Viral heterogeneity of the hepatitis C virus

Peter Simmonds
Department of Medical Microbiology, University of Edinburgh, Edinburgh, UK

This review summarises the classification of hepatitis C virus as a flavivirus, the identification and detection of HCV genotypes, and reviews the current information concerning the geographical and risk group associations of the common genotypes in Europe.

Key words: Viral heterogeneity.

Classification of hepatitis C as a flavivirus

Hepatitis C (HCV) has a positive sense RNA genome of approximately 9400 bases. This codes for a single polyprotein that is cleaved into structural proteins (capsid, envelope glycoproteins) and six non-structural proteins (NS2-NS5B) (1, 2). These features are shared with the vector-borne flaviviruses such as yellow fever virus and dengue viruses. However, HCV has an even closer relationship with pestiviruses, such as bovine viral diarrhoea virus, with which it shares a similarly structured internal ribosomal entry site in the 5′ non-coding region (5′NCR).

Amino acid sequence similarity between HCV, pestiviruses and flaviviruses has been detected in the region surrounding the active site of the RNA polymerase (3). Phylogenetic analysis of this region provides a useful means to visualise the relationship between HCV and its relatives (Fig. 1). GB virus B (GBV-B), a hepatotropic virus that infects new world primates is the most closely related to HCV, and shares a 5′NCR structure similar to that of HCV and pestiviruses. More distant is hepatitis G virus or GB virus C (PNF2161) that infects humans (but which is not associated with hepatitis) and homologous viruses in chimpanzees and new world primates (GBV-C, GBV-A, SM 70047 and AT 1122).

Based on these sequence relationships, the International Committee on Taxonomy of Viruses (ICTV) has classified HCV as member of the flaviviridae. They have further proposed that HCV, along with GBV-B, GBV-A and HGV/GBV-C, constitute a genus within the flaviviridae named “hepacivirus” (the other two genera remain as pestivirus and flavivirus) (4). While this is a useful interim proposal, it may be subject to future revision for two reasons. First, the hepacivirus genus includes GBV-A and HGV/GBV-C, which do not appear to code for a capsid protein, and have a 5′NCR, which produces a differently structured IRES resembling that of picornaviruses. Conversely, pestiviruses resemble HCV in genome structure, coding and IRES structure, but are placed into a separate genus. Another problem with the proposal is the name “hepacivirus”. This name was chosen to indicate the tropism of viruses in the genus for the liver, but this has not been demonstrated and is unlikely to be true for HGV/GBV-C or GBV-A.

A forum for discussion and contributions to these recommendations is available at the site http://s2as02.genes.nig.ac.jp.

Identification of HCV genotypes

The genetic heterogeneity of HCV was originally described when the sequences of HCV variants from Japan were compared with the prototype HCV sequence (HCV-PT) cloned from a chimpanzee experimentally infected with the non-A non-B agent (1). While the complete genome sequences of HCV-J (2) and -BK (5) from Japan were 92% similar, they showed less than 80% similarity with HCV-PT. Far more divergent variants of HCV have since been found in Japan (6, 7) and elsewhere (8, 9), leading to the adoption of an extended classification of HCV into types and subtypes. Using sequence comparisons as a method of virus classification, currently known variants of HCV collected from different parts of the world can be divided into six main “genotypes”, many of which contain more closely related variants (Fig. 2). For nomenclature, we
proposed that HCV be classified into types, corresponding to the main branches in the phylogenetic tree, and subtypes corresponding to the more closely related sequences within some of the major groups (10, 11). The types have been numbered 1 to 6, and the subtypes α, β and γ, in both cases in order of discovery. Therefore, the sequence cloned by Chiron is assigned type 1α, HCV-J and -BK are 1β, HC-J6 is type 2α and HC-J8 is 2β.

The variability of HCV is distributed throughout the genome with sequences coding for the non-structural proteins such as proteinase (NS3) and RNA polymerase (NS5B) showing comparable degrees of variability to the mean values over the complete genome. However, the core gene and particularly the non-coding regions at the ends of the genome are more conserved; the lack of variability of the 5′NCR makes this region the most suitable for virus detection by PCR (12, 13). The genes coding for the E1 and E2 glycoproteins are more variable, with E2 containing one or two “hypervariable regions” (HVR) of 20–30 amino acids that vary considerably not only between genotypes, but also within the same subtype. For example, epidemiologically unlinked variants of type 1b differ on average by 13.2 amino acids over a region of 27 residues (49% divergence) compared with 9% divergence over the rest of the genome (14). Even amongst individuals infected 17 years from a common source of HCV, over 36% divergence develops (14). As the envelope proteins are the likely targets of the humoral immune response to HCV infection, it has been suggested that the rapid rate of sequence change in HVR is driven by the immune system. Amino acid changes may alter the antigenic properties of the proteins, and allow the virus “escape” from neutralising antibodies (15–17).

Even though complete genomic sequences of each of the six major genotypes have been used for identification of genotypes, relatively short subgenomic regions, such as E1 (18) or NS5B (19) can also be used for reliable
Fig. 2. Sequence relationships between genotypes 1 to 6 using comparisons of currently available complete genomic sequences (listed in ref. 47; tree generated using the programs DNADIST and NEIGHBOR in the PHYLIP package; (48)). Nomenclature of HCV variants follows the consensus proposal for classification of HCV (11), i.e. the sequence cloned by Chiron is assigned type 1a, HCV-J and -BK are 1b, HC-J6 is type 2a and HC-J8 is 2b. The classification of JK049, which groups with type 3 variants, and D84262-5 and JK046, which group with type 6a, is discussed in the text.

Fig. 3. Phylogenetic analysis of nucleotide sequences from part of the HCV NSSB region amplified from samples of HCV-infected blood donors and hepatitis patients from several countries. Six main groups of sequence variants are found corresponding to types 1-6; each group contains a number of more closely related subtypes. Several variants from SE Asia cluster with type 6a but are labelled according to the original publications (21-23, 49) (see text).

HCV variants appears to me unlikely to resolve the classification issue described above. It also seems unnecessary for the description of the common variants of HCV, while the use of the word “clade” may cause confusion to those accustomed to the current literature on HCV genotypes.

HCV genotyping assays

Although essential for genotype discovery, nucleotide sequence determination is not currently practical for the routine laboratory identification of HCV genotypes. Many of the published methods for genotype identification are based upon the amplification of virus sequences in clinical specimens by polymerase chain reaction (PCR; Table 1). One commonly used method is a modified PCR that uses several sets of primers, each of which selectively amplifies sequences from different HCV genotypes. For example, DNA amplified by primers specific for the core region of type 1a are a different size from those produced from the amplification of types 1b, 2a, 2b and 3a and can therefore be differentiated by agarose gel electrophoresis of the PCR product (28-30). Other assays, such as the Innolipa assay (Innogenetics n.v. Gent, Belgium), are based upon the analysis of amplified DNA sequences. For example, the genotype of an amplified sequence
can be determined by hybridisation to genotype-specific probes (31–33) or through restriction fragment length polymorphisms in sequences amplified from the 5′NCR (34).

Serological typing methods are based upon the detection of type-specific antibody to epitopes of HCV that differ between genotypes (19, 35, 36). They have advantages over PCR-based methods in terms of speed and simplicity of sample preparation and in the use of simple equipment that would be found in any diagnostic virology laboratory. By careful optimisation of reagents, such assays may show high sensitivity and reproducibility. For example, type-specific antibody to NS4 peptides of genotypes 1 to 6 can be detected in approximately 85–90% of NANB patients (19) using the HC1–6 serotyping assay (Murex Biotech, Dartford, UK). However, antigenic similarity currently precludes the separate identification of subtypes such as 1a and 1b or 2a and 2b using NS4 peptides, or antigens from other parts of the genome.

A crucial assumption of all genotyping assays is that the region analysed (5′NCR, core, NS4, NS5B) is representative of the genome as a whole. This assumption would break down if recombination between HCV genotypes occurred during virus replication, so producing hybrid viruses containing contributions from different genotypes in different parts of the genome (37). Consistent results have been obtained by comparing results of genotyping assays based upon sequence analysis of different regions of the genome, and between PCR-based and serological typing assays (19, 38–42). This concordance provides no evidence for frequent recombination between genotypes of HCV. The comparisons demonstrate that currently available genotyping assays reliably identify infection with those HCV genotypes likely to be encountered in clinical practice in Europe, other Western countries and Japan (la, lb, 2, 3a and 4).

**HCV origins and epidemiology**

The RNA genome of HCV is replicated using a virally encoded RNA polymerase. RNA copying mechanisms lack the proofreading activity of DNA polymerase complexes and mutations arising from nucleotide misincor-

---

**TABLE 1**

Comparison of typing methods for HCV

<table>
<thead>
<tr>
<th>Method</th>
<th>Region</th>
<th>Genotypes detected*</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PCR based methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type-specific primers</td>
<td>Core</td>
<td>1a, 1b, 2a, 2b, 3a</td>
<td>(28, 29)</td>
</tr>
<tr>
<td></td>
<td>NS5</td>
<td>1a, 1b, 2a, 2b, 3b</td>
<td>(52)</td>
</tr>
<tr>
<td>RFLP</td>
<td>5′NCR</td>
<td>1*, 2a, 2b</td>
<td>(53)</td>
</tr>
<tr>
<td></td>
<td>S′NCR</td>
<td>1*, 2a, 2b, 3, 4, 5, 6</td>
<td>(34, 54)</td>
</tr>
<tr>
<td>Type-specific probes</td>
<td>NS5</td>
<td>1a, 1b, 2a, 2b</td>
<td>(10)</td>
</tr>
<tr>
<td></td>
<td>NS5</td>
<td>1a, 1b, 2a, 2b, 2c, 3a, 3b, 4, 5, 6</td>
<td>(30)</td>
</tr>
<tr>
<td></td>
<td>5′NCR</td>
<td>1*, 2a, 2b, 3, 4, 5</td>
<td>(35)</td>
</tr>
<tr>
<td></td>
<td>S′NCR</td>
<td>1*, 2, 3</td>
<td>(32)</td>
</tr>
<tr>
<td></td>
<td>Core</td>
<td>1a, 1b, 2a, 2b, 3a</td>
<td>(33)</td>
</tr>
<tr>
<td><strong>Serological methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peptide ELISA</td>
<td>NS4</td>
<td>1, 2, 3</td>
<td>(38)</td>
</tr>
<tr>
<td></td>
<td>NS4</td>
<td>1, 2, 3, 4, 5, 6</td>
<td>(19)</td>
</tr>
<tr>
<td></td>
<td>Core</td>
<td>1, 2</td>
<td>(35)</td>
</tr>
<tr>
<td></td>
<td>NS4/core</td>
<td>1, 2, 3</td>
<td>(56, 57)</td>
</tr>
<tr>
<td>Recombinant protein ELISA</td>
<td>NS4-4</td>
<td>1, 2</td>
<td>(58)</td>
</tr>
</tbody>
</table>

*Sequences of type 1a and 1b may be identical in the S′NCR, so in principle these two subtypes cannot be separately identified by any method based upon this region (see text).”

**TABLE 2**

Epidemiology of HCV genotypes in Europe

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Principal location</th>
<th>Typical age range</th>
<th>Risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1a</td>
<td>Northern Europe, (USA)</td>
<td>&lt;50</td>
<td>IVDUs, unknown</td>
</tr>
<tr>
<td>Type 1b</td>
<td>Global, frequent S. Europe</td>
<td>&gt;50</td>
<td>Unknown, PPE*</td>
</tr>
<tr>
<td>Type 2*</td>
<td>Italy, North Africa, Spain</td>
<td>&gt;50</td>
<td>Unknown, PPE</td>
</tr>
<tr>
<td>Type 3a</td>
<td>Northern Europe</td>
<td>&lt;50</td>
<td>IVDUs</td>
</tr>
<tr>
<td>Type 4a</td>
<td>Egypt, Middle East</td>
<td>all</td>
<td>Unknown, PPE?</td>
</tr>
</tbody>
</table>

* PPE: Past parenteral exposure.
† Type 2b in Northern Europe, types 2a and 2c in Southern Europe; type 2c prevalent in Italy.
portion during copying are not corrected. The high rate of sequence change of HCV observed experimentally is consistent with its presumed mechanism of replication, and confers upon HCV a considerable adaptive ability. However, most nucleotide substitutions that accumulate over time occur at synonymous sites (i.e. those do not change the sequence of the encoded proteins), presumably because their occurrence is on average less deleterious to the viral phenotype than those that produce amino acid changes. The accumulation of synonymous substitutions provides a useful "molecular clock" that provides estimates of times of divergence of HCV variants within a population. Determining the rate with which synonymous substitutions occur has been carried out by sequence comparisons over a period of several years of variants recovered from chronically infected humans or chimpanzees (43, 44). More recently we have determined the nucleotide sequence of parts of the E1 and NS2B genes of HCV variants infecting women originally infected through exposure to a single batch of HCV-contaminated anti-D immunoglobulin (Fig. 4) (45). As the source of infection was relatively homogeneous in sequence and samples from a large number of anti-D recipients were analysed 17 years after infection, an accurate estimate of the synonymous substitution rate was possible. The rate of 1.4–2.0×10⁻³ substitutions per site per year (in NS2B and E1 respectively) is within the range expected from the estimated error rate of RNA polymerases and the generation time of HCV during chronic infection, and is comparable to that measured in other RNA viruses (e.g. HGV/GBV-C (46)).

The inclusion of epidemiologically unlinked sequences of type 1b originating from Japan, Europe and the USA revealed an approximately four-fold greater degree of sequence divergence at synonymous sites than observed amongst variants from the Irish anti-D exposure. From this it might be inferred that the current worldwide distribution of type 1b originated from a common ancestor around 76 years ago, clearly an example of global epidemic spread. Using the calibrated molecular clock, the spread of other genotypes may have been even more dramatic and recent. For example, type 1a and 3a variants from Europe show mean divergences at synonymous sites of 0.163 and 0.154 in NS5B, predicting a common ancestor around 50 years ago. In contrast, type 2c variants are highly divergent in sequence from each other and may have been originated and been maintained in certain populations for at least 150 to 200 years.

There is a remarkable concordance between population diversity of the common HCV genotypes in Europe with their epidemiology. Different genotypes vary markedly in their association with specific risk factors, such as intravenous drug abuse, the age distribution of infected patients and frequencies of occurrence in different countries (Table 2). There is a particular con-
Contrast between types 1b and 2, which tend to infect individuals in an age range of 50–70 and which are predominantly distributed in Eastern (type 1b), and Southern Europe (type 1b and subtypes of type 2) with types 1a and 3a, which infect a younger population and are principally associated with intravenous drug abuse (Table 2); their predominant distribution in Northern Europe possibly reflects the absence of the older substratum of HCV-infected individuals found elsewhere in Europe. The restricted variability of type 1a and 3a is consistent with their spread in a different risk group from those likely to have been infected earlier with type 1b and type 2 by different routes of transmission.

If the rate of sequence change of HCV was maintained over long periods, it might be possible to calculate the times of divergence of different subtypes (e.g. between type 1a and 1b), or even between the genotypes of HCV (e.g. between genotype 1 and 2). Unfortunately, the extensive sequence differences make evolutionary reconstruction difficult; in particular it is difficult to estimate sequence distances between highly divergent variants because of the difficulty in estimating the frequency of unseen multiple substitution at individual nucleotide sites. Minimum estimates range from 600 years for the time of divergence of HCV subtypes to 1000–2000 years for the divergence of the six HCV genotypes; these are likely to be considerable underestimates but provide at least a preliminary indication of the great time depth of the current HCV population. How HCV could have been maintained in human populations for such lengths of time without frequent parenteral exposure is unclear. Similarly, the absence of any physical contact between human populations in different continents until relatively recently makes the current geographical distribution of HCV difficult to explain. Without long range travel there would seem to be no mechanism by which the common ancestor of all HCV variants could have been disseminated into all human populations and differentiated into region-specific genotypes over a period of 2000 years.

References
3. Miller RH, Purcell RH. Hepatitis C virus shares amino acid sequence similarity with pestiviruses and flaviviruses as well as members of two plant virus supergroups. Proc Natl Acad Sci USA 1990: 87: 2057–2061.
23. Tokita H, Okamoto H, Luengrojanakul R, Varesangthip K, Chainuvatti T, Lizuka H et al. Hepatitis C virus variants from Thailand classifiable into five novel genotypes in the sixth (6b), Viral heterogeneity
seventh (7c, 7d) and ninth (9b, 9c) major genetic groups. J Gen Virol 1995: 76: 2329–2335.


