Intensive temporal mapping of hepatitis C hypervariable region 1 quasispecies provides novel insights into hepatitis C virus evolution in chronic infection

Daniel Schmidt-Martin,1 Orla Crosbie,2 Elizabeth Kenny-Walsh2 and Liam J. Fanning1

Correspondence
Daniel Schmidt-Martin
danscma@yahoo.com

1Molecular Virology Diagnostic and Research Laboratory, Department of Medicine, National University of Ireland Cork, Clinical Sciences Building, Cork University Hospital, Cork, Ireland
2Department of Gastroenterology, Cork University Hospital, Cork, Ireland

Hepatitis C virus (HCV) is an RNA virus which exists as swarms of closely related viruses known as quasispecies (QS). A number of studies have demonstrated associations between QS hypervariable region 1 (HVR1) characteristics (diversity and complexity) and treatment success. We investigated HCV QS change in chronic infection over intervals of 2–4 weeks in 23 chronically infected individuals to describe the natural history of virus evolution and establish whether HCV QS characteristics could be used to individualize treatment regimens at a molecular level. HVR1 QS diversity, complexity and divergence continue to change in an unpredictable fashion in chronic infection even where there is little phylogenetic change, which is likely to preclude the use of these features in treatment individualization. Our phylogenetic analysis identified no change in the HVR1 QS in 12 subjects, minor change in four subjects and we describe a time-ordered phylogeny for the first time over a period as short as 16 weeks in seven subjects. We identified the existence of multiple subpopulation infections using partitioned analysis of QS and illustrated how subpopulations were sequentially replaced in a number of subjects. We illustrated marked variation in the nucleotide substitution per codon position between patients with sequence change and those without change in the phylogenetic tree. Analysis of codon-specific selection pressures identified a number of codons under purifying selection, suggesting that these code for structurally conserved amino acids. We also identified sections of the HVR1 under positive selection with marked sequence heterogeneity, suggesting that these may be potential epitope-binding sites.

INTRODUCTION

Hepatitis C virus (HCV) is a positive-stranded RNA virus of the family Flaviviridae and is estimated to chronically infect 120–170 million people worldwide (Alter, 2007; Pockros, 2002). Associated with significant morbidity and mortality, chronic hepatitis C is the leading indication for orthoptic liver transplantation in the USA. The cost of managing hepatitis C over the past decade has been estimated at $10 billion, but the emergence of new agents has resulted in the drug budget for HCV rising 10-fold to $1 billion per month (Szabo´ et al., 2003; Wong et al., 2000).

The HCV RNA polymerase lacks a proofreading mechanism, which due to high levels of mutation generates a collection of related genomes known as quasispecies (QS) (Steinhauer et al., 1992). A number of studies investigating hypervariable region 1 (HVR1), a 27 aa segment encoding part of the virus envelope, have demonstrated an association between low HVR1 QS complexity, or high QS diversity, and sustained virologic response (SVR) with dual pegylated IFN and ribavirin treatment. Although HVR1 has been extensively studied in a number of differing conditions [acute HCV infection (Chen & Wang, 2002; Farci et al., 2000; Kuntzen et al., 2007; Liu et al., 2010), human immunodeficiency virus (HIV) co-infection (Mao et al., 2001; Shuhart et al., 2006), alcoholic liver disease (Sherman et al., 1999), chronic infection (Brambilla et al., 1998; Brown et al., 2005; Cabot et al., 2001; Kumagai et al., 2007; McAllister et al., 1998; Qin et al., 2005; Ray et al., 2005; Sullivan et al., 2007; Wang et al., 2007), during treatment (Abbate et al., 2004; Grahovac et al., 2000) and...
following transplantation (Doughty et al., 2000; Lyra et al., 2002; Pessoa et al., 1999; Sánchez-Fueyo et al., 2001), little is known of the natural history of HVR1 QS evolution over short time intervals in subjects with chronic infection. Studies investigating HVR1 in chronic infection contain intervals of years if not decades between samples (Bailey et al., 2012; McAllister et al., 1998; Palmer et al., 2012; Ray et al., 2005), with the notable exception of Brambilla et al. (1998) who studied samples in three subjects on a six-monthly basis. It has been suggested, however, that the mean number of clones per sample may not have been sufficient for robust analysis of QS complexity (Fishman & Branch, 2009).

The emergence of triple therapy with protease inhibitors and the development of highly efficacious direct-acting antiviral drugs has resulted in a paradigm shift in expected treatment outcomes for those in a position to meet the high cost of these medications. Most patients with HCV, however, live in countries where the cost associated with treatment is likely to preclude access. We investigated HVR1 QS change in chronically infected subjects over shorter, standardized periods than had previously been undertaken in order to give a clearer picture of mechanisms of HVR1 QS change and to establish whether treatment could potentially be individualized based on HVR1 QS dynamics.

RESULTS

Study sample

We prospectively studied 23 chronically infected subjects aged 20–75 years with either genotype 1b (n=6) or 3a (n=17), six of whom had evidence of cirrhosis on liver biopsy (n=5 genotype 1b, n=1 with genotype 3a) (Table 1). There was no history of current or previous alcohol excess, current intravenous drug use and no evidence of co-infection with either hepatitis B or HIV, and all subjects met criteria for chronic infection (two PCR-positive serum samples a minimum of 6 months apart). For clarity, we include figureative illustrations using five of the subjects who were representative of the patterns of QS change we identified. The corresponding figures for the remaining 18 subjects are included in the online Supplementary Material.

Cloning results

In total, 3057 clones, a mean of 18 clones per sample (range 10–24), were successfully sequenced and analysed, once defective sequences had been eliminated. The clones analysed for each individual were monophyletic and mapped to the equivalent genotypes when reference sequences from the Los Alamos National Laboratory database were included (data not shown), indicating no cross-contamination of samples or undiagnosed mixed genotype infection.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sex</th>
<th>Country</th>
<th>Age (years)</th>
<th>Genotype</th>
<th>Mode of transmission</th>
<th>Cirrhosis</th>
<th>Samples</th>
<th>Total clones</th>
<th>Clones per sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>M</td>
<td>Ireland</td>
<td>75</td>
<td>1b</td>
<td>Blood transfusion</td>
<td>Yes</td>
<td>7</td>
<td>144</td>
<td>20.6</td>
</tr>
<tr>
<td>B</td>
<td>M</td>
<td>New Zealand</td>
<td>49</td>
<td>1b</td>
<td>Intravenous drug user</td>
<td>Yes</td>
<td>6</td>
<td>91</td>
<td>15.2</td>
</tr>
<tr>
<td>C</td>
<td>F</td>
<td>Ireland</td>
<td>59</td>
<td>1b</td>
<td>Blood product</td>
<td>Yes</td>
<td>5</td>
<td>88</td>
<td>17.6</td>
</tr>
<tr>
<td>D</td>
<td>M</td>
<td>Lithuania</td>
<td>61</td>
<td>1b</td>
<td>Unknown</td>
<td>Yes</td>
<td>8</td>
<td>121</td>
<td>15.1</td>
</tr>
<tr>
<td>E</td>
<td>M</td>
<td>Ireland</td>
<td>48</td>
<td>1b</td>
<td>Unknown</td>
<td>Yes</td>
<td>7</td>
<td>117</td>
<td>16.7</td>
</tr>
<tr>
<td>F</td>
<td>M</td>
<td>Australia</td>
<td>43</td>
<td>1b</td>
<td>Unknown</td>
<td>No</td>
<td>8</td>
<td>125</td>
<td>15.6</td>
</tr>
<tr>
<td>G</td>
<td>F</td>
<td>Latvia</td>
<td>27</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>8</td>
<td>154</td>
<td>19.3</td>
</tr>
<tr>
<td>H</td>
<td>M</td>
<td>Pakistan</td>
<td>40</td>
<td>3a</td>
<td>Iatrogenic</td>
<td>No</td>
<td>9</td>
<td>156</td>
<td>17.3</td>
</tr>
<tr>
<td>I</td>
<td>M</td>
<td>Poland</td>
<td>43</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>9</td>
<td>162</td>
<td>18.0</td>
</tr>
<tr>
<td>J</td>
<td>F</td>
<td>Poland</td>
<td>30</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>9</td>
<td>152</td>
<td>16.9</td>
</tr>
<tr>
<td>K</td>
<td>M</td>
<td>Ireland</td>
<td>42</td>
<td>3a</td>
<td>Iatrogenic</td>
<td>No</td>
<td>6</td>
<td>113</td>
<td>18.8</td>
</tr>
<tr>
<td>L</td>
<td>M</td>
<td>England</td>
<td>43</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>8</td>
<td>129</td>
<td>16.1</td>
</tr>
<tr>
<td>M</td>
<td>F</td>
<td>Ireland</td>
<td>36</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>7</td>
<td>125</td>
<td>17.9</td>
</tr>
<tr>
<td>N</td>
<td>M</td>
<td>Australia</td>
<td>52</td>
<td>3a</td>
<td>Unknown</td>
<td>No</td>
<td>3</td>
<td>59</td>
<td>19.7</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>Pakistan</td>
<td>21</td>
<td>3a</td>
<td>Iatrogenic</td>
<td>No</td>
<td>7</td>
<td>149</td>
<td>21.3</td>
</tr>
<tr>
<td>P</td>
<td>M</td>
<td>Poland</td>
<td>32</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>8</td>
<td>143</td>
<td>17.9</td>
</tr>
<tr>
<td>Q</td>
<td>F</td>
<td>Poland</td>
<td>23</td>
<td>3a</td>
<td>Unknown</td>
<td>No</td>
<td>8</td>
<td>123</td>
<td>15.4</td>
</tr>
<tr>
<td>R</td>
<td>M</td>
<td>Georgia</td>
<td>45</td>
<td>3a</td>
<td>Blood product</td>
<td>No</td>
<td>8</td>
<td>156</td>
<td>19.5</td>
</tr>
<tr>
<td>S</td>
<td>M</td>
<td>Portugal</td>
<td>47</td>
<td>3a</td>
<td>Unknown</td>
<td>No</td>
<td>7</td>
<td>100</td>
<td>14.3</td>
</tr>
<tr>
<td>T</td>
<td>M</td>
<td>Pakistan</td>
<td>45</td>
<td>3a</td>
<td>Unknown</td>
<td>Yes</td>
<td>8</td>
<td>158</td>
<td>19.8</td>
</tr>
<tr>
<td>U</td>
<td>M</td>
<td>Lithuania</td>
<td>39</td>
<td>3a</td>
<td>Intravenous drug user</td>
<td>No</td>
<td>8</td>
<td>175</td>
<td>21.9</td>
</tr>
<tr>
<td>V</td>
<td>M</td>
<td>Pakistan</td>
<td>19</td>
<td>3a</td>
<td>Unknown</td>
<td>No</td>
<td>7</td>
<td>153</td>
<td>21.9</td>
</tr>
<tr>
<td>W</td>
<td>M</td>
<td>Lithuania</td>
<td>32</td>
<td>3a</td>
<td>Unknown</td>
<td>No</td>
<td>8</td>
<td>164</td>
<td>20.5</td>
</tr>
</tbody>
</table>
**QS diversity, complexity and divergence**

We found marked variation in HVR1 QS diversity and complexity over periods as short as 2 weeks (Fig. 1, Table S1a, available in the online Supplementary Material). High QS diversity (above the third quartile \( \times 10^{-3} \)), which has previously been associated with successful response to treatment, had a positive predictive value of 64% in the subsequent sample (Fig. 1a). Only one subject demonstrated high diversity in all samples and no subjects had low diversity in all samples; 22% of subjects had both high and low diversity during the study period.

Low complexity (first quartile), which has also been associated with treatment success, was demonstrated in 13 subjects, but predicted low complexity in the subsequent sample in 51% of cases (Fig. 1b). No subject demonstrated high complexity throughout the study period and a single subject (subject T) demonstrated low complexity throughout the study period; five subjects demonstrated both low and high QS complexity during the study period.

Divergence from the original clonal set identified 10 subjects with little divergence during the study period (Fig. 1c).

The remaining 13 subjects demonstrated continuing, almost linear, divergence of varying magnitude from the original sample, leading to an ever-more distantly related QS population.

**Time-ordered phylogeny**

We evaluated all subjects using phylogenetics, median joining networks (MJNs) and partitioned analysis of quasispecies (PAQ). The findings for all subjects are summarized in Table 2.

We reconstructed phylogenetic trees for each subject using all unique HVR1 amino acid sequences for each sample, to investigate patterns of QS change with scale bars included (Figs. 2a and S1).

The phylogenetic analysis was characterized by two patterns: (i) 52% of subjects had compact trees suggesting little change in QS profile (subjects T and C; Figs 2a and S1) and (ii) in the remaining 48% of subjects, the trees demonstrated sequence change during the study period.
Table 2. Features of HVR1 QS change during the study for each subject as identified by phylogenetics, MJNs, PAQ and investigation for sequence-wide selection using PARRIS (www.datamonkey.org)

The presence of a time-ordered phylogeny, a single dominant subpopulation throughout the study period, replacement of the original dominant subpopulation during the study or evidence of sequence-wide selection is designated (+). MJN change is identified using ‘+’ where the change is limited to synonymous nucleotide change or changes in a single amino acid; changes in >2 aa are designated ‘++’. The total number of subpopulations reflects the number of HVR1 subpopulations identified for each subject during the study. Nucleotide substitutions per nucleotide position per year as calculated for each subject using MCMC analysis on BEAST.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Genotype</th>
<th>Cirrhosis</th>
<th>SVR</th>
<th>Phylogenetic change</th>
<th>Time-ordered phylogeny</th>
<th>Network change</th>
<th>Single dominant subpopulation</th>
<th>No. of subpopulations</th>
<th>Replacement of dominant subpopulation</th>
<th>Selection Nucleotide substitutions per site per year ($\times 10^{-2}$)</th>
<th>Substitutions codon position 1+2 : position 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1b</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>1.27</td>
<td>0.7245</td>
</tr>
<tr>
<td>B</td>
<td>1b</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>+</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td>1.522</td>
<td>0.5081</td>
</tr>
<tr>
<td>C</td>
<td>1b</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>1</td>
<td></td>
<td>1.942</td>
<td>0.0973</td>
</tr>
<tr>
<td>D</td>
<td>1b</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>3</td>
<td>+</td>
<td>1.66</td>
<td>0.5667</td>
</tr>
<tr>
<td>E</td>
<td>1b</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>4</td>
<td></td>
<td>2.098</td>
<td>0.441</td>
</tr>
<tr>
<td>F</td>
<td>1b</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>5</td>
<td></td>
<td>0.7122</td>
<td>0.7442</td>
</tr>
<tr>
<td>G</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>10</td>
<td>+</td>
<td>0.8423</td>
<td>1.113</td>
</tr>
<tr>
<td>H</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>8</td>
<td></td>
<td>0.747</td>
<td>1.028</td>
</tr>
<tr>
<td>I</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>3</td>
<td></td>
<td>0.3026</td>
<td>1.1</td>
</tr>
<tr>
<td>J</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>7</td>
<td>+</td>
<td>3.48</td>
<td>0.8627</td>
</tr>
<tr>
<td>K</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>3</td>
<td></td>
<td>2.069</td>
<td>1.186</td>
</tr>
<tr>
<td>L</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>2</td>
<td></td>
<td>3.342</td>
<td>0.4566</td>
</tr>
<tr>
<td>M</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>3</td>
<td></td>
<td>1.056</td>
<td>0.7984</td>
</tr>
<tr>
<td>N</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>3</td>
<td></td>
<td>0.4049</td>
<td>0.5109</td>
</tr>
<tr>
<td>O</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>5</td>
<td>+</td>
<td>1.756</td>
<td>2.626</td>
</tr>
<tr>
<td>P</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>2</td>
<td></td>
<td>1.347</td>
<td>0.8431</td>
</tr>
<tr>
<td>Q</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>7</td>
<td>+</td>
<td>6.164</td>
<td>1.3876</td>
</tr>
<tr>
<td>R</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>4</td>
<td></td>
<td>1.886</td>
<td>0.38647</td>
</tr>
<tr>
<td>S</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>5</td>
<td></td>
<td>1.238</td>
<td>0.621</td>
</tr>
<tr>
<td>T</td>
<td>3a</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>1</td>
<td></td>
<td>0.4957</td>
<td>0.3185</td>
</tr>
<tr>
<td>U</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>2</td>
<td></td>
<td>2.179</td>
<td>0.536</td>
</tr>
<tr>
<td>V</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>+</td>
<td>+</td>
<td></td>
<td>6</td>
<td></td>
<td>1.142</td>
<td>0.8573</td>
</tr>
<tr>
<td>W</td>
<td>3a</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>+</td>
<td>+</td>
<td></td>
<td>2</td>
<td></td>
<td>1.113</td>
<td>0.4541</td>
</tr>
</tbody>
</table>
HVR1 evolution in chronic HCV infection

(a) (b)

All Clones Week 0 Week16

C-1b

T-3a

Q-3a

F-1b

H-3a

http://vir.sgmjournals.org
Fig. 2. (a) Phylogenetic analysis of five subjects representative of HVR1 QS change. Trees were reconstructed using all unique HVR1 sequences at the nucleotide level for each sample using maximum composite likelihood (bootstrap 10 000 replicates). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree and scale bars included for each individual. The sequences from each sample are designated with unique colours. (b) MJN analysis of HVR1 QS change using all HVR1 sequences. Nodes are to scale and represent the prevalence of each sequence within the sample. Three graphical representations of the network are included: blue, all samples taken for the subject during the study; red and green, sequences present in the first and last samples, respectively.

In 30 % of subjects, the sequences from each sample clustered in distinct clades of the tree, with sequential migration of the sequences consistent with that described previously as a time-ordered phylogeny (subjects F, H and Q; Fig. 2a).

**Network analysis**

Phylogenetic analysis of HVR1 sequences is constrained to bifurcating trees and may underestimate the potential for multiple evolutionary connections between groups of sequences (Posada & Crandall, 2001). We therefore explored network associations between the HVR1 sequences using MJNs, which also allowed us to demonstrate the prevalence of each QS sequence within the sample studied (Figs. 2b and S2). MJN analysis identified no change in the HVR1 QS in 30 % of subjects, all of whom had no phylogenetic changes (Table 2). In 22 % of subjects, MJN analysis identified subtle changes in the QS milieu which were not immediately apparent on inspection of the phylogenetic tree, thus identifying alterations in the prevalence of the dominant sequence or the emergence of a new QS, as illustrated by subject C (Fig. 2b).

In the remaining 11 subjects, MJN analysis clearly illustrated the change in the HVR1 QS with the emergence of new sequences and disappearance of their predecessors, as illustrated by subjects H and Q or the diminishing prevalence of the original sequence in subject F (Fig. 2b).

**HVR1 subpopulations**

In order to evaluate evidence of HVR1 QS communities or subpopulations, we performed a PAQ. Subpopulations are a recently described QS phenomenon and are defined as groups or communities of closely related sequences which are remotely related to each other (Baccam et al., 2001; Ramachandran et al., 2011). We defined subpopulations as groups of sequences that differed within the group by <4 aa substitutions (15 %), but differed from all other sequences by ≥4 aa substitutions (Figs 3a and S3). In all cases this analysis correctly identified discrete closely grouped sequences on the subjects phylogenetic tree (data not shown).

We found evidence of multiple subpopulation infection in 21 subjects, whilst two subjects had a single subpopulation throughout the study (subjects C and T).

Amongst the 21 subjects with multiple subpopulations, the dominant subpopulation (subpopulation containing >50 % of sequences) in the initial sample had maintained this dominance at study completion in 11 subjects (Table 2). Nine subjects demonstrated a multiple subpopulation infection, where the initially dominant subpopulation had been replaced and was no longer present in the final sample (as seen in subjects H and Q) (Fig. 3a). Subject F had multiple subpopulations with incremental decrease in the prevalence of the initially dominant subpopulation as the study progressed (Fig. 3a).

**Persistence of individual sequences**

We also include an illustration of the persistence of individual amino acid sequences during the study period (Figs. 3b and S4) which highlights within-subpopulation QS change. This is clearly seen in subject C (a single-subpopulation infection) where the initial dominant sequence had been displaced at the end of the study. Equally, the degree of sequence change in our multiple subpopulation subjects becomes more apparent, with the most dramatic changes seen in subject Q, where each sample is characterized by a new dominant sequence with the disappearance of the preceding sequence, suggesting rapid virus evolution that is most likely the result of host immune-mediated clearance.

**Nucleotide substitution rates**

In order to quantify the rate of QS change, we calculated substitution rates per nucleotide position per year for each individual (Table 2). This highlighted significant differences in the tempo of intra-host QS change. By calculating the ratio of changes in codon positions 1 + 2 (which are likely to be non-synonymous) to position 3 (which are likely synonymous), we were able to highlight a reversal in the normal ratio (which is <1) amongst those subjects where QS change was the most dramatic. In all subjects where we described significant QS phylogenetic, subpopulation, or sequence change, we found that either the substitution rate was increased or the ratio of substitution in codon positions 1 + 2 : position 3 had increased to >0.85.
Selection

We found evidence of both positive and negative selective pressures on HVR1 at the codon level (Fig. 4.). Codons 8, 11, 12, 14, 15, 16, 18, 21 and 22 showed evidence for positive selection in >20% of subjects, with positive selection identified at codons 21 and 22, in 30

Fig. 3. (a) Subpopulations represented as a proportion of all sequences per sample for representative subjects. Subpopulations identified using PAQ and identifying subpopulations where all sequences within each subpopulation differed by <15% at the amino acid level and differed from all other sequences by ≥15%. (b) Prevalence of all unique amino acid sequences for representative subjects throughout the study period. *Asterisks denote patients with cirrhosis.
and 35% of subjects, respectively. Codon positions 23 and 25 were under negative selection in >20% of cases, suggesting that these may code for important conserved structural motifs of HVR1. We included the amino acids for each codon position which highlight cross-genotypic conservation of particular amino acids, most notably at positions 2 (T), 6 (G), 7 (G), 20 (F), 23 (G) and 26 (Q).

**PARRIS** analysis demonstrated evidence of sequence-wide positive selection in 17% of subjects (n=4) (Fig. 4). All four subjects with evidence of sequence-wide positive selection had multiple subpopulation infection, characterized by an identifiable time-ordered phylogeny.

### Cirrhotic versus non-cirrhotic subjects

A number of studies demonstrated differences in the QS behaviour between cirrhotic and non-cirrhotic subjects. We evaluated our cohort for differences, and found that QS diversity and complexity did not differ between cirrhotic and non-cirrhotic subjects, with \( P \) values of 0.14 and 0.08 respectively (Mann–Whitney test difference of medians). When selective pressures were evaluated, the sequences derived from cirrhotic subjects were more likely to be under negative pressure at the codon level (\( P<0.05 \)) and those derived from non-cirrhotic subjects were more likely to be under positive selection (\( P<0.05 \)). Although all four of the subjects where evidence of sequence-wide positive selection was found were not cirrhotic, this also did not reach statistical significance (\( P=0.19 \)) (subjects G, J, O and Q; Fig. 4).

**Non-cirrhotic subjects** did, however, demonstrate more subpopulations during the study period compared with cirrhotic subjects (\( P<0.05 \)). When selective pressures were evaluated, the sequences derived from cirrhotic subjects were more likely to be under negative pressure at the codon level (\( P<0.05 \)) and those derived from non-cirrhotic subjects were more likely to be under positive selection (\( P<0.05 \)). Although all four of the subjects where evidence of sequence-wide positive selection was found were not cirrhotic, this also did not reach statistical significance (\( P=0.19 \)) (subjects G, J, O and Q; Fig. 4).

## DISCUSSION

Our study aimed to investigate, by describing chronic HCV HVR1 QS evolution over short intervals, whether previously described associations between low QS complexity and high QS diversity could be exploited, in order to optimize the timing of treatment (Fan et al., 2009; Moreau et al., 2008). Additionally, our 2–4 week sample intervals allowed us to provide novel insights into short-interval HVR1 QS change in chronic infection.

The surreptitious nature of initial HCV infection limits our ability to establish the timing of transmission, but all subjects met criteria for chronic infection and all had confirmed infection for >3 years.

Although HVR1 has been described across many ethnic groups, we acknowledge that variations in the genetic background of our multi-ethnic cohort may result in altered immunological responses and, as a result, variation in virus sequence evolutionary patterns.
QS diversity, complexity and divergence, following the transition to chronic infection, continue to change rapidly in an unpredictable fashion. Diversity and/or complexity in the lowest or highest quartile were replicated in the subsequent sample in 40–64% of cases; 35% of subjects had both high and low diversity or complexity during the 16-week study period. This has important implications for the use of complexity and diversity in the individualization of HCV treatment, as the time required to generate these results is likely to render them obsolete given the magnitude of change we have described.

Phylogenetic change ranges from complete stasis to rapid sequence change, with sequential replacement of the entire QS population. Whilst HVR1 QS stasis has previously been described in many studies, the phenomenon of time-ordered phylogeny has previously only been described in samples spanning many years of chronic infection (Ramachandran et al., 2011). We describe this pattern in 30% of our cohort over 16 weeks, with subject Q providing a remarkable illustration of this phenomenon, to the extent that each sample generates a new discrete branch in the phylogenetic tree.

The transition to chronic infection has previously been associated with an acceleration of envelope evolution, which it has been argued suggests active envelope targeting antibody-mediated clearance, following the activation of the humoral immune system (Liu et al., 2010). Certainly, this would seem the most likely explanation for the pattern of widespread HVR1 QS change seen in our subjects. Interestingly, however, we demonstrated a similar time-ordered phylogeny in one subject (subject D; see Supplementary Material) with chronic HCV-related cirrhosis, following infection for >20 years. This contrasts with the lack of phylogenetic change we have seen in the remainder of our cirrhotic subjects. We postulate that this represents an illustration of transient immune activation, once the antigenic threshold had been reached (Sallie, 2005), which is likely to occur less frequently in patients with long-term infection as the virus adapts to exploit gaps in the adaptive immune response.

The identification of subpopulations and description of subpopulation prevalences during the study using partitioned analysis techniques provide a novel in-depth illustration of HVR1 change, which is difficult to appreciate using phylogenetics or MJNs alone. Many subjects were characterized by single-subpopulation infection with the transient presence of low-prevalence minor subpopulations. These data support the supposition that a virus well-adapted to the host immune environment exists within a local fitness optimum, with ongoing active competitive exclusion of minor competitor subpopulations. The converse pattern, whereby sequential samples demonstrate significant change in the subpopulation composition, suggests either intensive immune-mediated clearance or rapid virus evolution on a rugged fitness landscape (Schmidt-Martin, 2012).

Our amino acid prevalence data provide further insight into HVR1 evolution within subpopulations, particularly amongst subjects with single-subpopulation infection. This is best seen in subject E (see Supplementary Material), but can also be appreciated in subject C (Fig. 3.). Subject C represents an important illustration of the ongoing change in HCV despite many years of infection, as this subject became infected as a result of contaminated Anti-D in 1977. This suggests genetic drift within the plateau of a local fitness optimum, but also illustrates continued sampling of the sequence space even nearly 40 years after initial infection (Schmidt-Martin, 2012).

The study includes a mean of 18 clones per sample, which should include all sequences making up 10% of the underlying QS. As a result, some of the subtle changes in HVR1 QS we describe could be distorted due to selection bias, which could be clarified. Use of next generation sequencing would not, however, be expected to affect the widespread phylogenetic change described in many of our subjects (Fishman & Branch, 2009; Polyak et al., 1997).

Although differences between HVR1 QS within tissues and serum have been described, our hypothesis was based on studies of serum HVR1 QS.

By performing a quantitative Markov chain Monte Carlo (MCMC) analysis, we illustrated marked variation in HVR1 substitution rates across our cohort, which highlights both the rapid nature of sequence change in this part of the envelope protein and the variability of the intra-host tempo of HVR1 change. Those subjects with high substitution rates or a ratio of codon position 1 + 2 : position 3 > 0.85 corresponded with the subjects where sequence change was greatest. The high substitution rates seen in a number of subjects could only, we argue, be explained by immune-mediated clearance of the preceding QS.

Selection

Codon-based selection analysis of selection pressures provides insight into the complexity of selection pressures on HVR1 between individuals. Negative selection is often interpreted as identifying structurally constrained protein motifs. A number of codons display evidence of both positive and negative selection in differing subjects, suggesting flexibility in the structurally constrained codons, possibly reflecting variations in the codons targeted by the host immune response.

We found evidence for sequence-wide positive selection in four subjects (G, J, O and Q), all of whom had multiple subpopulation infections and had phylogenetic features consistent with a time-ordered phylogeny (Figs 2 and S2, Table 2). This strongly suggests humoral immune-mediated change is driving QS change. Short interval rapid sequential episodes of what is likely to be antibody-mediated clearance have, to the best of our knowledge, not previously been reported in chronic HCV infection.
Chronic HCV infection is associated with a 20% risk of developing cirrhosis at 20 years (Alter, 2007). Cirrhosis has been associated with reduced or impaired humoral immune response. Although our study includes limited numbers (n=23), when analysed by cirrhosis status we found less divergence of HVR1 QS amongst our cirrhotic cohort. Although not statistically significant, our data suggest a potential link between low HVR1 QS diversity and, more particularly, QS complexity and cirrhosis. It is particularly notable that even amongst the cirrhotic cohort only a single subject demonstrated high diversity and this occurred in a single sample. The cirrhotic subjects had fewer subpopulations (P<0.05) and were more likely to demonstrate negative selection at the codon level than our non-cirrhotic subjects (P<0.05), which themselves demonstrated a marked preponderance for positive selection (P<0.05).

We postulate that the data suggest that with the advent of cirrhosis, virus adaptation is more likely to result in a single dominant QS. Further evolution of the virus in the absence of the activation of antibody-mediated clearance is likely to result from sampling of the sequence space within the subpopulation, resulting in incremental rather than exponential increases in virus fitness. This process competitively excludes other subpopulations which serve as a reservoir for replenishing the QS in the event of a selective sweep, thus preserving chronic infection.

Amongst non-cirrhotic subjects, virus evolution, we argue, occurs more rapidly, is more extensive and is dependent on the emergence of subpopulations remotely related to the antibody-targeted QS. This feature of QS change in non-cirrhotic subjects results in time-ordered phylogenetic change.

**CONCLUSION**

Our study highlights the unpredictable nature of HVR1 QS dynamics in chronically HCV infection. HVR1 continues to evolve quickly as evidenced by the identification of time-ordered phylogenetic change in 30% of the cohort.

Change in HVR1 QS amongst non-cirrhotic patients occurred mostly as a result of the emergence of a new subpopulation, often remotely related to the preceding sequences, with the complete disappearance of the preceding subpopulation in most cases. HVR1 QS stasis in our cohort is characterized by single-subpopulation infection and is more commonly seen in cirrhotic subjects, suggesting a virus well-adapted to a local fitness optimum.

Cirrhosis was also associated with reduced QS divergence, fewer subpopulations and purifying selection. The unpredictable nature of change in HVR1 complexity and diversity over short time intervals is likely to preclude their use for pre-treatment individualization of therapy.

**METHODS**

Subjects attended for venepuncture every 2 weeks for a period of 16 weeks prior to commencing treatment with pegylated IFN and ribavirin. All subjects were prospectively recruited in an unselected fashion with inclusion criteria confined to chronic HCV infection between the ages of 16 and 75 years. Entry into the study was voluntary and no compensation financial, or otherwise, was provided to study participants. All participants were attending outpatients in a tertiary referral centre in the Republic of Ireland for ongoing management of chronic hepatitis C.

Subjects provided written informed consent and the study was undertaken under the governance of the Clinical Research Ethics Committee of the Cork Teaching Hospitals. Samples were centrifuged within 2 h and stored at −70 °C within 6 h of collection.

**RNA extraction and amplification.** The extraction of HCV RNA was performed by use of the Total Nucleic Acid Isolation protocol on a MagNA Pure LC (Roche Diagnostics) automated platform.

Reverse transcription followed by amplification of a 320 bp fragment corresponding to nt 1254–1572 of reference strain HCVGENS1 genotype 3a (GenBank accession number X76918) was performed using previously described protocols (Palmer et al., 2012). PCR amplicon of the correct size was confirmed using gel electrophoresis. Negative controls were included in parallel in order to observe for possible cross-contamination of samples. All sequences were subsequently analysed on phylogenetic trees using the neighbour-joining algorithm in order to ensure no inter-subject contamination had occurred.

**Cloning.** Cloning was performed using a Zero Blunt TOPO PCR Cloning kit using chemically competent cells (TOP10) (Invitrogen). Colonies were isolated and amplified using an illustra TempliPhi Amplification kit (GE Healthcare) and the sequencing of the dsDNA product was outsourced to EurofinsDNA MWG Operon.

**GenBank.** The sequences generated in this project were previously reported to GenBank (accession numbers HQ661384–HQ661791).

**Sequence analysis.** Sequence similarity was compared using the BLAST web-based tool http://blast.ncbi.nlm.nih.gov/Blast.cgi. The sequences were aligned using CLUSTAL W and analysis was performed following the exclusion of sequences which were either incomplete or contained stop codons. The sequences were screened for evidence of recombination using SBP analysis on the www.datamonkey.org server. The optimum evolutionary model for analysis was determined using jModeltest. intra-sample genetic diversity was calculated using the generalized time reversible model (GTR+I+F). Sequence divergence was calculated using a G distributed maximum-likelihood evaluation pairwise genetic distance between the groups of clones from each sample. Genetic complexity was described using normalized Shannon entropy (S_h), which was calculated as S_h = ∑[(π_i × ln(π_i))/ln(π)], where π_i is the number of times each particular sequence appeared in the QS and n is the number of sequences in the sample. Phylogenetic analysis was performed using MEPAGS maximum composite likelihood with a bootstrap value of 10000. Codon-specific selection pressures were estimated using random effects likelihood and evidence for sequence-wide selection was established using PARRIS (PARtitioning approach for Robust Inference of Selection) through the www.datamonkey.org server.

Intra-host virus population evolution was further visualized in the form of MJNs using SplitsTree4. PAQ was used to identify subpopulations with a minimum difference of 15% at the amino acid level between all subpopulations. Nucleotide substitution rates per nucleotide site per year and the ratio of substitutions at codon
position 1 + 2: position 3 were calculated using MCMC analysis on BEAST as described previously (Gray et al., 2011).

Analysis of QS continuous variable metrics was performed using the Mann-Whitney U test as the data were not normally distributed and categorical variables were analysed using χ² analysis.

ACKNOWLEDGEMENTS

We gratefully acknowledge that funding for this project was provided by Molecular Medicine Ireland through the Clinician Scientist Structured Training Programme.

REFERENCES


