TITLE

Analysis of the evolution and structure of a complex intra-host viral population in chronic Hepatitis C mapped by ultra-deep pyrosequencing.

Running Title

Informed analysis of HCV using UDPS

Authors

Brendan A. Palmer\textsuperscript{a}, Zoya Dimitrova\textsuperscript{b}, Pavel Skums\textsuperscript{b}, Orla Crosbie\textsuperscript{c}, Elizabeth Kenny-Walsh\textsuperscript{c} and Liam J. Fanning\textsuperscript{a#}.

Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland\textsuperscript{a}; Division of Viral Hepatitis, Centers of Disease Control and Prevention, Atlanta, Georgia, USA\textsuperscript{b}; Department of Gastroenterology, Cork University Hospital, Cork, Ireland\textsuperscript{c}.

Abstract word count: 195

Word count: 5735

Figures: 5

Tables: 5

# Corresponding author

Dr. Liam J. Fanning, l.fanning@ucc.ie
ABSTRACT
Hepatitis C virus (HCV) causes chronic infection in up to 50-80% of infected individuals. Hypervariable region 1 (HVR1) variability is frequently studied to gain an insight into the mechanisms of HCV adaptation during chronic infection, but the changes to and persistence of HCV subpopulations during intra-host evolution are poorly understood. In this study, we have used ultra-deep pyrosequencing (UDPS) to map viral heterogeneity of a single patient over 9.6 years of chronic HCV genotype 4a infection. Informed error correction of the raw UDPS data was performed using a temporally matched clonal data set. The resultant data set reported the detection of low frequency recombinants throughout the study period implying that recombination is an active mechanism through which HCV can explore novel sequence space. The data indicates that poly-virus infection of hepatocytes has occurred but that the fitness quotients of recombinant daughter virions are too low to compete against the parental genomes. The subpopulations of parental genomes contributing to the recombination events highlighted a dynamic virome where subpopulations of variants are in competition. In addition, we provide direct evidence that demonstrates the growth of subdominant populations to dominance in the absence of a detectable humoral response.

IMPORTANCE
Analysis of ultra-deep pyrosequencing data sets derived from virus amplicons frequently relies on software tools that are not optimised for amplicon analysis, assume random incorporation of sequencing errors and are focused on achieving higher specificity at the expense of sensitivity. Such analysis is further complicated by the presence of hypervariable regions. In this study, we make use of a temporally matched reference
sequence data set to inform error correction algorithms. Using this methodology we were able to (1) detect multiple instances of Hepatitis C virus intra-subtype recombination at the E1/E2 junction (a phenomenon rarely reported in the literature) and (2) interrogate the longitudinal quasispecies complexity of the virome. Parallel to the UDPS, isolation of IgG-bound virions was found to be co-incident with the collapse of specific viral subpopulations.
INTRODUCTION

Hepatitis C virus (HCV) is a single-stranded positive-sense RNA virus belonging to the Flaviviridae family. Chronic HCV infection can remain asymptomatic for decades. HCV infection is ultimately associated with gradual loss of liver function making it the leading global etiological cause of liver-specific morbidity and mortality (1). The small RNA genome (~9.6 kb) encodes two envelope glycoproteins E1 and E2 which form non-covalent heterodimers with E2 specifically identified as playing important roles in host cell recognition and humoral immune evasion (2-4). During the acute stages of infection, transmission bottlenecks contribute to an initial collapse of genomic diversity, yet rapid evolution and adaptation results in numerous distinct variants or quasispecies (5, 6). The causative protein for this mutational change is the low fidelity RNA dependent RNA polymerase, which incorporates approximately one mutation per genome replication event (7).

The first neutralising epitopes were described for HCV within hypervariable region 1 (HVR1) (8). The HVR1 comprises the 27 N-terminal amino acids of E2 and displays the greatest heterogeneity of the entire HCV genome. Mutational change at the HVR1 over time averts epitope recognition and contributes towards immune escape (4, 8, 9). While linear and conformational sensitive epitopes outside the HVR1 have also been identified as antigenic targets, it is the HVR1 that is immunodominant (10, 11). HVR1 variability is frequently studied to gain an insight into the adaptation of HCV to external selection pressures (12-14). There is evidence to suggest that humoral immune evasion is host-specific. Analysis of a cohort of women, 20 years after infection from a common source outbreak, identified patient specific evolution at the HVR1 (15). More recent longitudinal
studies of HVR1 variability have identified collapses in sequence heterogeneity, with latter sampling points dominated by a handful of host-adapted variants (16-18). Humoral immune evasion is not confined to amino acid modulation of the antigenic epitopes. Evidence is accumulating that gain/loss of glycosylation acceptor sites also contribute to concealment from neutralising antibodies (nAb) (19, 20). Additionally, the observation that HCV can infect neighbouring cells by direct cell-to-cell transmission identifies an innovative mechanism to avoid nAb targeting (21, 22). Collectively, these observations suggest mechanisms through which highly refined, host-adapted sequences can evolve. The stochastic exploration of novel sequence space can also be facilitated through recombination, yet reports of recombination in HCV are rare (Reviewed in 23). To date a single recombination breakpoint spanning the NS2/NS3 gene junction for HCV inter-genotypic recombinants has been identified (23, 24). This contrasts with intra-genotype and intra-subtype recombinants that can accommodate recombination breakpoints at multiple sites along the genome, including the E1/E2 junction (23-26). Furthermore, intra-genotypic recombination has been observed to occur at greater frequencies in vitro than inter-genotypic recombination (27). Taken together, this implies a greater acceptance towards genetic exchange where there is an inherent shared homology between donor sequences. The sample set used in this study is from a treatment naïve individual chronically infected with HCV genotype 4a that has been reported previously at clonal level (16). The virome contained two readily distinguishable lineages (L1 and L2). The presence of an in-frame 3 bp insertion at the 5′-end of E2 in a proportion of L1 haplotypes, giving
The initial complexity of the sequences observed gave way to a monophyletic population after 8.6 years. The presence of such well-defined populations facilitated the identification of rarely reported intra-subtype recombinants (16). In the present study, we take advantage ultra-deep pyrosequencing (UDPS), the availability of additional samples and the phylogenetic diversity to broaden our investigation into the basal recombination potential of HCV in vivo. Additionally, we chronicle the temporal dominance of a third sublineage of L1, which had previously been indiscernible. Finally, the rise of the monophyletic population to dominance is demonstrated to occur in the absence of a detectable, specific antibody response. This was observed to coincide with the collapse of the competing lineage that was subject to humoral targeting. These observations were supported by the underlying patterns of HVR1 evolution.
MATERIALS AND METHODS

Sample set. Ten samples (RL1-10), encompassing 9.6 years of chronic HCV genotype 4a infection from a single, treatment naïve patient, and a homogenous plasmid control template of known sequence (RL11, GQ985374) were subjected to pyrosequencing analysis in this study. RL5 has not been previously reported and was selected for analysis as it transected the 2.7 year gap between RL4 and RL6. RL12 (available subsequent to UDPS) was analysed clonally extending the study timeframe to 10.6 years. A waiver of consent was provided by Clinical Research Ethics Committee of the Cork Teaching Hospitals as the samples used in this study were surplus to requirements following diagnostic investigations.

Pyrosequencing of the HVR1 region. RNA from patient serum samples was prepared and a 321 bp fragment spanning the E1/E2 region was amplified as previously described (16). The amplified fragment corresponds to positions 1209-1526 of a reference genotype 4a strain (Accession number DQ418782). The extension times used were threefold longer than those recommended for the polymerase (45 s/kb for PWO, Roche) to reduce in vitro artefacts during PCR (28). Measures to protect against inter-sample contamination were employed (29). The mean viral titer was 5.9 HCV RNA log_{10} IU/ml (range 5.1 - 6.6 log_{10} IU/ml). To ensure that the initial amount of template was not limiting we performed a 1:100 dilution of the viral RNA which yielded an amplicon visualised by gel electrophoresis for each sample. Amplicons were purified using a PCR purification kit (Qiagen) and quantified using a Biophotometer (Eppendorf). Samples were prepared in equimolar concentrations and diluted to a final concentration of $1 \times 10^7$ molecules/ml. Pyrosequencing was performed using the 454 GS FLX titanium platform with sample...
specific multiplex identifier sequence adapted libraries for Lib-1 sequencing (Roche 454 Life Sciences, Branford, CT, USA).

**Extraction of IgG-bound virions from serum.** Separation of IgG-bound virions was performed using a Qproteome Albumin/IgG Depletion kit (Qiagen) as previously described (30). Briefly, 25 µl of serum (the carrying capacity of the Albumin/IgG depletion column) was diluted with PBS to a final volume of 100 µl and incubated with end over end mixing for 5 min at RT. The flow-through was collected by centrifugation and the column was subjected to serial washes with PBS (n=4) to remove any residual unbound virus followed by on-column lysis of virions. Where the above protocol did not yield viable sequences, the result was confirmed by repetition of the full procedure with precipitation of extracted RNA, followed by nested PCR.

**Data handling and error correction.** The raw sff data files were managed using SFFFile tools (Roche). Low quality reads and reads shorter than 90% of the expected amplicon lengths were removed. The obtained data sets were processed using a multi-step sequencing error correction and local haplotype reconstruction pipeline described below.

Previous characterisation of the quasispecies population within the sample set revealed the presence of complex mix of variants with fluctuating lineage and sublineage composition over time (16). Preliminary phylogenetic analysis of the UDPs data revealed the existence of a clearly distinguishable third sublineage of L1 (hereafter L1c) in RL1 and RL5. Eight 24 bp HVR1 motifs, that defined the haplotype groups L1a, L1b, L1c and L2, were subsequently generated (Table 1). To increase the sensitivity of sequencing
error correction algorithms, we partitioned the data by (sub)lineage according to the presence of corresponding motifs and corrected sequencing errors in each (sub)lineage separately. Moreover, in order to ensure the quality of the analysed data and the absence of PCR and sequencing chimeras, reads that had more than 3 bp difference from the best matching sequence from this motif set were removed. It should be noted that the removed reads have a very low quality alignment with the reference sequences, which may indicate that the majority of these sequences are indeed PCR or sequencing chimeras. The obtained data sets were processed by the sequential application of algorithms k-mer error correction (KEC) and a customized version of empirical threshold (ET) (31). Skums et al. (2012) have previously demonstrated this process to be highly accurate in finding true haplotypes and removing false haplotypes.

KEC consists of the three stages. Stage 1: The set of k-mers (substring of fixed length k) of reads from the processed data set are calculated and the distribution of frequencies of k-mers is analysed (31). It was observed, that the frequencies of erroneous and correct k-mers follow different distributions (32-34). Based on this fact, the error threshold is calculated as the minimal frequency of k-mers separating two different distributions.

Stage 2: k-mers with frequencies lower than the error threshold are considered erroneous and used to identify and correct the errors. The corrections are based on an analysis of different factors, including the length of a segment of consecutive erroneous k-mers, sequences of nucleotides at the end of that segment and the frequencies of the similar correct k-mers. The procedure of error correction is repeated iteratively i times. Stage 3: The reads containing k-mers that were not corrected in stage 2 are discarded. The following parameters of KEC were used: k=25, i=3. To further identify and correct
homopolymer errors the data was post-processed using ET, using a reference clone (GQ985371) as an external reference and 166 unique clonal sequences as internal references (16). All haplotypes retained following application of this pipeline were preserved in the final data set and analysed unless otherwise stated. Of the 166 clonal nucleotide sequences, 54 were re-isolated during this process. This frequency (0.33) reflects previous comparisons of clonal versus UDPS sequence recovery (35).

Detection of recombination. The presence of intra-subtype recombinants identified through clonal analysis suggested that informed exploration of the UDPS data would also reveal recombinants (16). We applied the pairwise homoplasy index (PHI) test as implemented in SplitsTree 4.13 to assign a probability that the aligned sequences within a sample set contained recombinants (36). P-values <0.05 were indicative of recombination (37). To better identify recombinants within sample sets, the Neighbour Net (NNet) algorithm was initially applied to the clonal data set using SplitsTree 4.13 with and without the inclusion of inter-lineage recombinant sequences generated in silico from the consensus sequences of the lineage subsets. The in silico breakpoint chosen was at position 214 of the amplicon sequence to broadly reflect the predicted breakpoint regions observed from the data (Fig. S1). Haplotypes with conflicting phylogenetic signals are ‘split’ away from the dominant population branches and it is this feature that is suggestive of the presence of recombinants (36). NNet was utilised on a sample by sample basis unless otherwise stated. Recombinants identified from NNet trees were tested against consensus sequences for each of the lineage subsets using Simplot to identify putative recombination breakpoint locations (38).
Bioinformatics analyses. MEGA5 was used to calculate synonymous and nonsynonymous mutation rates at the HVR1 (39). Initial grouping of clonal sequences was performed by phylogenetic analysis using a general time reversible model (GTR+G+I). Calculations were performed separately on the four haplotype groups using nucleotide sequences present at a frequency >0.001.

Evidence of site specific directional selection was performing using the directional evolution of protein sequences (DEPS) method implemented in conjunction with the Jones Taylor Thornton amino acid substitution model (40). Co-evolving amino acid residues were identified using a baysian graphical model (BGM) through the Spidermonkey algorithm (41, 42). A one parent undirected network was applied. Only those sites with a posterior probability threshold of >0.9 were reported. For both sets of analysis a reduced data set, containing amino acid sequences present at a frequency >0.001, was used.

Sequence information and accession numbers. Amino acid residues were numbered according to the polyprotein of the H77 reference sequence (AF009606). As a subpopulation of variants (L1b) contained a single amino acid insertion, we identify the wild type residue as 387, whereas the amino acid insertion is identified as 387*. UDPS data sets used in this study are available at http://www.ucc.ie/liamfanning/hcv. Clonal nucleotide sequences derived from this study were deposited with GenBank and assigned accession numbers KF41792 -38 (RL5) and KC689336-42 (RL12).
RESULTS

Lineage overview. Original phylogenetic analysis of the clonal data set reported two lineages (L1 and L2) with L1 partitioned into two sublineages (L1a and L1b) (16). Following UDPs and initial phylogenetic analysis it was evident that a third L1 sublineage existed (L1c). The haplotypes described were partitioned into four groups based these phylogenetic observations; namely L1a, L1b, L1c and L2. Each of these groups had periods of dominance interspersed with periods of marked decline over the 9.6 year study period (Fig. 1). L1c is notable in this respect. L1c haplotypes initially represented >10% of the total number of sequences recovered from RL1. In the three subsequent samples this decreased to <0.07%, 0.03% and 0.1%, respectively. L1c dominates RL5 accounting for >60% of the individual reads in this sample and coincided with an equally significant decline in L1a sequences (the frequency of which fell from >0.58 at RL4 to <0.07 in RL5). Only seven L1c sequences were recovered in RL6 and RL7 combined, with none thereafter. Subsequent to this, L1b briefly dominated the population from RL6 to RL7. From RL8 to RL10, L1 sequence frequencies declined overall allowing L2 to rise to total domination in the latter samples. The results clearly demonstrate a virome structure in flux that is only readily observed when analysed over an expanded timeframe. The L1c phenotype is indicative of a serendipitous return to temporal dominance following the re-establishment of a host environment spanning RL5 that favoured the expansion of this group of variants.

Intrinsic population dynamics. Qualitatively the mutual relationship between the four subpopulations sees dominance of L1 over L2 from RL1-7. The transition to L2 dominance coincides with the suppression of all L1 populations. An inter-correlated
variable test using the four viral groupings explains the fractional competition between all populations (Table 2). The correlation between any two variables based on their respective frequencies within a sample was initially determined (Pearson correlation) and then recalculated where the effects of a third variable on this interaction is removed (1st order partial correlation). It can be seen from Table 2 that the third variable typically suppresses a larger correlation that would have been observed if this variable were not present. First order partial correlations identified significant suppressor variables; for example, the extent that the population dynamics of L1b can explain 68% of the variation in L2 when controlled for L1a (Table 2). Of note, L1a and L1b covariance rose from 0.6% to 42% when the supressing effects of L2 were accounted for (Table 2). The proportion of variation that cannot be explained by the inter-correlated variable analysis is liable to be shaped by the extent of error prone replication, the activity of the adaptive immune system and intrinsic immune-mediated clearance mechanisms.

A major perturbation to the virome occurs 4.6 years into the study timeline (Fig. 1, RL5). At this point L1c re-emerges as a dominant group within the sample space coincident with a L1a shift from neutral to positive selection with a sharp increase in nonsynonymous mutations (Fig. 2). L1b becomes established during this window by expanding its presence into the available sequence space. The number of unique haplotypes recovered for L1b dramatically increased between RL4 to RL5 without any significant alteration to either the synonymous or nonsynonymous mutation rates. The L1b phenotype is subject to frequency dependent selection whereby population expansion is linked to a contraction in sequence diversity. This was followed by a parallel selective sweep of both L1a and L1b. L1b dominates the entire virome at RL7 (Fig. 1), the
structure of which comprises of two equi-dominant E2 haplotypes (frequency of 0.48 and
0.49 respectively) differing by a single amino acid (T391A). The first of these haplotypes
was originally isolated from RL2 (frequency 0.16, preceding RL7 by 5.5 years), while the
second haplotype was identified in RL3 (frequency 0.11, preceding RL7 by 4.5 years).
Ultimately, all L1 groups decline in number, with only three L1a haplotypes still
detectable in RL10 at a combined frequency of <0.0006.

L2 haplotypes follow a divergent trajectory to that of L1 haplotypes. L2 haplotypes
exhibit a higher rate of synonymous mutations early in the study that is accounted for by
a subset of minor variants which have a distinct HVR1 to the dominant L2 population.
During this longitudinal study, the dominant L2 haplotypes exhibit low diversity and a
high degree of genetic stability. The dominant L2 haplotypes recovered from samples
RL4, RL7, RL8 and RL10 (frequencies of 0.24, 0.02, 0.14 and 0.87, respectively) only
differ by a single synonymous polymorphism. The pressure to maintain such an exact
replicate in spite of an error prone polymerase indicates the existence of a narrow host-
specific sequence space within which this variant can maintain sufficient fitness to
survive. Despite expanding into the replication space vacated by L1 in the latter time
points, L2 HVR1 consensus sequences have not altered significantly from those
recovered at earlier time points when L2 was a minor species. Indeed, only one unique L2
HVR1 amino acid variant was isolated from RL9 (n=12571 reads). We know from clonal
analysis of RL12 that a novel L2 HVR1 amino acid motif emerges to dominate the L2
HVR1 clonal profile. This haplotype was present in the raw UDPS data at a frequency of
0.0002. The resultant amino acid sequence again differs by a single nonsynonymous
change within the HVR1 (N395H). An L1a haplotype, present at a frequency of 0.0006,
was recovered from RL10 in the final data set through implementation of the motif hunter strategy, emphasising the sensitivity of the methodology used.

The conservative sequence structure documented here is not unique to this individual. We have also observed similar sequence stability for a monophyletic infection over 6 years in a genotype 4e chronic infection (data not shown). Together this evidence indicates that the development of host-specific virome adaptation is ongoing in established chronic infection.

UDPS reveals multiple instances of recombination. Following KEC-ET cleaning of the raw data, the final data set contained approximately 66,000 individual reads available for analysis. This represented a 170-fold increase over the clonal method. The identification of recombinants within the clonal data set directed our investigation towards elucidating the propensity for HCV to recombine under basal conditions in this patient.

The clonal data set of 166 unique sequences, which included two documented intra-subtype recombinants FJ744095 and JQ743309, was used to inform downstream applications of PHI and NNet (16). A statistically significant p-value for recombination (p < 0.002) was first obtained using the PHI test. Secondly, split-decomposition networks for the clonal data, supplemented with 12 in silico-derived recombinant sequences representing the four haplotype groups, were obtained with the NNet algorithm (Fig. 3A and B). Inclusion of these 12 sequences resulted in a marked increase in the p-value significance (p < 10^-4) with a concurrent increase in number and size of signal 'splits'
observed in the NNet graphs. FJ744095 and JQ743309 clustered with three of the in silico-derived recombinant sequences (Fig. 3B).

To test the robustness of our approach, we omitted both FJ744095 and JQ743309 from the clonal data and recomputed the PHI statistic for recombination. Unexpectedly, the recombination p-value remained significant (p < 0.015). Further inspection of the split-decomposition graphs identified a clonal sequence (HM363402) that mirrored the split signal of an L1a-L1c in silico recombinant sequence (Fig. 3B, green circle). Analysis with Simplot confirmed HM363402 as an L1a-L1c recombinant with the anticipated crossover region identified just inside the HVR1 (Fig. S1A). Exclusion of HM363402 together with FJ744095 and JQ743309 resulted in a nonsignificant recombination p-value (p < 0.07) for the clonal data.

We proceeded to test all the UDPS samples consecutively using the PHI statistic and each sample point was associated with significant p-values (range 0.034 - <10^{-4}). No significant correlation between the sample specific PHI statistic and (1) the sample number, (2) percentage recovery of reads, (3) number of haplotypes per (sub)lineage or (4) the (sub)lineage sample proportion were observed (r^2 value range 0.005 - 0.353). Representative examples of a high diversity population (RL1) and low diversity population (RL8) were selected for further analysis (Fig. 3C and D). Haplotypes exhibiting conflicting phylogenetic profiles were found to be five inter-lineage recombinants and two putative intra-lineage recombinants from within the RL1 data (Fig. 3C). L1c sequences contributed to all five inter-lineage recombinants identified in RL1. In each instance the recombinant sequence mosaic contained an L1c E1 region together
with the E2/HVR1 element of L1a or L1b (Fig. S1D-H). We note that the predicted recombination breakpoint regions were limited to the E1/E2 gene junction which is in agreement with independent reports of HCV intra-subtype recombinants (25). There was no evidence of recombinant sequences identified in RL1 containing elements of L2 acting as parental donor. This profile altered in RL8, where L2 sequences were identified as donors to each of the 7 inter-lineage recombinant sequences identified (Figs. 3D and S2I-O).

The formation of recombination artefacts in vitro during PCR has been shown to reflect the frequency of the constituent parental haplotypes (28). Neither in RL1 nor RL8 were recombinants found that contained sequence elements from parental donors L1a and L1b together, despite their relatively high combined frequencies in these samples (0.87 and 0.77, respectively). Our data argues in favour of authentic recombinants derived in situ within the liver given that L1c occupies <12% of the sample space in RL1 yet serves as parental donor to all of the recombinants (excluding the two putative intra-lineage recombinants, Fig. 3C). For RL8, L2 comprises <23% of the sample space yet L2 sequence signatures were observed in all the reported recombinants from this sample.

Furthermore, our previous report describing the clonal recombinants FJ744095 and JQ743309 documented the parental donor sequences from distinct samples to those from which the recombinant sequences were isolated (16). In the context of the work reported here it is notable that an amino acid homolog to the clonal recombinant JQ743309 was identified in the UDPS data from RL6. This haplotype contained four synonymous mutations across the full sequence length. This preceded the original clonal isolation of
It was possible to identify inter-sample maintenance and evolution of recombinant sequences from the UDPS data (Fig. 4). Such an observation further supports the contention that these recombinants are derived in situ within the host.

Finally, we repeated our analysis using a well characterised clonal data set where the patient cohort comprised of 22 women exposed to HCV genotype 1b contaminated blood products from a single source (15, 43). No evidence for recombination was observed in this instance.

**Evolutionary evaluation of (sub)lineage specific HVR1 populations.** The presence of functional micro-domains with the HVR1 has been established for the H77 sequence (44). In broad terms, the first 13 residues of the HVR1 (aa positions 384-396) regulate virus infectivity, coupled with putative masking of the CD81 binding site. A downstream region of nine residues (aa positions 400-408) is believed to contain the main neutralisation epitope (44, 45). In the absence of humoral immune selection pressure, the accumulation of mutations at random positions across this amplicon sequence during replication should in principle be selected on the basis of positive or negative effects on replicative potential of a given genome. It was of interest therefore to assess whether there exists a preferential evolution towards specific HVR1 residues within the variant subpopulations over time.

DEPS analysis revealed 5, 7 and 4 HVR1 position(s) under directional selection for L1a, L1b and L1c respectively (Table 3). The DEPS data is indicative of selective sweeps governing the haplotype profile. There was no evidence to suggest that L2 HVR1 sites were under directional selection. This data is consistent with L2 sequences occupying a
host-specific niche. By contrast L1b had the highest level of directional mutational change within the neutralisation epitope (Table 3).

The conservative maintenance of the overall physio-chemical amino acid composition of the HVR1 is required for protein function (3, 16, 46). Mutational change at one site may require compensatory changes at a second to achieve this preservation of function. In all, five pairs of co-dependent sites were identified from L1 where the cumulative posterior probability of site 1 and site 2 being conditionally dependent was >0.9 (Table 4). Only sites within the HVR1 exhibited co-dependencies and these sites were predominantly also under directional selection (Tables 3 and 4). In agreement with the previous observations of L2 HVR1 stability, no co-dependent sites were observed for the L2 haplotype group above the posterior probability threshold of 0.9.

Lineage specific humoral immune targeting. HCV RNA originating from IgG-bound virions was recovered from samples RL1-3, RL6 and RL7. In total, 12 unique nucleotide sequences (nine unique E2 amino acid sequences) were recovered. Phylogenetic analysis identified all 12 sequences as L1 isolates (Table 5). Eight of the 12 IgG-bound nucleotide sequences (five of the nine unique E2 amino acid sequences) were present in the UDPS data set. While each sequence ultimately demonstrated a marked decline post-isolation, the timeframe between detection on the column and non-detection in the UDPS data ranged from immediate (Fig. 5B) to 6 years (Fig. 5A). One IgG-bound variant isolated at RL6 remained dominant for a further 2.6 years before decreasing below the detectable threshold (Fig. 5D).
The collapse in L1a and L1b diversity post-RL5 (Fig. 2) is consistent with the exclusive capture of IgG-bound virions from these groups in RL6 and RL7. In contrast, the non-detection of L2 sequences from the IgG depletion columns tallies with the minimal L2 sequence diversity together with a HVR1 without detectable directional selection or epistatic evolution. This latter result may suggest weak immune targeting of L2 virions expressing this phenotype or immune evasion by other means (e.g. intracellular transmission) (21, 22). Despite occupying >95% of the sample space at RL9, nonsynonymous mutations were not observed in the HVR1 of L2 sequences, suggesting that these motifs are immune ‘invisible’ within this host during the period of infection analysed here. This sequence stability points to an active mechanism of maintenance and that the pervasive purifying selection observed over the preceding 10 years is the output of selection that preserves relative sequence homogeneity.
To interrogate complex UDPS data from a mixed lineage HCV genotype 4a infection we utilised the KEC-ET algorithm in conjunction with a previously published set of clonal data from temporally matched samples (16). From the data, it was evident that intra-subtype recombination events at the E1/E2 gene junction is ongoing during HCV natural infection. Recombinant haplotypes were present at low frequencies suggesting that, under basal conditions, parental haplotypes retain competitive fitness advantages. Overt competition between lineages for the shared replication space was evidenced at discrete times during the study period with demonstrable correlations between the viral groups (Fig. 1 and Table 2).

Recombination is a mechanism through which viruses can explore novel genomic space. The negligible frequency of verifiable recombinants (globally for HCV) supports the hypothesis that fitness costs associated with de novo chimeric genomes likely impinge on the recombinant’s ability to compete against dominant parental strains for replication space within the quasispecies swarm. Previously, we reported the identification of intra-subtype recombinants by clonal analysis (16). The expectation was that UDPS would therefore reveal a more populous set of recombinants. Retrospective analysis of HIV data sets has demonstrated that recombinants, which were previously overlooked can be identified by use of the NNet algorithm in conjunction with the PHI statistic (36, 47). Our analysis supports this and demonstrates the suitability of our strategy to the study of HCV UDPS data. Most of the recombinants reported here were present at frequencies <0.001. As such we cannot ignore the possibility that in vitro recombination artefacts may also be present in the final data set. The acceptance of low frequency recombinants requires...
critical evaluation weighed against the documented prevalence and characteristics of in vitro artefacts (28, 48-50). To that end we submit four main arguments in support of our broader recombination findings. Firstly, the clonal recombinant JQ743309 and an UDPS amino acid homolog were independently isolated from separate samples (RL6 and RL8) 1.6 years apart. Secondly, the demonstrable longevity observed with JQ743309 extends to other recombinants within the UDPS data set whereby homologous inter-sample recombinants (up to 3 years apart) were present (Fig. 4). This, coupled with observed evolution of recombinant sequences (RL7-10) within the UDPS data set, argues against the chance occurrence of recombinants significantly contributing to our findings (28).

Thirdly, the collapse of L1 sequences in latter samples (RL8-10) was not reflected in the recombinants identified from these samples, where genetic signatures from L1 haplotypes were present in recombinant haplotypes phylogenetically classified as arising from L2. Finally, for RL1 and RL8, where the UDPS data was examined in depth, the recombinant profile did not reflect the lineage frequencies present (28). Our results argue in favour of recombinants being rare but real in the context of this chronic infection.

Both replicative (strand switching by the polymerase) and non-replicative (ligation of genomic fragments by host ligases) mechanisms of recombination in HCV are postulated to occur (Reviewed in 24). In vitro studies have both confirmed the ability of HCV to recombine and demonstrated the subsequent viability of these recombinants (27, 51). Additionally, intra-subtype E2 gene exchange recombinants have exhibited comparable infectivity titers to wild type controls, whereas inter-subtype recombinants required compensatory mutations to facilitate establishment in cell culture (52). Following the seminal description of a 2k/1b recombinant in 2002 (53), only a handful of in vivo
recombinants have since been reported even when high risk populations have been specifically investigated (Reviewed in 54). To our knowledge, this is the first report documenting multiple intra-subtype HCV recombinants from UDPS data. We suggest that studies designed to identify recombinants, do so with a focus on inter as well as intra-genotypic recombinants, as this may better reflect the natural history of HCV.

The sample specific recombination profiles observed may also be a feature of related variants competing for the same replication space. L1a and L1c demonstrated competitive profiles between RL4-6, which were mirrored by L1b and L2 thereafter (Fig. 2 and Table 2). Collectively, the data documents a haplotype group (L1c) struggling to gain purchase within a competitive replicative space. A sudden shift in environmental landscape favoured the rapid expansion of L1c from the viral reservoir, albeit finitely (55, 56). The outgrowth of L1c at RL5 coincides with the mass haplotype extinction event of L1a and L1b at RL4 (16). This extinction may have provided a temporary competitive reprieve for the L1c group that, prior to RL4, were only detectable through implementation of the motif hunter strategy. In agreement with work by Ruiz-Jarabo and colleagues (2002), this period of L1c restoration was concomitant with a temporal growth in fitness as evidenced by an increase in frequency from <0.001 in RL4 to >0.6 in RL5 (57).

It has been demonstrated that competition within a complex quasispecies mosaic can result in the suppression of high fitness variants (58). *In vitro* mechanisms of intracellular competition have been identified that facilitate the preferential expansion of one genome over another (59). Competitive lineage suppression of L2 by L1 is evident during RL1-7 (Table 2). The architectural organisation of the liver can additionally provide spatial
segregation between the different viral groups, with clusters of infected hepatocytes surrounded by areas of fibrosis, cirrhotic tissue or uninfected hepatocytes accessible by cell-to-cell transmission (21, 60, 61). These phenomena may, in part, account initially for the continued persistence of L2 virions and (more generally) fringe populations of minor variants until such time as the environmental landscape again alters.

Eventually, all L1 groups succumb to mass haplotype extinction events coupled with convergent evolution and frequency dependent selection, that is in part linked with a nAb directed response. A moving paradigm of cyclical competition-directed by population collapse, establishment of a new population order and immune pressure resulted in the progressive rise of L2. L2 haplotypes remained predominantly under purifying or negative selection, irrespective of its positioning in the sample space (Fig. 2). In spite of the gross perturbation to the virome, the viral load increased between RL9 to RL12 (HCV RNA 5.1 log_{10} IU/ml - 6.6 log_{10} IU/ml, respectively). The 321 bp amplicon used in this study does not allow us to account for selection pressures outside of the E1/E2 gene junction, such as MHC class I-restricted alleles in NS3 (15, 62). However, given the continued maintenance for the L2 HVR1 epitope following the expansion of this lineage, it is likely that L2 had similar replicative fitness to L1 (yet was competitively outcompeted for replication space in the presence of a dominant L1 population) and/or the humoral immune response targeting de novo L2 virions was less effective than that for L1 virions (17, 58). In reality, both competition and immune selection pressures are likely to contribute to varying degrees overtime given (1) the rise in viral load and (2) the lineage profile of virions under humoral immune selection pressure through IgG fractionation. The 12 unique nucleotide sequences recovered by IgG purification
phylogenetically clustered with L1 haplotypes (Table 5). Of the five IgG-bound predicted E2 amino acid sequences identified in the UDPS data, two (both L1b haplotypes) were present in their respective samples >30%. It has been suggested that such high frequency variants may elicit a strong neutralising response against themselves and closely related variants (17).

Few studies have documented longitudinal quasispecies dynamics in treatment naïve chronically HCV infected individuals. Evidence is accumulating to suggest that host-adapted monophyletic convergence of the virome is a feature of long term infection (16, 17, 63). The observation that our sample set contains multiple instances of recombination is suggestive of an evolutionary strategy positioned to alleviate structural and functional constraints developed by the HVR1 over time (64). It has been argued that recombination would be detrimental to beneficial mutations accumulated through co-dependent epistatic evolution (65). The restriction of recombination to gene junctions alleviates this problem (23, 25). Of the five co-dependent epistatic sites identified here, there exists the potential for all five to be transferred to the daughter virions intact, based on the predicted crossover regions (Table 4 and Fig. S1).

Judicious application of the KEC and ET algorithms facilitated the formulation of a data set that more accurately represented the true quasispecies population dynamics. However, it is essential to recognise that the context of an infection predicates whether or not recombination will firstly occur and secondly, be detectable. Mixed lineage infections of one subtype, such as the infection described here, may uniquely meet all the prerequisites for measurable recombination, i.e., (1) multi-variant infection of single hepatocytes, (2)
sufficiently divergent intra-host communities and (3) informed analysis of complex data sets that account for hypervariable heterogeneity.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. David S. Campo (Division of Viral Hepatitis, Centers for Disease Control and Prevention) for his critical reading of this manuscript. We would also like to thank John Levis and Dr. Kevin Hegarty for initial characterisation of patient sera.
REFERENCES


Evidence of recombination in intrapatient populations of hepatitis C virus.  


FIGURE LEGENDS

**Fig. 1.** The area graph represents the percentage of overall sequences contributing to the subpopulations of L1a (dark grey), L1b (black), L1c (white) and L2 (light grey) per sample over time.

**Fig. 2.** Summary of (A) the number of unique full length nucleotide haplotypes for each group (L1a, L1b, L1c, L2 and combined total) recovered per time point. (B) d_s (solid line) and d_N (dashed line) measurements across the HVR1 are given for those nucleotide sequences with frequencies >0.001. The HVR1 is defined between positions 196-276/279 of the amplicon sequence. The number of (non)synonymous substitutions per (non)synonymous site from averaging over all sequence pairs within each group per time point are shown. The importance of separating haplotypes into defined groups is illustrated by the deceptively high d_s and d_N values recovered when the data set is not segregated.

**Fig. 3.** Phylogenetic networks for the (A) complete clonal data set and (B) supplemented with 12 *in silico* generated inter-lineage recombinants (blue circles). Yellow circles denote recombinants FJ744095 and JQ743309 identified previously through clonal analysis, while the green circle denotes the novel clonal recombinant HM363402 identified in this study. UDPS data subsets (C) RL1 and (D) RL8 are given as representative images for samples containing a complex and a less diverse population of haplotypes (355 versus 176 unique nucleotide sequences, respectively). Red circles identify inter-(sub)lineage recombinants. Black circles identify putative intra-lineage
recombinants. Genetic distance of 0.01 nucleotide substitutions per site is given by the scale bar.

Fig. 4. Inter-sample maintenance of recombinants. (A) Composite phylogenetic network of the RL7-10 subset. Haplotype clusters displaying split decomposition signals are identified (i-iii). (B-D) Phylogenetic analysis of clusters i-iii, consensus sequences of the (sub)lineages and inter-lineage recombinant sequences generated in silico using a rooted general time reversible model (GTR+G+I). The data demonstrates inter-sample maintenance and evolution of recombinant haplotypes arguing in favour of the recombinant sequences arising in situ within the host. Sequences are identified by time point as black circles (RL7), open squares (RL8) and black triangles (RL9). Sequences arising from RL10 did not feature in groups i-iii. A genotype 4a strain (Y11604) was used as the reference out group. Bootstrap values (of 1,000 re-samplings) above 70 are shown. Genetic distance of 0.01 nucleotide substitutions per site is given by the scale bar.

Fig. 5. Detection of IgG-bound virus within the UDPS data set. Five of nine unique E2 amino acid sequences translated from RNA obtained following Albumin/IgG column extraction of virions were present within the UDPS data set (A-E). The percentage sample proportion of these sequences at nucleotide (grey lines) and amino acid (black lines) level within the overall sample population are given. The respective sequence accession numbers are (A) GQ985332, (B) GQ985336, (C) GQ985372, (D) HM363384 and (E) GQ985333. Black arrows indicate the sampling point.
Fig. 1
Fig. 2

Haplotypes for different samples and their mean distance.
Fig. 3
Fig. 4
Fig. 5
<table>
<thead>
<tr>
<th>Motif</th>
<th>HVR1</th>
<th>Sequence</th>
<th>Amplicon position</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1a_i</td>
<td>Typical</td>
<td>GCAGGGGTTGACGCCGACCAC</td>
<td>181 - 204</td>
</tr>
<tr>
<td>L1a_ii</td>
<td>Typical</td>
<td>GCAGGGGTTGACGCCAGACC</td>
<td>181 - 204</td>
</tr>
<tr>
<td>L1b_i</td>
<td>Atypical</td>
<td>GCAGGAGTCGACGCGGACCAC</td>
<td>181 - 204</td>
</tr>
<tr>
<td>L1b_ii</td>
<td>Atypical</td>
<td>GCAGGAGTTGACGCGACCAC</td>
<td>181 - 204</td>
</tr>
<tr>
<td>L1c_i</td>
<td>Typical</td>
<td>GCAGGGGTCGACGCCAGCCAC</td>
<td>181 - 204</td>
</tr>
<tr>
<td>L1c_ii</td>
<td>Typical</td>
<td>GCAGGGGTTGACGCCAGCCAC</td>
<td>181 - 204</td>
</tr>
<tr>
<td>L2_i</td>
<td>Typical</td>
<td>GGGGCTGTAGCGTCCAGCAGCCT</td>
<td>211 - 234</td>
</tr>
<tr>
<td>L2_ii</td>
<td>Typical</td>
<td>GGGGCTGTGGCATCCAGCAAACC</td>
<td>211 - 234</td>
</tr>
<tr>
<td>Input variables$^a$</td>
<td>Pearson Correlation$^b$</td>
<td>Inter-correlated variables$^c$</td>
<td>1$^{	ext{st}}$ order Partial Correlation</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------</td>
<td>---------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>WX</td>
<td>-0.079</td>
<td>WX.Y</td>
<td>-0.093</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WX.Z</td>
<td>-0.648</td>
</tr>
<tr>
<td>WY</td>
<td>-0.239</td>
<td>WY.X</td>
<td>-0.243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WY.Z</td>
<td>-0.478</td>
</tr>
<tr>
<td>WZ</td>
<td>-0.502</td>
<td>WZ.X</td>
<td>-0.750</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WZ.Y</td>
<td>-0.623</td>
</tr>
<tr>
<td>XY</td>
<td>-0.047</td>
<td>XY.W</td>
<td>-0.068</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XY.Z</td>
<td>-0.360</td>
</tr>
<tr>
<td>XZ</td>
<td>-0.671</td>
<td>XZ.W</td>
<td>-0.824</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XZ.Y</td>
<td>-0.721</td>
</tr>
<tr>
<td>YZ</td>
<td>-0.309</td>
<td>YZ.W</td>
<td>-0.510</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YZ.X</td>
<td>-0.459</td>
</tr>
</tbody>
</table>

$^a$W=L1a, X=L1b, Y=L1c, Z=L2

$^b$Lineage frequencies by sample were used to build correlations

$^c$The degree of association between two (sub)lineages where the effects of a third (sub)lineage on this interaction is accounted for

$^d$Significant correlations are identified in bold
<table>
<thead>
<tr>
<th>Lineage</th>
<th>Site</th>
<th>Overall Site Composition</th>
<th>MRCA Residue</th>
<th>Inferred Substitutions</th>
<th>EBF$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1a</td>
<td>384</td>
<td>G$<em>{25}$K$</em>{22}$</td>
<td>G</td>
<td>G$<em>{1}$$\leftrightarrow$G$</em>{22}$; K$<em>{1}$$\leftrightarrow$K$</em>{22}$</td>
<td>G: $&gt;10^5$; K: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>386</td>
<td>I$<em>{25}$T$</em>{22}$</td>
<td>I</td>
<td>I$<em>{1}$$\leftrightarrow$T$</em>{22}$</td>
<td>T: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>391</td>
<td>V$<em>{26}$A$</em>{21}$</td>
<td>V</td>
<td>A$<em>{21}$$\leftrightarrow$V$</em>{26}$</td>
<td>A: $&gt;10^9$</td>
</tr>
<tr>
<td></td>
<td>397</td>
<td>S$<em>{31}$N$</em>{14}$K$_{22}$</td>
<td>N</td>
<td>N$<em>{0}$$\leftrightarrow$N$</em>{14}$; S$<em>{31}$$\leftrightarrow$S$</em>{14}$</td>
<td>S: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>403</td>
<td>L$<em>{24}$F$</em>{23}$</td>
<td>F</td>
<td>F$<em>{0}$$\leftrightarrow$F$</em>{23}$</td>
<td>F: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>408</td>
<td>K$<em>{22}$S$</em>{15}$R$_{8}$</td>
<td>K</td>
<td>K$<em>{1}$$\leftrightarrow$R$</em>{8}$; S$<em>{22}$$\leftrightarrow$S$</em>{15}$</td>
<td>S: $&gt;10^5$</td>
</tr>
<tr>
<td>L1b</td>
<td>349</td>
<td>T$<em>{44}$A$</em>{26}$S$_{3}$</td>
<td>T</td>
<td>A$<em>{25}$$\leftrightarrow$T$</em>{44}$</td>
<td>A: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>384</td>
<td>G$<em>{43}$R$</em>{22}$K$_{2}$</td>
<td>G</td>
<td>G$<em>{0}$$\leftrightarrow$G$</em>{22}$; R$<em>{43}$$\leftrightarrow$R$</em>{22}$</td>
<td>R: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>385</td>
<td>T$<em>{43}$A$</em>{23}$P$_{7}$</td>
<td>T</td>
<td>A$<em>{23}$$\leftrightarrow$T$</em>{43}$; P$<em>{7}$$\leftrightarrow$P$</em>{23}$</td>
<td>A: $&gt;10^9$</td>
</tr>
<tr>
<td></td>
<td>392</td>
<td>A$<em>{66}$P$</em>{3}$V$_{2}$</td>
<td>A</td>
<td>A$<em>{0}$$\leftrightarrow$P$</em>{3}$; V$<em>{66}$$\leftrightarrow$V$</em>{2}$</td>
<td>P: 238.3</td>
</tr>
<tr>
<td></td>
<td>397</td>
<td>S$<em>{29}$K$</em>{25}$N$<em>{17}$R$</em>{2}$</td>
<td>N</td>
<td>K$<em>{24}$$\leftrightarrow$K$</em>{17}$; N$<em>{0}$$\leftrightarrow$N$</em>{22}$</td>
<td>K: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>398</td>
<td>H$<em>{54}$R$</em>{12}$N$<em>{10}$Q$</em>{1}$</td>
<td>H</td>
<td>H$<em>{0}$$\leftrightarrow$H$</em>{12}$; R$<em>{54}$$\leftrightarrow$R$</em>{12}$; N$<em>{10}$$\leftrightarrow$N$</em>{54}$</td>
<td>R: 601.7</td>
</tr>
<tr>
<td></td>
<td>404</td>
<td>D$<em>{43}$S$</em>{15}$H$<em>{7}$T$</em>{3}$Q$<em>{1}$A$</em>{1}$N$_{1}$</td>
<td>D</td>
<td>D$<em>{1}$$\leftrightarrow$D$</em>{15}$; S$<em>{43}$$\leftrightarrow$S$</em>{15}$; H$<em>{7}$$\leftrightarrow$H$</em>{43}$; T$<em>{3}$$\leftrightarrow$T$</em>{43}$; Q$<em>{1}$$\leftrightarrow$Q$</em>{43}$; A$<em>{1}$$\leftrightarrow$A$</em>{43}$; N$<em>{1}$$\leftrightarrow$N$</em>{43}$</td>
<td>S: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>405</td>
<td>S$<em>{32}$Y$</em>{12}$F$<em>{11}$P$</em>{6}$A$<em>{3}$V$</em>{4}$T$_{3}$</td>
<td>S</td>
<td>P$<em>{6}$$\leftrightarrow$P$</em>{12}$; S$<em>{32}$$\leftrightarrow$S$</em>{12}$; Y$<em>{12}$$\leftrightarrow$Y$</em>{32}$; F$<em>{11}$$\leftrightarrow$F$</em>{32}$; A$<em>{3}$$\leftrightarrow$A$</em>{11}$; V$<em>{4}$$\leftrightarrow$V$</em>{14}$; T$<em>{3}$$\leftrightarrow$T$</em>{13}$; S$<em>{0}$$\leftrightarrow$S$</em>{32}$; Y$<em>{12}$$\leftrightarrow$Y$</em>{32}$</td>
<td>Y: $&gt;10^5$</td>
</tr>
<tr>
<td>L1c</td>
<td>363</td>
<td>C$<em>{12}$S$</em>{6}$</td>
<td>C</td>
<td>C$<em>{0}$$\leftrightarrow$C$</em>{6}$; S$<em>{12}$$\leftrightarrow$S$</em>{6}$</td>
<td>S: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>395</td>
<td>D$<em>{12}$N$</em>{6}$</td>
<td>D</td>
<td>D$<em>{1}$$\leftrightarrow$D$</em>{12}$; N$<em>{1}$$\leftrightarrow$N$</em>{12}$; S$<em>{6}$$\leftrightarrow$S$</em>{12}$</td>
<td>N: 355.8</td>
</tr>
<tr>
<td></td>
<td>397</td>
<td>K$<em>{12}$N$</em>{4}$</td>
<td>K</td>
<td>K$<em>{0}$$\leftrightarrow$K$</em>{12}$; N$<em>{4}$$\leftrightarrow$N$</em>{12}$</td>
<td>N: 301.6</td>
</tr>
<tr>
<td></td>
<td>398</td>
<td>H$<em>{13}$S$</em>{4}$Q$_{1}$</td>
<td>H</td>
<td>H$<em>{0}$$\leftrightarrow$H$</em>{13}$; S$<em>{4}$$\leftrightarrow$S$</em>{13}$</td>
<td>S: $&gt;10^5$</td>
</tr>
<tr>
<td></td>
<td>404</td>
<td>A$<em>{1}$S$</em>{4}$</td>
<td>A</td>
<td>A$<em>{0}$$\leftrightarrow$A$</em>{13}$; S$<em>{4}$$\leftrightarrow$S$</em>{13}$</td>
<td>S: 205.9</td>
</tr>
<tr>
<td>L2</td>
<td>342</td>
<td>S$<em>{15}$G$</em>{8}$</td>
<td>S</td>
<td>G$<em>{8}$$\leftrightarrow$G$</em>{15}$; S$<em>{0}$$\leftrightarrow$S$</em>{15}$</td>
<td>G: $&gt;10^5$</td>
</tr>
</tbody>
</table>

$^a$Most recent common ancestor

$^b$A$_{C}$$\leftrightarrow$B$_{A}$ indicates C substitutions from A to B and D substitutions from B to A

$^c$Sites with an Empirical Bayes Factor $>100$ are reported with preferential evolution towards the given residue
TABLE 4. Co-dependent epistatic evolution

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Site1</th>
<th>Site2</th>
<th>$P{S_1 \rightarrow S_2}$&lt;sup&gt;a&lt;/sup&gt;</th>
<th>$P{S_1 \leftarrow S_2}$&lt;sup&gt;b&lt;/sup&gt;</th>
<th>$P{S_1 \leftrightarrow S_2}$&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1a</td>
<td>G384K</td>
<td>I386T</td>
<td>0.41</td>
<td>0.53</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>L403F</td>
<td>K408R</td>
<td>0.35</td>
<td>0.61</td>
<td>0.96</td>
</tr>
<tr>
<td>L1b</td>
<td>T385A</td>
<td>G384R</td>
<td>0.86</td>
<td>0.13</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>T385A</td>
<td>S405Y</td>
<td>0.76</td>
<td>0.24</td>
<td>1.00</td>
</tr>
<tr>
<td>L1c</td>
<td>T396A</td>
<td>D395N</td>
<td>0.44</td>
<td>0.52</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<sup>a</sup>Posterior probability for site 2 being conditionally dependent on site 1
<sup>b</sup>Posterior probability for site 1 being conditionally dependent on site 2
<sup>c</sup>Posterior probability for site 1 and site 2 being conditionally dependent
TABLE 5. Haplotypes targeted by humoral immune activity

<table>
<thead>
<tr>
<th>Sample</th>
<th>IgG-bound virus</th>
<th>Accession Number</th>
<th>Detectable in pyro data</th>
<th>Lineage</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL1</td>
<td>+</td>
<td>GQ985330</td>
<td>-</td>
<td>1b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GQ985332</td>
<td>+</td>
<td>1c</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GQ985333</td>
<td>+</td>
<td>1a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GQ985336</td>
<td>+</td>
<td>1b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GQ985337</td>
<td>-</td>
<td>1b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GQ985350</td>
<td>+</td>
<td>1a</td>
</tr>
<tr>
<td>RL2</td>
<td>+</td>
<td>FJ744071</td>
<td>-</td>
<td>1a</td>
</tr>
<tr>
<td>RL3</td>
<td>+</td>
<td>EU482135</td>
<td>-</td>
<td>1a</td>
</tr>
<tr>
<td>RL4</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>RL5</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>RL6</td>
<td>+</td>
<td>HM363384</td>
<td>+</td>
<td>1b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HM363385</td>
<td>+</td>
<td>1b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HM363386</td>
<td>+</td>
<td>1b</td>
</tr>
<tr>
<td>RL7</td>
<td>+</td>
<td>GQ985372</td>
<td>+</td>
<td>1a</td>
</tr>
<tr>
<td>RL8</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>RL9</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>RL10</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Abbreviation: N/A, not applicable