Detection of Hepatitis C Virus Transmission by Use of DNA Mass Spectrometry

Lilia M. Ganova-Raeva,1 Zoya E. Dimitrova,1 David S. Campo,1 Yulin Lin,1 Sumathi Ramachandran,1 Guo-liang Xia,1 Christiane Honisch,2 Charles R. Cantor,2 and Yury E. Khudyakov1

1Molecular Epidemiology and Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia; and 2Sequenom, San Diego, California

The molecular detection of transmission of rapidly mutating pathogens such as hepatitis C virus (HCV) is commonly achieved by assessing the genetic relatedness of strains among infected patients. We describe the development of a novel mass spectrometry (MS)–based approach to identify HCV transmission. MS was used to detect products of base-specific cleavage of RNA molecules obtained from HCV polymerase chain reaction fragments. The MS-peak profiles were found to reflect variation in the HCV genomic sequence and the intrahost composition of the HCV population. Serum specimens originating from 60 case patients from 14 epidemiologically confirmed outbreaks and 25 unrelated controls were tested. Neighbor-joining trees constructed using MS-peak profile–based Hamming distances showed 100% accuracy, and linkage networks constructed using a threshold established from the Hamming distances between epidemiologically unrelated cases showed 100% sensitivity and 99.93% specificity in transmission detection. This MS-based approach is rapid, robust, reproducible, cost-effective, and applicable to investigating transmissions of other pathogens.

Keywords. viral transmission; genetic distance; RNA Mass-spectrometry; relatedness; network.

Hepatitis C virus (HCV) is a human pathogen belonging to the Flaviviridae family [1]. Its prevalence is about 2%–3% globally but can be >20% in regions of Africa and Asia where HCV is highly endemic [2–6]. In the United States, 3.2 million people are estimated to be infected with HCV [7, 8]. Infection is associated with the development of cirrhosis and hepatocellular carcinoma and is a major indication for liver transplantation [9]. Widespread use of injectable agents, whether medically approved or illicit, is a leading risk factor for HCV transmission [10, 11]; patients in healthcare settings are of special concern [12, 13]. The current standard-of-care therapy, which involves interferon in combination with ribavirin, is efficacious in approximately 50% of chronically infected patients [14]. As HCV infection is often asymptomatic, many infected people remain unidentified and, therefore, are not offered treatment [15]. There is no vaccine against hepatitis C.

Genetic heterogeneity is a hallmark of HCV. The virus is classified into 6 major genotypes and >50 subtypes [16]. In each infected individual, HCV exists as multiple variants or quasi-species [17–20]. The HCV genome, which encodes a single polyprotein, contains 2 highly heterogeneous regions: the hypervariable region 1 (HVR1), located at the 5' end of the envelope gene E2, and a region located in the NS5a gene [21–24]. Consensus sequencing of these 2 regions, and also of NS5b [25, 26], is commonly used to determine genetic relatedness among HCV strains and to identify HCV transmission [27]. However, a consensus sequence cannot adequately represent the entire HCV population present in the host, particularly in chronically infected patients, in whom the viral genetic heterogeneity can be extensive [28, 29]. Additionally, the extensive intrahost HCV evolution results in considerable changes in the structure of the HCV population in
patients over time, thus impeding genetic identification of HCV transmission that occurred in the distant past [30]. In an outbreak setting, consensus sequencing may prevent the identification of HCV variants directly involved in transmission among genetically related variants that are unassociated with the outbreak [31].

Accurate identification of HCV strains involved in transmission can be achieved by matching the genetic compositions of viral populations sampled from infected hosts. Since HVR1 is one of the most variable regions of the HCV genome, analysis of intrahost HVR1 variants is frequently used for identification of and tracking HCV transmissions [32–37]. Such analysis involves (1) separation of individual HVR1 variants either by genetic cloning of PCR amplicons [38] or by polymerase chain reaction (PCR) cloning, using end-point limiting-dilution of complementary DNA [32, 39]; and (2) sequencing of these variants. Next-generation sequencing technologies couple the separation of genetic variants with sequencing, thus simplifying significantly the assessment of viral heterogeneity. However, the high rate of sequencing errors generated per DNA read [40] and the problematic representation of intrahost viral heterogeneity [41] are potential hindrances to adopting next-generation sequencing approaches to the detection of transmissions.

DNA mass spectrometry (MS) provides a different platform for genetic analysis of viral strains. It has successfully been adopted for resequencing, microbial typing, and single-nucleotide polymorphism discovery [42, 43]. Here, we describe the development of an MS approach for identification of genetic equivalence among HCV strains involved in outbreaks. It does not entail separation and sequencing of viral variants; rather, it uses patterns of MS peaks generated from the entire intrahost viral population. We found the MS approach to be as accurate as previously established molecular technologies [33, 36] in the detection of HCV transmission but significantly more cost-effective and less time-consuming.

**MATERIALS AND METHODS**

**Specimens**

Sixty serum specimens were collected from 14 epidemiologically confirmed HCV outbreaks that occurred in the United States during 2003–2008 [35, 39–43]. A total of 25 serum specimens from patients infected with epidemiologically unrelated HCV subtypes 1a and 1b, identified from the National Health and Nutrition Examination Survey III, were geographically matched to the outbreak specimens. The collections are listed and described under SD1 in the Supplementary Data file. Genotypes and subtypes of each case were previously determined by phylogenetic analysis of the NS5B sequences.

**MS-Peak Profiles**

Data on MS peaks were extracted from runs for all 4 nucleotides after evaluation using an internal-system quality cutoff. The data were collected as lists of detected peaks for every single nucleotide–specific reaction, along with their corresponding mass and normalized peak intensity, which reflects the number of fragments found with the same mass [43]. To allow data collection, we used a small reference set compiled from GenBank entries representing HVR1 sequences of genotypes 1a and 1b (SD5). Peak lists were exported from iSEQ and merged according to an existing algorithm provided by Sequenom (San Diego, CA).

---

**Nucleic Acid Isolation, Reverse Transcription PCR (RT-PCR), Nested PCR, and MassCLEAVE™**

Nucleic acids were isolated with MagnaPure LC (Roche, SD2). The extracted nucleic acid underwent RT-PCR. HVR1 was amplified by PCR by using previously described protocol [30] (SD2). To assess heterogeneity, we repeatedly (8–12 times) sampled the first-round PCR products of HVR1 without limiting dilution to generate templates for the second-round amplification (SD2), using tagged nested primers. The second-round PCR product, a 308-nucleotide fragment encompassing the HCV genomic region at nucleotide positions 1302–1610, was then subject to the MassCLEAVE™ protocol [42] (SD4). The process is illustrated in Figure 1. All primers used for the first- and second-round PCRs are listed in SD3. All samples were processed in parallel in 384-well microtiter plates. PCR setup, SAP, post-PCR base-specific cleavage reactions (MassCLEAVE™, Sequenom, San Diego, CA), and pre- and postsequencing treatments were performed using an automated liquid handler Biomek 3000 (Beckman Coulter, Fullerton, CA).

---

**Figure 1.** Principle of transcription coupled RNaseA cleavage. Cleavage products are printed on 384-spot SpectroCHIPs and analyzed by mass-assisted laser desorption/ionization time-of-flight mass spectrometry.

---
The MS-peak list was concatenated in the order T(TF), A (TR), C(CF), and G(CR) to generate 1 MS-peak profile per sample. Each MS-peak profile was presented as a numeric vector of the detected masses and their corresponding signal intensity, which was used to calculate the distances between the MS-peak profiles (SD6). MS-peak profile distances between all subsamples with 4 eligible reactions were measured using the Hamming distance (presence or absence of a peak), Euclidean distance (taking into consideration peak intensity), and modified Euclidean distance (with a limit in the maximum intensity). Since all distances showed similar results, the Hamming distance was used throughout the analysis.

The MS-peak profile- and sequence-based distance matrices were used to construct the neighbor-joining trees [44]. The sequence-based distances were calculated using the Kimura 2-parameter model [45]. All calculations were performed using Matlab (Mathworks, Natick, MA).

**Statistical Analyses**

Analysis of molecular variance (AMOVA) was used to quantify the level of differentiation between cases. AMOVA partitions the variance of the MS-peak profile- or sequence-based distances into 3 sources: among the replicas for each HCV strain, among HCV variants from a single outbreak, and among epidemiologically unrelated HCV strains. AMOVA was calculated using Arlequin [46]. Significance levels of the genetic variance components were estimated using a permutation test (n = 10 000). In this test, the group label of each population sample is randomly permuted to simulate the null hypothesis that there is no difference between the groups. It must be noted that the pairwise comparisons among unrelated strains include the following: (1) pairwise distances among epidemiologically unrelated strains, (2) pairwise distances between epidemiologically unrelated and outbreak strains, and (3) pairwise distances between strains from different outbreaks. Pearson correlation was used to establish the degree to which sample heterogeneity as identified by MS-peak profile is related to the heterogeneity observed by sequencing data. The Mantel test was used to find correlation between the MS-peak profile- and sequence-based distance matrices.

**RESULTS**

**MS-Peak Profile Variability of HCV Amplicons With Known Genetic Complexity**

PCR is a stochastic process and, when applied to a complex viral population, can produce DNA amplicons that differentially represent intrahost viral variants. Experimental conditions may additionally bias representation of viral populations in PCR fragments [43]. These considerations imply that repeat amplification from heterogeneous viral populations can result in variable representation of intrahost viral variants in PCR amplicons, thus potentially reducing reproducibility of the genetic evaluation based on a single PCR amplification.

To establish the reproducibility of MS-peak profiles, 2 groups of samples were tested: (1) 16 amplicons from HCV HVR1 clones that have been separated by PCR cloning (S) and (2) amplicons from 23 artificially generated mixtures (M), each of which comprised 8 HVR1 clones (SD1). Each of those 39 samples was PCR amplified 6–12 times, and MS-peak profiles from the generated replicas were acquired after MassCLEAVE™ (see Material and Methods). The mean difference (±standard error of the mean [SEM]) in MS-peak profiles generated from the S sample replicas was 1.11%±0.06%, indicating reproducibility under the experimental conditions used for detecting MS-peak profiles. However, the mean difference (±SEM) among the M sample replicas was 3.85%±0.35%, which was significantly greater than that among S sample replicas (P < .0001). The lower MS-peak profile reproducibility for complex mixtures of viral variants suggests that MS-peak profiles generated from repeated rounds of PCR can more accurately represent the intrahost HCV HVR1 complexity.

Each base-specific MassCLEAVE™ reaction generates a set of short RNA fragments composed of k nucleotides (k-mers) originating from transcripts generated from PCR amplicons. The k-mers are detected as mass peaks by MS (49). The number of the mass peaks reflects the number of different k-mers in the amplicon mixture. There are 2 sources of k-mer heterogeneity: (1) k-mers derived from a single PCR target-molecule and (2) k-mers created by nucleotide variations from different molecules. Genetic diversity is associated with the latter source. Thus, the number of k-mers can be used as a measure of the intrahost viral diversity represented in the PCR amplicons. Hence, the number of mass peaks derived from PCR fragments obtained from S and M samples should be different. Indeed, the mean number of peaks was significantly higher (P = .0001, by the multiresponse permutation procedure) in the M samples (mean, 124.31) than in the S samples (mean, 86.36), indicating that viral genetic diversity is reflected in the MS-peak profiles (Figure 2).

**Intersample MS-Peak Profile Distances**

To assess the usefulness of MS-peak profiles for estimating genetic relatedness among HCV strains, we tested 60 serum samples (OB samples) obtained from 14 HCV outbreaks that occurred in the United States during 2003–2008 [32–37]. Additionally, 25 serum samples (NR samples) were selected from cases of infection with HCV subtypes 1a and 1b that were epidemiologically not related to the outbreaks (SD1). On average, 9 replicas (range, 6–12 replicas) of HVR1 were amplified by PCR from each of the specimens (SD2).

The Hamming distance between MS-peak profiles from the OB and NR samples was used to measure genetic relatedness among HCV strains. The matrix of distances among all 793 PCR replicas obtained from samples of the 85 cases allowed
for a qualitative identification of each outbreak (Figure 3). In
the OB group, interspecimen MS-peak profile distances ob-
tained for cases in the same outbreak were similar to distances
among PCR replicas and were smaller than among the NR
group (Figure 4). The frequency distribution of pairwise dis-
tances among samples (Figure 4) shows that the NR MS-peak
profiles are bimodal, reflecting distances within the subtypes
and between different subtypes. The use of the minimum dis-
tances between PCR replicas from any pair of specimens
allowed for a greater, albeit incomplete, separation between
distributions of intraoutbreak distances in the OB group and
distances in the NR group (Figure 4B).

Comparison With Sequence-Based Analysis
The agreement between sequence- and MS-peak profile-based
measurements of HVR1 nucleotide-sequence diversity was low
but significant (Pearson correlation, \( r = 0.2637; \ P < .0188 \)).
In contrast, the correlation between the sequence-based and MS-
peak profile distances was very high (Mantel test, \( r = 0.8361; \ P < .0001 \)). Scatterplots of sequence-based [45] and MS-peak
profile-based distances (Figure 5) show clear differentiation
among pairwise comparisons of OB samples from each out-
break, NR samples that carried HCV from the same subtype,
and NR samples carrying HCV from different subtypes.

Neighbor-joining trees [44] were built using either se-
quence- or MS-peak profile-based distances. Both MS-peak
profile- and sequence-based trees clustered correctly all speci-
mens that were epidemiologically linked (Figure 6). The MS-
peak profile–based clustering completely agreed with the se-
quence-based clustering and entirely resolved the overlap in
the distance frequencies observed in Figure 4B, as none of the
OB samples were found to group with NR cases.

AMOVA was used to quantify the level of differentiation
between the OB and NR samples [46]. AMOVA partitioned
the variance of the MS-peak profile– or sequence-based dis-
tances into 3 sources: (1) among the replicas of each speci-
men, (2) among OB specimens within an outbreak, and (3)
among NR specimens (Figure 6C). It was found that 60.8% of
all variance in the MS-peak profile distances was due to signif-
icant differences among the nonrelated samples from the OB
groups and samples from the NR groups (\( P < .0001 \)). The se-
quence-based method showed higher differences among non-
related cases (73.73%), as can be observed in the smaller
distances between outbreaks in the MS-peak profile similarity
tree (Figure 6B). It also showed lower differences among the
cases belonging to the same outbreak, which is reflected in the
tight clusters in the sequence-based tree (Figure 6A).

Graphical Linkages of Related Cases
Visualization of genetic relatedness by use of tree models
allows for qualitative clustering of HCV variants. However,
genetic detection of transmissions does not require assessment
of the degree of genetic differences among strains beyond es-
ablishing their genetic equivalence. Genetically related and
unrelated HCV strains can be distinguished using a distance
threshold. Analysis of the 300 pairwise MS-peak profile–based
distances among the NR samples established 0.0506 as the
minimal distance (Figure 4B and 4C), which was therefore es-
ablished as threshold. Figure 7 shows the links among cases
with a distance below the threshold. This simple method cor-
rectly identified all the samples involved in individual out-
breaks, indicating 100% sensitivity in the genetic detection of
transmissions. There were 1553 pairwise distances among cases
belonging to different outbreaks, all of which were below
the threshold. When the 1500 pairwise distances between OB
and NR cases were considered, only 1 NR case had a MS-peak
profile-distance to 1 OB case below the threshold, indicating
99.93% specificity of transmission detection.

DISCUSSION
Transmission is a fundamental viral property, knowledge of
which is essential for understanding dissemination of infection
and disease. As such, transmission detection is key to the sur-
veillance of infectious diseases. However, molecular detection
of transmissions is a very complex task. It involves not only
identification of the virus, but, for virus that is highly mutable,
also assessment of the genetic association among its variants.
The major assumption of genome-based detection of transmissions is that the genetic composition of the viral strain that has been passed from one patient to another remains approximately “the same.” Thus, the task of genome-based detection of transmission can be reduced to the evaluation of the genetic identity of viral strains without consideration of genetic relatedness beyond genetic equivalence. For HCV, this limited evaluation of genetic relatedness among variants can be best achieved by analyzing a sample of sequences obtained from the intrahost viral population, using short genomic regions [31, 32]. Consensus sequences of the regions can inadequately represent viral variants involved in transmission [30], especially when minority variants from the source establish the new infections in the recipients [36].

In this study, we explored the use of MS-peak profiles for identification of HCV transmission. MS-peak profiles generated using base-specific cleavage of RNA transcripts derived from PCR fragments [42, 43] are a very rich source of information about the nucleotide sequence and structure of the intrahost viral population. The approach used here detects a pattern of short RNA fragments, or k-mers. The k-mer structure of the MS data closely reflects the heterogeneity of viral populations. The number and variety of k-mers that can be derived from a short amplicon obtained from a single sequence variant is limited. Hence, mutations within the amplified region should generate additional k-mers that will likely result in independent MS peaks. This supposition is supported by the observation that the mean number of peaks was significantly higher in the M samples than in the S samples (Figure 2). Therefore, the genetic composition of the intrahost HCV population significantly contributes to MS-peak profile complexity. The genetic distances estimated using MS-peak profiles should be distinctly affected by sequence heterogeneity and composition of intrahost HCV populations and by their representation in the PCR products. Indeed, the sequence-based method showed average differences that were lower than MS-peak profile-based distances among cases belonging to the same outbreak, as shown by AMOVA and as evident from

Figure 3. Heat map of the mass spectrometry–peak profile distance matrix. Distances between all subsamples (n = 793; average, 9.3 per patient) from OB (n = 60) and NR patients (n = 25) were plotted. The diagonal represents the comparison of each sample to itself. The color scale represents observed range of distances, for which colder colors signify shorter distances and warmer colors connote longer distances. Samples are grouped by specific outbreaks, denoted as capital letters (SD1) on both axes. Detailed information about the exact number of repeats per individual patient can be found in SD7. See Materials and Methods for descriptions of the OB and NR groups.
comparing the sequence-based tree with the MS-based tree (Figure 6A and 6B).

Since the k-mers are derived from numerous RNA molecules representing the heterogeneous intrahost HCV population, an MS-peak profile reflects the primary structure of the genomic region and nucleotide variations among HCV variants in this region, thus combining all necessary genetic information for accurate molecular detection of HCV transmissions. Indeed, the MS-peak profiles obtained using the...
HCV HVR1-containing PCR fragments were shown here to be efficient in establishing genetic equivalence among HCV variants involved in outbreaks. A small variation among PCR replicas from a single clone template (Figure 2) indicates significant reproducibility of the MS-peak profiles produced using the applied MS technology.

Owing to the stochastic nature of PCR, complex mixtures of the intrahost HCV variants may be inadequately represented in a single PCR product. The observed variation of peak numbers in the MS-peak profile among repeat PCR products obtained from mixtures of HVR1 clones (Figure 2) strongly supports this suggestion. The disproportionate representation of certain sequences in amplified fragments should result in a measurable variation of the k-mer composition in the MS-peak profile. Such deviations may potentially contribute to false-negative results, thus reducing sensitivity of the transmission detection. The data obtained in this study show that a more adequate representation of the intrahost HCV variants can be achieved using several PCR replicas derived from each test specimen.

Analyses conducted here reveal a fundamental similarity between MS-peak profile– and sequence-based distances for the genetic detection of transmissions (Figure 6). The separation between MS-peak profile–distances among genetically related and unrelated cases is clear and permits HCV variants involved in outbreaks to be discriminated from all nonoutbreak variants (Figure 4). Such genome-based detection of transmission does not require complete assessment of phylogenetic relationships among HCV variants and, therefore, can be achieved using a threshold and visualized using simple linkage graphs (Figure 7).

In conclusion, MS-peak profile–based detection of HCV transmissions as described here has been found to match the accuracy of sequence-based approaches. Moreover, whereas next-generation sequencing technologies require an extensive data preprocessing for identification and repair of sequence errors [40], the data generated from MS-peak profiles can be analyzed immediately upon acquisition. Because the approach developed in this study is more facile, less time-consuming, and considerably less costly than conventional sequencing, it is applicable to other pathogens and suitable for use in laboratories involved in routine surveillance of infectious diseases.

Notes

Financial support. This work was supported by intramural funding from the Centers for Disease Control and Prevention.

Potential conflicts of interest. C. C. is an officer and holds equity in Sequenom, a commercial entity that manufactures the platform used in this study. All other authors report no potential conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References