Analysis of the Evolution and Structure of a Complex Intrahost Viral Population in Chronic Hepatitis C Virus Mapped by Ultradeep Pyrosequencing

Brendan A. Palmer,^a Zoya Dimitrova,^b Pavel Skums,^b Orla Crosbie,^c Elizabeth Kenny-Walsh,^c Liam J. Fanning^a

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

ABSTRACT

Hepatitis C virus (HCV) causes chronic infection in up to 50% to 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 intrahost evolution are poorly understood. In this study, we used ultradeep pyrosequencing (UDPS) to map the 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 indicate that polyvirus infection of hepatocytes has occurred but that the fitness quotients of recombinant daughter virions are too low for the daughter virions 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 ultradeep pyrosequencing data sets derived from virus amplicons frequently relies on software tools that are not optimized 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 made use of a temporally matched reference sequence data set to inform error correction algorithms. Using this methodology, we were able to (i) detect multiple instances of hepatitis C virus intrasubtype recombination at the E1/E2 junction (a phenomenon rarely reported in the literature) and (ii) interrogate the longitudinal quasispecies complexity of the virome. Parallel to the UDPS, isolation of IgG-bound virions was found to coincide with the collapse of specific viral subpopulations.

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 (~9.6-kb) RNA genome encodes two envelope glycoproteins, E1 and E2, which form noncovalent 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, and yet rapid evolution and adaptation result 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 neutralizing epitopes were described for HCV within hypervariable region 1 (HVR1) (8). HVR1 comprises the 27 N-terminal amino acids (aa) of E2 and displays the greatest heterogeneity of the entire HCV genome. Mutational change at HVR1 over time averts epitope recognition and contributes to immune escape (4, 8, 9). While linear and conformational sensitive epitopes outside HVR1 have also been identified as antigenic targets, it is 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 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 HVR1 (15). More-recent longitudinal studies of HVR1 variability have identified collapses in sequence heterogeneity, with latter sampling points dominated by a few host-adapted variants (16–18). Humoral immune evasion is not confined to amino acid modulation of the antigenic epitopes. Evidence is accumulating that gain or loss of glycosylation acceptor sites also contributes to concealment from neutralizing antibodies (nAb) (19, 20). Additionally, the observation that HCV can infect neighboring cells by direct cell-to-cell transmission identifies an innovative mechanism to avoid nAb targeting (21, 22). Collectively, these observations suggest mech-

Received 17 July 2014 Accepted 10 September 2014

Published ahead of print 17 September 2014

Editor: M. S. Diamond

Address correspondence to Liam J. Fanning, lfanning@ucc.ie.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/JVI.01732-14>.

Copyright © 2014, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JVI.01732-14

anisms through which highly refined, host-adapted sequences can evolve.

The stochastic exploration of novel sequence space can also be facilitated through recombination, and yet reports of recombination in HCV are rare (reviewed in reference 23). To date, a single recombination breakpoint spanning the NS2/NS3 gene junction for HCV intergenotypic recombinants has been identified (23, 24). This contrasts with intragenotype and intrasubtype recombinants, which can accommodate recombination breakpoints at multiple sites along the genome, including the E1/E2 junction (23–26). Furthermore, intragenotypic recombination has been observed to occur at greater frequencies *in vitro* than intergenotypic recombination (27). Taken together, this implies a greater acceptance of genetic exchange where there is an inherent shared homology between donor sequences.

The sample set used in this study was from a treatment-naïve individual chronically infected with HCV genotype 4a that has been reported previously at the 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 rise to an atypical 28-aa HVR1, facilitated further categorization into two sublineages (L1a and L1b). 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 intrasubtype recombinants (16). In the present study, we took advantage of ultradeep 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 chronicled the temporal dominance of a third sublineage of L1 which had previously been indiscernible. Finally, the rise of the monophyletic population to dominance was 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 to RL10), 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 had not been previously reported and was selected for analysis because it transected the 2.7-year gap between the times of collection of RL4 and RL6. RL12 (available subsequently to UDPS) was analyzed clonally, extending the study time frame 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 HVR1. 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 to 1526 of a reference genotype 4a strain (GenBank accession number DQ418782). The extension times used were 3-fold longer than those recommended for the polymerase (45 s/kb for Pwo; Roche) to reduce *in vitro* artifacts during PCR (28). Measures to protect against inter-sample contamination were employed (29). The mean viral titer was 5.9 HCV RNA log₁₀ IU/ml (range, 5.1 to 6.6 log₁₀ IU/ml). To ensure that the initial amount of the template was not limiting, we performed a 1:100 dilution of the viral RNA, which yielded an amplicon visualized by gel electrophoresis for each sample. Amplicons were purified using a PCR

TABLE 1 Motif sequences defining the quasispecies mosaic

Motif	HVR1	Sequence	Amplicon position
L1a_i	Typical	GCAGGGGTTGACGCCGGGACCATC	181–204
L1a_ii	Typical	GCAGGGGTTGACGCCAAGACCACC	181–204
L1b_i	Atypical	GCAGGAGTCGACGCCGGGACCCAC	181–204
L1b_ii	Atypical	GCAGGAGTTGACGCCAGGGCCAC	181–204
L1c_i	Typical	GCAGGGGTCGACGCCAAGACCAC	181–204
L1c_ii	Typical	GCAGGGGTTGACGCCAGGATCCCC	181–204
L2_i	Typical	GGGGCTGTAGCGTCCAGCAGCACT	211–234
L2_ii	Typical	GGGGCTGTGGCATCCAGCAACGCC	211–234

purification kit (Qiagen) and quantified using a Biophotometer (Eppendorf). Samples were prepared in equimolar concentrations and diluted to a final concentration of 1 × 10⁷ molecules/ml. Pyrosequencing was performed using a 454 GS FLX titanium platform with sample-specific multiplex identifier sequence-adapted libraries for Lib-1 sequencing (Roche 454 Life Sciences, Branford, CT).

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 phosphate-buffered saline (PBS) to a final volume of 100 µl and incubated with end-over-end mixing for 5 min at room temperature (RT). The flowthrough was collected by centrifugation, and the column was subjected to serial washes (*n* = 4) with PBS to remove any residual unbound virus followed by on-column lysis of virions. Where the protocol described above 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 multistep sequencing error correction and local haplotype reconstruction pipeline as described below.

Previous characterization of the quasispecies population within the sample set revealed the presence of a complex mix of variants with fluctuating lineage and sublineage compositions over time (16). Preliminary phylogenetic analysis of the UDPS data revealed the existence of a clearly distinguishable third sublineage of L1 (here L1c) in RL1 and RL5. Eight 24-bp HVR1 motifs, which defined haplotype groups L1a, L1b, L1c, and L2, were subsequently generated (Table 1). To increase the sensitivity of the 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 analyzed data and the absence of PCR and sequencing chimeras, reads that had more than a 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 did indeed represent PCR or sequencing chimeras. The obtained data sets were processed by the sequential application of the algorithms k-mer error correction (KEC) and a customized version of empirical threshold (ET) (31). Skums et al. (31) have previously demonstrated this process to be highly accurate in finding true haplotypes and removing false haplotypes.

KEC consists of the three stages. In stage 1, the set of k-mers (substring of fixed length *k*) of reads from the processed data set is calculated and the distribution of frequencies of k-mers is analyzed (31). It was previously 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. In stage 2, k-mers with frequencies lower than the error

threshold are considered erroneous and are 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, the 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. In stage 3, the reads containing k-mers that were not corrected in stage 2 are discarded. The following parameters of KEC were used: $k = 25$ and $i = 3$. To further identify and correct homopolymer errors, the data were postprocessed using ET and 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 analyzed unless otherwise stated. Of the 166 clonal nucleotide sequences, 54 were reisolated during this process. This frequency (0.33) reflects previous comparisons of clonal sequence recovery versus UDPS sequence recovery (35).

Detection of recombination. The presence of intrasubtype 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 of <0.05 were indicative of recombination (37). To better identify recombinants within sample sets, the Neighbor Net (NNet) algorithm was initially applied to the clonal data set using SplitsTree 4.13 with and without the inclusion of interlineage 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 (see Fig. S1 in the supplemental material). 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 utilized 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 HVR1 (39). Initial grouping of clonal sequences was performed by phylogenetic analysis using a general time-reversible model with gamma-distributed and invariant sites (GTR + G + I). Calculations were performed separately on the four haplotype groups using nucleotide sequences present at a frequency of >0.001 .

Evidence of site-specific directional selection was found using the directional evolution of protein sequences (DEPS) method implemented in conjunction with the Jones-Taylor-Thornton amino acid substitution model (40). Coevolving amino acid residues were identified using a Bayesian 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 analyses, a reduced data set, containing amino acid sequences present at a frequency of >0.001 , was used.

Sequence information and nucleotide sequence 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 KF417927 to KF417938 (RL5) and KC689336 to KC689341 (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 sublin-

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

age (L1c) existed. The haplotypes described were partitioned into four groups, namely, L1a, L1b, L1c, and L2, on the basis of these phylogenetic observations. 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 the corresponding data coincided with an equally significant decline in the number of L1a sequences (the frequency of which fell from >0.58 in RL4 to <0.07 in RL5). Only seven L1c sequences were recovered in RL6 and RL7 combined, with none recovered 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 flux in the virome structure that is readily observed only when analyzed over an expanded time frame. The L1c phenotype is indicative of a serendipitous return to temporal dominance following the reestablishment of a host environment spanning RL5 that favored the expansion of this group of variants.

Intrinsic population dynamics. Qualitatively, the mutual relationships between the four subpopulations revealed dominance of L1 over L2 from RL1 to RL7. The transition to L2 dominance coincided with the suppression of all L1 populations. Results of an intercorrelated variable test using the four viral groupings explain 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 such that the effects of a third variable on this interaction were removed (first-order partial correlation). It can be seen from Table 2 that the third variable typically suppressed 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 to which the population dynamics of L1b can explain 68% of the variation in L2 when the calculations are controlled for L1a (Table 2). Of note, L1a and L1b covariance rose

TABLE 2 Inter(sub)lineage correlations

Input variables ^a	Pearson correlation ^b		Intercorrelated variables ^c	First-order partial correlation	
	<i>r</i>	<i>r</i> ²		<i>r</i>	<i>r</i> ^{2d}
WX	−0.079	0.006	WX.Y	−0.093	0.009
			WX.Z	−0.648	0.420
WY	−0.239	0.057	WY.X	−0.243	0.059
			WY.Z	−0.478	0.228
WZ	−0.502	0.252	WZ.X	−0.750	0.563
			WZ.Y	−0.623	0.388
XY	−0.047	0.002	XY.W	−0.068	0.005
			XY.Z	−0.360	0.130
XZ	−0.671	0.450	XZ.W	−0.824	0.679
			XZ.Y	−0.721	0.520
YZ	−0.309	0.095	YZ.W	−0.510	0.260
			YZ.X	−0.459	0.211

^a W, L1a; X, L1b; Y, L1c; Z, L2.^b Lineage frequencies corresponding to each sample were used to build correlations.^c Data represent the degree of association between two (sub)lineages where the effects of a third (sub)lineage on this interaction are accounted for.^d Significant correlations are identified in bold.

from 0.6% to 42% when the suppressing effects of L2 were accounted for (Table 2). The proportion of variation that cannot be explained by the intercorrelated 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 occurred 4.6 years into the study timeline (Fig. 1; RL5). At that point, L1c reemerged as a dominant group within the sample space coincident with a L1a shift from neutral to positive selection and 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 and RL5 without any significant alteration to either the synonymous or nonsynonymous mutation rates. The L1b phenotype is subject to a 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 dominated the entire virome at RL7 (Fig. 1), the structure of which comprised of two equidominant E2 haplotypes (frequencies 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 of 0.16, preceding RL7 by 5.5 years), while the second haplotype was identified in RL3 (frequency of 0.11, preceding RL7 by 4.5 years). Ultimately, all L1 groups declined in number, with only three L1a haplotypes still detectable in RL10 at a combined frequency of <0.0006.

L2 haplotypes followed a trajectory divergent from that of L1 haplotypes. L2 haplotypes exhibited a higher rate of synonymous mutations early in the study that was accounted for by a subset of minor variants which had a HVR1 distinct from that seen in the dominant L2 population. During this longitudinal study, the dominant L2 haplotypes exhibited 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) differed by only a single synonymous polymorphism. The pressure to maintain such an exact replicate in

spite of the presence 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 had 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 ($n = 12,571$ reads). We know from clonal analysis of RL12 that a novel L2 HVR1 amino acid motif emerged 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 differed by a single nonsynonymous change (N395H) within HVR1. 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, emphasizing the sensitivity of the methodology used.

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

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 toward 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 intrasubtype 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. Second, 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 the 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 performed with Simplot confirmed **HM363402** to be an L1a-L1c recombinant with the anticipated crossover region identified just inside HVR1 (see Fig. S1A in the supplemental material). 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 to $<10^{-4}$). No significant correlation between the sample-specific PHI statistic and (i) the sample number, (ii) the percentage recovery of reads, (iii) the number of

FIG 2 (A) Number of unique full-length nucleotide haplotypes for each group (L1a, L1b, L1c, and L2 and the combined total) recovered per time point. (B) d_N (solid line) and d_S (dashed line) measurements across HVR1 are given for those nucleotide sequences with frequencies of >0.001 . HVR1 is defined between position 196 and position 276 or 279 of the amplicon sequence. The numbers 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_N and d_S values recovered when the elements in the data set were not segregated.

haplotypes per (sub)lineage, or (iv) the (sub)lineage sample proportion was observed (r^2 value range, 0.005 to 0.353). Representative examples of a high-diversity population (RL1) and a low-diversity population (RL8) were selected for further analysis (Fig. 3C and D). Haplotypes exhibiting conflicting phylogenetic profiles were found to be five interlineage recombinants and two pu-

tative intralineage recombinants from within the RL1 data (Fig. 3C). L1c sequences contributed to all five interlineage 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 (see Fig. S1D to H in the supplemental material). We note that the predicted recombi-

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

tion breakpoint regions were limited to the E1/E2 gene junction, which is in agreement with independent reports of HCV intrasubtype recombinants (25). There was no evidence of recombinant sequences identified in RL1 containing elements of L2 acting as a parental donor. This profile altered in RL8, where L2 sequences were identified as donors to each of the 7 interlineage recombinant sequences identified (Fig. 3D; see also Fig. S1I to O).

The formation of recombination artifacts *in vitro* during PCR has been shown to reflect the frequency of the constituent parental haplotypes (28). No recombinants were found in RL1 or RL8 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 argue in favor of authentic recombinants derived *in situ* within the liver,

given that L1c occupied <12% of the sample space in RL1 and yet served as a parental donor to all of the recombinants (excluding the two putative intralineage recombinants; Fig. 3C). For RL8, L2 comprised <23% of the sample space and 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 arising 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 JQ743309 clonal recombinant 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 JQ743309 (RL8) by 1.6 years. It was possible to identify intersample 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-characterized clonal data set where the patient cohort comprised 22 women exposed to HCV genotype 1b-contaminated blood products from a single source (15, 43). No evidence of recombination was observed in this instance.

Evolutionary evaluation of (sub)lineage-specific HVR1 populations. The presence of functional microdomains with HVR1 has been established for the H77 sequence (44). In broad terms, the first 13 residues of HVR1 (aa positions 384 to 396) regulate virus infectivity, coupled with putative masking of the CD81 binding site. A downstream region of nine residues (aa positions 400 to 408) is believed to contain the main neutralization 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 the replicative potential of a given genome. It was therefore of interest to assess whether there existed a preferential evolution toward specific HVR1 residues within the variant subpopulations over time.

DEPS analysis revealed 6, 7, and 4 HVR1 positions under directional selection for L1a, L1b, and L1c, respectively (Table 3). The DEPS data are indicative of selective sweeps governing the haplotype profile. There was no evidence to suggest that L2 HVR1 sites were under directional selection. These data are consistent with L2 sequences occupying a host-specific niche. In contrast, L1b had the highest level of directional mutational change within the neutralization epitope (Table 3).

The conservative maintenance of the overall physiochemical amino acid composition of 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 codependent 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 HVR1 exhibited codependencies, and these sites were predominantly also under directional selection (Tables 3 and 4). In agreement with the previous observations of L2 HVR1 stability, no codependent 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 to RL3, RL6, and RL7. In total, 12 unique nucleotide se-

quences (9 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 (5 of the 9 unique E2 amino acid sequences) were present in the UDPS data set. While each sequence ultimately demonstrated a marked decline postisolation, the time between detection on the column and nondetection 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 its level decreased to 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 nondetection of L2 sequences from the IgG depletion columns tallies with the minimal L2 sequence diversity together with the presence of HVR1 without detectable directional selection or epistatic evolution. The 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 HVR1 of L2 sequences, suggesting that these motifs are “invisible” to the immune system within this host during the period of infection analyzed here. This sequence stability points to an active mechanism of maintenance and indicates that the pervasive purifying selection observed over the preceding 10 years was the output of selection that preserved relative sequence homogeneity.

DISCUSSION

To interrogate complex UDPS data from a mixed-lineage HCV genotype 4a infection, we utilized 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 intrasubtype recombination events at the E1/E2 gene junction are 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 intrasubtype 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 observation and demonstrates the suitability of our strategy to the study of HCV UDPS data. Most of the recombinants reported here were present at frequencies of <0.001. As such, we cannot ignore the possibility that *in vitro* recombination artifacts may also be present in the final data set. The acceptance of low-frequency recombinants requires critical evaluation weighed against the documented preva-

FIG 4 Intersample maintenance of recombinants. (A) Composite phylogenetic network of the RL7-to-RL10 subset. Haplotype clusters displaying split decomposition signals are identified (i to iii). (B to D) Phylogenetic analysis of clusters i to iii, consensus sequences of the (sub)lineages, and interlineage recombinant sequences generated *in silico* using a rooted general time-reversible model (GTR + G + I). The data demonstrate intersample maintenance and evolution of recombinant haplotypes arguing in favor of the hypothesis of the recombinant sequences arising *in situ* within the host. Sequences are identified by time points as black circles (RL7), open squares (RL8), and black triangles (RL9). Sequences arising from RL10 did not feature in groups i to iii. A genotype 4a strain (Y11604) was used as the reference outgroup. Bootstrap values (of 1,000 resamplings) above 70 are shown. The genetic distance of 0.01 nucleotide substitutions per site is given by the scale bar.

TABLE 3 Sites under directional selection

Lineage	Site	Overall site composition	MRCA residue ^a	Inferred substitution ^b	EBF ^c
L1a	384	G ₂₅ K ₂₂	G	G ₁ ↔ ₂₂ K	G: >10 ⁵ ; K: >10 ⁵
	386	I ₂₅ T ₂₂	I	I ₁ ↔ ₂₂ T	T: >10 ⁵
	391	V ₂₆ A ₂₁	V	A ₂₁ ↔ ₁ V	A: >10 ⁵
	397	S ₃₁ N ₁₄ K ₂	N	N ₀ ↔ ₃₀ S	S: >10 ⁵
	403	L ₂₄ F ₂₃	F	F ₀ ↔ ₂₄ L	F: >10 ⁵
	408	K ₂₄ S ₁₅ R ₈	K	K ₁ ↔ ₈ R	
	408	K ₂₄ S ₁₅ R ₈	K	K ₀ ↔ ₁₅ S	S: >10 ⁵
L1b	349	T ₄₄ A ₂₆ S ₃	T	A ₂₅ ↔ ₁ T	A: >10 ⁵
	384	G ₄₈ R ₂₃ K ₂	G	G ₀ ↔ ₂₃ R	R: >10 ⁵
	385	T ₄₃ A ₂₃ P ₇	T	A ₂₃ ↔ ₀ T	A: >10 ⁵
	385	T ₄₃ A ₂₃ P ₇	T	P ₆ ↔ ₁ T	
	392	A ₆₆ P ₅ V ₂	A	A ₀ ↔ ₅ P	P: 238.3
	397	S ₂₉ K ₂₅ N ₁₇ R ₂	N	K ₂₄ ↔ ₀ N	K: >10 ⁵
	397	S ₂₉ K ₂₅ N ₁₇ R ₂	N	N ₀ ↔ ₂₈ S	
	398	H ₅₄ R ₁₂ N ₆ Q ₁	H	H ₀ ↔ ₆ N	
	398	H ₅₄ R ₁₂ N ₆ Q ₁	H	H ₀ ↔ ₁₂ R	R: 601.7
	404	D ₄₃ S ₁₅ H ₇ T ₅ Q ₁ A ₁ N ₁	D	D ₀ ↔ ₇ H	
	404	D ₄₃ S ₁₅ H ₇ T ₅ Q ₁ A ₁ N ₁	D	D ₁ ↔ ₁₅ S	S: >10 ⁵
	404	D ₄₃ S ₁₅ H ₇ T ₅ Q ₁ A ₁ N ₁	D	D ₀ ↔ ₅ T	
	405	S ₃₂ Y ₁₂ F ₁₁ P ₆ A ₅ V ₄ T ₃	S	A ₅ ↔ ₀ S	
	405	S ₃₂ Y ₁₂ F ₁₁ P ₆ A ₅ V ₄ T ₃	S	F ₁₁ ↔ ₀ S	
	405	S ₃₂ Y ₁₂ F ₁₁ P ₆ A ₅ V ₄ T ₃	S	P ₆ ↔ ₀ S	
	405	S ₃₂ Y ₁₂ F ₁₁ P ₆ A ₅ V ₄ T ₃	S	S ₀ ↔ ₄ V	
	405	S ₃₂ Y ₁₂ F ₁₁ P ₆ A ₅ V ₄ T ₃	S	S ₀ ↔ ₁₂ Y	Y: >10 ⁵
L1c	363	C ₁₂ S ₆	C	C ₀ ↔ ₆ S	S: >10 ⁵
	395	D ₁₂ N ₆	D	D ₁ ↔ ₆ N	N: 355.8
	397	K ₁₄ N ₄	K	K ₀ ↔ ₄ N	N: 301.6
	398	H ₁₃ S ₄ Q ₁	H	H ₀ ↔ ₄ S	S: >10 ⁵
	404	A ₁₄ S ₄	A	A ₀ ↔ ₄ S	S: 205.9
L2	342	S ₁₅ G ₈	S	G ₈ ↔ ₀ S	G: >10 ⁵

^a MRCA, most recent common ancestor.^b A_C↔_DB indicates C substitutions from A to B and D substitutions from B to A.^c Sites with an empirical Bayes factor (EBF) > 100 with preferential evolution toward the given residue are reported.

lence and characteristics of *in vitro* artifacts (28, 48–50). To that end, we submit four main arguments in support of our broader recombination findings. First, the clonal recombinant JQ743309 and an UDPS amino acid homolog were independently isolated from separate samples (RL6 and RL8) 1.6 years apart. Second, the demonstrable longevity observed with JQ743309 extends to other recombinants within the UDPS data set whereby homologous intersample recombinants (collected on occasions up to 3 years apart) were present (Fig. 4). This, coupled with observed evolu-

tion of recombinant sequences (RL7 to RL9) within the UDPS data set, argues against the chance occurrence of recombinants significantly contributing to our findings (28). Third, the collapse of L1 sequences in the latter samples (RL8 to RL10) 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 were examined in depth, the recombinant profile did not reflect the lineage frequencies present (28). Our results argue in favor of recombinants being rare but real in the context of this chronic infection.

Both replicative (strand switching by the polymerase) and nonreplicative (ligation of genomic fragments by host ligases) mechanisms of recombination in HCV are postulated to occur (reviewed in reference 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 infectivity titers comparable to those of 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 few *in vivo* recombinants have since been reported even when high-risk pop-

TABLE 4 Codependent epistatic evolution

Lineage	Site 1	Site 2	P {S1→S2} ^a	P {S1←S2} ^b	P {S1↔S2} ^c
L1a	G384K	I386T	0.41	0.53	0.94
	L403F	K408R	0.35	0.61	0.96
L1b	T385A	G384R	0.86	0.13	0.99
	T385A	S405Y	0.76	0.24	1.00
L1c	T396A	D395N	0.44	0.52	0.96

^a The posterior probability (P) for site 2 was conditionally dependent on site 1.^b The posterior probability for site 1 was conditionally dependent on site 2.^c The posterior probabilities for site 1 and site 2 were conditionally dependent.

TABLE 5 Haplotypes targeted by humoral immune activity

Sample	IgG-bound virus presence	GenBank accession no.	Detectable in pyro data	Lineage
RL1	+	GQ985330	—	1b
	+	GQ985332	+	1c
	+	GQ985333	+	1a
	+	GQ985336	+	1b
	+	GQ985337	—	1b
	+	GQ985350	+	1a
RL2	+	FJ744071	—	1a
RL3	+	EU482135	—	1a
RL4	—	N/A	N/A ^a	N/A
RL5	—	N/A	N/A	N/A
RL6	+	HM363384	+	1b
	+	HM363385	+	1b
	+	HM363386	+	1b
RL7	+	GQ985372	+	1a
RL8	—	N/A	N/A	N/A
RL9	—	N/A	N/A	N/A
RL10	—	N/A	N/A	N/A

^a N/A, not applicable.

ulations have been specifically investigated (reviewed in reference 54). To our knowledge, this is the first report documenting multiple intrasubtype HCV recombinants from UDPS data. We suggest that studies designed to identify recombinants should do so with a focus on inter- as well as intragenotypic 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 and RL6 that were mirrored by L1b and L2 thereafter (Fig. 2 and Table 2). Collectively, the data document a haplotype group (L1c) struggling to gain purchase within a competitive replicative space. A sudden shift in the environmental landscape favored 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 members of the L1c group, which, prior to RL4, were detectable only 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 in the data corresponding to the time from RL1 to RL7 (Table 2). The architectural organization 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 the positioning of L2 in the sample space (Fig. 2). In spite of the gross perturbation of the virome, the viral load increased from RL9 to RL12 (HCV RNA, 5.1 log₁₀ IU/ml to 6.6 log₁₀ IU/ml, respectively). The 321-bp amplicon used in this study does not allow us to account for selection pressures outside the E1/E2 gene junction, such as major histocompatibility complex (MHC) class I-restricted alleles in NS3 (15, 62). However, given the continued maintenance of the L2 HVR1 epitope following the expansion of this lineage, it is likely that L2 had replicative fitness similar to that of L1 (and yet was 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 various degrees over time, given (i) the rise in viral load and (ii) 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 at >30%. It has been suggested that such high-frequency variants may elicit a strong neutralizing response to themselves and closely related variants (17).

Few studies have documented longitudinal quasispecies dynamics in treatment-naïve individuals chronically infected with HCV. 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 HVR1 over time (64). It has been argued that recombination would be detrimental to beneficial mutations accumulated through codependent epistatic evolution (65). The restriction of recombination to gene junctions alleviates this problem (23, 25). Of the five codependent epistatic sites identified here, there exists the potential for all five to be transferred intact to the daughter virions, based on the predicted crossover regions (Table 4; see also Fig. S1 in the supplemental material).

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 recognize that the context of an infection determines, first, whether or not recombination will occur and, second, whether or not it will be detectable. Mixed-lineage infections of one subtype, such as the infection described here, may uniquely meet all the

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. The percentages (sample proportion) of these sequences at the nucleotide (gray lines) and amino acid (black lines) levels within the overall sample population are given. The respective sequence accession numbers are [GQ985332](#) (A), [GQ985336](#) (B), [GQ985372](#) (C), [HM363384](#) (D), and [GQ985333](#) (E). Black arrows indicate the sampling point.

prerequisites for measurable recombination, i.e., (i) multivariant infection of single hepatocytes, (ii) sufficient divergence of intra-host communities, and (iii) informed analysis of complex data sets that account for hypervariable heterogeneity.

ACKNOWLEDGMENTS

We thank David S. Campo (Division of Viral Hepatitis, Centers for Disease Control and Prevention) for his critical reading of the manuscript. We also thank John Levis and Kevin Hegarty for initial characterization of patient sera.

REFERENCES

- Yamane D, McGivern DR, Masaki T, Lemon SM. 2013. Liver injury and disease pathogenesis in chronic hepatitis C. *Curr. Top. Microbiol. Immunol.* 369:263–288. http://dx.doi.org/10.1007/978-3-642-27340-7_11.
- Deleersnyder V, Pillez A, Wychowski C, Blight K, Xu J, Hahn YS, Rice CM, Dubuisson J. 1997. Formation of native hepatitis C virus glycoprotein complexes. *J. Virol.* 71:697–704.
- Penin F, Combet C, Germanidis G, Frainais PO, Deleage G, Pawlotsky JM. 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. <http://dx.doi.org/10.1128/JVI.75.12.5703-5710.2001>.
- von Hahn T, Yoon JC, Alter H, Rice CM, Rehmann B, Balfe P, McKeating JA. 2007. Hepatitis C virus continuously escapes from neutralizing antibody and T-cell responses during chronic infection in vivo. *Gastroenterology* 132:667–678. <http://dx.doi.org/10.1053/j.gastro.2006.12.008>.
- Wang GP, Sherrill-Mix SA, Chang KM, Quince C, Bushman FD. 2010. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *J. Virol.* 84:6218–6228. <http://dx.doi.org/10.1128/JVI.02271-09>.
- Martell M, Esteban JI, Quer J, Genesca J, Weiner A, Esteban R, Guardia J, Gomez 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.
- Duffy S, Shackleton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9:267–276. <http://dx.doi.org/10.1038/nrg2323>.
- Kato N, Sekiya H, Ootsuyama Y, Nakazawa T, Hijikata M, Ohkoshi S, Shimotohno K. 1993. Humoral immune response to hypervariable region 1 of the putative envelope glycoprotein (gp70) of hepatitis C virus. *J. Virol.* 67:3923–3930.

9. Farci P, Shimoda A, Wong D, Cabezon T, De Gioannis D, Strazzer A, Shimizu Y, Shapiro M, Alter HJ, Purcell RH. 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. <http://dx.doi.org/10.1073/pnas.93.26.15394>.
10. Keck ZY, Sung VM, Perkins S, Rowe J, Paul S, Liang TJ, Lai MM, Fong SK. 2004. Human monoclonal antibody to hepatitis C virus E1 glycoprotein that blocks virus attachment and viral infectivity. *J. Virol.* 78:7257–7263. <http://dx.doi.org/10.1128/JVI.78.13.7257-7263.2004>.
11. Broering TJ, Garrity KA, Boatright NK, Sloan SE, Sandor F, Thomas WD, Jr, Szabo G, Finberg RW, Ambrosino DM, Babcock GJ. 2009. Identification and characterization of broadly neutralizing human monoclonal antibodies directed against the E2 envelope glycoprotein of hepatitis C virus. *J. Virol.* 83:12473–12482. <http://dx.doi.org/10.1128/JVI.01138-09>.
12. Farci P, Shimoda A, Coiana A, Diaz G, Peddis G, Melpolder JC, Strazzer A, Chien DY, Munoz SJ, Balestrieri A, Purcell RH, Alter HJ. 2000. The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* 288:339–344. <http://dx.doi.org/10.1126/science.288.5464.339>.
13. Farci P, Strazzer R, Alter HJ, Farci S, Degioannis D, Coiana A, Peddis G, Usai F, Serra G, Chessa L, Diaz G, Balestrieri A, Purcell RH. 2002. Early changes in hepatitis C viral quasispecies during interferon therapy predict the therapeutic outcome. *Proc. Natl. Acad. Sci. U. S. A.* 99:3081–3086. <http://dx.doi.org/10.1073/pnas.052712599>.
14. Li H, McMahon BJ, McArdle S, Bruden D, Sullivan DG, Shelton D, Deubner H, Gretsch DR. 2008. Hepatitis C virus envelope glycoprotein co-evolutionary dynamics during chronic hepatitis C. *Virology* 375:580–591. <http://dx.doi.org/10.1016/j.virol.2008.02.012>.
15. Ray SC, Fanning L, Wang XH, Netski DM, Kenny-Walsh E, Thomas DL. 2005. Divergent and convergent evolution after a common-source outbreak of hepatitis C virus. *J. Exp. Med.* 201:1753–1759. <http://dx.doi.org/10.1084/jem.20050122>.
16. Palmer BA, Moreau I, Levis J, Harty C, Crosbie O, Kenny-Walsh E, Fanning LJ. 2012. Insertion and recombination events at hypervariable region 1 over 9.6 years of hepatitis C virus chronic infection. *J. Gen. Virol.* 93:2614–2624. <http://dx.doi.org/10.1099/vir.0.045344-0>.
17. Ramachandran S, Campo DS, Dimitrova ZE, Xia GL, Purdy MA, Khudyakov YE. 2011. Temporal variations in the hepatitis C virus intra-host population during chronic infection. *J. Virol.* 85:6369–6380. <http://dx.doi.org/10.1128/JVI.02204-10>.
18. Li H, Hughes AL, Bano N, McArdle S, Livingston S, Deubner H, McMahon BJ, Townsend-Bulson L, McMahon R, Rosen HR, Gretsch DR. 2011. Genetic diversity of near genome-wide hepatitis C virus sequences during chronic infection: evidence for protein structural conservation over time. *PLoS One* 6:e19562. <http://dx.doi.org/10.1371/journal.pone.0019562>.
19. Helle F, Duverlie G, Dubuisson J. 2011. The hepatitis C virus glycan shield and evasion of the humoral immune response. *Viruses* 3:1909–1932. <http://dx.doi.org/10.3390/v3101909>.
20. Pantua H, Diao J, Ullsch M, Hazen M, Mathieu M, McCutcheon K, Takeda K, Date S, Cheung TK, Phung Q, Hass P, Arnott D, Hongo JA, Matthews DJ, Brown A, Patel AH, Kelley RF, Eigenbrot C, Kapadia SB. 2013. Glycan shifting on hepatitis C virus (HCV) E2 glycoprotein is a mechanism for escape from broadly neutralizing antibodies. *J. Mol. Biol.* 425:1899–1914. <http://dx.doi.org/10.1016/j.jmb.2013.02.025>.
21. Brimacombe CL, Grove J, Meredith LW, Hu K, Syder AJ, Flores MV, Timpe JM, Krieger SE, Baumert TF, Tellinghuisen TL, Wong-Staal F, Balfe P, McKeating JA. 2011. Neutralizing antibody-resistant hepatitis C virus cell-to-cell transmission. *J. Virol.* 85:596–605. <http://dx.doi.org/10.1128/JVI.01592-10>.
22. Timpe JM, Stamatakis Z, Jennings A, Hu K, Farquhar MJ, Harris HJ, Schwarz A, Desombere I, Roels GL, Balfe P, McKeating JA. 2008. Hepatitis C virus cell-cell transmission in hepatoma cells in the presence of neutralizing antibodies. *Hepatology* 47:17–24. <http://dx.doi.org/10.1002/hep.21959>.
23. Morel V, Fournier C, Francois C, Brochot E, Helle F, Duverlie G, Castelain S. 2011. Genetic recombination of the hepatitis C virus: clinical implications. *J. Viral Hepat.* 18:77–83. <http://dx.doi.org/10.1111/j.1365-2893.2010.01367.x>.
24. Galli A, Bukh J. 2014. Comparative analysis of the molecular mechanisms of recombination in hepatitis C virus. *Trends Microbiol* 22:354–364. <http://dx.doi.org/10.1016/j.tim.2014.02.005>.
25. Sentandreu V, Jimenez-Hernandez N, Torres-Puente M, Bracho MA, Valero A, Gosalbes MJ, Ortega E, Moya A, Gonzalez-Candelas F. 2008. Evidence of recombination in inpatient populations of hepatitis C virus. *PLoS One* 3:e3239. <http://dx.doi.org/10.1371/journal.pone.0003239>.
26. Gao F, Nainan OV, Khudyakov Y, Li J, Hong Y, Gonzales AC, Spelbring J, Margolis HS. 2007. Recombinant hepatitis C virus in experimentally infected chimpanzees. *J. Gen. Virol.* 88:143–147. <http://dx.doi.org/10.1099/vir.0.82263-0>.
27. Scheel TK, Galli A, Li YP, Mikkelsen LS, Gottwein JM, Bukh J. 2013. Productive homologous and non-homologous recombination of hepatitis C virus in cell culture. *PLoS Pathog.* 9:e1003228. <http://dx.doi.org/10.1371/journal.ppat.1003228>.
28. Lahr DJ, Katz LA. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47:857–866.
29. Kwok S, Higuchi R. 1989. Avoiding false positives with PCR. *Nature* 339:237–238. <http://dx.doi.org/10.1038/339237a0>.
30. Moreau I, O'Sullivan H, Murray C, Levis J, Crosbie O, Kenny-Walsh E, Fanning LJ. 2008. Separation of hepatitis C genotype 4a into IgG-depleted and IgG-enriched fractions reveals a unique quasispecies profile. *Virol. J.* 5:103. <http://dx.doi.org/10.1186/1743-422X-5-103>.
31. Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, Yokosawa J, Zelikovsky A, Khudyakov Y. 2012. Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* 13(Suppl 10):S6. <http://dx.doi.org/10.1186/1471-2105-13-S10-S6>.
32. Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: does the read length matter? *Genome Res.* 19:336–346. <http://dx.doi.org/10.1101/gr.079053.108>.
33. Chaisson MJ, Pevzner PA. 2008. Short read fragment assembly of bacterial genomes. *Genome Res.* 18:324–330. <http://dx.doi.org/10.1101/gr.7088808>.
34. Zhao X, Palmer LE, Bolanos R, Mircean C, Fasulo D, Wittenberg GM. 2010. EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol.* 17:1549–1560. <http://dx.doi.org/10.1089/cmb.2010.0127>.
35. Dimitrova Z, Campo DS, Ramachandran S, Vaughan G, Ganova-Raeva L, Lin Y, Forbi JC, Xia G, Skums P, Pearlman B, Khudyakov Y. 2011. Evaluation of viral heterogeneity using next-generation sequencing, endpoint limiting-dilution and mass spectrometry. *In Silico Biol.* 11:183–192. <http://dx.doi.org/10.3233/ISB-2012-0453>.
36. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254–267. <http://dx.doi.org/10.1093/molbev/msj030>.
37. Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681. <http://dx.doi.org/10.1534/genetics.105.048975>.
38. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73:152–160.
39. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28:2731–2739. <http://dx.doi.org/10.1093/molbev/msr121>.
40. Kosakovsky Pond SL, Poon AF, Leigh Brown AJ, Frost SD. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol. Biol. Evol.* 25:1809–1824. <http://dx.doi.org/10.1093/molbev/msn123>.
41. Poon AF, Lewis FI, Frost SD, Kosakovsky Pond SL. 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* 24:1949–1950. <http://dx.doi.org/10.1093/bioinformatics/btn313>.
42. Poon AF, Lewis FI, Pond SL, Frost SD. 2007. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput. Biol.* 3:e231. <http://dx.doi.org/10.1371/journal.pcbi.0030231>.
43. Kenny-Walsh E. 1999. Clinical outcomes after hepatitis C infection from contaminated anti-D immune globulin. *N. Engl. J. Med.* 340:1228–1233.
44. Guan M, Wang W, Liu X, Tong Y, Liu Y, Ren H, Zhu S, Dubuisson J, Baumert TF, Zhu Y, Peng H, Aurelian L, Zhao P, Qi Z. 2012. Three different functional microdomains in the hepatitis C virus hypervariable

- region 1 (HVR1) mediate entry and immune evasion. *J. Biol. Chem.* 287: 35631–35645. <http://dx.doi.org/10.1074/jbc.M112.382341>.
45. Bankwitz D, Steinmann E, Bitzegeio J, Ciesek S, Friesland M, Herrmann E, Zeisel MB, Baumert TF, Keck ZY, Fong SK, Pecheur EI, Pietschmann T. 2010. Hepatitis C virus hypervariable region 1 modulates receptor interactions, conceals the CD81 binding site, and protects conserved neutralizing epitopes. *J. Virol.* 84:5751–5763. <http://dx.doi.org/10.1128/JVI.02200-09>.
 46. Hino K, Korenaga M, Orito E, Katoh Y, Yamaguchi Y, Ren F, Kitase A, Satoh Y, Fujiwara D, Okita K. 2002. Constrained genomic and conformational variability of the hypervariable region 1 of hepatitis C virus in chronically infected patients. *J. Viral Hepat.* 9:194–201. <http://dx.doi.org/10.1046/j.1365-2893.2002.00349.x>.
 47. Salemi M, Gray RR, Goodenow MM. 2008. An exploratory algorithm to identify intra-host recombinant viral sequences. *Mol. Phylogenet. Evol.* 49:618–628. <http://dx.doi.org/10.1016/j.ympev.2008.08.017>.
 48. Gregori J, Esteban JI, Cubero M, García-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. 2013. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. *PLoS One* 8:e83361. <http://dx.doi.org/10.1371/journal.pone.0083361>.
 49. Görzer I, Guelly C, Trajanoski S, Puchhammer-Stöckl E. 2010. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. *J. Virol. Methods* 169:248–252. <http://dx.doi.org/10.1016/j.jviromet.2010.07.040>.
 50. Shao W, Boltz VF, Spindler JE, Kearney MF, Maldarelli F, Mellors JW, Stewart C, Volfovsky N, Levitsky A, Stephens RM, Coffin JM. 2013. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology* 10:18. <http://dx.doi.org/10.1186/1742-4690-10-18>.
 51. Reiter J, Perez-Vilaro G, Scheller N, Mina LB, Diez J, Meyerhans A. 2011. Hepatitis C virus RNA recombination in cell culture. *J. Hepatol.* 55:777–783. <http://dx.doi.org/10.1016/j.jhep.2010.12.038>.
 52. Carlsen TH, Scheel TK, Ramirez S, Fong SK, Bukh J. 2013. Characterization of hepatitis C virus recombinants with chimeric E1/E2 envelope proteins and identification of single amino acids in the E2 stem region important for entry. *J. Virol.* 87:1385–1399. <http://dx.doi.org/10.1128/JVI.00684-12>.
 53. Kalinina O, Norder H, Mukomolov S, Magnusius LO. 2002. A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *J. Virol.* 76:4034–4043. <http://dx.doi.org/10.1128/JVI.76.8.4034-4043.2002>.
 54. González-Candelas F, López-Labrador FX, Bracho MA. 2011. Recombination in hepatitis C virus. *Viruses* 3:2006–2024. <http://dx.doi.org/10.3390/v3102006>.
 55. Ruiz-Jarabo CM, Arias A, Baranowski E, Escarmis C, Domingo E. 2000. Memory in viral quasispecies. *J. Virol.* 74:3543–3547. <http://dx.doi.org/10.1128/JVI.74.8.3543-3547.2000>.
 56. Llaure AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog.* 6:e1001005. <http://dx.doi.org/10.1371/journal.ppat.1001005>.
 57. Ruiz-Jarabo CM, Arias A, Molina-París C, Briones C, Baranowski E, Escarmis C, Domingo E. 2002. Duration and fitness dependence of quasispecies memory. *J. Mol. Biol.* 315:285–296. <http://dx.doi.org/10.1006/jmbi.2001.5232>.
 58. de la Torre JC, Holland JJ. 1990. RNA virus quasispecies populations can suppress vastly superior mutant progeny. *J. Virol.* 64:6278–6281.
 59. Webster B, Wissing S, Herker E, Ott M, Greene WC. 2013. Rapid intracellular competition between hepatitis C viral genomes as a result of mitosis. *J. Virol.* 87:581–596. <http://dx.doi.org/10.1128/JVI.01047-12>.
 60. Gismondi MI, Díaz Carrasco JM, Valva P, Becker PD, Guzmán CA, Campos RH, Preciado MV. 2013. Dynamic changes in viral population structure and compartmentalization during chronic hepatitis C virus infection in children. *Virology* 447:187–196. <http://dx.doi.org/10.1016/j.virol.2013.09.002>.
 61. Kandathil AJ, Graw F, Quinn J, Hwang HS, Torbenson M, Perelson AS, Ray SC, Thomas DL, Ribeiro RM, Balagopal A. 2013. Use of laser capture microdissection to map hepatitis C virus-positive hepatocytes in human liver. *Gastroenterology* 145:1404–1413.e10. <http://dx.doi.org/10.1053/j.gastro.2013.08.034>.
 62. Bailey JR, Laskey S, Wasilewski LN, Munshaw S, Fanning LJ, Kenny-Walsh E, Ray SC. 2012. Constraints on viral evolution during chronic hepatitis C virus infection arising from a common-source exposure. *J. Virol.* 86:12582–12590. <http://dx.doi.org/10.1128/JVI.01440-12>.
 63. Sullivan DG, Bruden D, Deubner H, McArdle S, Chung M, Christensen C, Hennessy T, Homan C, Williams J, McMahon BJ, Gretch DR. 2007. Hepatitis C virus dynamics during natural infection are associated with long-term histological outcome of chronic hepatitis C disease. *J. Infect. Dis.* 196:239–248. <http://dx.doi.org/10.1086/518895>.
 64. Campo DS, Dimitrova Z, Yokosawa J, Hoang D, Perez NO, Ramachandran S, Khudyakov Y. 2012. Hepatitis C virus antigenic convergence. *Sci. Rep.* 2:267. <http://dx.doi.org/10.1038/srep00267>.
 65. Sanjuán R, Moya A, Elena SF. 2004. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc. Natl. Acad. Sci. U. S. A.* 101:15376–15379. <http://dx.doi.org/10.1073/pnas.0404125101>.