INTRODUCTION

Hepatitis C virus (HCV), is a positive-sense RNA virus in the family Flaviviridae, is a newly emerged human pathogen estimated to infect about 2.2% of the world’s population (Alter, 2007). HCV is not successfully cleared by the immune system in a substantial proportion of patients, and persistent HCV infection can result in severe liver disease, including cirrhosis and hepatocellular carcinoma (Lauer & Walker, 2001; Muller, 1996). HCV is highly polymorphic, and six major genotypes, including approximately 21 subtypes, have been identified on the basis of phylogenetic analyses (Simmonds et al., 2005; Timm & Roggendorf, 2007). Worldwide, the most common genotype is HCV-1b, while HCV-1a is widely distributed in Northern Europe and in North America (Simmonds et al., 2005). Because the polymorphism of HCV reflects a high mutation rate, it represents a potential challenge for the development of both vaccines and new drug treatments, given the apparent capacity of this virus to evolve evasive strategies (Timm & Roggendorf, 2007).

Both experimental infection of chimpanzees (Weiner et al., 1995; Erickson et al., 2001) and analyses of viral sequences derived from human patients (Guo et al., 2004; Seifert et al., 2004; Timm et al., 2004, 2007; Cox et al., 2005; Guglietta et al., 2005; Ray et al., 2005; Tester et al., 2005; Gaudieri et al., 2006; Poon et al., 2007; Neumann-Haefelin et al., 2008) have provided evidence that binding by the host class I major histocompatibility complex (MHC) and recognition by CD8+ cytotoxic T cells (CTL) can select for escape mutants in CTL epitopes of HCV. As in other flaviviruses, the genome of HCV encodes a polyprotein that is cleaved to form 10 proteins: the three structural proteins core, E1 and E2; p7, an integral membrane protein; and the six non-structural proteins, NS2, NS3, NS4A, NS4B, NS5A and NS5B. CTL epitopes have been reported from each of the 10 proteins, but the majority of reported cases of CTL escape mutants have involved NS3.

Functionally, NS3 plays a role in two proteases that process the non-structural proteins: the NS2-3 protease, consisting of NS2 and the N-terminal domain of NS3, and a serine protease constituted by the N-terminal 180 residues of NS3 and E2; p7, an integral membrane protein; and the six non-structural proteins, NS2, NS3, NS4A, NS4B, NS5A and NS5B. CTL epitopes have been reported from each of the 10 proteins, but the majority of reported cases of CTL escape mutants have involved NS3.

In spite of evidence that escape mutants in CTL epitopes may be selectively favoured within infected hosts, CTL...
epitopes show evidence of strong purifying selection at the population level, acting to eliminate most non-synonymous mutations (Hughes et al., 2007). This paradoxical finding might be explained by the hypothesis that CTL escape mutants, although providing a benefit to the virus in terms of immune evasion, also impose a cost to the virus in terms of decreased viability or transmissibility. Consistent with this hypothesis are observations that CTL epitopes at which escape mutations had previously occurred may, in the absence of the presenting class I MHC allele, revert to the population consensus sequence (Timm et al., 2004; Ray et al., 2005; Kuntzen et al., 2007). In addition, there is evidence that certain CTL escape mutants can be harmful to the virus; for example, mutants in the NS3 1073–1081 CTL epitope diminish the protease activity of NS3 (Söderholm et al., 2006).

Comparison of synonymous and non-synonymous nucleotide polymorphism is a major source of insight into the action of natural selection, and different aspects of the pattern of nucleotide polymorphism reveal different aspects of selection (Hughes, 1999; Hughes & Hughes, 2007). In the case of most protein-coding genes, the mean number of synonymous substitutions per synonymous site \( (d_s) \) exceeds the mean number of non-synonymous substitutions per non-synonymous site \( (d_N) \). Because most non-synonymous mutations are deleterious, this pattern is evidence of past purifying selection, acting to reduce the frequency of deleterious variants (Kimura, 1977).

Slightly deleterious mutations may not be eliminated immediately by natural selection but can persist in populations (Ohta, 1973). The existence of an excess of relatively rare non-synonymous polymorphisms is evidence that deleterious variants are currently in the process of being eliminated by natural selection (Tajima, 1989; Hughes et al., 2003; Hughes, 2005, 2007a; Hughes & Hughes, 2007; Hughes & Piontkivska, 2008). In many populations, such slightly deleterious variants may drift to relatively high frequencies during a bottleneck, when natural selection cannot efficiently remove them (Ohta, 1973). Subsequently, when population size increases, the frequency of such slightly deleterious alleles will decrease as a result of purifying selection (Hughes et al., 2003). In the case of CTL epitopes of HCV and other viruses, positive selection favouring the elimination of presentation by host class I MHC might similarly cause the increase of otherwise slightly deleterious mutations, which will then be subjected to purifying selection in the absence of the presenting class I allele.

Here, we analysed coding sequence polymorphism in complete polyprotein-encoding sequences of HCV-1b, the most common genotype of HCV worldwide, in order to test the hypothesis that CTL epitopes are subjected to conflicting evolutionary pressures. We tested for evidence of conflicting pressures by examining the patterns of synonymous and non-synonymous polymorphism in known CTL epitopes and in other regions of the 10 proteins. Because of the importance of the CD4+ (helper) T-cell response in clearing HCV infection (Gerlach et al., 2005), there has also been interest in natural selection acting on HCV epitopes presented to CD4+ T cells by class II MHC (Wang et al., 2002). Therefore, we compared the patterns of polymorphism in CTL epitopes with that in epitopes presented to CD4+ T cells.

**METHODS**

**Sequences analysed and phylogenetic analysis.** We analysed six complete polyprotein-encoding sequences of HCV-1a and 91 complete polyprotein-encoding sequences of HCV-1b (Supplementary Table S1 available in JGV Online) from the Los Alamos HCV Sequence Database, available at http://hcv.lanl.gov/content/sequence/hcvTools/Outlines.html (Kuiken et al., 2005). Sequences from the database containing undetermined nucleotides, premature stop codons and/or gaps were excluded from the analysis; recombinant sequences were also excluded from the analysis. We designate predicted epitopes presented by HLA class I MHC to CD8+ T cells and those presented by HLA class II to CD4+ T cells as CTL and CD4 epitopes, respectively. Both categories of epitopes in the HCV proteins were identified following the Los Alamos Hepatitis C Immunology Database (http://hcv.lanl.gov/content/immuno/immuno-main.html). Only epitopes from human with known HLA presenting antigens were included (Yusim et al., 2005; Supplementary Tables S2 and S3 available in JGV Online).

Using the PAUP* program, version 4.0b10 (Swofford, 2003), we constructed a neighbour-joining (NJ) tree (Saitou & Nei, 1987) of HCV-1b sequences based on the proportion of nucleotide distance, rooted with HCV-1a sequences (Supplementary Fig. S1 available in JGV Online). The bootstrap method was used to assess the reliability of internal branches in the phylogenetic tree (Felsenstein, 1985), 1000 bootstrap replicates were used (Supplementary Fig. S1). The phylogenetic tree was used to identify 28 phylogenetically independent ‘sister pairs’ of closely related sequences (Supplementary Fig. S1). Assuming this tree, we used the maximum-parsimony method (Swofford, 2003) to reconstruct the last common ancestor for HCV-1b sequences. We compared the reconstructed ancestral sequence with the consensus sequence of our 91 HCV-1b sequences computed using the consensus tool available from the Los Alamos Hepatitis C Immunology Database.

Using the MEGA 4.0 program (Tamura et al., 2007), we reconstructed the phylogeny using a number of additional distances, including the maximum composite nucleotide distance, the Poisson amino acid distance and the JTT amino acid distance. All of these distances yielded results very similar to that based on the proportion of nucleotide difference (data not shown).

We computed the consistency index (CI; Swofford, 2003) for each site in the NJ tree and in the 95% bootstrap consensus tree (a tree in which all branches receiving less than 95% bootstrap support were collapsed). CI is defined for a given site as the ratio of the minimum possible number of changes at a site to the number of changes at that site inferred assuming a given tree (Supplementary Fig. S2 available in JGV Online). CI is thus inversely related to the degree of homoplasy (forward-and-backward or parallel evolutionary changes) occurring at a site assuming a given tree. In other words, as the number of forward-and-backward and parallel changes increase, CI decreases. In computation of CI, we excluded singleton sites (i.e. sites at which a difference from the most abundant nucleotide was found in only one of the sequences analysed) because CI at such sites is trivial.

**Nucleotide substitution and diversity.** For the epitope and non-epitope domains of each protein, we estimated the number of...
synonymous substitutions per synonymous site \((d_S)\) and the number of non-synonymous substitutions per non-synonymous site \((d_K)\) between the two members of each sister pair by the Nei and Gojobori’s method (Nei & Gojobori, 1986) using the MEGA 3.1 software (Kumar et al., 2004). In preliminary analyses, we also estimated \(d_S\) and \(d_K\) by two more complex methods: Li’s (Li, 1993) method and the Yang and Nielsen’s method (Yang & Nielsen, 2000). The results of all three methods were almost identical, as is expected when sequences are closely related and thus a simple substitution model is expected to perform as well as a more complicated one (Nei & Kumar, 2000; Hughes & French, 2007).

We estimated gene diversity at individual nucleotide sites by the formula:

\[
1 - \sum_{i=1}^{n} \chi_i^2
\]

where \(n\) is the number of alleles and \(\chi_i\) is the population frequency of the \(i\)th allele (Nei, 1987; p. 177). Single nucleotide polymorphisms were classified either as synonymous or non-synonymous depending on their effect on the encoded nucleotide sequence. We excluded ambiguous sites at which both synonymous and non-synonymous variants occurred or at which the polymorphism could be considered synonymous or non-synonymous depending on the pathway taken by evolution.

In preliminary analyses, no significant differences in gene diversity were seen among CTL epitopes presented by different HLA class I alleles (data not shown). Likewise, no significant differences in gene diversity were seen among CD4 epitopes presented by different HLA class II alleles (data not shown). Therefore, CTL epitopes presented by different alleles and loci were combined for purposes of statistical analysis. In this connection, it is worth noting that several amino acid residues formed part of epitopes presented by more than one allele and sometimes by alleles at different loci (Supplementary Tables S2 and S3). Likewise, CTL and CD4 epitopes sometimes overlapped.

In comparing epitope and non-epitope regions, we defined as epitopes all those listed in Supplementary Tables S2 and S3. Although most of these epitopes were originally discovered in HCV-1a rather than in HCV-1b, many were conserved in at least some HCV-1b sequences. In our data, of 938 sites in CTL-epitope regions that were polymorphic in our HCV-1b sample, 882 (94.0 %) were in codon positions at which the amino acid reported for the original epitope was found in at least some of the HCV-1b sequences in our population. Moreover, 574 of 938 (61.2 %) polymorphic sites in CTL epitopes occurred in sites within reported CTL epitopes that were 100 % conserved (at all amino acid sites) in at least some of the HCV-1b sequences in our sample. Similarly, 331 of 545 (73.7 %) polymorphic sites in CD4 epitopes were in codon positions at which the amino acid reported for the original epitope was found in at least some of the sequences in our population. In addition, 398 of 545 (73.0 %) polymorphic sites in CD4 epitopes occurred at sites within reported CD4 epitopes that were 100 % conserved (at all amino acid sites) in at least some of the HCV-1b sequences in our sample.

In preliminary analyses, we defined as epitopes only those that were 100 % conserved in at least one sequence in our dataset. The results, when we used this restricted dataset, were essentially identical to those obtained using all epitopes listed in Supplementary Tables S2 and S3. The definition of epitopes based on Supplementary Tables S2 and S3 is more conservative since, in the absence of experimental data, it is uncertain whether any given fixed amino acid difference in these regions between HCV-1a and HCV-1b removes the epitope. Therefore, we report below only the results using all epitopes listed in Supplementary Tables S2 and S3.

In statistical analyses we used robust methods that avoid the statistically undesirable properties of model-dependence (Hughes et al., 2006). Because gene diversity at polymorphic sites was not normally distributed, non-parametric methods were used to analyse gene diversity (Hollander & Wolfe, 1973). All statistical analyses were conducted using the Minitab statistical package, release 13 (http://www.minitab.com/). We did not use the so called codon-based methods of analysis because they depend on several questionable assumptions, most notably the unwarranted assumption that the existence of one or more codons with \(d_S > d_K\) implies positive selection (Hughes, 2007b).

**RESULTS**

**Nucleotide substitution between sister pairs**

In comparisons between the members of 28 phylogenetically independent sister pairs of HCV-1b genomes, the median number of synonymous substitutions per synonymous site \((d_S)\) was significantly greater than the median number of non-synonymous substitutions per non-synonymous site \((d_K)\) in both CTL-epitope and non-CTL-epitope regions of each of the 10 proteins (Sign test, \(P<0.001\) in each case, Table 1). Likewise, mean \(d_S\) was significantly greater than mean \(d_K\) in both CTL-epitope and non-CTL-epitope regions of each of the 10 proteins (paired \(t\)-test, \(P<0.001\) in each case, Table 1). The pattern of \(d_S > d_K\) is evidence of purifying selection on both CTL-epitope and non-CTL-epitope regions.

Median \(d_S\) in CTL-epitope regions differed significantly from that in non-CTL-epitope regions in the case of three of the 10 proteins: E2, NS3 and NS5A (Table 1). In the case of E2, median \(d_S\) in the non-CTL-epitope regions was significantly greater than that in the CTL-epitope regions, whereas in NS3 and NS5A, median \(d_S\) in the CTL-epitope regions was significantly greater than that in the non-CTL-epitope regions (Table 1). E2 was thus the only protein for which there was evidence of stronger purifying selection on the CTL-epitope than on non-CTL-epitope regions, while in NS3 and NS5A there was evidence that purifying selection was stronger on non-CTL-epitope regions than on CTL-epitope regions. In non-epitope regions, NS3 had the lowest mean \(d_N\) and the lowest median \(d_N\) of all 10 proteins (Table 1).

**Synonymous and non-synonymous polymorphism**

The ratio of non-synonymous to synonymous polymorphic sites in the complete polyprotein \((1509:2172 or 0.69)\) differed significantly from the ratio of non-synonymous to synonymous changes \((411:900 or 0.46)\) reconstructed by maximum-parsimony as having been fixed in the ancestor of HCV-1b \((\chi^2=38.0, 1\ \text{d.f.}, P<0.001)\). Thus, HCV-1b showed a relative excess of non-synonymous polymorphic sites. In order to test for ongoing purifying selection at these sites, we estimated gene diversity at individual polymorphic sites. For the complete polyprotein, median gene diversity at synonymous
HCV-1b between 28 phylogenetically independent pairs of genomes. However, there was evidence that this process did not occur with equal likelihood at synonymous and non-synonymous polymorphic sites (0.160, n=2172) was significantly greater than that at non-synonymous polymorphic sites (0.043, n=1509, Mann–Whitney test P<0.001), as expected in the case of ongoing purifying selection against many non-synonymous variants.

**Reconstructed ancestral sequence**

We applied the maximum-parsimony method to reconstruct the last common ancestor of our sample of HCV-1b sequences, assuming the NJ tree. The reconstructed ancestral sequence was not identical to the consensus sequence of the HCV-1b sequences. In fact, 300 of 9018 sites (3.3%) differed between the reconstructed ancestral sequence and the consensus sequence. The per cent difference at synonymous sites was 10.1%, whereas that at non-synonymous sites was 1.0%.

When we plotted gene diversity against the frequency in our sample of the nucleotide found in the reconstructed ancestor of HCV-1b (the ancestral nucleotide), we found a parabolic relationship for both synonymous and non-synonymous polymorphic sites (Fig. 1). Gene diversities were highest at sites where the ancestral nucleotide occurred at an intermediate frequency, but low at sites where the ancestral nucleotide was either very common or very rare (Fig. 1). This pattern is consistent with a pattern of nucleotide turnover at many sites, with the ancestral nucleotide being replaced by a new nucleotide, which in some cases is increasing near to fixation.

However, there was evidence that this process did not occur with equal likelihood at synonymous and non-synonymous sites. Median frequency of the ancestral nucleotide was significantly greater in the case of non-synonymous polymorphic sites (0.978) than in the case of synonymous polymorphic sites (0.912, Mann–Whitney test, P<0.001). This difference in median frequency evidently reflected the fact that there were fewer non-synonymous than synonymous sites at which the ancestral nucleotide occurred with very low frequency (Fig. 1). At 182 of 2172 (8.4%) synonymous polymorphic sites, the

![Fig. 1. Gene diversity at individual polymorphic nucleotide sites in the HCV-1b polyprotein coding region plotted against the frequency of the reconstructed ancestral nucleotide. Each point corresponds to a single polymorphic site, either synonymous or non-synonymous.](image)

<table>
<thead>
<tr>
<th>Protein</th>
<th>CTL epitope</th>
<th>Non-CTL epitope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(d_k \pm \text{SEM})</td>
<td>(d_N \pm \text{SEM})</td>
</tr>
<tr>
<td>Core</td>
<td>0.0684 ± 0.0107 (0.0649)</td>
<td>0.0056 ± 0.0012 (0.0055)</td>
</tr>
<tr>
<td>E1</td>
<td>0.1585 ± 0.0210 (0.1341)</td>
<td>0.0153 ± 0.0026 (0.0144)</td>
</tr>
<tr>
<td>E2</td>
<td>0.2335 ± 0.0260 (0.2511)</td>
<td>0.0488 ± 0.0049 (0.0055)</td>
</tr>
<tr>
<td>p7</td>
<td>0.2011 ± 0.0380 (0.1367)</td>
<td>0.0236 ± 0.0066 (0.0000)</td>
</tr>
<tr>
<td>NS2</td>
<td>0.1773 ± 0.0201 (0.1895)</td>
<td>0.0304 ± 0.0012 (0.0055)</td>
</tr>
<tr>
<td>NS3</td>
<td>0.1909 ± 0.0213 (0.2047)</td>
<td>0.0181 ± 0.0020 (0.0193)</td>
</tr>
<tr>
<td>NS4A</td>
<td>0.2307 ± 0.0275 (0.2132)</td>
<td>0.0125 ± 0.0029 (0.0000)</td>
</tr>
<tr>
<td>NS4B</td>
<td>0.1928 ± 0.0224 (0.1861)</td>
<td>0.0123 ± 0.0020 (0.0116)</td>
</tr>
<tr>
<td>NS5A</td>
<td>0.1770 ± 0.0224 (0.1710)</td>
<td>0.0295 ± 0.0033 (0.0296)</td>
</tr>
<tr>
<td>NS5B</td>
<td>0.1244 ± 0.0134 (0.1203)</td>
<td>0.0094 ± 0.0018 (0.0069)</td>
</tr>
</tbody>
</table>

*Median \(d_N\) in non-CTL-epitope regions differed significantly from that in CTL-epitope regions (Sign test, two-tailed P<0.01). Mean \(d_N\) in non-CTL-epitope regions differed significantly from that in CTL-epitope regions (paired t-test, two-tailed P<0.001).
frequency of the ancestral nucleotide was less than 0.5, whereas in only 49 of 1509 non-synonymous sites (3.2%) the frequency of the ancestral allele was less than 0.5. The difference in proportions between synonymous and non-synonymous sites was highly significant ($\chi^2 = 39.9$, 1 d.f., $P<0.001$). Thus non-ancestral nucleotides were more likely to reach high frequencies at synonymous sites than at non-synonymous sites.

**Polymorphic sites in epitopes**

In all proteins and in both epitopes and non-epitope regions, gene diversities at synonymous polymorphic sites exceeded those at non-synonymous polymorphic sites and the difference was highly significant in the case of most of the proteins (Supplementary Fig. S3 available in JGV Online). However, in NS3, median gene diversity at non-synonymous polymorphic sites in non-CTL-epitope regions (0.022) was significantly lower than that at non-synonymous polymorphic sites in CTL-epitope regions (0.054, Fig. 2a). This pattern was not seen in any other protein (Supplementary Fig. S3). In fact, for all proteins except NS3, the median gene diversity was identical (0.043) at non-synonymous polymorphic sites in CTL-epitope and non-CTL-epitope regions (Fig. 2b). Median gene diversity at non-synonymous polymorphic sites in CTL epitopes in NS3 was not significantly different from that in the other proteins, but median gene diversity at non-synonymous polymorphic sites in non-CTL-epitope regions of NS3 was significantly lower than that in non-CTL-epitope regions of the other proteins (Mann–Whitney test, $P<0.001$).

In order to compare patterns of non-synonymous polymorphism in CD4 epitopes with that in CTL epitopes, we categorized polymorphic sites as follows: (i) not in either a CD4 or a CTL epitope; (ii) in a CD4 epitope only; (iii) in a CTL epitope only; and (iv) in both CD4 and CTL epitopes (Fig. 3). In the case of NS3, there was a highly significant difference ($P<0.001$, Kruskal–Wallis test) in median gene diversity among the four categories (Fig. 3a). The highest median gene diversity (0.064) was in sites in CTL epitopes only, while median gene diversity in sites in CD4 epitopes only (0.022) was identical to that in sites in neither kind of epitope (Fig. 3a).

On the other hand, in all other proteins except NS3, there was not a significant difference in median gene diversity among the four categories of sites (Fig. 3b). Median gene diversity at sites in neither kind of epitope was significantly less in NS3 (0.022) than in the other proteins (0.043, $P<0.01$, Mann–Whitney test, Fig. 3). Similarly, median gene diversity at sites in CD4 epitopes only was significantly less in NS3 (0.022) than in the other proteins (0.054, $P<0.01$, Mann–Whitney test, Fig. 3). On the other hand, median gene diversity at sites in CTL epitopes only was significantly greater in NS3 (0.064) than in the other proteins (0.043, $P<0.05$, Mann–Whitney test, Fig. 3). At sites included in both CD4 and CTL epitopes, NS3 and the other proteins were not significantly different and in fact showed essentially identical median values of gene diversity (0.043).

**CI**

The CI was measured at individual non-synonymous polymorphic sites, excluding singleton sites, based on the NJ tree (Supplementary Fig. S1). In NS3, median CI values differed significantly among non-synonymous sites categorized by their occurrence in CD4 epitopes and/or CTL epitopes ($P<0.01$, Kruskal–Wallis test, Fig. 4a). Similarly, in the other proteins, median CI values differed significantly among non-synonymous sites categorized by their occurrence in CD4 epitopes and/or CTL epitopes ($P<0.01$, Supplementary Fig. S1).
Kruskal–Wallis test, Fig. 4b). However, NS3 showed a distinctly different pattern than that seen in the other proteins. In sites in CD4 epitopes but not in CTL epitopes, median CI in NS3 (0.450) was significantly greater than that in other proteins (0.143, \(P < 0.01\), Mann–Whitney test, Fig. 4). Conversely, in sites in CTL epitopes but not in CD4 epitopes, median CI in NS3 (0.211) was significantly less than that in other proteins (0.333, \(P < 0.01\), Mann–Whitney test, Fig. 4). Essentially the same pattern was seen when CI was calculated based on the 95% bootstrap consensus tree, rather than the NJ tree (data not shown).

The low median CI in NS3 at non-synonymous polymorphic sites in CTL epitopes, but not in CD4 epitopes, was evidently due to an excess of sites with very low CI. In NS3, the 40 non-synonymous polymorphic sites in CTL epitopes, but not in CD4 epitopes, included 11 (27.5%) with CI \(\leq 0.1\) on the basis of the NJ tree. By contrast, in the other proteins, only 13 of 133 (9.7%) such sites had CI \(\leq 0.1\). The difference between the two proportions was highly significant (\(\chi^2 = 8.1, 1\) d.f., \(P < 0.01\)). There was a similar excess of sites with very low CI among non-synonymous polymorphic sites in CD4 epitopes, but not in CTL epitopes in proteins other than NS3. Excluding NS3, the 19 non-synonymous polymorphic sites in CD4 epitopes, but not in CTL epitopes, included 13 (33.3%) with CI \(\leq 0.1\). By contrast, in NS3, only 2 of 18 (11.1%) such sites had CI \(\leq 0.1\). However, in the latter case, the difference was not statistically significant (\(\chi^2 = 3.8, 1\) d.f., \(P = 0.051\)). The results using CI based on the 95% bootstrap consensus tree were essentially the same (data not shown).
DISCUSSION

Analysis of complete genome sequences of HCV-1b showed strong evidence of purifying selection on protein-coding regions, including those that encode CTL epitopes. In comparisons between members of phylogenetically independent pairs of genomes, the number of synonymous substitutions per synonymous site ($d_s$) significantly exceeded the number of non-synonymous substitutions per non-synonymous site ($d_{ns}$) in each of the 10 proteins making up the viral polyprotein. Moreover, the gene diversity within the HCV-1b population was significantly lower at non-synonymous polymorphic sites than at synonymous polymorphic sites, evidence that there are abundant slightly deleterious non-synonymous variants subject to ongoing purifying selection.

Both synonymous and non-synonymous polymorphic sites showed a pattern in which certain new mutations had increased in frequency, in some cases eventually becoming more frequent than the ancestral nucleotide. However, this pattern of nucleotide turnover was much more evident in the case of synonymous polymorphic sites than in the case of non-synonymous polymorphic sites. This difference between synonymous and non-synonymous variants reflects the effect of purifying selection on many of the latter, limiting a free turnover of alleles.

In discussing the evolution of HCV, a number of authors have used the consensus sequence as a reference, referring to both evolution away from the consensus and reversion to the consensus (e.g. Timm et al., 2004; Ray et al., 2005; Kuntzen et al., 2007). However, our ancestral sequence reconstruction showed that the ancestral sequence and the consensus sequence are not necessarily highly similar. Indeed, in the case of the HCV-1b genomes we analysed, they differed at 300 nt sites. Our evidence of allelic turnover provides a biological explanation for this difference. Turnover of the ancestral nucleotide at a number of sites will cause the consensus sequence, which is after all merely a statistical abstraction that may not correspond to any real sequence, to diverge from the ancestral sequence.

Nucleotide sequence polymorphism in the NS3 protein, while consistent with the pattern of purifying selection observed in the other proteins, revealed certain differences. In the case of NS3, median gene diversity at non-synonymous polymorphic sites outside either CD4 epitopes or CTL epitopes was unusually low. Moreover, gene diversities at non-synonymous polymorphic sites in CTL epitopes but not in CD4 epitopes were unusually high (Figs 2 and 3). In NS3, median $d_{ns}$ in comparisons between independent pairs of genomes was significantly greater in the CTL epitopes than in the non-epitope regions of NS3 (Table 1). Also, NS3 showed a lower $d_s$ in non-epitope regions than any other protein (Table 1). Thus, the results identified NS3 as a highly conserved protein, with strong ongoing purifying selection, but also revealed less evidence of purifying selection acting on the CTL epitopes of NS3 than in the other proteins.

In addition to a relative relaxation of purifying selection, non-synonymous polymorphic sites in the CTL epitopes, but not in CD4 epitopes, of NS3 showed an unusually low median CI, whether CI was based on the NJ tree or the 95% bootstrap consensus tree (Fig. 4). Moreover, sites in CTL epitopes, but not CD4 epitopes, of NS3 showed a relative excess of non-synonymous polymorphic sites with very low CI; in other words, sites with a high incidence of forward-and-backward or parallel non-synonymous nucleotide changes. The high incidence of forward-and-backward or parallel non-synonymous changes in CTL epitopes of NS3 indicates the frequent re-occurrence of the same non-synonymous mutations in the codons encoding these epitopes. Interestingly, in proteins other than NS3, non-synonymous polymorphic sites in CD4 epitopes, but not in CTL epitopes, were characterized by low median CI (Fig. 4). Although the numbers of such sites were small, this pattern suggests a similar high incidence of forward-and-backward or parallel non-synonymous nucleotide changes at CD4 epitopes in proteins other than NS3.

A pattern of numerous forward-and-backward or parallel non-synonymous nucleotide changes is most easily explained by the occurrence of escape mutations, which are selectively favoured within certain hosts but are subject to purifying selection at the population level. Consistent with this interpretation was the observation that the non-synonymous polymorphic sites with very low CI included certain sites in CTL epitopes at which apparent escape mutations have been reported in virus samples from patients. For example, the Y→F replacement at aa 1444 in NS3 was reported to occur independently in different patients and to reduce CTL response (Cox et al., 2005). The corresponding nucleotide site was polymorphic in our population of HCV-1b genomic sequences, with low CI (0.067 based on the NJ tree and 0.033 based on the 95% bootstrap consensus tree). Neumann-Haefelin et al. (2008) similarly reported evidence of frequent reversals of this same mutation throughout a phylogeny of HCV-1b. Likewise, a G→S replacement at aa 1409 in the CTL epitope NS3 1406–1415 was associated in a population study with the presence in the host of the presenting class I MHC molecule A*02 (Ray et al., 2005). The two polymorphic non-synonymous sites contributing to this amino acid change had low CI (0.143 and 0.167, respectively, based on the NJ tree; and 0.133 and 0.087 based on the 95% bootstrap consensus tree).

In our dataset, the CTL epitope NS3 1073–1081 included two non-singular polymorphic non-synonymous sites with low CI: (i) a site causing amino acid change I→V at residue 1074 (CI=0.053 based on the NJ tree and 0.043 based on the 95% bootstrap consensus tree); and (ii) a site causing amino acid change V→A at residue 1077 (CI=0.143 based on both trees). Both of these amino acid replacements caused reduced recognition by human HLA-A2-restricted
CTLs; the latter reduced HLA-A2-binding avidity (Söderholm et al., 2006).

The low CI values at sites causing possible escape mutants indicate that mutations at these sites are frequently re-occurring but are not fixed. Purifying selection is the most likely force acting to prevent population fixation of such mutations, which are presumably advantageous to the virus in hosts possessing the presenting class I MHC molecule. Thus, our results point to a conflict between positive selection favouring escape mutants and purifying selection acting to remove them.

In order to understand the population genetic processes involved in such a conflict, it is helpful to consider the case of a hypothetical CTL epitope presented by a given class I HLA molecule. Suppose that, in a host expressing an HLA molecule, an escape mutation occurs in the virus; if that escape mutation is in no way harmful to the virus but provides an advantage in preventing recognition by a frequently encountered host HLA allele, population genetics theory predicts that it is likely to become fixed in the entire viral population, as long as the virus continues to encounter hosts expressing that HLA molecule. The speed of fixation will be greater if the HLA molecule has a high frequency in the host population, but, assuming a large viral population, fixation is likely to occur eventually even if the HLA molecule is relatively uncommon. Likewise, it should be expected that escape mutants in an epitope more commonly targeted by host T cells will be more advantageous and thus more likely to be fixed than in the case where the epitope is rarely targeted.

On the other hand, if the escape mutation imposes some fitness cost on the virus, purifying selection will constantly act to decrease the frequency of the escape mutant, whenever the presenting HLA molecule is absent. Such purifying selection would be expected to be particularly effective if the escape mutant confers relatively little advantage; for example, if it is in an epitope rarely targeted by host T cells. However, if the escape mutant confers a substantial advantage in the presence of the presenting HLA allele, the escape mutant will be continually reintroduced to the population and then reduced in frequency by purifying selection. The result will be a continual process of forward-and-backward or parallel non-synonymous nucleotide substitutions in the CTL epitope in question.

In summary, our results support the hypothesis that certain mutants in CTL epitopes of HCV-1b are subjected to conflicting evolutionary pressures: positive selection within hosts expressing the presenting class I MHC molecule and purifying selection in the population at large. The fact that this pattern is most strikingly observed in the case of NS3 suggests that the evolutionary conflict between immune escape and functional constraint on the protein is more acute in the case of NS3 than any of the other proteins of HCV-1b. In future studies, it will be important to test this hypothesis further, using population data from both HCV-1b and other HCV genotypes, particularly HCV-1a.

ACKNOWLEDGEMENTS

This research was supported by the grant GM43940 from the National Institutes of Health to A.L.H.

REFERENCES


