Evolutionary dynamics of hepatitis C virus envelope genes during chronic infection

Richard J. P. Brown,1† Vicky S. Juttla,1† Alexander W. Tarr,1 Rebecca Finnis,1 William L. Irving,1 Shelley Hemsley,2 Darren R. Flower,2 Persephone Borrow2 and Jonathan K. Ball1

1Microbiology and Infectious Diseases, Institute of Infection, Immunity and Inflammation, The University of Nottingham, Queen’s Medical Centre, Nottingham NG7 2UH, UK
2The Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire RG20 7NN, UK

Hepatitis C virus (HCV) envelope glycoproteins E1 and E2 are important targets for the host immune response. The genes encoding these proteins exhibit a high degree of variability that gives rise to differing phenotypic traits, including alteration in receptor-binding affinity and immune recognition and escape. In order to elucidate patterns of adaptive evolution during chronic infection, a panel of full-length E1E2 clones was generated from sequential serum samples obtained from four chronically infected individuals. By using likelihood-based methods for phylogenetic inference, the evolutionary dynamics of circulating HCV quasispecies populations were assessed and a site-by-site analysis of the $d_N/d_S$ ratio was performed, to identify specific codons undergoing diversifying positive selection. HCV phylogenies, coupled with the number and distribution of selected sites, differed markedly between patients, highlighting that HCV evolution during chronic infection is a patient-specific phenomenon. This analysis shows that purifying selection is the major force acting on HCV populations in chronic infection. Whilst no significant evidence for positive selection was observed in E1, a number of sites under positive selection were identified within the ectodomain of the E2 protein. All of these sites were located in regions hypothesized to be exposed to the selective environment of the host, including a number of functionally defined domains that have been reported to be involved in immune evasion and receptor binding. Dated-tip methods for estimation of underlying HCV mutation rates were also applied to the data, enabling prediction of the most recent common ancestor for each patient’s quasispecies.

INTRODUCTION

Hepatitis C virus (HCV) is a positive-sense RNA virus and is the sole member of the genus Hepacivirus of the family Flaviviridae (Lindenbach & Rice, 2001). Globally, approximately 170 million people are at risk of liver disease due to chronic HCV infection, with an estimated 3 million individuals newly infected per annum (WHO, 1999). Of those infected, a reported 80% fail to clear the virus, a significant number of whom will go on to develop severe liver disease, including cirrhosis and hepatocellular carcinoma (Alter et al., 1992; Muller, 1996; Saito et al., 1990).

HCV circulates within an infected host as a heterogeneous viral population containing genetically distinct, but closely related variants, known as quasispecies (Bukh et al., 1995; Martell et al., 1992). The propensity for genetic change is associated primarily with the error-prone nature of the RNA-dependent RNA polymerase, together with the high HCV replicative rate in vivo (Fukumoto et al., 1996; Neumann et al., 1998; Ramratnam et al., 1999; Zeuzem, 2000). Chronic infection arises, at least in part, through the outgrowth of immune-escape mutants (Farcì et al., 2000; Frasca et al., 1999; Majid et al., 1999; Ray et al., 1999; Wang & Eckels, 1999). The envelope glycoprotein genes display some of the highest levels of HCV genetic heterogeneity, with E2 exhibiting greater variability than E1. A hypervariable region (HVR1) is located at the N terminus of E2 and this region is the major determinant for strain-specific neutralizing-antibody responses (Bartosch et al., 2003; Farcì et al., 1994, 1996; Rosa et al., 1996; Shimizu et al., 1994). The rate and nature of nucleotide substitutions within HVR1 during the early stages of infection appear to be correlated with outcome: patients harbouring a stable HVR1
quasispecies frequently resolve infection, whilst those with evidence of a rapidly evolving population develop chronic infection (Farci et al., 2000; Ray et al., 1999).

Evolution of the viral quasispecies continues during the chronic phase and differences in evolutionary rates and disease severity in individuals with differing levels of immunocompetency highlight the importance of antibody responses in controlling the infection (Booth et al., 1998; Kumar et al., 1994). Our current knowledge of adaptive evolution within the envelope genes during HCV chronic infection is based on estimates of synonymous (\(d_s\)) and non-synonymous (\(d_\text{NS}\)) nucleotide-substitution rates, averaged across very small regions of the envelope genes, including HVR1 (Curran et al., 2002; Gretch et al., 1996; Honda et al., 1994; McAllister et al., 1998; Smith, 1999). Unfortunately, such analyses are unable to provide insight into the evolution of a number of regions that are critical in envelope glycoprotein function, such as receptor-binding regions. In addition, previous methods utilized average \(d_\text{NS}/d_s\) ratios across the entire region under study. This is a highly conservative criterion for detecting positive selection, as only a few codons within the protein may be under diversifying selection. The signal could therefore be diluted in a background of purifying selection, maintained via strong functional constraint. To overcome analytical problems associated with differential selection across a region, the distribution of the \(d_\text{NS}/d_s\) ratio (\(\omega\)) can now be estimated for individual amino acids by assessing competing models of codon substitution within a maximum-likelihood (ML) framework (Yang & Bielawski, 2000). These ML methods have recently been applied to the identification of site-specific adaptive mutations in human immunodeficiency virus (HIV) env genes (Choisy et al., 2004) and partial E1E2 sequence datasets from individuals undergoing the acute phase of HCV infection (Sheridan et al., 2004). The latter study extended earlier findings of Ray et al. (1999) and Farci et al. (2000), revealing a statistically significant association between disease outcome and the number of positively selected sites (Sheridan et al., 2004).

In this report, we assess the evolutionary dynamics of chronic HCV infection by using temporally spaced, full-length E1E2 sequences generated from patient sera. These novel datasets are utilized for high-resolution phylogenetic reconstruction, identification of codon sites undergoing positive Darwinian selection and estimation of dates of their most recent common ancestor (MRCA), derived from patient-specific HCV mutation rates.

**METHODS**

**Source of samples.** Patient samples were obtained from the Trent HCV Study Cohort (Mohsen, 2001). All patients were chronically infected and were HCV treatment-naive (Table 1). Disease status was inferred from liver biopsies assessed by a single pathologist and scored according to both the Ishak (Ishak et al., 1995) and Knodell (Knodell et al., 1981) systems for derivation of a histological-activity index. Paired biopsies were read by the same pathologist, but without knowledge of the order of the biopsies. Two of the patients were defined as having severe progressive disease (SP) and two were defined as having mild, non-progressive disease (MN) (Table 1). Patient human leukocyte antigen (HLA) typing was performed by the Blood Transfusion Service and HCV genotyping was conducted via the InnoLiPA reverse-hybridization assay (Bayer Diagnostics). Sequential serum samples were collected from each of the four individuals and stored at \(-80^\circ\text{C}\) prior to RNA extraction.

**Amplification of E1E2.** RNA was recovered from 100 \(\mu\)l aliquots of serum by using a commercially available RNA-extraction kit (Fluka) and resuspended in 20 \(\mu\)l \(\text{dH}_2\text{O}\). Four microlitres of RNA was used in a 15 \(\mu\)l volume reverse-transcription step by using a commercially available first-strand kit (Pharmacia) containing 5 pmol antiense primer ASO (5’-CAGCAGGACCGCTTCAGCG-3’; positions 2619–2639 of the HCV clone H genome; GenBank accession no. M67463). Aliquots (1 \(\mu\)l) of resulting cDNA were used as template in a first-round full-length E1E2 PCR, containing 5 pmol primer E1OS (5’-GGACGGGTTAATCTGAC-AACAGG-3’, outer sense, positions 818–840) and primer ASO, 200 mM dNTPs and 0–5 U Expand High Fidelity polymerase (Roche) in a 25 \(\mu\)l reaction volume containing 1 x Expand buffer B. The PCR-cycling parameters were 25 cycles of 94 °C for 15 s, 50°C for 30 s and 72°C for 90 s, with a 5 s increase to the extension time following each cycle. One microlitre of the first-round product was then used in second-round reactions with primers 170 (5’-ATGGTTCCTCTTTCCTATC-3’, inner sense, positions 852–869) and 746 (5’-TTATGCTCTGTTGAT-3’, inner antisense, positions 2582–2599), using identical conditions to the first-round amplification.

**Cloning and sequence analysis.** E1E2 amplification products were ligated into a pcDNA3.1 V5 DTOPO expression vector (Invitrogen) and five clones representative of each sequential time point (TP) for each patient were sequenced by using BigDye Terminator chemistry (Perkin Elmer). Nucleotide sequences were aligned by using CLUSTAL_X (Thompson et al., 1997) with manual adjustment. Codon triplets containing gaps, ambiguous nucleotides or premature stop mutations were removed from each alignment prior to evolutionary analysis. Primer sequences at the 5’ and 3’ ends of the 1752 bp E1E2 amplicons were also removed to prevent any experimentally introduced bias to the phylogenetic analyses. Amino acid translations were performed by using MEGA version 2.1 (Kumar et al., 2001).

**Identification of recombinant sequences.** Individual patient datasets were checked for the presence of recombinant sequences prior to any analysis, as the models utilized for subsequent analyses assume that recombination has not taken place. Patient-specific alignments were divided into three segments of approximately 600 bp and simple neighbour-joining (NJ) (Saitou & Nei, 1987) trees were generated for each segment, utilizing the distance criterion implemented by PAUP* version 4.0b10 (Swofford, 2003) under a K80 model of nucleotide substitution (Kimura, 1980). Resultant reconstructed topologies were checked by eye for maintenance of a consistent branching order to identify any possible mosaic sequences. By using this method, a number of putative E1E2 recombinants were identified in datasets SP-1 and MN-2. Suspect sequences were then subjected to an informative-site test (Robertson et al., 1995) and sequences that demonstrated statistically significant evidence for recombination (\(P<0.05\)) were omitted from all subsequent phylogenetic analyses (GenBank accession nos: SP-1, AY957986/AY957997/AY957998/AY958002; MN-2, AY958048/AY958051/AY958059). This analysis is available from the authors on request.

**Phylogenetic reconstruction.** Molecular phylogenetic reconstructions were generated for each individual dataset (minus recombinants) by utilizing the likelihood criterion implemented by PAUP*
Table 1. Summary of patient data

<table>
<thead>
<tr>
<th>Patient genotype</th>
<th>Serum-isolation date</th>
<th>No. sequences</th>
<th>GenBank clone designations</th>
<th>Biopsy date</th>
<th>Fibrosis score</th>
<th>Ishak score</th>
<th>Knodell score</th>
<th>HLA type I</th>
<th>HLA type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP-1/3a</td>
<td>TP-A, 14 Oct 1996</td>
<td>5 (A1–A5)</td>
<td>UKN3A2.35–39</td>
<td>Sep 1995</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>A</td>
<td>26/30</td>
</tr>
<tr>
<td></td>
<td>TP-B, 9 May 1997</td>
<td>5 (B1–B5)</td>
<td>UKN3A2.1–5</td>
<td>Jul 1997</td>
<td>3</td>
<td>12</td>
<td>10</td>
<td>B</td>
<td>13/38</td>
</tr>
<tr>
<td></td>
<td>TP-C, 2 July 1998</td>
<td>5 (C1–C5)</td>
<td>UKN3A2.10–14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TP-D, 9 Feb 1999</td>
<td>5 (D1–D5)</td>
<td>UKN3A2.22–28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP-2/3a</td>
<td>TP-A, 16 Jan 1995</td>
<td>5 (A1–A5)</td>
<td>UKN3A4.2–6</td>
<td>Mar 1995</td>
<td>0</td>
<td>NA</td>
<td>4</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UKN3A4.15–19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MN-1/3a</td>
<td>TP-A, 28 Oct 1994</td>
<td>5 (A1–A5)</td>
<td>UKN3A1.1–6</td>
<td>Jan 1995</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>A</td>
<td>29/32</td>
</tr>
<tr>
<td></td>
<td>TP-B, 4 Nov 1996</td>
<td>5 (B1–B5)</td>
<td>UKN3A1.11–17</td>
<td>Jul 1997</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>B</td>
<td>27/56</td>
</tr>
<tr>
<td></td>
<td>TP-C, 30 Jun 1997</td>
<td>5 (C1–C5)</td>
<td>UKN3A1.21–25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TP-D, 23 Aug 1999</td>
<td>5 (D1–D5)</td>
<td>UKN3A1.28–32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TP-B, 19 Dec 1994</td>
<td>5 (B1–B5)</td>
<td>UKN1A14.14–18</td>
<td>Mar 1995</td>
<td>0</td>
<td>NA</td>
<td>3</td>
<td>B</td>
<td>62/35</td>
</tr>
<tr>
<td></td>
<td>TP-C, 22 May 1995</td>
<td>5 (C1–C5)</td>
<td>UKN1A14.27–32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TP-D, 20 Jan 1997</td>
<td>5 (D1–D5)</td>
<td>UKN1A14.38–44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Identification of positively selected sites. Patient-specific E1E2 sequence alignments and their corresponding unrooted phylogenetic trees were subjected to ML methods for identifying specific codon sites under diversifying selection by using the CODEML program of the PAML package, version 3.14 (Yang, 1997). ML methods implemented in CODEML employ competing models of codon substitution that incorporate various statistical distributions to allow for variable selection. M7beta and M8beta adequated via implementation of only two models of codon substitution: M7beta and M8beta. The M7beta null model incorporates a beta distribution, \( \beta(p,q) \), approximated by 10 discrete categories. Variable \( \omega \) ratios are allowed, depending on the values of \( p \) and \( q \), but are always between 0 and 1. Thus, M7beta does not permit positive selection. The M8beta model is identical to M7beta except that there is an additional class of sites possessing a free parameter, \( \omega_1 \), that is unconstrained, permitting a class of sites with \( \omega > 1 \) if selection is occurring. M7beta can then be compared with M8beta via a likelihood-ratio test (LRT). When M8beta suggests the occurrence of sites under diversifying selection, an empirical Bayes method is used to calculate the posterior probabilities of the assignment of \( \omega \) ratios to sites. When sites are identified as being under positive selection (\( \omega > 1 \)) with significant Bayesian posterior probabilities (>95%), this is indicative of the action of diversifying selection.

Analysis of the potential of HCV peptide sequences to act as class I-restricted T-cell epitopes. A database of HCV peptides identified as epitopes recognized by HLA class I-restricted T cells was created by using data reported in the literature. Peptides within patient HCV sequences that were predicted to bind with high affinity to the patient’s HLA alleles were determined by using the BIMAS site (www-bimas.dcrt.nih.gov/molbio/hla_bind/index.html) (Parker et al., 1994), as were the predicted HLA-binding affinities of variant versions of peptide sequences.

Estimation of mutation rate and MRCA. For rapidly evolving viruses, it is possible to estimate their rates of evolution via comparison of sequences isolated at different TPs. Patient-specific HCV mutation rates were inferred by using the dated-tips method (Rambaut, 2000) implemented in the BASEML program of PAML, version 3.14 (Yang, 1997). The single-rate dated-tips (SRDT) model allows the estimation of the underlying rate of molecular evolution from sequences with different, non-contemporaneous dates of isolation under a constant rate of substitution (molecular clock) enforced at each TP. If the SRDT model is significantly better than the single-rate (SR) model at describing the data (via an LRT), the ML estimates of substitution rates may be considered valid, even if the molecular-clock hypothesis is rejected (Jenkins et al., 2002). This rate can then be used to date the MRCA.

RESULTS

Assessment of intrapatient E1E2 evolution

The four patients yielded nucleotide sequence alignments of 16–20 continuous 1683–1704 bp fragments encompassing...
the E1E2 genes. These alignments were then subjected to ML phylogenetic reconstruction in order to assess the specific patterns of HCV evolution apparent in each chronically infected individual. The majority of E1E2 sequences derived from patient SP-1 (Fig. 1a) fall into one of three well-defined clades: one clade contains two highly divergent sequences and one clade contains all of the sequences present at the second TP only, whilst the third main group contains sequences obtained in all but the second TP. Apart from the cluster representing TP-B sequences, there was little evidence of grouping according to TP. The topology reconstructed from the SP-2 E1E2 sequences (Fig. 1b) shows a number of distinct viral populations, and grouping according to TP is more evident. Viruses derived from the first TP fall into one of two distinct clusters: one well-defined lineage is composed of monophyletic sequences derived from TP-C, with an antecedent sequence found at TP-B. The other lineage is composed of the majority of sequences derived from TP-B, which are antecedent to the monophyletic population derived from TP-D. The phylogeny reconstructed from the MN-1 sequences (Fig. 1c) shows a number of discrete viral populations that cluster on the basis of sampling date. Sequences derived from TP-A, -B and -C show progressive divergent evolution. The lineage observed at TP-D appears to be unrelated to the previous, progressively evolving lineage, suggesting that a selective sweep has driven a low-level viral variant to fixation between TP-C and -D. Finally, the MN-2 sequences (Fig. 1d) also exhibit progressive diversifying evolution. Sequences derived from TP-A form a monophyletic group, with the remaining sequences forming a number of discrete viral lineages, congruent with several distinct viral populations circulating at each sampling at TP-B and -C. However, a single monophyletic cluster is observed at TP-D, containing an isolate that was present at TP-B. Presumably, this variant becomes selectively advantageous between TP-C and -D and so is swept to fixation as the major viral population at the final sampling date.

**Fig. 1.** Rooted GTR +I+Γ ML trees. Asterisks assigned to internal nodes indicate the number of times that nodes were present after bootstrap resampling of the data and are derived from 1000 replications. Values > 70% are represented by *, values > 95% are represented by **. Branch lengths are in accordance with the scale bar and are proportional to genetic distance. (a) Patient SP-1 phylogeny; (b) patient SP-2 phylogeny; (c) patient MN-1 phylogeny; (d) patient MN-2 phylogeny.

○, TP-A; ●, TP-B; □, TP-C; ■, TP-D.
Evolutionary dynamics of HCV in chronic infection

Table 2. Positively selected sites in E1E2 glycoproteins

<table>
<thead>
<tr>
<th>Patient</th>
<th>Model</th>
<th>Inl</th>
<th>Positively selected codons</th>
<th>$2\Delta l$</th>
<th>df</th>
<th>$\chi^2$</th>
<th>Mean $\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP-1</td>
<td>M7 beta</td>
<td>-3780-29</td>
<td>Not allowed</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.4000</td>
</tr>
<tr>
<td></td>
<td>M8 beta + $\omega$</td>
<td>-3678-85</td>
<td>384N, 395G, 401S, 405P, 416S, 496V, 4995</td>
<td>22-88</td>
<td>2</td>
<td>$P&lt;0.001$</td>
<td>0.5363</td>
</tr>
<tr>
<td>SP-2</td>
<td>M7 beta</td>
<td>-3181-11</td>
<td>Not allowed</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.2251</td>
</tr>
<tr>
<td></td>
<td>M8 beta + $\omega$</td>
<td>-3178-67</td>
<td>404S</td>
<td>4-88</td>
<td>2</td>
<td>NS</td>
<td>0.2388</td>
</tr>
<tr>
<td>MN-1</td>
<td>M7 beta</td>
<td>-3053-40</td>
<td>Not allowed</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.2296</td>
</tr>
<tr>
<td></td>
<td>M8 beta + $\omega$</td>
<td>-3051-30</td>
<td>479I</td>
<td>4-20</td>
<td>2</td>
<td>NS</td>
<td>0.2778</td>
</tr>
<tr>
<td>MN-2</td>
<td>M7 beta</td>
<td>-3316-81</td>
<td>Not allowed</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.2000</td>
</tr>
<tr>
<td></td>
<td>M8 beta + $\omega$</td>
<td>-3309-19</td>
<td>384V, 582D</td>
<td>15-24</td>
<td>2</td>
<td>$P&lt;0.001$</td>
<td>0.2939</td>
</tr>
</tbody>
</table>

Identification of sites under positive selection

To elucidate patterns of adaptive evolution in chronic HCV infections, a site-by-site analysis of $\omega$ ratios was performed (see Table 2). For the SP-1 dataset, the M8 beta + $\omega$ analysis identified a total of seven sites that had a $\omega$ value greater than 1, and were thus subject to positive selection. A further 19 sites were identified with $\omega > 1$, but with non-significant posterior probabilities (data not shown). Only one codon in each of the patient SP-2 and MN-1 datasets was under positive selection, although the M7 beta–M8 beta comparison was not statistically significant for either of these. M8 beta + $\omega$ identifies a total of two sites that are predicted to be subject to positive selective pressure in the analysis of patient MN-2 virus sequences, and is significantly better than the null model at describing the observed data. A further 22 sites are identified as having $\omega$ values elevated above 1 in this dataset, although these assignments were not statistically significant (data not shown). Whilst the site-specific model M8 beta + $\omega$ identifies individual sites under positive Darwinian selection in all patient datasets, average $\omega$ values for the entire E1E2-coding region are all less than 1, suggesting that purifying selection due to functional constraint is the main force acting on E1E2 (Table 2). This observation confirms that pairwise-averaging methods lack the power to detect diversifying selection in these patient viral sequences, with specific sites having elevated $\omega$ ratios being diluted in a background of purifying selection.

Mapping selected sites onto functional regions of E1E2

To assess the potential impact of these adaptive mutations, sites under selection were superimposed onto an E1E2 alignment, highlighting regions of known function (Fig. 2). Although the number and distribution of identified adaptive mutations appears to be patient-specific, a number of broad observations can be made. All of the selected sites resided in the E2 gene; there was no evidence for positive selection in E1. In total, five selected sites were clustered within HVR1. Crucially, a further three selected sites were located proximal to a CD81-binding domain, and one within a region recognized by antibodies that neutralize CD81 binding (Flint et al., 1999; Owsianka et al., 2001) and infectivity of pseudovirus carrying genotype 1 HCV envelope glycoproteins (Hsu et al., 2003).

Evaluation of the potential rate of CD8 T-cell responses in driving amino acid change at the selected sites

In HIV-1 infection, selective pressure exerted by the virus-specific CD8 T-cell response represents an important driving force for viral sequence variation (Goulder & Watkins, 2004). The possibility that HLA class I-restricted T-cell responses may have constituted one of the forces underlying the observed diversifying sequence change in the HCV glycoproteins was thus also considered. Lack of availability of patient peripheral blood mononuclear cells precluded mapping of the epitopes in E1E2 against which each patient’s T-cell response was directed; however, two of the patients (SP-2 and MN-2) possessed commonly studied HLA class I alleles (Table 1), enabling prediction of potential T-cell epitopes by comparison with previously described epitopes in E1E2 and via prediction of peptides in each patient’s autologous virus sequence that would be able to bind with high affinity to their HLA class I alleles. Methods for prediction of peptide binding to major histocompatibility complex (MHC) molecules are used widely for the identification of potential T-cell epitopes. Their efficiency depends on the MHC allele studied, being highest for alleles where most peptide-binding data are available, such as HLA-A2 (Yu et al., 2002).

The single selected site in SP-2’s E2 sequence was found to lie within a region of sequence where overlapping HLA-A2-restricted epitopes have been identified in HCV gt1 viruses (Grüner et al., 2000; Shirai et al., 1995; Tsai et al., 1998; Urbani et al., 2001), although the disparity between the patient’s autologous virus (gt3a) sequence and that of the reported epitopes raises doubts as to whether the same mutations would have constituted T-cell epitopes in patient SP-2. This site is also contained within an autologous virus peptide (LFSQGARQNL) that is predicted to bind with
very high affinity to HLA-Cw4, another of the patient's HLA alleles, which may thus have constituted an epitope recognized by the CD8 T-cell response in this patient. However, it is unclear whether the amino acid variation that occurred within this sequence in patient SP-2's HCV quasispecies may have represented escape from a T-cell response to this epitope: the observed sequence changes did not reduce the predicted affinity of binding of the putative epitope peptide to HLA-Cw4, although they could potentially have affected the T-cell response via other mechanisms, e.g. by altering T-cell receptor recognition of the peptide–MHC complex.

Neither of the selected sites in patient MN-2's E2 sequence falls within previously reported T-cell epitopes; however, the site where particularly strong selection pressure was observed (384V) forms the C-terminal residue of an autologous virus peptide (LLFAGVDA) that is predicted to bind with very high affinity to HLA-A2, which may have been one of the epitopes targeted by patient MN-2's HCV-specific CD8 T-cell response. The fact that this sequence spans the E1E2 cleavage site would not preclude generation of the putative epitope peptide, as proteasomal processing of proteins within the cytoplasm constitutes a major source of peptide generation for presentation to T cells. Notably, although all viral clones sequenced from this patient at TP-A contained the LLFAGVDA sequence, 100 % of clones sequenced at all subsequent TPs bore amino acid changes at the C terminus of the putative epitope (a residue typically involved in peptide anchoring to the HLA-A2 molecule) that were predicted to effectively ablate binding of this peptide to HLA-A2. The estimated half-time of dissociation

Fig. 2. Distribution of amino acid sites under diversifying selection in an E1E2 sequence alignment. Each dataset analysed by CODEML is represented by a single sequence and aligned against the HCV-H reference sequence from the Los Alamos National Laboratory (LANL) HCV database. Sites identified by codon-substitution model M8 beta+ as having ω > 1, with a Bayesian posterior probability in excess of 95 %, are underlined and highlighted in bold typeface. Their amino acid position relative to the AUG start codon of HCV-H (LANL) is also shown above the alignment. Functional regions within E1E2 are identified below the alignment: E1 TM, E1 transmembrane domain; E2 TM, E2 transmembrane domain; HVR1, hypervariable region 1; CD81imAbs, CD81-inhibiting mAbs (Flint et al., 1999; Owsianka et al., 2001); HVR2, hypervariable region 2; CD81-1–3, CD81-binding domains: CD81-1 (Yagnik et al., 2000), CD81-2a (Yagnik et al., 2000), CD81-2b (Flint et al., 1999), CD81-3 (Roccasceca et al., 2003); #, predicted N-linked glycosylation site (NXT or NXS).
of a complex between HLA-A0201 and the index peptide LLFAGVDAV has a score of 493±042, whilst the residue 9 E, N and H variants subsequently selected for in the patient HCV quasispecies have scores only of 0±016, 0±528 and 0±528, respectively. Likewise, experimental data that we have obtained previously show that E and H, and, to a lesser extent, N, are extremely poor P9 anchors for HLA-A0201 (Doytchinova et al., 2004). These findings are thus consistent with the hypothesis that the strong selection pressure at this site may have been provided by CD8 T cells directed against the A2-restricted epitope LLFAGVDAV, which drove selection for amino acid changes that conferred escape from this response by ablating binding of the epitope peptide to HLA-A2.

**Dated-tip estimations of mutation rates and MRCA dates**

By using the dated-tips method (Rambaut, 2000), ML predictions of HCV mutation rates (\(\mu\)) for each patient’s viral population were obtained (Table 3). For patient SP-1, two sequences corresponding to the highly divergent clade (A4 and C2) were omitted from the analysis. Although the data are not strictly clock-like (a differential-rate model fits the data better than SRDT: data not shown), the SRDT model provides a significantly better fit than the SR null model for data obtained from MN-1 and MN-2; therefore, these data can be used to estimate \(\mu\) and hence the MRCA (Jenkins et al., 2002). Similarly, for SP-2, the comparison of SRDT with SR is significant, so \(\mu\) and the MRCA date can be estimated. However, for SP-1, SRDT does not fit the data significantly better than SR, inferring that \(\mu\), and subsequently the MRCA, cannot be estimated for this patient with any degree of accuracy. There is no relationship between time and the number of observed nucleotide substitutions for this patient’s virus. The estimated dates for the MRCA sequence, obtained for all of the patient-specific quasi-species, fall within the period between the recorded date of initial risk of exposure to HCV and the first HCV PCR-positive sample. The rate of nucleotide substitution ranged from \(1±39\times10^{-4}\) to \(3±95\times10^{-3}\) substitutions site\(^{-1}\) year\(^{-1}\). TipDate analysis was also performed on patient-specific trees rooted with an outgroup strain. Whilst this analysis gave similar results for patient-specific mean mutation rates and MRCA, the 95% confidence intervals around the mean were considerably larger.

**DISCUSSION**

This report describes the evolution of the HCV envelope genes over several years during chronic infection. Patients were recruited for this study on the basis of knowledge of histologically defined liver-disease status, an absence of antiviral therapy and the availability of sequential serum samples. In total, 80 full-length E1E2 clones were generated for the four patients and subjected to various phylogenetic analyses. The generated topologies, rates of evolution and number and location of selected sites differed markedly between the four patients. To our knowledge, this is the first reported example of site-by-site analysis of diversifying selection applied to full-length E1E2 genes during chronic infection.

**Phylogenetic analysis of E1E2 during chronic infection**

Reconstruction of accurate phylogenies is influenced by the phylogenetic ‘depth’, which is linked directly to the choice of gene used for a specific analysis. The HVR1 region of E2, which is considered to be the most rapidly diverging portion of the HCV genome, has been used extensively in recent molecular epidemiological studies for the recovery of

---

**Table 3. Patient infection data and likelihood estimates**

<table>
<thead>
<tr>
<th>Patient</th>
<th>Estimated date of infection</th>
<th>PCR(^{+})</th>
<th>Model</th>
<th>lnL</th>
<th>Parameters</th>
<th>2(\Delta)L</th>
<th>df</th>
<th>(\chi^2)</th>
<th>IDT(\mu)</th>
<th>IDTMRCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP-2</td>
<td>1969(^{IVDU}) 16 Jan 1995</td>
<td>SR</td>
<td>–3202-59</td>
<td>14</td>
<td>1·38</td>
<td>NS</td>
<td>1</td>
<td>1±07±8-86±10-4</td>
<td>1983±77±11-55</td>
<td></td>
</tr>
<tr>
<td>MN-1</td>
<td>1976(^{IVDU}) 18 Oct 1994</td>
<td>SR</td>
<td>–3323-00</td>
<td>19</td>
<td>4·90</td>
<td>NS</td>
<td>1</td>
<td>1±07±8-86±10-4</td>
<td>1983±77±11-55</td>
<td></td>
</tr>
</tbody>
</table>

*Date of infection was estimated from epidemiological data as being the first date when the individual was probably exposed to HCV.*
‘shallow’ intrapatient HCV phylogenies (Alfonso et al., 2004; Allain et al., 2000; Curran et al., 2002). However, HVR1 in isolation may not be an ideal candidate for such investigations, given its high level of sequence diversity and relatively short length. Phylogenetic analyses conducted on HVR1 may result in erroneous or misleading data, due to an inability to distinguish between synapomorphic and homoplastic substitutions (McCormack & Clewley, 2002). By extending the analysis to the entire E1 and E2 glycoprotein genes, we aimed to achieve a more robust and representative phylogenetic analysis. These loci exhibit both variable and conserved regions and constitute a larger dataset on which to perform the analysis, resulting in a more accurate assessment of patient-specific evolutionary trends. The phylogenetic reconstructions presented here suggest that HCV E1E2 evolution is patient-specific. One key observation was the identification of a number of putative recombinant sequences in two of our patient-specific datasets. Whilst it is impossible to discern whether these sequences represent true recombinant lineages suggests that these are chimeric products derived from in vitro template switching during reverse transcription (Zaphiropoulos, 2002) or true recombinant events in E1E2, the absence of any observed recombinant lineages suggests that these are mosaic sequences in a selected-site analysis can erroneously inflate the number of positively selected codons observed.

**Selected sites in E1E2**

The distribution of selected sites between these functional regions was patient-dependent and diversifying selection within the E1E2-coding region was confined to a relatively small number of sites. A high degree of conservation was observed within E1E2, indicating that the majority of amino acids are functionally constrained. No sites within E1 exhibited positive selection. One possible explanation for this finding is that E1 is hidden and is therefore not a strong target for host antibody responses. Indeed, E1 is reported to be a poor natural immunogen for humoral responses (Fournillier et al., 2001). Computational analysis predicts that HCV E1 is a truncated, class II membrane-fusion protein, homologous to those observed in other members of the family Flaviviridae (Garry & Dash, 2003), and is unlikely to be surface-exposed (Allison et al., 2001). Similarly, the transmembrane domains of E1 and E2, which are also occluded from antibody responses, also lacked sites that were under positive selection.

The positively selected sites were located within regions of E2 that are thought to be surface-exposed (Yagnik et al., 2000) and therefore prime targets for host antibody responses (Wack et al., 2001). Three of the four patients’ selected sites mapped to the HVR1 region. Unsurprisingly, none of these HVR1 mutations mapped to residues previously proposed to be functionally constrained (McAllister et al., 1998; Penin et al., 2001; Puntoriero et al., 1998; Smith, 1999). HVR1 is known to contain potent, strain-specific, neutralizing-antibody determinants (Farci et al., 1994, 1996; Shimizu et al., 1994, 1996) and our data support the concept of immune-driven evolution in this region (Booth et al., 1998; Kumar et al., 1994; Okamoto et al., 1992; Ray et al., 1999). HVR1 is implicated in scavenger receptor BI (SR-BI) binding (Scarselli et al., 2002), although the precise residues involved have yet to be reported. Whether or not these mutations arise to escape SR-BI-blocking antibodies and whether they affect SR-BI-binding affinity are key questions that are currently being investigated.

The absence of sites under selective pressure in non-exposed regions of the viral glycoproteins [which, although not targeted by antibodies, do contain T-cell epitopes (Ward et al., 2002)] suggests that the humoral response may exert more selective pressure on HCV replication than cell-mediated responses, at least during the chronic phase of infection. Nonetheless, we did find one example of selective change that was highly suggestive of escape from an epitope-specific CD8 T-cell response. Although comprehensive studies of the extent and kinetics of escape from the virus-specific CD8 T-cell response during human HCV infection are currently lacking, work in the chimpanzee model supports the hypothesis that escape may be among the mechanisms by which HCV evades CD8 T-cell control in this infection (Shoukry et al., 2004).

Positively selected sites were not confined to HVR1, with a number observed downstream, highlighting the importance of analysing the complete E2-coding region. Patient SP-1 possessed an adaptive mutation in the E2 region 412–447, a region that contains epitopes recognized by antibodies that inhibit CD81 binding (Flint et al., 1999; Owsianka et al., 2001) and neutralize infectivity of retroviral pseudotypes complemented by HCV genotype 1 envelope glycoproteins (Hsu et al., 2003). In addition, two of the four patients’ quasispecies possessed adaptive mutations proximal to the CD81-1 binding domain. Again, whether or not these mutations correlate with immune escape and altered CD81 affinity is under investigation.

E1 and E2 are highly glycosylated proteins, with five to six potential N-linked glycosylation sites in E1 and 11 potential sites in E2 (Goffard & Dubuisson, 2003; Meunier et al., 1999) (Fig. 2). Glycosylation in HCV envelope proteins is necessary for correct glycoprotein processing and folding (Goffard & Dubuisson, 2003; Huang et al., 1997; Li et al., 1993; Wu et al., 1995). Some variability in the location and number of N-linked glycosylation sites in our dataset was apparent. However, most sites were highly conserved and whilst two selected sites (416 SP-1; 582 MN-2) were located within N-linked glycosylation motifs (NXT or NXS), neither mutation altered the predicted glycosylation pattern. Glycosylation might mask important epitopes from host antibody responses (Schonning et al., 1996; Wei et al., 2003) and, as such, undergo positive selection (Choisy et al., 2004), but our data show that, in HCV infection, ensuring correct
conformation is probably more important than immune shielding.

Consequences of adaptive mutation

Considering the location of the sites under selection, the most likely consequence of the adaptive mutations is escape from antibodies that either block or interfere with CD81 and SR-B1 binding or another, as-yet-unidentified component of the entry process. This escape may be at the cost of reduced receptor-binding affinity. Interplay between host immunity and evolution of receptor-binding sites is not unprecedented; numerous studies have highlighted the role of mutations in and around the CD4-binding domain of HIV-1 gp120 and escape from antibodies that block CD4 interaction (Beaumont et al., 2004; Pinter et al., 2004; Pugach et al., 2004). The development and implementation of robust retrovirus pseudotype cell-entry assays will allow functional analysis of the E1E2 clone panel, to further elucidate the relationship between E1E2 evolution, host antibody responses and receptor-binding affinities.

Estimated mutation rates and infection dates

Dated-tip estimations of μ for each patient-specific HCV population showed evidence for a disparity in mutation rates between histologically defined liver-disease status. Mutation rates were patient-specific rather than being correlated with disease status, with estimated rates of 1.39 × 10⁻⁴ to 3.95 × 10⁻⁵ substitutions site⁻¹ year⁻¹, which are comparable to previous estimates (Allain et al., 2000; Curran et al., 2002; Smith, 1999). These estimates were then used to calculate the date for the quasispecies MRCA sequence. The Trent HCV Study Cohort (Molisen, 2001) patients complete a detailed risk-factor questionnaire to provide patient demographic details, in conjunction with information on the spread of HCV (Ryder et al., 2004). The recorded dates of HCV infection were taken to be when either intravenous drug use (IVDU) commenced or other activity with associated risk (ear piercing) first occurred (Table 3). Duration of infection was then estimated heuristically by using the first date of exposure to risk. Epidemiological data suggested that all four patients acquired their HCV infection between 1969 and 1976. However, SRDT MRCA estimations are considerably nearer the present day for both sets of patients. MRCA sequence estimates for the severe progressors appear to extend further back in time than the estimates derived from the mild non-progressors, suggesting a longer period of HCV infection in these patients. Indeed, HCV-induced fibrosis progression is known to be influenced by increased duration of infection, as well as numerous other factors (Serra et al., 2003). Our data show that quasispecies divergence from the founder viral sequence/population generally increases through time, whilst diversity remains stable. Genetic bottlenecks, selective sweeps and random genetic drift all reduce population diversity and, therefore, time to the MRCA. It is therefore probable that our MRCA estimates will be more recent than the date of transmission.

In conclusion, the evolutionary forces driving the diversity of HCV quasispecies in chronic infection are likely to be dependent on a plethora of factors. The host immune system definitely plays a significant role, but factors such as duration of infection, route of transmission (Gordon et al., 1993), size of original inoculum (Lau et al., 1993), age at infection, sex, alcohol consumption (Brechot et al., 1996), HLA type (Isaguliants & Ozeretskovskaya, 2003) and genotype (Marrone & Sallie, 1996), as well as hepatitis B virus (Weltman et al., 1995) and HIV (Martin et al., 1989) co-infection, all contribute to and affect intrapatient viral evolution. The cumulative effect of all these variables is likely to result in the patient-specific molecular evolution of HCV that we observe in chronic infection. Most importantly, this study has shown that previous studies of envelope adaptive evolution, which utilized methods that rely on average estimates of positive selection or were restricted to HVR1, were unable to highlight important sites under selection. It is significant that, in all patients’ quasispecies, we observed a strong association between sites under selection and those regions known, or thought, to be targeted by neutralizing antibodies and cell-mediated immunity, as well as regions involved in receptor binding.

ACKNOWLEDGEMENTS

This work was supported EU FP5 contract QLK2-CT-2001-01120 and by grants from the University of Nottingham Research Committee and the University Hospital Special Trustees. This is manuscript no. 101 from The Edward Jenner Institute for Vaccine Research. The authors would like to thank Paul Sharp, Elizabeth Bailes and Arvind Patel for useful discussion.

REFERENCES


http://vir.sgmjournals.org
Hypervariable approach using the bootstrap.


Puntoriero, G., Meola, A., Lahm, A. & 9 other authors (1998). Towards a solution for hepatitis C virus hypervariability: mimotopes of the hypervariable region 1 can induce antibodies cross-reacting with a large number of viral variants. *EMBO J* 17, 3521–3533.


Roccasecca, R., Ansuini, H., Vitelli, A. & 11 other authors (2003). Binding of the hepatitis C virus E2 glycoprotein to CD81 is strain specific and is modulated by a complex interplay between hypervariable regions 1 and 2. *J Virol* 77, 1856–1867.


