be deter-
dined. When entropy of sequences from the 5 subjects who
responded to alpha interferon therapy was compared to that of
8 subjects selected from the 40 subjects who did not respond,
those who responded had lower entropy values (31). Further
study with more extensive sampling is required to determine
whether Shannon entropy is truly superior to the number of
clonotypes as a predictor of diversity and whether it has bio-
logical significance.

We report a method for measuring HCV quasispecies com-
plexity that combines HDA and SSCP in a single gel visualized
with UV light. The method was sensitive and specific for de-
tecting clonotypes, and the number of clonotypes detected
correlated strongly with sequence diversity when 33 cloned

cDNAs are assessed. We introduce the use of entropy as a
measure of complexity, incorporating the distribution of vari-
ants as well as the number of clonotypes, but suggest that a
larger number of cloned cDNAs needs to be assessed when this
measure is to be used. Our approach is expected to facilitate
accurate analysis of the large number of cross-sectional and
longitudinal samples that are now becoming available and
could lead to clinical-laboratory assays for diagnosis or moni-
toring of HCV patients.

![Graph](image_url)

**FIG. 4. Measures of complexity.** The clonotype distributions for samples D and G were plotted, and the normalized Shannon entropy values ($H^9$) are indicated in the inset.

**TABLE 2. Measures of complexity: sensitivity to sampling**

<table>
<thead>
<tr>
<th>No. of cloned cDNAs examined</th>
<th>Correlation coefficient ($R^2$)</th>
<th>Clonotypes' vs $D_w$</th>
<th>Entropy' vs $D_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.46</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.70</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.78</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.93</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>

*Least-squares linear regression $R^2$ for the indicated relationship.
*The leftmost $n$ lanes were examined in rows for which $n$ was $<33$.
*The number of clonotypes and degree of entropy were calculated from the
cloned cDNAs examined.
ACKNOWLEDGMENTS

This study was supported by National Institutes of Health grants AI-40035, DA-04334, and DA-08004. We are grateful to the subjects in the ALIVE cohort for their generosity and time. We also thank Stephen Villano for identifying appropriate specimens and Amy Weiner for assistance in the development of the HDA+SSCP method.

REFERENCES


