Hepatitis C Virus Genotyping: Clinical Implications and Methods

FREDERICK S. NOLTE, PhD
Atlanta, Georgia

Hepatitis C virus (HCV) chronically infects at least 1% of the world's population and is a leading cause of end-stage liver disease. HCV displays a remarkable degree of genomic diversity, with the six major genotypes and numerous subtypes differing in geographic distribution. The ability of this virus to cause persistent infections is a direct result of its genomic plasticity and the evolution of quasispecies within an infected individual. HCV genotype has emerged as an important factor both in predicting a sustained response to, and in determining the duration of, antiviral therapy. Although a variety of methods have been used for genotyping HCV, nucleotide sequencing of a phylogenetically informative region remains the gold standard.

Key words: hepatitis C virus, genotyping, nucleic acid sequencing.

Hepatitis C virus (HCV) was discovered in 1989 by the use of molecular cloning techniques through the joint efforts of the Centers for Disease Control and Prevention (CDC) and the Chiron Corporation [1]. HCV chronically infects at least 1% of the world's population. Chronically infected individuals are at increased risk for developing liver cirrhosis and hepatocellular carcinoma. In the United States, end-stage liver disease caused by chronic HCV infection is the leading reason for liver transplantation.

The best available therapy for HCV, interferon (IFN) α in combination with ribavirin, is not highly effective. The sustained virologic response rates in treated patients range from 30% to 70% and depend on several important clinical and virologic factors [2,3]. HCV genotype is among the factors that influence response to therapy.

HCV is a positive sense single-stranded RNA virus that is most similar to viruses belonging to the genera Pestivirus and Flavivirus. Shortly after its discovery, it became clear that the virus showed substantial nucleotide sequence heterogeneity throughout the viral genome [4]. This genomic heterogeneity may account for some of the differences in disease outcome and response to therapy observed in HCV-infected individuals. Clearly, the heterogeneity among isolates has complicated the development of diagnostic tests and vaccines.

The clinical implications of HCV genomic heterogeneity and the different HCV genotyping methods are described in this review. In addition, the relative strengths and limitations of the genotyping methods are addressed, and their relevance to the treatment of patients with HCV infection is highlighted. The methods for HCV genotyping and the clinical significance of HCV genotypes have been the subjects of two previous review articles [5,6].
Genomic Organization and Diversity

HCV is an RNA virus with a positive sense, single-stranded genome of approximately 9,400 nucleotides (nt) encoding a single polyprotein of approximately 3,000 amino acids (aa). The long open reading frame (ORF) is flanked at each end by a short untranslated region (UTR). The genome structure is most similar to viruses of the family Flaviviridae, which includes many of the arthropod-borne viruses. As in other flaviviruses, the three N-terminal proteins of HCV (core, envelope 1, and envelope 2) are probably structural, and the four C-terminal proteins (nonstructural 2, 3, 4, and 5) are thought to function in viral replication (Fig. 1).

The 5' UTR is a highly conserved region of 341 nt and has a complex secondary structure. It contains an internal ribosome entry site and presumably is important in the translation of the long ORF. The 3' UTR contains a short region that varies in sequence and length, followed by a polypyrimidine stretch of variable length, and finally a highly conserved sequence of 98 nt, which constitutes the terminus of the genome. The function of the 3' UTR is not known, but it is thought to be essential for viral replication.

The envelope 1 (E1) and 2 (E2) regions of HCV are the most variable regions within the genome at both the nt and aa levels. A 27 aa hypervariable region of the E2 protein (HVR1) has been described that may be the mechanism by which the virus evades the host’s immune system.

The nonstructural regions 2 (NS2) and 3 (NS3) contain a Zn-dependent protease that cleaves the polyprotein at the NS2–NS3 junction. The amino-terminal portion of the NS3 protein also has protease activity and cleaves the polyprotein at several sites. The carboxyterminal portion of the NS3 protein has helicase activity, which is important for HCV replication. The NS4A protein is a cofactor for NS3 protease activity. The NS5B region encodes the RNA-dependent RNA polymerase, which replicates the viral genome [7].

Choo et al. [8] determined the first complete HCV genome sequence in 1991. As additional genome sequences from isolates from different parts of the world were determined and compared, it was evident that HCV exists as distinct genotypes with as much as 35% sequence diversity over the whole viral genome [9]. Much of the early literature on genotyping is confusing because investigators developed and used their own classification schemes. However, a consensus nomenclature system was developed in 1994. In this system, the genotypes are numbered using Arabic numerals in order of their discovery, and the more closely related strains within some types are designated as subtypes with lowercase letters. The complex of genetic variants found within an individual isolate is termed the quasispecies. The quasispecies results from the accumulation of mutations that occur during viral replication in the host. The terminology and degree of nucleotide similarity that define the relationships of HCV variants are shown in Table 1.

Six major genotypes were originally identified. Sequence analysis of the E1 region suggested that HCV could be grouped into 6 major genotypes and 12 subtypes [10]. The same investigators sequenced 573 nt of the core region of the same isolates to confirm this classification scheme [11]. Simmonds et al. [12] were also able to classify HCV isolates into the same six major genotypes and numerous subtypes using sequence analysis of the NS5B region. Analyses of full-length ORF sequences have

![Fig. 1. Organization of the HCV genome.](image_url)

### Table 1. HCV Genomic Heterogeneity

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Nucleotide Similarity*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
<td>Heterogeneity among different viruses</td>
<td>66% to 69%</td>
</tr>
<tr>
<td>Subtype</td>
<td>Closely related viruses within each genotype</td>
<td>77% to 80%</td>
</tr>
<tr>
<td>Quasispecies</td>
<td>Complex of genetic variants within individual viruses</td>
<td>91% to 99%</td>
</tr>
</tbody>
</table>

Data from Zein NN [6].

*Full-length genome sequence identity.