*Journal of General Virology* (2005), **86**, 1931–1942 DOI 10.1099/vir.0.80957-0

# 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 Borrow[2] and Jonathan K. Ball[1]

[1]Microbiology and Infectious Diseases, Institute of Infection, Immunity and Inflammation, The University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK

[2]The Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire RG20 7NN, UK

Correspondence
Jonathan K. Ball
jonathan.ball@nottingham.ac.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 alterations 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.

Received 9 February 2005
Accepted 21 March 2005

## 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 (Farci *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; Farci *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

†These authors contributed equally to this work.

The GenBank/EMBL/DDBJ accession numbers for the sequence data reported here are AY957985–AY958064.

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 ($d_N$) 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 $d_N/d_S$ 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 $d_N/d_S$ 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 °C prior to RNA extraction.

**Amplification of E1E2.** RNA was recovered from 100 µl aliquots of serum by using a commercially available RNA-extraction kit (Fluka) and resuspended in 20 µl $dH_2O$. Four microlitres of RNA was used in a 15 µl volume reverse-transcription step by using a commercially available first-strand kit (Pharmacia) containing 5 pmol antisense primer ASO (5′-CAGCAGCGACGGCGTTCA-GCG-3′; positions 2619–2639 of the HCV clone H genome; GenBank accession no. M67463). Aliquots (1 µl) of resulting cDNA were used as template in a first-round full-length E1E2 PCR, containing 5 pmol primer E1OS (5′-GGACGGGGTAAACTATGC-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 µl reaction volume containing 1 × 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′-ATGGGTTCCTCTTTTTCTATC-3′, inner sense, positions 852–869) and 746 (5′-TTATGCTTCTGCTTGTGATAT-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 ($P<0{\cdot}05$) 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

NA, Not available.

| Patient/ genotype | Serum-isolation date | No. sequences | GenBank clone designations | Biopsy date | Fibrosis score | Ishak score | Knodell score | HLA type I | HLA type II |
|---|---|---|---|---|---|---|---|---|---|
| SP-1/3a | TP-A, 14 Oct 1996 | 5 (A1–A5) | UKN3A2.35–39 | Sep 1995 | 1 | 7 | 7 | A 26/30 | DRB1 0102 |
| | TP-B, 9 May 1997 | 5 (B1–B5) | UKN3A2.1–5 | Jul 1997 | 3 | 12 | 10 | B 13/38 | DRB1 0701 |
| | TP-C, 2 July 1998 | 5 (C1–C5) | UKN3A2.10–14 | | | | | Cw 6 | DQB1 0501 |
| | TP-D, 9 Feb 1999 | 5 (D1–D5) | UKN3A2.22–28 | | | | | | DQB1 0201 |
| SP-2/3a | TP-A, 16 Jan 1995 | 5 (A1–A5) | UKN3A4.2–6 | Mar 1995 | 0 | NA | 4 | A 2 | DRB1 0101 |
| | TP-B, 25 Mar 1996 | 5 (B1–B5) | UKN3A4.34–38 | Dec 1996 | 1 | 8 | 7 | B 62/35 | DRB1 0103 |
| | TP-C, 23 Dec 1996 | 5 (C1–C5) | UKN3A4.15–19 | | | | | Cw 3/4 | DQB1 0501 |
| | TP-D, 30 Jun 1997 | 5 (D1–D5) | UKN3A4.26–30 | | | | | | |
| MN-1/3a | TP-A, 28 Oct 1994 | 5 (A1–A5) | UKN3A1.1–6 | Jan 1995 | 0 | 1 | 1 | A 29/32 | DRB1 0301 |
| | TP-B, 4 Nov 1996 | 5 (B1–B5) | UKN3A1.11–17 | Jul 1997 | 0 | 1 | 1 | B 27/56 | DRB1 0308 |
| | TP-C, 30 Jun 1997 | 5 (C1–C5) | UKN3A1.21–25 | | | | | Cw – | DQB1 0201 |
| | TP-D, 23 Aug 1999 | 5 (D1–D5) | UKN3A1.28–32 | | | | | | DQB1 0204 |
| MN-2/1a | TP-A, 21 Oct 1992 | 5 (A1–A5) | UKN1A14.1–5 | Dec 1992 | 0 | 3 | 4 | A 2/3 | DRB1 1501 |
| | TP-B, 19 Dec 1994 | 5 (B1–B5) | UKN1A14.14–18 | Mar 1995 | 0 | NA | 3 | B 62/35 | DQB1 0602 |
| | TP-C, 22 May 1995 | 5 (C1–C5) | UKN1A14.27–32 | | | | | Cw 3/4 | |
| | TP-D, 20 Jan 1997 | 5 (D1–D5) | UKN1A14.38–44 | | | | | | |

version 4.0b10 (Swofford, 2003) under a GTR+I+Γ model of nucleotide substitution (Sullivan *et al.*, 1999; Yang, 1994a, 1994b). The proportion of invariant sites (I) and the α shape parameter of the gamma distribution (Γ) were estimated (with eight discrete categories) from an initial NJ tree (Saitou & Nei, 1987) and subsequently fixed during heuristic optimization, which used the TBR branch-swapping algorithm with nucleotide frequencies and base-exchangeability parameters estimated from the data under the general time-reversible (GTR) model. Statistical confidence limits to infer the robustness of internal nodes were estimated by using the bootstrap approach (Felsenstein, 1985). Bootstrap values assigned to ML tree nodes were estimated from bootstrap consensus trees and are given as percentages derived from 1000 replicate NJ trees estimated under the GTR substitution model. For MN viral sequences, ML trees were rooted at the monophyletic population observed at TP-A. For SP viral sequences, where the TP-A sequences were not monophyletic, the root was positioned at the mid-point of the TP-A sequences.

**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 $\omega$ ratios across codon sites (Yang & Bielawski, 2000). Identification of specific codon sites under diversifying selection can be assessed adequately via implementation of only two models of codon substitution: $M7_{beta}$ and $M8_{beta+\omega}$. The $M7_{beta}$ null model incorporates a beta distribution, $\beta(p,q)$, approximated by 10 discrete categories. Variable $\omega$ rates are allowed, depending on the values of $p$ and $q$, but are always between 0 and 1. Thus, $M7_{beta}$ does not permit positive selection. The $M8_{beta+\omega}$ model is identical to $M7_{beta}$ 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. $M7_{beta}$ can then be compared with $M8_{beta+\omega}$ via a likelihood-ratio test (LRT). When $M8_{beta+\omega}$ 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.

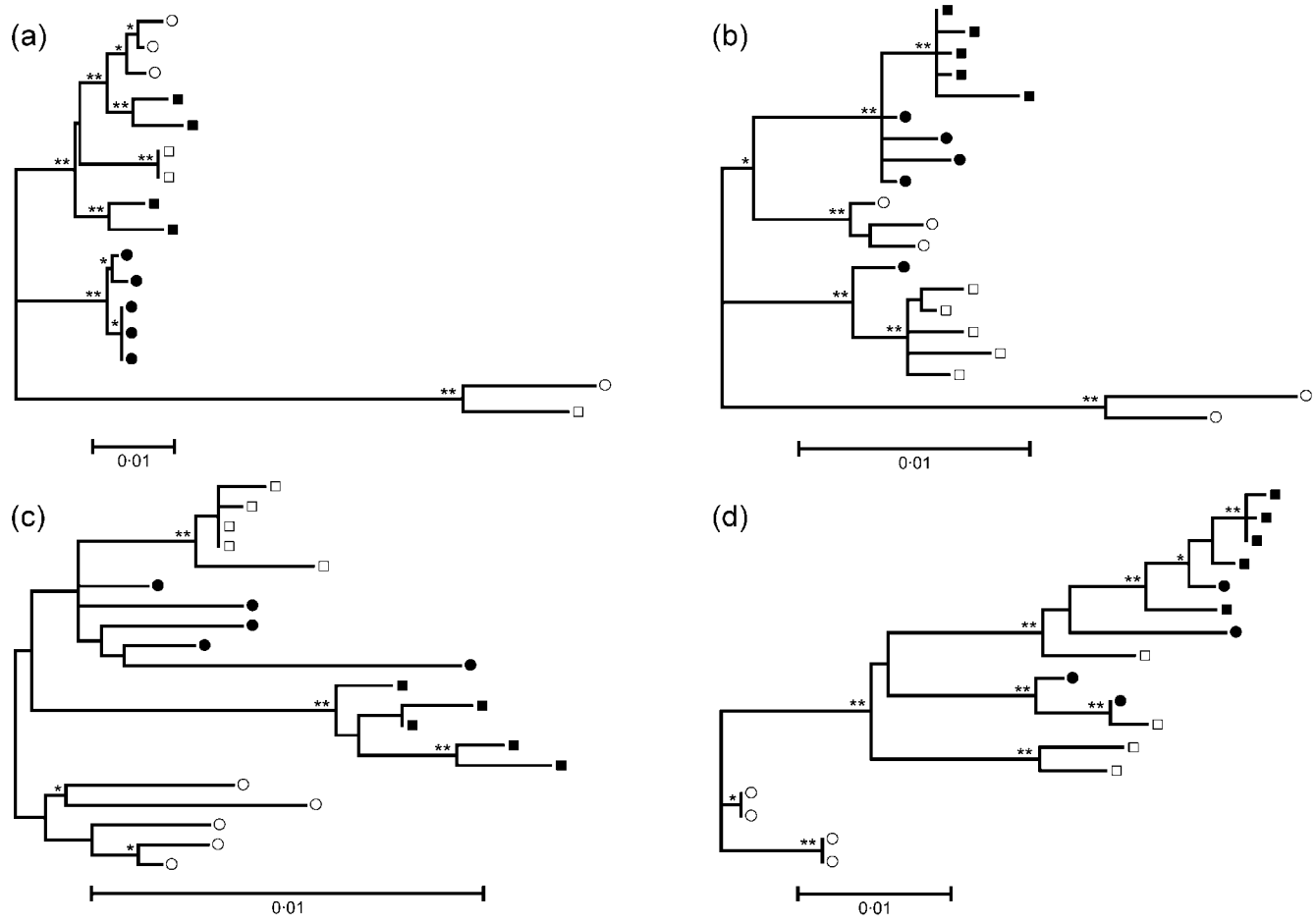


**Fig. 1.** Rooted GTR+I+Γ ML trees. Asterisks assigned to internal nodes indicate the number of times that nodes were present after bootstrap resampling of the data and are derived from 1000 replications. Values >70% are represented by *, values >95% are represented by **. Branch lengths are in accordance 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.

**Table 2.** Positively selected sites in E1E2 glycoproteins

Codons with a Bayesian posterior probability greater than 0·95 of belonging to the positively selected class ($\omega > 1$) are shown in normal typeface; codons at which $P > 0{\cdot}99$ are shown in bold. Sequences SP-1.A1, SP-2.A4, MN-1.A2 and MN-2.A1 are used as references for amino acids; ln*l*, log-likelihood score; Δ, likelihood-ratio statistic; df, degrees of freedom between nested models; NS, not significant.

| Patient | Model | ln*l* | Positively selected codons | 2Δ*l* | df | $\chi^2$ | Mean $\omega$ |
|---|---|---|---|---|---|---|---|
| SP-1 | $M7_{beta}$ | −3780·29 | Not allowed | – | – | – | 0·4000 |
| | $M8_{beta+\omega}$ | −3768·85 | 384N, 395G, **401S**, 405P, 416S, 496V, 499S | 22·88 | 2 | $P<0{\cdot}001$ | 0·5363 |
| SP-2 | $M7_{beta}$ | −3181·11 | Not allowed | – | – | – | 0·2251 |
| | $M8_{beta+\omega}$ | −3178·67 | 404S | 4·88 | 2 | NS | 0·2388 |
| MN-1 | $M7_{beta}$ | −3053·40 | Not allowed | – | – | – | 0·2296 |
| | $M8_{beta+\omega}$ | −3051·30 | 479I | 4·20 | 2 | NS | 0·2778 |
| MN-2 | $M7_{beta}$ | −3316·81 | Not allowed | – | – | – | 0·2000 |
| | $M8_{beta+\omega}$ | −3309·19 | **384V**, 582D | 15·24 | 2 | $P<0{\cdot}001$ | 0·2939 |

## 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 $M7_{beta}$–$M8_{beta+\omega}$ 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 sequences 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

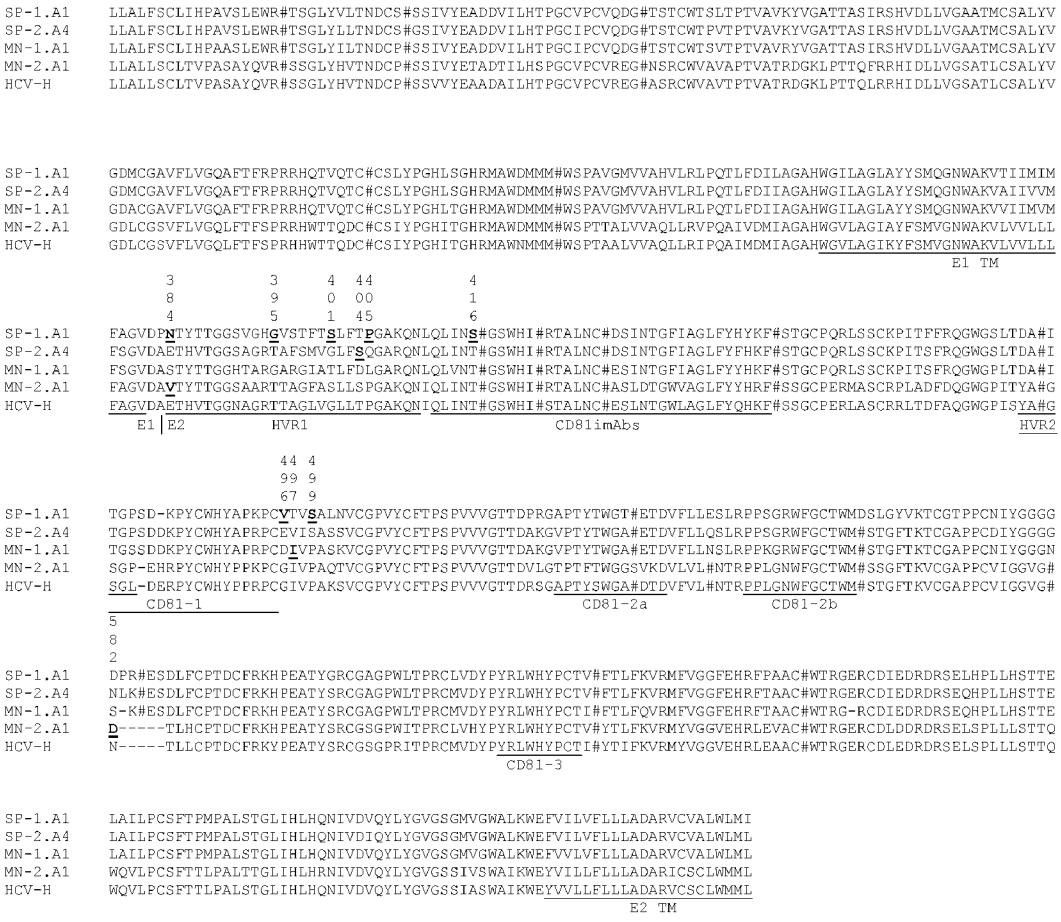


**Fig. 2.** Distribution of amino acid sites under diversifying selection in an E1E2 sequence alignment. Each dataset analysed by CODEML is represented by a single sequence and aligned against the HCV-H reference sequence from the Los Alamos National Laboratory (LANL) HCV database. Sites identified by codon-substitution model $M8_{beta+\omega}$ as having $\omega > 1$, 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 (Roccasecca *et al*., 2003); #, predicted *N*-linked glycosolation site (NXT or NXS).

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**V**) 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**V** 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

of a complex between HLA-A0201 and the index peptide LLFAGVDA**V** 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 0·106, 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 quasispecies, 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 $1{\cdot}39 \times 10^{-4}$ to $3{\cdot}95 \times 10^{-3}$ 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 MRCAs, 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

IVDU, Intravenous drug use; EP, ear piercing; PCR$^+$, date of initial HCV PCR-positive sample; ln*l*, log-likelihood score; $\Delta$, likelihood-ratio statistic; df, degrees of freedom between nested models; *l*DT*e*$\mu$, maximum-likelihood dated-tip estimation of underlying HCV mutation rate (substitutions site$^{-1}$ year$^{-1}$); *l*DT*e*MRCA, maximum-likelihood dated-tip estimation of MRCA of patient-specific quasispecies; NS, not significant.

| Patient | Estimated date of infection* | PCR$^+$ | Model | ln*l* | Parameters | 2$\Delta$*l* | df | $\chi^2$ | *l*DT*e*$\mu$ | *l*DT*e*MRCA |
|---|---|---|---|---|---|---|---|---|---|---|
| SP-1 | 1976$^{IVDU}$ | 19 Jul 1995 | SR | −3203·28 | 13 | – | – | – | | |
| | | | SRDT | −3202·59 | 14 | 1·38 | 1 | NS | $1{\cdot}07 \times 10^{-3} \pm 8{\cdot}68 \times 10^{-4}$ | 1983·77 ± 11·55 |
| SP-2 | 1969$^{IVDU}$ | 16 Jan 1995 | SR | −3323·00 | 19 | – | – | – | | |
| | | | SRDT | −3320·61 | 20 | 4·80 | 1 | $P<0{\cdot}05$ | $1{\cdot}61 \times 10^{-3} \pm 8{\cdot}23 \times 10^{-4}$ | 1985·95 ± 5·01 |
| MN-1 | 1976$^{IVDU}$ | 18 Oct 1994 | SR | −3169·72 | 19 | – | – | – | | |
| | | | SRDT | −3164·99 | 20 | 9·46 | 1 | $P<0{\cdot}005$ | $1{\cdot}39 \times 10^{-3} \pm 5{\cdot}19 \times 10^{-4}$ | 1991·46 ± 1·65 |
| MN-2 | 1974$^{EP}$ | 1 Oct 1992 | SR | −3499·06 | 16 | – | – | – | | |
| | | | SRDT | −3484·87 | 17 | 28·38 | 1 | $P<0{\cdot}0005$ | $3{\cdot}95 \times 10^{-3} \pm 8{\cdot}81 \times 10^{-4}$ | 1990·58 ± 0·90 |

*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 *in vivo* recombination events in E1E2, the absence of any observed recombinant lineages suggests that these are chimeric products derived from *in vitro* template switching during reverse transcription (Zaphiropoulos, 2002) or cDNA amplification (Meyerhans *et al.*, 1989). Irrespective of origin, this highlights the importance of checking the robustness of sequence data prior to performing the analyses detailed here. Indeed, the models used in the various analyses detailed assume no recombination, and inclusion of 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 (Schønning *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-BI 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 $\mu$ for each patient-specific HCV population show limited 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 $1{\cdot}39 \times 10^{-4}$ to $3{\cdot}95 \times 10^{-3}$ 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 (Mohsen, 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

**Alfonso, V., Flichman, D. M., Sookoian, S., Mbayed, V. A. & Campos, R. H. (2004).** Evolutionary study of HVR1 of E2 in chronic hepatitis C virus infection. *J Gen Virol* **85**, 39–46.

**Allain, J.-P., Dong, Y., Vandamme, A.-M., Moulton, V. & Salemi, M. (2000).** Evolutionary rate and genetic drift of hepatitis C virus are not correlated with the host immune response: studies of infected donor-recipient clusters. *J Virol* **74**, 2541–2549.

**Allison, S. L., Schalich, J., Stiasny, K., Mandl, C. W. & Heinz, F. X. (2001).** Mutational evidence for an internal fusion peptide in flavivirus envelope protein E. *J Virol* **75**, 4268–4275.

**Alter, M. J., Margolis, H. S., Krawczynski, K. & other authors (1992).** The natural history of community-acquired hepatitis C in the United States. The Sentinel Counties Chronic non-A, non-B Hepatitis Study Team. *N Engl J Med* **327**, 1899–1905.

**Bartosch, B., Vitelli, A., Granier, C. & 7 other authors (2003).** Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor. *J Biol Chem* **278**, 41624–41630.

**Beaumont, T., Quakkelaar, E., van Nuenen, A., Pantophlet, R. & Schuitemaker, H. (2004).** Increased sensitivity to CD4 binding site-directed neutralization following in vitro propagation on primary lymphocytes of a neutralization-resistant human immunodeficiency virus IIIB strain isolated from an accidentally infected laboratory worker. *J Virol* **78**, 5651–5657.

**Booth, J. C., Kumar, U., Webster, D., Monjardino, J. & Thomas, H. C. (1998).** Comparison of the rate of sequence variation in the hypervariable region of E2/NS1 region of hepatitis C virus in normal and hypogammaglobulinemic patients. *Hepatology* **27**, 223–227.

**Brechot, C., Nalpas, B. & Feitelson, M. A. (1996).** Interactions between alcohol and hepatitis viruses in the liver. *Clin Lab Med* **16**, 273–287.

**Bukh, J., Miller, R. H. & Purcell, R. H. (1995).** Genetic heterogeneity of hepatitis C virus: quasispecies and genotypes. *Semin Liver Dis* **15**, 41–63.

**Choisy, M., Woelk, C. H., Guégan, J.-F. & Robertson, D. L. (2004).** Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J Virol* **78**, 1962–1970.

**Curran, R., Jameson, C. L., Craggs, J. K., Grabowska, A. M., Thomson, B. J., Robins, A., Irving, W. L. & Ball, J. K. (2002).** Evolutionary trends of the first hypervariable region of the hepatitis C virus E2 protein in individuals with differing liver disease severity. *J Gen Virol* **83**, 11–23.

**Doytchinova, I. A., Walshe, V. A., Jones, N. A., Gloster, S. E., Borrow, P. & Flower, D. R. (2004).** Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J Immunol* **172**, 7495–7502.

**Farci, P., Alter, H. J., Wong, D. C., Miller, R. H., Govindarajan, S., Engle, R., Shapiro, M. & Purcell, R. H. (1994).** Prevention of hepatitis C virus infection in chimpanzees after antibody-mediated *in vitro* neutralization. *Proc Natl Acad Sci U S A* **91**, 7792–7796.

**Farci, P., Shimoda, A., Wong, D. & 7 other authors (1996).** Prevention of hepatitis C virus infection in chimpanzees by hyperimmune serum against the hypervariable region 1 of the envelope 2 protein. *Proc Natl Acad Sci U S A* **93**, 15394–15399.

**Farci, P., Shimoda, A., Coiana, A. & 9 other authors (2000).** The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* **288**, 339–344.

**Felsenstein, J. (1985).** Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

**Flint, M., Maidens, C., Loomis-Price, L. D., Shotton, C., Dubuisson, J., Monk, P., Higginbottom, A., Levy, S. & McKeating, J. A. (1999).** Characterization of hepatitis C virus E2 glycoprotein interaction with a putative cellular receptor, CD81. *J Virol* **73**, 6235–6244.

**Fournillier, A., Wychowski, C., Boucreux, D., Baumert, T. F., Meunier, J.-C., Jacobs, D., Muguet, S., Depla, E. & Inchauspé, G. (2001).** Induction of hepatitis C virus E1 envelope protein-specific immune response can be enhanced by mutation of N-glycosylation sites. *J Virol* **75**, 12088–12097.

**Frasca, L., Del Porto, P., Tuosto, L., Marinari, B., Scottà, C., Carbonari, M., Nicosia, A. & Piccolella, E. (1999).** Hypervariable region 1 variants act as TCR antagonists for hepatitis C virus-specific $CD4^+$ T cells. *J Immunol* **163**, 650–658.

**Fukumoto, T., Berg, T., Ku, Y., Bechstein, W. O., Knoop, M., Lemmens, H., Lobeck, H., Hopf, U. & Neuhaus, P. (1996).** Viral dynamics of hepatitis C early after orthotopic liver transplantation: evidence for rapid turnover of serum virions. *Hepatology* **24**, 1351–1354.

**Garry, R. F. & Dash, S. (2003).** Proteomics computational analyses suggest that hepatitis C virus E1 and pestivirus E2 envelope glycoproteins are truncated class II fusion proteins. *Virology* **307**, 255–265.

**Goffard, A. & Dubuisson, J. (2003).** Glycosylation of hepatitis C virus envelope proteins. *Biochimie* **85**, 295–301.

**Gordon, S. C., Elloway, R. S., Long, J. C. & Dmuchowski, C. F. (1993).** The pathology of hepatitis C as a function of mode of transmission: blood transfusion vs. intravenous drug use. *Hepatology* **18**, 1338–1343.

**Goulder, P. J. R. & Watkins, D. I. (2004).** HIV and SIV CTL escape: implications for vaccine design. *Nat Rev Immunol* **4**, 630–640.

**Gretch, D. R., Polyak, S. J., Wilson, J. J., Carithers, R. L., Jr, Perkins, J. D. & Corey, L. (1996).** Tracking hepatitis C virus quasispecies major and minor variants in symptomatic and asymptomatic liver transplant recipients. *J Virol* **70**, 7622–7631.

**Grüner, N. H., Gerlach, T. J., Jung, M.-C. & 9 other authors (2000).** Association of hepatitis C virus-specific $CD8^+$ T cells with viral clearance in acute hepatitis C. *J Infect Dis* **181**, 1528–1536.

**Honda, M., Kaneko, S., Sakai, A., Unoura, M., Murakami, S. & Kobayashi, K. (1994).** Degree of diversity of hepatitis C virus quasispecies and progression of liver disease. *Hepatology* **20**, 1144–1151.

**Hsu, M., Zhang, J., Flint, M., Logvinoff, C., Cheng-Mayer, C., Rice, C. M. & McKeating, J. A. (2003).** Hepatitis C virus glycoproteins mediate pH-dependent cell entry of pseudotyped retroviral particles. *Proc Natl Acad Sci U S A* **100**, 7271–7276.

**Huang, X., Barchi, J. J., Jr, Lung, F.-D. T., Roller, P. P., Nara, P. L., Muschik, J. & Garrity, R. R. (1997).** Glycosylation affects both the three-dimensional structure and antibody binding properties of the HIV-$1_{IIIB}$ GP120 peptide RP135. *Biochemistry* **36**, 10846–10856.

**Isaguliants, M. G. & Ozeretskovskaya, N. N. (2003).** Host background factors contributing to hepatitis C virus clearance. *Curr Pharm Biotechnol* **4**, 185–193.

**Ishak, K., Baptista, A., Bianchi, L. & 13 other authors (1995).** Histological grading and staging of chronic hepatitis. *J Hepatol* **22**, 696–699.

**Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. (2002).** Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol* **54**, 156–165.

**Kimura, M. (1980).** A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111–120.

**Knodell, R. G., Ishak, K. G., Black, W. C., Chen, T. S., Craig, R., Kaplowitz, N., Kiernan, T. W. & Wollman, J. (1981).** Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis. *Hepatology* **1**, 431–435.

**Kumar, U., Monjardino, J. & Thomas, H. C. (1994).** Hypervariable region of hepatitis C virus envelope glycoprotein (E2/NS1) in an agammaglobulinemic patient. *Gastroenterology* **106**, 1072–1075.

**Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001).** MEGA2: molecular evolutionary genetic analysis software. *Bioinformatics* **17**, 1244–1245.

**Lau, J. Y., Davis, G. L., Kniffen, J., Qian, K. P., Urdea, M. S., Chan, C. S., Mizokami, M., Neuwald, P. D. & Wilber, J. C. (1993).** Significance of serum hepatitis C virus RNA levels in chronic hepatitis C. *Lancet* **341**, 1501–1504.

**Li, Y., Luo, L., Rasool, N. & Kang, C. Y. (1993).** Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. *J Virol* **67**, 584–588.

**Lindenbach, B. D. & Rice, C. M. (2001).** *Flaviviridae*: the viruses and their replication. In *Fields Virology*, 4th edn, pp. 991–1041. Edited by D. M. Knipe & P. M. Howley. Philadelphia, PA: Lippincott Williams & Wilkins.

**Majid, A., Jackson, P., Lawal, Z., Pearson, G. M. J., Parker, H., Alexander, G. J. M., Allain, J.-P. & Petrik, J. (1999).** Ontogeny of hepatitis C virus (HCV) hypervariable region 1 (HVR1) heterogeneity and HVR1 antibody responses over a 3 year period in a patient infected with HCV type 2b. *J Gen Virol* **80**, 317–325.

**Marrone, A. & Sallie, R. (1996).** Genetic heterogeneity of hepatitis C virus. The clinical significance of genotypes and quasispecies behavior. *Clin Lab Med* **16**, 429–449.

**Martell, M., Esteban, J. I., Quer, J., Genescà, J., Weiner, A., Esteban, R., Guardia, J. & Gómez, J. (1992).** Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J Virol* **66**, 3225–3229.

**Martin, P., Di Bisceglie, A. M., Kassianides, C., Lisker-Melman, M. & Hoofnagle, J. H. (1989).** Rapidly progressive non-A, non-B hepatitis in patients with human immunodeficiency virus infection. *Gastroenterology* **97**, 1559–1561.

**McAllister, J., Casino, C., Davidson, F., Power, J., Lawlor, E., Peng, L. Y., Simmonds, P. & Smith, D. B. (1998).** Long-term evolution of the hypervariable region of hepatitis C virus in a common-source-infected cohort. *J Virol* **72**, 4893–4905.

**McCormack, G. P. & Clewley, J. P. (2002).** The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Rev Med Virol* **12**, 221–238.

**Meunier, J.-C., Fournillier, A., Choukhi, A., Cahour, A., Cocquerel, L., Dubuisson, J. & Wychowski, C. (1999).** Analysis of the glycosylation sites of hepatitis C virus (HCV) glycoprotein E1 and the influence of E1 glycans on the formation of the HCV glycoprotein complex. *J Gen Virol* **80**, 887–896.

**Meyerhans, A., Cheynier, R., Albert, J., Seth, M., Kwok, S., Sninsky, J., Morfeldt-Månson, L., Asjö, B. & Wain-Hobson, S. (1989).** Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. *Cell* **58**, 901–910.

**Mohsen, A. H. (2001).** The epidemiology of hepatitis C in a UK health regional population of 5·12 million. *Gut* **48**, 707–713.

**Muller, R. (1996).** The natural history of hepatitis C: clinical experiences. *J Hepatol* **24**, 52–54.

**Neumann, A. U., Lam, N. P., Dahari, H., Gretch, D. R., Wiley, T. E., Layden, T. J. & Perelson, A. S. (1998).** Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy. *Science* **282**, 103–107.

**Okamoto, H., Kojima, M., Okada, S. I. & 7 other authors (1992).** Genetic drift of hepatitis C virus during an 8·2-year infection in a chimpanzee: variability and stability. *Virology* **190**, 894–899.

**Owsianka, A., Clayton, R. F., Loomis-Price, L. D., McKeating, J. A. & Patel, A. H. (2001).** Functional analysis of hepatitis C virus E2 glycoproteins and virus-like particles reveals structural dissimilarities between different forms of E2. *J Gen Virol* **82**, 1877–1883.

**Parker, K. C., Bednarek, M. A. & Coligan, J. E. (1994).** Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* **152**, 163–175.

**Penin, F., Combet, C., Germanidis, G., Frainais, P.-O., Deléage, G. & Pawlotsky, J.-M. (2001).** Conservation of the conformation and positive charges of hepatitis C virus E2 envelope glycoprotein hypervariable region 1 points to a role in cell attachment. *J Virol* **75**, 5703–5710.

**Pinter, A., Honnen, W. J., He, Y., Gorny, M. K., Zolla-Pazner, S. & Kayman, S. C. (2004).** The V1/V2 domain of gp120 is a global regulator of the sensitivity of primary human immunodeficiency virus type 1 isolates to neutralization by antibodies commonly induced upon infection. *J Virol* **78**, 5205–5215.

**Pugach, P., Kuhmann, S. E., Taylor, J., Marozsan, A. J., Snyder, A., Ketas, T., Wolinsky, S. M., Korber, B. T. & Moore, J. P. (2004).** The prolonged culture of human immunodeficiency virus type 1 in primary lymphocytes increases its sensitivity to neutralization by soluble CD4. *Virology* **321**, 8–22.

**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.

**Rambaut, A. (2000).** Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399.

**Ramratnam, B., Bonhoeffer, S., Binley, J. & 7 other authors (1999).** Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet* **354**, 1782–1785.

**Ray, S. C., Wang, Y.-M., Laeyendecker, O., Ticehurst, J. R., Villano, S. A. & Thomas, D. L. (1999).** Acute hepatitis C virus structural gene sequences as predictors of persistent viremia: hypervariable region 1 as a decoy. *J Virol* **73**, 2938–2946.

**Robertson, D. L., Hahn, B. H. & Sharp, P. M. (1995).** Recombination in AIDS viruses. *J Mol Evol* **40**, 249–259.

**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.

**Rosa, D., Campagnoli, S., Moretto, C. & 11 other authors (1996).** A quantitative test to estimate neutralizing antibodies to the hepatitis C virus: cytofluorimetric assessment of envelope glycoprotein 2 binding to target cells. *Proc Natl Acad Sci U S A* **93**, 1759–1763.

**Ryder, S. D., Irving, W. L., Jones, D. A., Neal, K. R. & Underwood, J. C. (2004).** Progression of hepatic fibrosis in patients with hepatitis C: a prospective repeat liver biopsy study. *Gut* **53**, 451–455.

**Saito, I., Miyamura, T., Ohbayashi, A. & 10 other authors (1990).** Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. *Proc Natl Acad Sci U S A* **87**, 6547–6549.

**Saitou, N. & Nei, M. (1987).** The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

**Scarselli, E., Ansuini, H., Cerino, R. & 7 other authors (2002).** The human scavenger receptor class B type I is a novel candidate receptor for the hepatitis C virus. *EMBO J* **21**, 5017–5025.

**Schønning, K., Jansson, B., Olofsson, S. & Hansen, J.-E. S. (1996).** Rapid selection for an *N*-linked oligosaccharide by monoclonal antibodies directed against the V3 loop of human immunodeficiency virus type 1. *J Gen Virol* **77**, 753–758.

**Serra, M. A., Rodríguez, F., del Olmo, J. A., Escudero, A. & Rodrigo, J. M. (2003).** Influence of age and date of infection on distribution of hepatitis C virus genotypes and fibrosis stage. *J Viral Hepat* **10**, 183–188.

**Sheridan, I., Pybus, O. G., Holmes, E. C. & Klenerman, P. (2004).** High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *J Virol* **78**, 3447–3454.

**Shimizu, Y. K., Hijikata, M., Iwamoto, A., Alter, H. J., Purcell, R. H. & Yoshikura, H. (1994).** Neutralizing antibodies against hepatitis C virus and the emergence of neutralization escape mutant viruses. *J Virol* **68**, 1494–1500.

**Shimizu, Y., Igarashi, H., Kiyohara, T., Cabezon, T., Farci, P., Purcell, R. H. & Yoshikura, H. (1996).** A hyperimmune serum against a synthetic peptide corresponding to the hypervariable region 1 of hepatitis C virus can prevent viral infection in cell cultures. *Virology* **223**, 409–412.

**Shirai, M., Arichi, T., Nishioka, M., Nomura, T., Ikeda, K., Kawanishi, K., Engelhard, V. H., Feinstone, S. M. & Berzofsky, J. A. (1995).** CTL responses of HLA-A2.1-transgenic mice specific for hepatitis C viral peptides predict epitopes for CTL of humans carrying HLA-A2.1. *J Immunol* **154**, 2733–2742.

**Shoukry, N. H., Cawthon, A. G. & Walker, C. M. (2004).** Cell-mediated immunity and the outcome of hepatitis C virus infection. *Annu Rev Microbiol* **58**, 391–424.

**Smith, D. B. (1999).** Evolution of the hypervariable region of hepatitis C virus. *J Viral Hepat* **6**, 41–46.

**Sullivan, J., Swofford, D. L. & Naylor, G. J. P. (1999).** The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol Biol Evol* **16**, 1347–1356.

**Swofford, D. L. (2003).** PAUP*: Phylogenetic Analysis Using Parsinomy (*and other methods), version 4. Sunderland. MA: Sinauer Associates.

**Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997).** The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.

**Tsai, S. L., Chen, Y. M., Chen, M. H., Huang, C. Y., Sheen, I. S., Yeh, C. T., Huang, J. H., Kuo, G. C. & Liaw, Y. F. (1998).** Hepatitis C virus variants circumventing cytotoxic T lymphocyte activity as a mechanism of chronicity. *Gastroenterology* **115**, 954–965.

**Urbani, S., Uggeri, J., Matsuura, Y., Miyamura, T., Penna, A., Boni, C. & Ferrari, C. (2001).** Identification of immunodominant hepatitis C virus (HCV)-specific cytotoxic T-cell epitopes by stimulation with endogenously synthesized HCV antigens. *Hepatology* **33**, 1533–1543.

**Wack, A., Soldaini, E., Tseng, C.-T. K., Nuti, S., Klimpel, G. R. & Abrignani, S. (2001).** Binding of the hepatitis C virus envelope protein E2 to CD81 provides a co-stimulatory signal for human T cells. *Eur J Immunol* **31**, 166–175.

**Wang, H. & Eckels, D. D. (1999).** Mutations in immunodominant T cell epitopes derived from the nonstructural 3 protein of hepatitis C virus have the potential for generating escape variants that may have important consequences for T cell recognition. *J Immunol* **162**, 4177–4183.

**Ward, S., Lauer, G., Isba, R., Walker, B. & Klenerman, P. (2002).** Cellular immune responses against hepatitis C virus: the evidence base 2002. *Clin Exp Immunol* **128**, 195–203.

**Wei, X., Decker, J. M., Wang, S. & 12 other authors (2003).** Antibody neutralization and escape by HIV-1. *Nature* **422**, 307–312.

**Weltman, M. D., Brotodihardjo, A., Crewe, E. B., Farrell, G. C., Bilous, M., Grierson, J. M. & Liddle, C. (1995).** Coinfection with hepatitis B and C or B, C and delta viruses results in severe chronic liver disease and responds poorly to interferon-alpha treatment. *J Viral Hepat* **2**, 39–45.

**WHO (1999).** Global surveillance and control of hepatitis C. *J Viral Hepat* **6**, 35–47.

**Wu, Z., Kayman, S. C., Honnen, W. & 7 other authors (1995).** Characterization of neutralization epitopes in the V2 region of human immunodeficiency virus type 1 gp120: role of glycosylation in the correct folding of the V1/V2 domain. *J Virol* **69**, 2271–2278.

**Yagnik, A. T., Lahm, A., Meola, A., Roccasecca, R. M., Ercole, B. B., Nicosia, A. & Tramontano, A. (2000).** A model for the hepatitis C virus envelope glycoprotein E2. *Proteins* **40**, 355–366.

**Yang, Z. (1994a).** Estimating the pattern of nucleotide substitution. *J Mol Evol* **39**, 105–111.

**Yang, Z. (1994b).** Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**, 306–314.

**Yang, Z. (1997).** PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555–556.

**Yang, Z. & Bielawski, J. P. (2000).** Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496–503.

**Yu, K., Petrovsky, N., Schonbach, C., Koh, J. Y. & Brusic, V. (2002).** Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med* **8**, 137–148.

**Zaphiropoulos, P. G. (2002).** Template switching generated during reverse transcription? *FEBS Lett* **527**, 326.

**Zeuzem, S. (2000).** Hepatitis C virus: kinetics and quasispecies evolution during anti-viral therapy. *Forum* **10**, 32–42.