Evolutionary Analysis Provides Insight Into the Origin and Adaptation of HCV

Diego Forni[1], Rachele Cagliani[1], Chiara Pontremoli[1], Uberto Pozzoli[1],
Jacopo Vertemara[2], Luca De Gioia[2], Mario Clerici[3,4] and Manuela Sironi[1]*

[1] Bioinformatics Laboratory, Scientific Institute IRCCS E.Medea, Bosisio Parini, Italy, [2] Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan, Italy, [3] Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy, [4] Don C. Gnocchi Foundation Onlus, IRCCS, Milan, Italy

Hepatitis C virus (HCV) belongs to the *Hepacivirus* genus and is genetically heterogeneous, with seven major genotypes further divided into several recognized subtypes. HCV origin was previously dated in a range between ~200 and 1000 years ago. Hepaciviruses have been identified in several domestic and wild mammals, the largest viral diversity being observed in bats and rodents. The closest relatives of HCV were found in horses/donkeys (equine hepaciviruses, EHV). However, the origin of HCV as a human pathogen is still an unsolved puzzle. Using a selection-informed evolutionary model, we show that the common ancestor of extant HCV genotypes existed at least 3000 years ago (CI: 3192–5221 years ago), with the oldest genotypes being endemic to Asia. EHV originated around 1100 CE (CI: 291–1640 CE). These time estimates exclude that EHV transmission was mainly sustained by widespread veterinary practices and suggest that HCV originated from a single zoonotic event with subsequent diversification in human populations. We also describe a number of biologically important sites in the major HCV genotypes that have been positively selected and indicate that drug resistance-associated variants are significantly enriched at positively selected sites. HCV exploits several cell-surface molecules for cell entry, but only two of these (CD81 and OCLN) determine the species-specificity of infection. Herein evolutionary analyses do not support a long-standing association between primates and hepaciviruses, and signals of positive selection at CD81 were only observed in Chiroptera. No evidence of selection was detected for OCLN in any mammalian order. These results shed light on the origin of HCV and provide a catalog of candidate genetic modulators of HCV phenotypic diversity.

Keywords: hepatitis C virus, equine hepacivirus, molecular dating, tMRCA, positive selection, resistance-associated amino acid variants, CD81

INTRODUCTION

Hepatitis C virus (HCV, genus *Hepacivirus*, family *Flaviviridae*) is a hepatotropic human pathogen with an estimated worldwide seroprevalence around 2.8% (Mohd Hanafiah et al., 2013).

The seven major HCV genotypes display remarkable antigenic variability and are classified into several recognized subtypes (Smith et al., 2014). A small number of "epidemic" subtypes (1a, 1b, 3a, and 2a) account for the overwhelming majority of infections worldwide (Simmonds, 2013;

Messina et al., 2015). Their spread occurred recently (in the last 50–100 years), following the development of practices that cause parenteral exposure (Pybus et al., 2001; Simmonds, 2013; Jackowiak et al., 2014). The epidemic subtypes represent a small fraction of HCV diversity. In sub-Saharan Africa and South-East Asia, the pattern of HCV diversity is characterized by highly divergent subtypes of the same genotype dominating transmissions across geographically contiguous areas ("endemic" transmission) (Simmonds, 2013; Jackowiak et al., 2014).

Hepaciviruses have been identified in several domestic and wild mammals, the largest viral diversity being observed in bats and rodents (Kapoor et al., 2011, 2013; Burbelo et al., 2012; Drexler et al., 2013; Quan et al., 2013; Corman et al., 2015; Scheel et al., 2015; Walter et al., 2016; Van Nguyen et al., 2018). The closest relatives of HCV were found in horses/donkeys and dogs (equine and canine hepaciviruses, EHV, and CHV) (Pybus and Gray, 2013; Pybus and Theze, 2016), but the origin of HCV as a human pathogen is still an unsolved puzzle. Some authors envisaged the possibility that HCV evolved from a horse-to-human transmission event (Pfaender et al., 2014; Scheel et al., 2015), others suggested that HCV originated in relatively recent times from one or multiple cross-species transmission events from a still to be defined species (Pybus and Gray, 2013; Scheel et al., 2015; Pybus and Theze, 2016). However, its strict species-specificity, as well as the ability of HCV to persist lifelong in humans, led to the alternative hypothesis whereby HCV-related viruses have been infecting humans and other primates throughout their evolutionary history (Simmonds, 2013; Scheel et al., 2015).

Another open question is whether HCV genotypes differ in terms of transmissibility or disease outcome. Conversely, the viral genotype is known to represent a predictive factor for antiviral treatment success. Treatment with pegylated interferon and ribavirin (pegIFN/RBV) is less likely to be successful in patients infected with genotypes 1 and 4 compared to those infected with genotypes 2 and 3 (Bartenschlager et al., 2013), and variants at the human *IFNL3/IFNL4* locus were reproducibly associated with pegIFN/RBV treatment outcome (O'Brien et al., 2014). Interestingly, the same polymorphisms are strongly associated with spontaneous HCV clearance (O'Brien et al., 2014). Recently, direct-acting antivirals (DAAs) have greatly improved the efficacy of treatment strategies for HCV. However, DAAs have limited pan-genotypic activity and tend to select for resistance-associated amino acid variants (RAVs) (Lontok et al., 2015). Most RAVs naturally occur in treatment-naive patients and in some instances RAV prevalence differs among genotypes or subtypes (Rai and Deval, 2011; Bartenschlager et al., 2013; Lontok et al., 2015; Cannalire et al., 2016; Chen et al., 2016; Patino-Galindo et al., 2016).

Herein, we apply an evolutionary approach to shed light into HCV origin, to analyze its adaptation to human populations, and to verify if the emergence of drug-resistant variants is a result of positive selection. We also analyzed the most important cell-surface molecules that mediate HCV cell entry to determine whether HCV or related hepaciviruses exerted a selective pressure on these receptors and whether selection was particularly strong in specific mammalian orders or superorders.

MATERIALS AND METHODS

Evolutionary Analysis of HCV Receptors

Coding sequences for mammalian *OCLN*, *CD81*, *CLDN1*, and *SCARB1* (encoding SRB1) were retrieved from the NCBI[1] database. A list of species is available in the Supplementary Table S1. DNA alignments were performed with the RevTrans 2.0 utility[2], MAFFT v6.240 as an aligner) (Wernersson and Pedersen, 2003), which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. All alignments were screened for the presence of recombination using Genetic Algorithm Recombination Detection (GARD) (Kosakovsky Pond et al., 2006) and two methods (RDP and GENECONV) (Sawyer, 1989; Martin and Rybicki, 2000) implemented in the RDP4 program (Martin et al., 2017). RDP and GENECONV were selected because they showed good power in previous simulation analyses (Posada and Crandall, 2001; Bay and Bielawski, 2011) and only breakpoints identified by both methods were accepted. The cutoff p-value was set to 0.01 in both GARD and RDP4. No method detected recombination in any alignment.

After running a codon model selection analysis in HYPHY (Delport et al., 2010), gene trees were generated by maximum-likelihood using phyML with the approximate likelihood-ratio test (aLRT) method (Guindon et al., 2010).

To search for positive selection, we applied the site models implemented in PAML (Yang, 2007). Specifically, a model (M8, positive selection model) that allows a class of sites to evolve with dN/dS > 1 was compared to two models (M7 and M8a, neutral models) that do not allow dN/dS > 1 (Yang, 2007). Positively selected sites were identified through Bayes Empirical Bayes (BEB) analysis (Anisimova et al., 2002; Yang et al., 2005) from model M8 (with a posterior probability cutoff of 0.90) and through Mixed Effects Model of Evolution (MEME), with a default p-value cutoff of 0.10) (Murrell et al., 2012). Only sites detected by both methods were considered as positively selected. Very similar results were obtained using the gene tree obtained with PhyML or the species tree as inputs for PAML analysis.

Time Estimates

To estimate the time to the most common ancestor (tMRCA) of the 7 HCV genotypes, we used alignments of 67 HCV sequences with known isolation dates (Supplementary Table S2). Viral sequences were retrieved from the NCBI[1] database. We used PRANK (Loytynoja and Goldman, 2005) to generate multiple sequence alignments and GUIDANCE (Sela et al., 2015) for filtering unreliably aligned codons (we masked codons with a score < 0.90), as suggested (Privman et al., 2012).

The NS5B region we used for dating is non-recombining, as assessed by GARD and RDP4 analyses of the entire non-structural portion of the genome. This region was selected because it is one of the most conserved across HCV genotypes (Kapoor et al., 2011) and required very minor filtering.

---

[1]http://www.ncbi.nlm.nih.gov/
[2]http://www.cbs.dtu.dk/services/RevTrans/

The tMRCA of EHV/CHV was estimated on a phylogeny of 18 sequences with complete coding sequence information: 17 EHV isolated from horses or donkeys and 1 CHV (Kapoor et al., 2011; Burbelo et al., 2012; Walter et al., 2016) (Supplementary Table S3). GUIDANCE detected no uncertainty in the alignments (all codons had a score > 0.9) and GARD detected a breakpoint at position 6729. In good agreement, RDP4 detected a breakpoint at position 6747. The terminal 2145 nucleotides were thus removed and the resulting alignment was used for tree construction and tMRCA estimate.

To determine whether there was sufficient temporal structure in the NS5B and EHV phylogenies to estimate divergence times, we calculate the correlation coefficients (r) of regressions of root-to-tip genetic distances against sequence sampling times (Murray et al., 2016). A method that minimizes the residual mean squares of the models, rather than one that maximizes r^2, was applied, as suggested (Murray et al., 2016). The p-values were calculated by performing 1,000 clustered permutations of the sampling dates (Duchene et al., 2015; Murray et al., 2016). Evidence for temporal structure was obtained for both phylogenies (NS5B: $r = 0.25$, $r^2 = 0.0625$, p-value $= 0.024$; EHV: $r = 0.35$, $r^2 = 0.1225$, p-value $= 0.013$).

The action of saturation and purifying selection can underestimate the tMRCA, and selection-informed models can improve branch length estimation (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013, 2014). We thus applied a branch-site test (aBS-REL, adaptive branch-site random effects likelihood) (Smith et al., 2015) to estimate branch lengths while taking into account the effect of different selective pressures among lineages in the phylogeny. Previous works showed that, in the presence of saturation and selection, aBS-REL estimates branch lengths more reliably than other commonly used models such as the general time-reversible substitution model with a four-bin gamma rate distribution (GTR+Γ_4) (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013).

Estimate of divergence times was performed by a penalized-likelihood (PL) method implemented in r8s (Sanderson, 2003), using the aBS-REL tree as the input tree. In the case of 3 extremely long (i.e., more than 50 substitution/site) terminal branches, most likely resulting from low precision in point estimates of dN/dS, MG94 branch lengths were used instead of aBS-REL lengths to avoid biases in the tMRCA estimates. MG94 (Muse and Gaut, 1994) is the baseline model used by aBS-REL and it models synonymous and nonsynonymous substitution rate variation across branches but not across sites (Muse and Gaut, 1994).

The PL method in r8s employs a smoothing parameter, which represents how much the assumption of a molecular clock has been relaxed. Cross-validation was run to determine the best smoothing value for the aBS-REL tree (Sanderson, 2003).

A latin hypercube sampling scheme (LHC) was used to sample from the aBS-REL parameter distributions so as to estimate the confidence interval, as previously suggested (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013, 2014). Briefly, 500 samples were drawn from aBS-REL analyses to estimate branch length variance, 500 trees were generated, and then used as input

trees for r8s. The upper and the lower 95% bounds are used as confidence intervals.

As a comparison, branch lengths were also estimated using phyML with a maximum-likelihood approach and a GTR+Γ_4 model (Guindon et al., 2009).

Detection of Positive Selection in HCV Genomes

We used coding sequence information for the 67 recognized HCV subtypes (Smith et al., 2014) (Supplementary Table S2). Viral sequences were retrieved from the NCBI[1] database. As above, we used PRANK (Loytynoja and Goldman, 2005) to generate multiple sequence alignments and GUIDANCE (Sela et al., 2015) for filtering unreliably aligned codons (Privman et al., 2012). GARD (Kosakovsky Pond et al., 2006) and the above-mentioned methods implemented in RDP4 (Sawyer, 1989; Martin and Rybicki, 2000) were used to screen for recombination.

Phylogenetic trees were generated by maximum-likelihood using phyML (Guindon et al., 2010) after codon model selection in HYPHY (Delport et al., 2010).

To investigate whether episodic positive selection acted on the internal branches of HCV phylogenies, we applied the branch-site tests from PAML (Zhang et al., 2005) and BUSTED (branch-site unrestricted statistical test for episodic diversification) (Murrell et al., 2015). The p-values from both methods were FDR-corrected to account for multiple tested branches. A branch was considered positively selected when statistically significant evidence was obtained with both methods.

Positively selected sites were then identified through the BEB analysis from model MA (with a posterior probability cutoff of 0.90) (Zhang et al., 2005) and with BUSTED (with a p-value cutoff of 0.05) (Murrell et al., 2015). Sites were called as positively selected if they were detected by both methods. BEB and BUSTED were in good general agreement. On selected branches, 63.07% of BEB sites were also detected by BUSTED and 65.68% of BUSTED sites were also identified by BEB.

We used Single-Likelihood Ancestor Counting (SLAC) and fixed effects likelihood (FEL) (Kosakovsky Pond and Frost, 2005) to calculate the rates of nonsynonymous and synonymous changes at each site in the structural and in the non-structural region alignments (analyses were performed on alignments split on the basis of the recombination breakpoint detected by GARD). Both SLAC and FEL estimate the probability of selection at each site in an alignment through the dN-dS metric (rate of nonsynonymous changes-rate of synonymous changes) (Kosakovsky Pond and Frost, 2005). This metric is preferred over the conventional dN/dS ratio as this latter is rendered to infinite for dS values equal to 0.

Although SLAC and FEL use different methodologies to estimate substitution rates (Kosakovsky Pond and Frost, 2005), they yielded very similar results (for all sites in the HCV genome, Spearman's correlation coefficient $= 0.73$, p-value $< 10^{-16}$). SLAC and FEL were used to estimate the evolutionary rate at RAV vs. non-RAV positions. The dN-dS statistics was also exploited to provide an overall view of HCV genome evolution. In this case,

only SLAC results are shown (those obtained with FEL were very similar).

RAV Analysis

A total of 127 RAVs in NS3, NS4B, NS5A, and NS5B were included in the study (Supplementary Table S4). RAVs were obtained by merging six recently compiled lists of naturally occurring variants associated with DAA resistance (Rai and Deval, 2011; Bartenschlager et al., 2013; Lontok et al., 2015; Cannalire et al., 2016; Chen et al., 2016; Patino-Galindo et al., 2016). Specifically, a list of positions where RAVs have been reported was compiled, irrespective of the viral genotype carrying the RAV. For statistical analysis, six RAVs that occur at codons pruned from the alignment by GUIDANCE were removed. The overall probability of RAV occurrence was calculated over the number of non-pruned codons in NS3, NS4B, NS5A, and NS5B (total length = 1877 codons).

Protein 3D Structures and Docking

For the NS5A-OAS1 docking, the crystal structure of the NS5A domain I of a subtype 1b strain (PDB: 1ZH1) and the structure of human OAS1 (PDB: 4IG8) were obtained from the Protein Data Bank (PDB) (Tellinghuisen et al., 2005; Donovan et al., 2013). The 1b subtype was used in the original study describing interaction between OAS1 and NS5A (Taguchi et al., 2004). Two docking methods were applied and their results compared for consistency. Specifically, we used the PatchDock server (Schneidman-Duhovny et al., 2005) and imposed that the F37 residue in NS5A is located at the binding interface, as described (Taguchi et al., 2004). We next used ClusPro (Comeau et al., 2004; Kozakov et al., 2017) without any assumption on the residues involved in binding. The best models from ClusPro and Patchdock were superimposed and revealed an almost identical binding pose. We next checked that the two OAS1 regions necessary for NS5A binding (Taguchi et al., 2004) were located at the interface: one of these two regions (amino acids 235-275) defines most of the contact area in the models. Overall, these observations indicate that the binding interaction pose obtained with the two docking methods is reliable.

The C19 sphingomyelin ligand structure was prepared for simulation using the LigPrep module of Maestro (Schrodinger Release, 2016-4: Maestro Schrodinger, LLC, New York, NY, United States, 2016) to determine the 3D structure and ionization states at pH 7.0 ± 0.2. The NS5B structure of a subtype 1b strain (PDB ID: 4MK7 chain A) was processed with Protein Preparation Wizard module of Maestro to remove crystal water molecules, add hydrogen atoms, assign bond orders, and optimize the orientation of hydroxyl groups. Docking simulation was carried out with IFD protocol (Sherman et al., 2006) in Extra Precision mode. The receptor grid was generated around the Sphingomyelin Binding Domain (residues E230 to G263). All the simulations were performed using the OPLS3 force field (Harder et al., 2016). The docked ligand-receptor poses were analyzed in terms of Glide Score, a scoring function that estimate protein-ligand binding affinities (Friesner et al., 2006).

The 3D structures were rendered using PyMOL (The PyMOL Molecular Graphics System, Version 1.8.4.0 Schrödinger, LLC).

RESULTS

CD81 Is a Target of Positive Selection in Bats

Host receptors that mediate virus entry often evolve under positive selection (a situation that favors amino acid replacements over silent substitutions) to avoid viral recognition (Sironi et al., 2015).

Four cell-surface molecules, CD81, occludin (OCLN), claudin 1 (CLDN1), and scavenger receptor class B member 1 (SRB1), are particularly important for HCV cell entry, although only CD81 and OCLN determine the species-specificity of infection (Bartenschlager et al., 2013; Lindenbach and Rice, 2013). We thus reasoned that if primates experienced long-term interactions with HCV or closely related hepaciviruses, signals of positive selection should be detectable at the genes encoding HCV receptors. Conversely, if selection is evident in other mammalian orders, these may represent the ancestral HCV reservoirs.

We retrieved coding sequences of the four genes for mammalian species belonging to different orders, superorders or clades (Supplementary Table S1). Due to the hypothesized role of bats as reservoir hosts for mammalian hepaciviruses, Chiroptera were analyzed separately from other species in the Laurasiatheria superorder. Evidence of positive selection was searched for using models that allow dN/dS to vary among sites in the alignment (Yang, 2007).

No evidence of positive selection was detected for *CLDN1* and *SRB1* (**Table 1**). Also, selection was not detected at *OCLN* or *CD81* in primates (**Table 1**), an observation that does not support a long-standing selective pressure exerted by hepaciviruses on these hosts. However, positive selection was not detected for rodents, either, although these mammals are known to host a wide diversity of hepaciviruses (Drexler et al., 2013; Kapoor et al., 2013; Scheel et al., 2015).

For Laurasiatherian *OCLN*, we detected one positively selected site located in the extracellular loop 1, which is not directly involved in HCV binding. Notably, evidence of positive selection for *CD81* was detected only in Chiroptera (**Figure 1** and **Table 1**). The five positively selected sites are located at the binding interface with the HCV glycoprotein. Single point mutations in the corresponding region of the human receptor were shown to reduce or abolish the interaction with the HCV E2 protein and/or HCV infectivity (**Figure 1** and **Table 1**) (Drummer et al., 2002; Bertaux and Dragic, 2006; Flint et al., 2006). Conversely, the CD81 region required for infection of hepatocytes by *Plasmodium yoelii* (a rodent parasite) sporozoites is unaffected by selection (**Figure 1**) (Yalaoui et al., 2008).

The observation that the positively selected sites are involved in HCV binding and infectivity suggests that hepaciviruses contributed to shape the genetic diversity of *CD81* in bats.

HCV Originated at Least 3000 Years Ago

Previous studies provided estimates of the time to the most recent common ancestor (tMRCA) of HCV genotypes in a range between ∼200 and 1000 years ago, with one single study indicating that HCV origin may date 2000 years back

(Smith et al., 1997; Pybus et al., 2001; Kapoor et al., 2011; Simmonds, 2013; Li et al., 2014; Lu et al., 2014). The tMRCA of equine/canine hepaciviruses (EHV/CHV) was estimated to be recent, dating around 1800 CE (Pybus and Theze, 2016).

It is well known that the temporal variation in rates of nucleotide substitutions often results in underestimation of the age of viral lineages (Duchene et al., 2014; Aiewsakun and Katzourakis, 2016). Purifying selection and substitution saturation are strongly associated with temporal rate variation (Duchene et al., 2014).

Simulations experiments indicated that "classic" models (e.g., GTR+Γ_4) tend to underestimate branch lengths in the presence of purifying selection (Wertheim and Kosakovsky Pond, 2011) and substitution saturation (Wertheim et al., 2013). Because both phenomena are more pronounced for internal branches, length underestimation is more severe for these branches and dating inferences are consequently affected (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013). The use of models that allow site- and branch-specific variation in selective pressure can improve branch length estimates in the presence of both purifying selection and substitution saturation (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013).

We thus applied the aBS-REL model (adaptive branch-site random effects likelihood) to calculate the tMRCA of
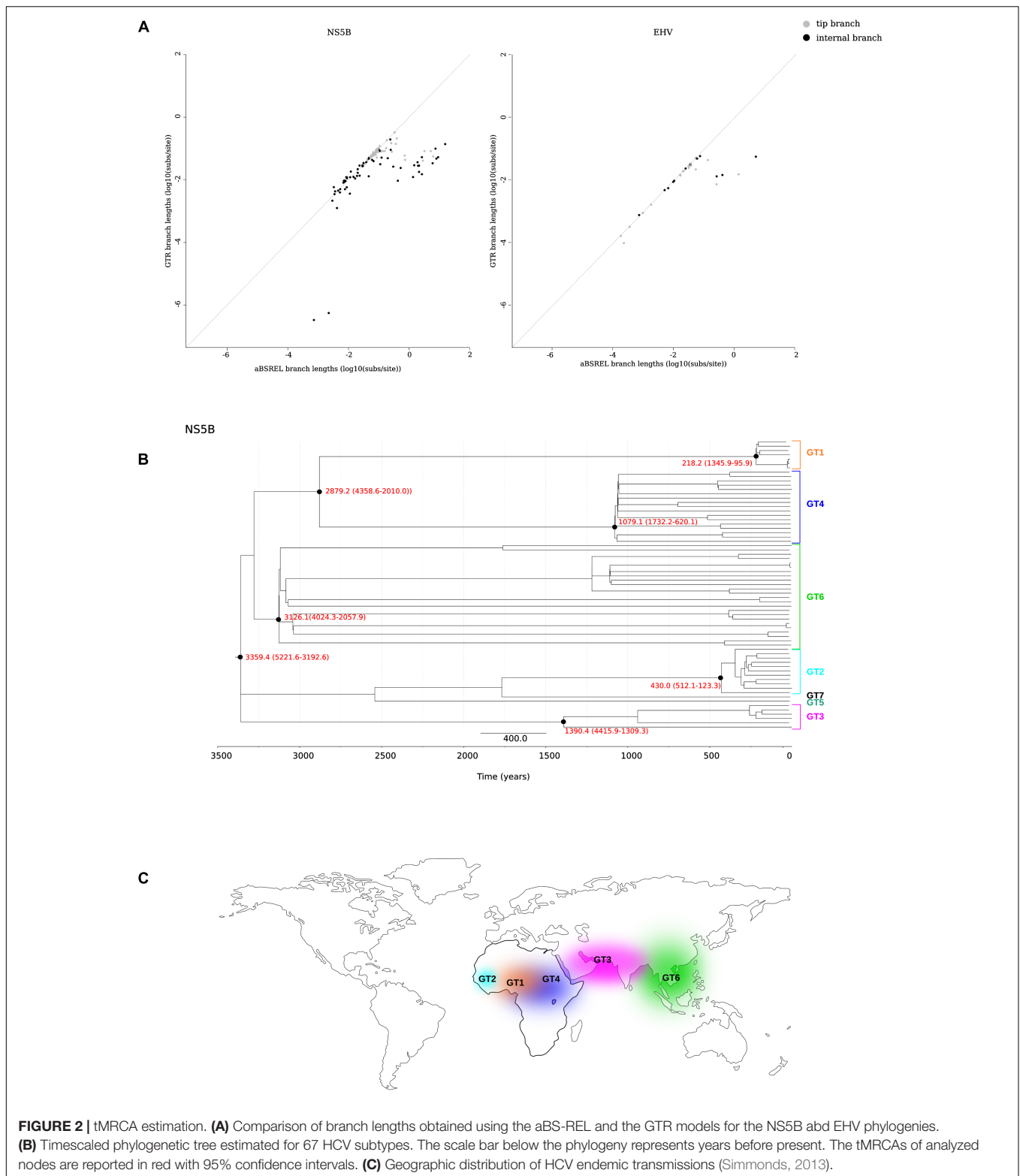
relevant nodes in a phylogeny of 67 HCV strains (Smith et al., 2014) (Supplementary Table S2). This approach was previously applied to revise the dating of other RNA viruses (Wertheim and Kosakovsky Pond, 2011; Wertheim et al., 2013). We constructed a phylogeny for the NS5B region and we found that only 0.7% of branches showed saturation of dS. As expected, branch lengths estimated with aBS-REL were generally longer than those obtained with a GTR+Γ_4 model, and this was particularly true for internal branches (**Figure 2A** and Supplementary Figure S1). Using the aBS-REL model, we estimated the tMRCA of the seven HCV genotypes to be 3359 years ago (95% CI: 5221–3192) (**Figure 2B**). The deepest tMRCA was obtained for genotype 6 (endemic in South-East Asia), with an origin dating at least 2000 years ago (**Figures 2B,C**). The tMRCAs of genotype 3 (endemic in South Asia) and of the ancestor of genotypes 1 and 4 (both widespread in Central Africa) were estimated to be in the 800 BCE–700 CE range (**Figures 2B,C**). Results for genotype 2 (endemic in West Africa) indicated an origin around 430 years ago (**Figures 2B,C**).

Horse-to-human hepaciviral transmission was hypothesized as the origin of HCV (Pfaender et al., 2014; Scheel et al., 2015). EHV isolates show limited divergence and high levels of purifying selection (Simmonds, 2013; Pfaender et al., 2014).

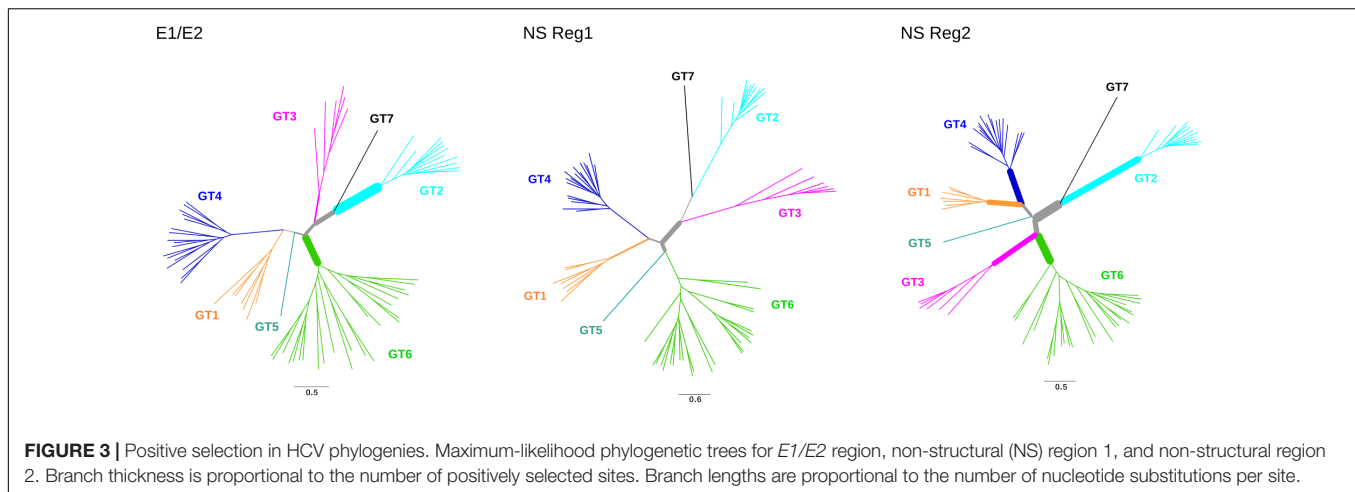**TABLE 1 |** Likelihood ratio test statistics for models of variable selective pressure among sites.

Gene	Genus or clade	*n*. species	Tree length	M8a vs. M8		M7 vs. M8		% of sites[c] (average dN/dS)	Positively selected sites
				−2ΔlnL[a]	*p*-value[b]	−2ΔlnL[a]	*p*-value[b]		
CD81									
NM_004356	Primates/Scandentia	27	1.719	2.259	0.133	0.808	0.668	6.313 (1.033)	
	Glires	20	2.648	1.242	0.265	1.238	0.539	6.374 (1.337)	
	Chiroptera	10	1.490	14.340	1.526×10^{-4}	15.170	5.079×10^{-4}	6.282 (3.515)	T163, L174, C175, G178, L185
	Laurasiatheria*	32	2.982	0.147	0.701	5.302	0.071	10.175 (1.107)	
	Afrotheria/Xenarthra	8	1.741	0.100	0.752	0.389	0.823	4.537 (1.104)	
CLDN1									
NM_021101	Primates/Scandentia	26	1.15	0.001	0.977	1.050	0.591	6.169 (1)	
	Glires	18	2.732	5.117	1	0.004	1	0.001 (3.418)	
	Chiroptera	8	0.512	0.689	0.406	0.152	0.927	6.490 (3.284)	
	Laurasiatheria*	34	1.968	0.225	0.635	5.813	0.055	3.362 (1.239)	
	Afrotheria/Xenarthra	7	1.312	2.460	0.117	4.733	0.094	1.697 (3.284)	
OCLN									
NM_002538	Primates/Scandentia	27	1.596	0.130	0.719	0.0616	0.970	4.844 (1)	
	Glires	21	4.069	1.147	0.284	4.396	0.111	1.479 (1.067)	
	Chiroptera	13	1.039	1.738	0.187	2.859	0.239	1.640 (2.205)	
	Laurasiatheria*	33	3.630	5.998	0.014	9.635	0.008	0.223 (4.253)	L98
	Afrotheria/Xenarthra	9	1.741	0	1	0.006	0.997	0.037 (2.536)	
SCARB1									
NM_005505	Primates/Scandentia	27	2.118	0.286	0.593	3.155	0.207	4.652 (1.317)	
	Glires	18	4.270	0	1	1.231	0.540	6.413 (1)	
	Chiroptera	11	2.168	3.809	0.051	11.584	0.003	17.344 (1.418)	
	Laurasiatheria*	28	3.333	2.723	0.099	5.753	0.056	1.158 (2.279)	
	Afrotheria/Xenarthra	7	1.833	0.091	1	0.009	0.995	0.093 (2.220)	

[a]Twice the difference of likelihood for the two models compared; [b]p-value of rejecting the neutral models (M8a and M7) in favor of the positive selection model (M8); [c]% of sites (average dN/dS) is the estimated percentage of sites evolving under positive selection by M8 (dN/dS for these codons); *Laurasiatheria excluding Chiroptera.

FIGURE 1 | Positively selected sites in CD81. The membrane topology of CD81 is shown. Positions involved in HCV binding and/or infectivity are highlighted in yellow, both on the structure (circle) and on the protein alignment. Positively selected sites Chiroptera are indicated in red. LEL: large extracellular loop; SEL: short extracellular loop. Positions refer to the human sequence (Accession ID: NM_004356).

We thus used the same approach described above to obtain the tMRCA of extant EHV/CHV strains (**Figure 2A**) (Supplementary Table S3), producing an estimate of 863 years ago (95% CI: 1726–377).

Positive Selection Shaped the Diversity of HCV Genotypes

A number of studies have explored the recent selective events that generated intra-genotype and within-host genetic diversity (Jackowiak et al., 2014), whereas the deeper evolutionary history of HCV has remained largely unexplored. We thus used complete sequence information for the 67 recognized HCV subtypes (Smith et al., 2014) to search for selective events that occurred during the radiation of the seven genotypes.

The structural and non-structural coding regions were analyzed separately, and the core region was excluded due to the presence of a putative alternative reading frame (Xu et al., 2001).

We next tested for the presence of recombination using GARD and RDP4 (see the section "Materials and Methods"). None of the RDP4 methods detected recombination either in the structural or in the non-structural regions. Conversely, GARD detected a possible breakpoint at the end of NS3. The lack of consistency among recombination detection methods is somehow expected as these approaches have distinct performances depending on the amount of recombination, the genetic diversity of analyzed sequences, the tree topology, and the rate variation among sites (Posada and Crandall, 2001; Bay and Bielawski, 2011). Because our aim was to avoid the inflation of positive

FIGURE 2 | tMRCA estimation. **(A)** Comparison of branch lengths obtained using the aBS-REL and the GTR models for the NS5B abd EHV phylogenies. **(B)** Timescaled phylogenetic tree estimated for 67 HCV subtypes. The scale bar below the phylogeny represents years before present. The tMRCAs of analyzed nodes are reported in red with 95% confidence intervals. **(C)** Geographic distribution of HCV endemic transmissions (Simmonds, 2013).

selection inference caused by unrecognized recombination, the alignment was split into two sub-alignments (based on the GARD-detected breakpoint) with slightly different topologies (**Figure 3**).

Evidence of episodic positive selection along the internal branches of the phylogenies was searched for using two different branch-site methods, which rely on different assumptions of dN/dS variation among branches (Zhang

FIGURE 3 | Positive selection in HCV phylogenies. Maximum-likelihood phylogenetic trees for *E1/E2* region, non-structural (NS) region 1, and non-structural region 2. Branch thickness is proportional to the number of positively selected sites. Branch lengths are proportional to the number of nucleotide substitutions per site.

et al., 2005; Murrell et al., 2015). The two methods provided evidence of positive selection on multiple branches in the structural and non-structural region alignments (**Figure 3**, Supplementary Tables S5, S6). A total of 102 sites were found to be targeted by positive selection, nine of them selected on more than one branch (Supplementary Table S6).

Positively Selected Sites Impinge on Central Processes of HCV Life Cycle

Analysis of positively selected sites was performed by inspection of literature reports describing the effect of specific mutations, as well as by performing docking analyses. Overall, positively selected sites could be categorized on the basis of the functional effect they are likely to entail, as follows: cell entry, interaction with the host immune system, and association with membrane/lipids. The details of these sites are reported below, as well as in **Figures 4**, **5**.

Cell Entry

As expected, none of the positively selected sites in E2 involved the conserved residues at the CD81 binding site (Lavie et al., 2014). Interestingly though, a mutation at residue 216 in E1 was previously obtained through *in vitro* adaptation of HCV to mouse cells (**Figure 4A**) (Bitzegeio et al., 2010). Specifically, the L216F mutation (as well as two other mutations in the hypervariable region 1, HVR1) allow infection of cells expressing CD81 of rodent origin (Bitzegeio et al., 2010), suggesting that HCV adaptation to new hosts does not necessarily entail changes at the direct binding interface with CD81.

Several positively selected sites were detected in the transmembrane (TM) regions of both E1 and E2 (**Figure 4A**). A positively selected site in the E1 TM domain (residue 374) was shown to reduce the dependency on SCARB1 for cell-to-cell spread *in vitro*, but increases viral susceptibility to antibody-mediated neutralization (Catanese et al., 2013). Interestingly, a similar phenotype was described for a T563V mutant in the JFH1 viral background (corresponding to the positively selected 561 site in strain H77) (**Figure 4A**) (Zuiani

et al., 2016). Using a site-directed saturation mutagenesis approach, the T563V mutation was shown to confer growth advantage *in vitro* via reduced dependency on SCARB1 and probably stronger binding to CD81 (Zuiani et al., 2016).

Interaction With the Host Immune System

*S*everal mechanisms for HCV resistance to IFN have been proposed. The NS5A protein displays a double-stranded RNA-activated protein kinase R (PKR) binding site that overlaps with the IFN sensitivity determinant (ISDR) region but includes 26 additional amino acids essential for PKR binding (Gale et al., 1998). Two positively selected sites were detected within this 26 amino acid region (**Figure 5A**). Mutations at the second selected site (position I302 in H77, corresponding to C298 in the JFH1 sequence) arise when HCV is passaged in the presence of IFN-alpha (Perales et al., 2013). The C298I change is positively selected on the genotypes 1/4 branch and both genotypes are relatively resistant to IFN therapy (Jackowiak et al., 2014).

The binding of NS5A to OAS1 also contributes to inhibition of IFN antiviral activities (Taguchi et al., 2004). Based on previous biochemical data (Taguchi et al., 2004), we performed protein–protein docking of NS5A domain I and human OAS1. Results showed that three of the positively selected sites in NS5A domain I, positions 54, 78, and 93 are at the direct binding interface with OAS1 (**Figure 5B**). Notably, besides being described as RAVs for DAAs, substitutions at the Y93 site in the subtype 1b background occur at different frequency depending on the host's genotype at the *IFNL3/IFNL4* locus (Akamatsu et al., 2015; Peiffer et al., 2016). A similar trend was observed for variants at NS5A positions 31 (not included in the crystal), 37 (which modulates OAS1 binding) (Taguchi et al., 2004), 52 (close to the binding interface), and 54 (a positively selected RAV close to the binding interface) (**Figure 5B**) (Akamatsu et al., 2015). Interestingly, sites 93 and 54 are positively selected on the genotype 6 and 3 branches, respectively (Supplementary Table S5). Both these
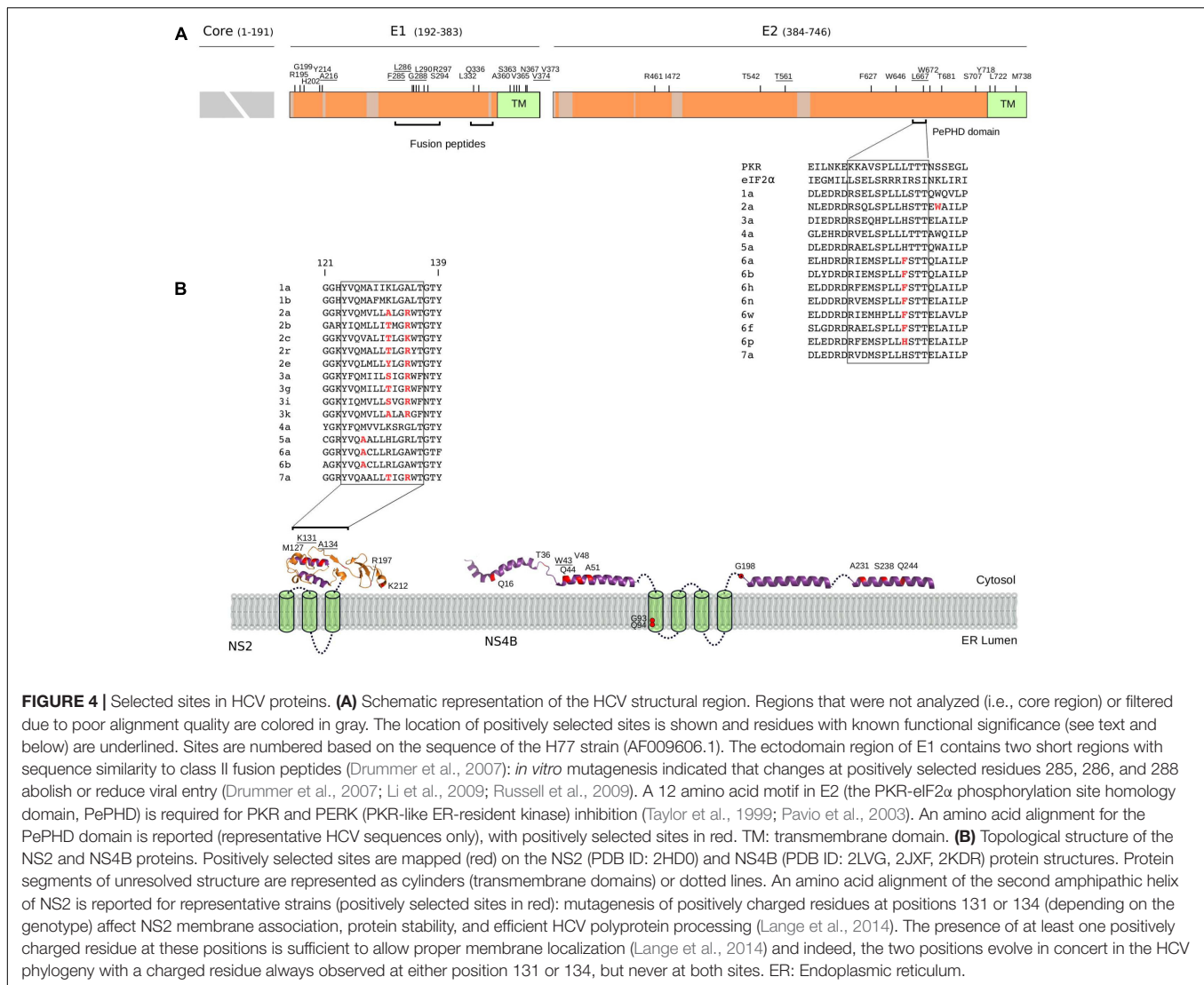
FIGURE 4 | Selected sites in HCV proteins. **(A)** Schematic representation of the HCV structural region. Regions that were not analyzed (i.e., core region) or filtered due to poor alignment quality are colored in gray. The location of positively selected sites is shown and residues with known functional significance (see text and below) are underlined. Sites are numbered based on the sequence of the H77 strain (AF009606.1). The ectodomain region of E1 contains two short regions with sequence similarity to class II fusion peptides (Drummer et al., 2007): *in vitro* mutagenesis indicated that changes at positively selected residues 285, 286, and 288 abolish or reduce viral entry (Drummer et al., 2007; Li et al., 2009; Russell et al., 2009). A 12 amino acid motif in E2 (the PKR-eIF2α phosphorylation site homology domain, PePHD) is required for PKR and PERK (PKR-like ER-resident kinase) inhibition (Taylor et al., 1999; Pavio et al., 2003). An amino acid alignment for the PePHD domain is reported (representative HCV sequences only), with positively selected sites in red. TM: transmembrane domain. **(B)** Topological structure of the NS2 and NS4B proteins. Positively selected sites are mapped (red) on the NS2 (PDB ID: 2HD0) and NS4B (PDB ID: 2LVG, 2JXF, 2KDR) protein structures. Protein segments of unresolved structure are represented as cylinders (transmembrane domains) or dotted lines. An amino acid alignment of the second amphipathic helix of NS2 is reported for representative strains (positively selected sites in red): mutagenesis of positively charged residues at positions 131 or 134 (depending on the genotype) affect NS2 membrane association, protein stability, and efficient HCV polyprotein processing (Lange et al., 2014). The presence of at least one positively charged residue at these positions is sufficient to allow proper membrane localization (Lange et al., 2014) and indeed, the two positions evolve in concert in the HCV phylogeny with a charged residue always observed at either position 131 or 134, but never at both sites. ER: Endoplasmic reticulum.

genotypes are endemic in Asia, where the frequency of the protective human *IFNL3/IFNL4* variant (rs12979860) is highest (**Figure 5C**).
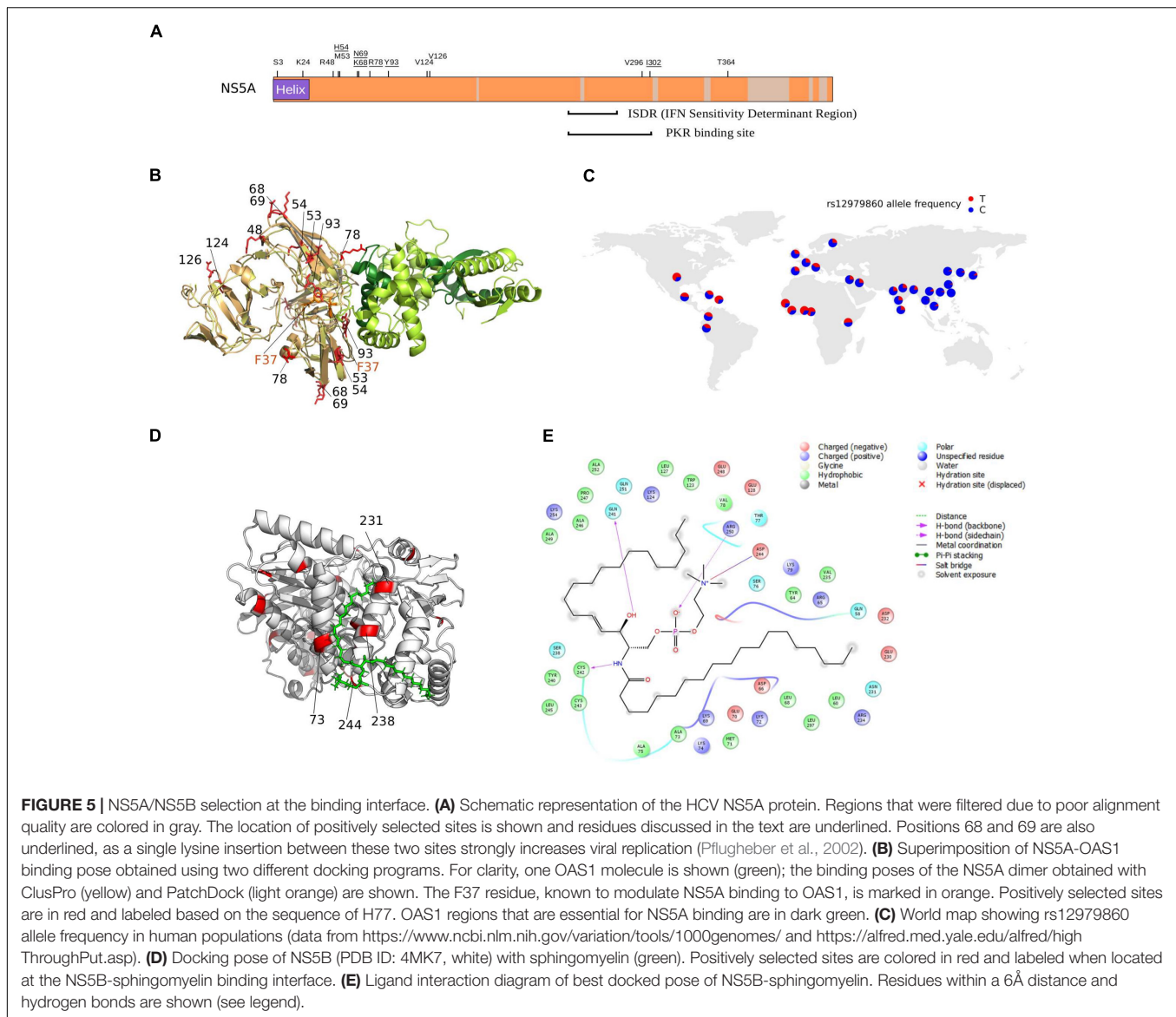
Recently, a large-scale analysis of patients mostly infected with HCV genotype 3a, identified 30 sites in the viral polyprotein showing association with one or more HLA alleles (Ansari et al., 2017). We found three of these sites (561 in E2, 620 in NS3, and 48 in NS4B) to be positively selected, with site 620 in NS3 (position 1646 in the polyprotein) representing one of the strongest association detected in the study (Ansari et al., 2017).

Association With Membranes and Lipids
All HCV proteins are tethered to intracellular membranes either via transmembrane domain or through amphipathic alpha helices or both (Bartenschlager et al., 2013). We identified several positively selected sites within amphipathic alpha helices (**Figure 4B**). In the case of NS4B, virtually all sites are located in these regions,

where several RAVs also map (Supplementary Table S4). Among these, W43 (selected on the genotype 4 branch) is essential for NS4B association to the membrane and to lipid droplets (Gouttenoire et al., 2009; Tanaka et al., 2013).

We found positively selected sites within the sphingomyelin binding domain (SBD) of the HCV RNA polymerase (NS5B). Binding of sphingomyelin to NS5B allows localization of the polymerase to lipid rafts and activates the enzymatic activity in a genotype-dependent manner (Sakamoto et al., 2005; Weng et al., 2010). Remarkably, mutagenesis experiments indicated that the positively selected sites 244 and 238 modulate sphingomyelin binding and activation, respectively (Weng et al., 2010). We thus docked a sphingomyelin molecule onto the 3D structure of subtype 1b NS5B (**Figure 5D**). Docking results confirmed a salt bridge interaction between residue 244 and the sphingomyelin (along with residues 241, 242, and 250) and the localization of residue 238 at the interaction surface. Moreover, analysis of atomic distances indicated that two additional selected sites

FIGURE 5 | NS5A/NS5B selection at the binding interface. **(A)** Schematic representation of the HCV NS5A protein. Regions that were filtered due to poor alignment quality are colored in gray. The location of positively selected sites is shown and residues discussed in the text are underlined. Positions 68 and 69 are also underlined, as a single lysine insertion between these two sites strongly increases viral replication (Pflugheber et al., 2002). **(B)** Superimposition of NS5A-OAS1 binding pose obtained using two different docking programs. For clarity, one OAS1 molecule is shown (green); the binding poses of the NS5A dimer obtained with ClusPro (yellow) and PatchDock (light orange) are shown. The F37 residue, known to modulate NS5A binding to OAS1, is marked in orange. Positively selected sites are in red and labeled based on the sequence of H77. OAS1 regions that are essential for NS5A binding are in dark green. **(C)** World map showing rs12979860 allele frequency in human populations (data from https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/ and https://alfred.med.yale.edu/alfred/high ThroughPut.asp). **(D)** Docking pose of NS5B (PDB ID: 4MK7, white) with sphingomyelin (green). Positively selected sites are colored in red and labeled when located at the NS5B-sphingomyelin binding interface. **(E)** Ligand interaction diagram of best docked pose of NS5B-sphingomyelin. Residues within a 6Å distance and hydrogen bonds are shown (see legend).

(231 and 73) are likely involved in the binding of sphingomyelin (**Figure 5E**).

RAVs Show Significant Overlap With Positively Selected Sites

A recent analysis indicated that several RAVs occurred *de novo* on external branches of HCV phylogenies, although a minority appeared on internal branches, indicating a common origin in multiple strains (Patino-Galindo et al., 2016). However, the evolutionary history of RAVs has remained a poorly investigated issue.

Up to now, direct-acting antivirals (DAAs) have been developed to inhibit the function of the NS3, NS4B, NS5A, and NS5B proteins. To obtain a codon-wise measure of selective pressure acting on these proteins across the HCV phylogeny, we calculated dN-dS at each site. Comparison of RAV and non-RAV sites revealed no statistically significant difference in

dN-dS calculated with either SLAC (Wilcoxon's rank-sum test p-value = 0.19) or FEL (Wilcoxon's rank-sum test p-value = 0.59) (**Figure 6A**). This suggests that RAVs do not originate and are not maintained in viral populations as a result of relaxed selective constraints. This overall trend does not imply that all RAV sites evolve at the same rate. Thus, we tested whether positions where RAVs occur are more likely to diversify through the action of natural selection. Nine RAV positions coincided with positively selected sites, a number higher than expected by chance (Binomial test, p-value = 0.012) (Supplementary Table S4). Moreover, three of these RAV positions were targeted by positive selection on two distinct branches of the phylogeny, a situation observed for only 8.8% of selected sites (**Figure 6A**).

To gain a comprehensive view of the action of selection, we plotted dN-dS along the HCV genome and we superimposed the location of positively selected sites and of RAVS (**Figure 6B**).
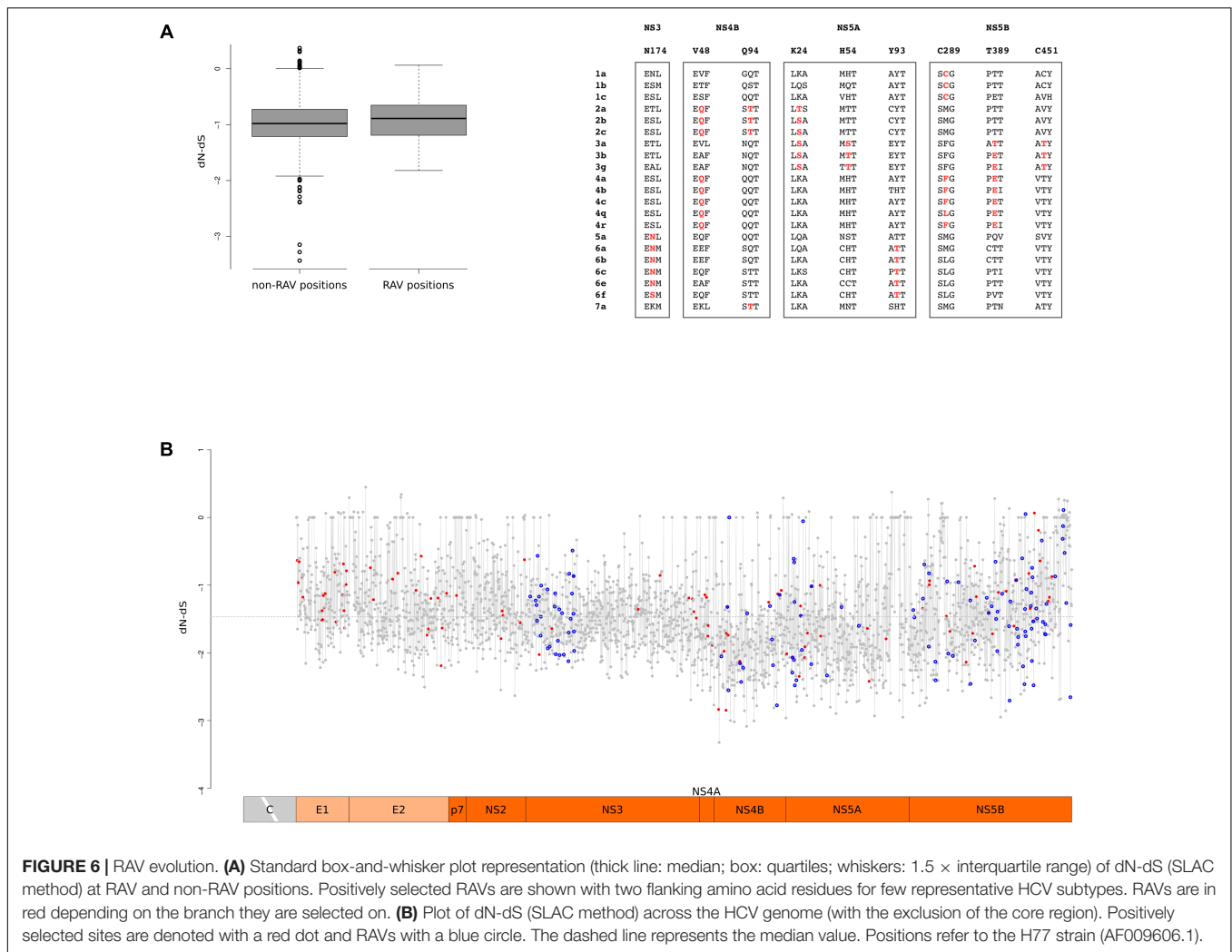
| | NS3 | NS4B | | NS5A | | | NS5B | | |
	N174	V48	Q94	K24	H54	Y93	C289	T389	C451
1a	ENL	EVF	GQT	LKA	MHT	AYT	SCG	PTT	ACY
1b	ESM	ETF	QST	LQS	MQT	AYT	SCG	PTT	ACY
1c	ESL	ESF	QQT	LKA	VHT	AYT	SCG	PET	AVH
2a	ETL	EQF	STT	LTS	MTT	CYT	SMG	PTT	AVY
2b	ESL	EQF	STT	LSA	MTT	CYT	SMG	PTT	AVY
2c	ESL	EQF	STT	LSA	MTT	CYT	SMG	PTT	AVY
3a	ETL	EVL	NQT	LSA	MST	EYT	SFG	ATT	ATY
3b	ETL	EAF	NQT	LSA	MTT	EYT	SFG	PET	ATY
3g	EAL	EAF	NQT	LSA	TTT	EYT	SFG	PEI	ATY
4a	ESL	EQF	QQT	LKA	MHT	AYT	SFG	PET	VTY
4b	ESL	EQF	QQT	LKA	MHT	THT	SFG	PEI	VTY
4c	ESL	EQF	QQT	LKA	MHT	AYT	SFG	PET	VTY
4q	ESL	EQF	QQT	LKA	MHT	AYT	SLG	PET	VTY
4r	ESL	EQF	QQT	LKA	MHT	AYT	SFG	PEI	VTY
5a	ENL	EQF	QQT	LQA	NST	ATT	SMG	PQV	SVY
6a	ENM	EEF	SQT	LQA	CHT	ATT	SMG	CTT	VTY
6b	ENM	EEF	SQT	LKA	CHT	ATT	SLG	CTT	VTY
6c	ENM	EQF	STT	LKS	CHT	PTT	SLG	PTI	VTY
6e	ENM	EAF	STT	LKA	CCT	ATT	SLG	PTT	VTY
6f	ESM	EQF	STT	LKA	CHT	ATT	SLG	PVT	VTY
7a	EKM	EKL	STT	LKA	MNT	SHT	SMG	PTN	ATY

FIGURE 6 | RAV evolution. **(A)** Standard box-and-whisker plot representation (thick line: median; box: quartiles; whiskers: 1.5 × interquartile range) of dN-dS (SLAC method) at RAV and non-RAV positions. Positively selected RAVs are shown with two flanking amino acid residues for few representative HCV subtypes. RAVs are in red depending on the branch they are selected on. **(B)** Plot of dN-dS (SLAC method) across the HCV genome (with the exclusion of the core region). Positively selected sites are denoted with a red dot and RAVs with a blue circle. The dashed line represents the median value. Positions refer to the H77 strain (AF009606.1).

In agreement with a previous work that analyzed conservation and selection in HCV-1a and HCV-1b genomes (Patino-Galindo and Gonzalez-Candelas, 2017), dN-dS tended to be higher in the E1 and E2 regions, compared to the non-structural portions. A non-negligible fraction of codons showed almost complete conservation: about 7% displayed dS = 0 and most of these (89%) also had dN = 0. Both RAV positions and positively selected sites displayed variable dN-dS values. We note that this is not unexpected for positively selected sites as they were identified using branch-site tests and are thus not selected across the entire HCV phylogeny.

DISCUSSION

The worldwide spread of HCV epidemic strains started recently, in the 1930s–1940s, but the existence of genetically diverse endemic HCV strains that circulate in specific geographic locations implies a long-standing association of the virus with human populations (Simmonds, 2013). Using an evolutionary model that accounts for variation in the pressure of natural selection across sites and branches, we show that the common ancestor of extant HCV genotypes existed at least 3000 years ago, with a lower bound estimate of ∼5200 years before present. The same approach placed the origin of EHV/CHV around 1100 CE, with a lower bound at 290 CE. Although we applied a selection-informed approach that accounts for the effect of purifying selection and is relatively robust in the presence of substitution saturation (Wertheim et al., 2013), we most likely failed to fully correct for time-dependent substitution rate variation (Ho et al., 2015). Thus, the time of HCV and EHV/CHV origin might be underestimated. However, these time estimates rule out the possibility that the use of horse serum to obtain therapeutic anti-toxins originated HCV and, more generally, seem to exclude the hypothesis that the human virus derived from a cross-species transmission event from horses (Pfaender et al., 2014; Scheel et al., 2015). Nonetheless, the sampling of EHV is still sparse. The isolation of additional EHV sequences, possibly from a wider geographic range, may date the origin of this virus further back.

Horse domestication begun around 5000 years ago in Central Asia (Outram et al., 2009): where EHV found to be older, close contacts between horses and humans may have resulted in

zoonotic cross-species transmission or in the acquisition of HCV and EHV from a common source (possibly bat or rodent). Active sampling of horse and donkeys worldwide will be required to address this important point.

The time frame of HCV origin and the geographic distribution of the endemic strains are not consistent with the possibility that HCV dispersed with humans following the major out-of-Africa colonization routes across the Old World (65000–45000 years ago) (Nielsen et al., 2017). The dating we estimated for HCV origin instead suggests that the (possible) zoonotic transmission and subsequent spread in human populations occurred in a time-frame when long-distance trade routes were being established. The deepest tMRCA for HCV genotypes was obtained for genotype 6, possibly indicating an Asian origin of extant strains. Several historical situations may help explain the diffusion of HCV in Asia and Africa.

Maritime trading routes across the South Chinese Sea were already developed in the first millennium BCE (Hung et al., 2013), and regular sea connections between South-East Asia and India were established by 1000–500 BCE (Pearson, 2016). More extensive human movements connecting Asia with Africa occurred starting around ∼300 BCE as a result of Alexander the Great expansion (336-325 BCE) and the full development of the Silk Road (since 200 BCE), this latter comprising maritime and land routes that reached Africa (Lawler, 2014). Around this time, Indian slaves, mostly women, were regularly imported to Egypt (Mark, 2002). Moreover, archaeological evidences suggest that trade circuits connected the East Coast of Africa with South Asia in the first millennium CE[3].

Whereas these represent possible transmission routes, it remains unclear how, in a time when parenteral exposure was limited, HCV spread and was maintained in human populations. Vertical transmission of HCV is rare and sexual contact is also thought to be a scarcely efficient route for HCV dissemination. It was thus proposed that cultural practices (e.g., circumcision, tattooing, or acupuncture) (Shepard et al., 2005; Simmonds, 2013) or natural routes such as arthropod biting (Pybus et al., 2007) might account for endemic HCV transmission. This issue is not addressed by the analyses herein. However, our time estimates place the origin of EHV in a time frame when invasive veterinary procedures and parenteral exposure were most likely rare. Although we cannot exclude that some forms of human intervention (e.g., the use of spurs and animal housing in crowded stables) facilitated viral transmission, natural routes are likely to account for the spread of EHV. Whether the same mechanisms are responsible for the endemic transmission of HCV and of EHV remains an open question. In this respect, we note that sexually transmitted infections (STIs) that disrupt mucosal integrity have been implicated in increased sexual transmission of HCV, at least in high-risk groups (Chan et al., 2016). STIs have probably been common throughout human history and several such diseases are known in horses (Dobson and Carper, 1996; Samper and Tibary, 2006). The association between HCV sexual transmission and specific STIs might help explain the geographic distribution

of endemic HCV strains. For instance, lymphogranuloma venereum and granuloma inguinale are confined to tropical and subtropical regions (Richens, 2006; White, 2009) and their presence may have facilitated the maintenance and diversification of endemic HCV strains in these areas. Although the present-day distribution of these STIs may not necessarily reflect the historical prevalence of the causative bacteria, the hypothesis of STI-facilitated HCV transmission warrants further consideration.

Although we found that the tMRCA for HCV is greater than the previously inferred time ranges (Smith et al., 1997; Pybus et al., 2001; Kapoor et al., 2011; Simmonds, 2013; Li et al., 2014; Lu et al., 2014), HCV can be thought of as a recently emerged pathogen, largely postdating the origin of our species. This observation, as well as the presently limited evidence of hepacivirus infection in non-human primates, does not support a long-standing association between these viruses and primates (Simmonds, 2013; Scheel et al., 2015). The absence of selection signatures at primate CD81 and OCLN, the two molecules that determine the species-specificity of HCV infection, further supports this view. However, as noted above, we also failed to detect positive selection for HCV receptors in glires, even though the wide variety of rodent hepaciviruses suggests that these mammals have been infected by hepaciviruses for a long time (Drexler et al., 2013; Kapoor et al., 2013; Scheel et al., 2015). Thus, results obtained for the HCV receptors are clearly not conclusive and should be regarded as preliminary. Indeed, the power of the positive selection tests depends not only on the strength of selection, but also on the number and divergence of the analyzed sequences (Anisimova et al., 2001, 2002). Considering these two parameters, we have most likely a power lower than 80% although accuracy is higher than 90% (Anisimova et al., 2001, 2002). Given this premise, it is nonetheless interesting that the only mammalian order showing evidence of selection at CD81 was Chiroptera, which host a wide diversity of hepaciviruses (Kapoor et al., 2011, 2013; Drexler et al., 2013; Quan et al., 2013; Corman et al., 2015; Scheel et al., 2015; Walter et al., 2016).

Known bat hepaciviruses share very little similarity to the HCV E2 protein, and this also applies to the E2 regions involved in CD81 binding. This raises the possibility that these animal viruses engage cellular receptors distinct from CD81, although previous studies on coronaviruses showed that the same region of the cellular receptor can be bound by viruses that share little sequence or structural similarity in their glycoprotein receptor binding domains (Forni et al., 2017). Also, recent studies have indicated that HCV adaptation to the usage of rodent CD81 molecules is not necessarily paralleled by changes at the direct binding interface (Bitzegeio et al., 2010). Thus, the receptor preferences of non-human hepaciviruses will require experimental investigation. Nonetheless, the observation that the CD81 positively selected sites in Chiroptera are located in a small region that corresponds to the E2 binding site is consistent with the view that CD81-binding hepaciviruses have been infecting bats for a long time. However, we note that CD81 was also shown to interact with other pathogens. For instance, CD81 is required for human *Plasmodium falciparum*

---

[3]http://www.sealinksproject.com/

and rodent *Plasmodium yoelii* sporozoite hepatocyte infection (Silvie et al., 2003). Even though the protein region necessary and sufficient for *P. yoelii* infection is distinct from that involved in HCV binding (Yalaoui et al., 2008), we cannot exclude that bat-infecting *Plasmodium* parasites (Schaer et al., 2013) bind different regions of CD81. Also, CD81 was shown to be important for influenza virus infection (He et al., 2013), but the molecular details of the interaction with this virus are unknown. Thus, we cannot rule out the possibility that pathogens other than hepaciviruses were responsible for the selection signatures we detected at CD81 in bats.

Zoonotic events are commonly associated with bursts of positive selection whereby the pathogen adapts to successfully infect and be transmitted in a new species (Longdon et al., 2014). A common trade-off of the adaptation to new hosts is the loss or reduction of infectivity in the original host (Longdon et al., 2014). Indeed, HCV infection is now restricted to humans (and to experimentally infected chimpanzees). Because the ancestor of HCV has not been identified, knowledge of the molecular events that allowed HCV adaptation to humans remains elusive. We identified several positively selected sites in the E1 and E2 regions and some of these were shown to modulate the viral requirement of specific cellular factors (i.e., CD81 and SCARB1). Whereas these changes are not expected to represent specific adaptations to the human host, they may confer a selective advantage by increasing viral titers or cell-to-cell spread, a process possibly implicated in the establishment of a persistent infection (Ding et al., 2014).

An interesting possibility is that, during its spread in Asia and Africa, HCV has adapted in response to the genetic background of distinct human populations, as expected given the distinctive geographic localization of viral genotypes for most of their evolutionary history. A paradigmatic example of this is the distribution of the *IFNL3/IFNL4* polymorphism most strongly associated with spontaneous clearance of HCV: the frequency of the protective genotype differs dramatically among populations and this variant explains part of the ethnic variance in the probability to clear HCV infection (O'Brien et al., 2014).

We identified two positively selected sites (H54 and Y93 in NS5A) on the branches of the Asian endemic genotypes (3 and 6, respectively). Analysis of patients infected with HCV subtype 1b showed that substitutions at the Y93 sites and, to a lesser extent at the H54 position, vary depending on the host *IFNL3/IFNL4* genotype (Akamatsu et al., 2015; Peiffer et al., 2016). This observation implies viral adaptation to the host depending on *IFNL3/IFNL4* allelic status. Although these findings are presently limited to genotype 1b viruses (Akamatsu et al., 2015; Pedergnana et al., 2016; Peiffer et al., 2016), they provide a proof of principle for the hypothesis that human genetic diversity exerted a selective pressure on HCV and possibly contributed to the radiation of the seven genotypes.

Clearly, it remains to be evaluated whether changes at positions Y93 and H54 exert their effects via modulation of OAS1 binding, as the docking analysis suggests, at least for position 93. Alternatively, other mechanisms may be at play: a lysine insertion between NS5A positively selected sites K68 and N69, which are

not at the binding interface with OAS1, regulates PKR and IRF-3 activity (Pflugheber et al., 2002; Sumpter et al., 2004).

In general, the positively selected sites we identified herein represent excellent candidates for future functional studies as they are expected to modulate viral phenotypes. For instance, selected sites in NS5B account for genotype differences in terms of sphingomyelin-driven polymerase activation (Sakamoto et al., 2005; Weng et al., 2010), and the C298I change that arose in the common ancestor of genotypes 1 and 4 may help explain the poor response of these genotypes to IFN therapy.

Exposure to treatment cannot be regarded as the selective force underlying the evolution of the HCV sequences analyzed herein, as all strains derive from treatment-naive subjects. It was recently reported that several RAVs have a recent origin and occurred independently on distinct viral lineages (Patino-Galindo et al., 2016). This observation is in line with the notion that substitutions at RAV positions can reduce viral fitness (Bartenschlager et al., 2013) and thus fail to be maintained in viral populations or transmission clusters. Indeed, by calculating the strength of selective pressure acting on HCV proteins targeted by DAAs, we observed that most RAV positions are subject to a degree of purifying selection similar to non-RAV positions. This suggests that RAVs do not arise via relaxed selection but rather that these positions are functionally constrained. However, we also found that RAV positions are targeted by positive selection more frequently than expected by chance. Variation at these sites must therefore be adaptive for the virus, raising the possibility that DAA treatment further selects for DAA-resistant viruses with high fitness. This pattern contrasts with observations in HIV-1 infection, as positive selection at codons that confer drug resistance was only observed in patients receiving antiretroviral therapy (de S Leal et al., 2004). We note, though, that most RAVs investigated here were described for genotype 1 and might confer little or no drug resistance to other HCV genotypes. Conversely, most RAV sites were positively selected on branches different than that leading to genotype 1. Their functional relevance will thus need to be analyzed further.

AUTHOR CONTRIBUTIONS

MS conceived the study, with inputs from DF and MC. MS and MC supervised the project. RC and CP performed the evolutionary analysis in mammals. DF performed the molecular dating analyses. RC, CP, and DF performed the evolutionary analysis of HCV phylogenies. MS and DF performed the RAV analyses. JV and LDG performed the docking analyses. RC and DF produced the figures, with input from all authors. UP provided support during the bioinformatic analyses. MS, MC, and DF wrote the manuscript, with critical input from all authors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00854/full#supplementary-material

REFERENCES

Aiewsakun, P., and Katzourakis, A. (2016). Time-dependent rate phenomenon in viruses. *J. Virol.* 90, 7184–7195. doi: 10.1128/JVI.00593-16

Akamatsu, S., Hayes, C. N., Ochi, H., Uchida, T., Kan, H., Murakami, E., et al. (2015). Association between variants in the interferon lambda 4 locus and substitutions in the hepatitis C virus non-structural protein 5A. *J. Hepatol.* 63, 554–563. doi: 10.1016/j.jhep.2015.03.033

Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18, 1585–1592. doi: 10.1093/oxfordjournals.molbev.a003945

Anisimova, M., Bielawski, J. P., and Yang, Z. (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* 19, 950–958. doi: 10.1093/oxfordjournals.molbev.a004152

Ansari, M. A., Pedergnana, V., L C Ip, C., Magri, A., Von Delft, A., Bonsall, D., et al. (2017). Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.* 49, 666–673. doi: 10.1038/ng.3835

Bartenschlager, R., Lohmann, V., and Penin, F. (2013). The molecular and structural basis of advanced antiviral therapy for hepatitis C virus infection. *Nat. Rev. Microbiol.* 11, 482–496. doi: 10.1038/nrmicro3046

Bay, R. A., and Bielawski, J. P. (2011). Recombination detection under evolutionary scenarios relevant to functional divergence. *J. Mol. Evol.* 73, 273–286. doi: 10.1007/s00239-011-9473-0

Bertaux, C., and Dragic, T. (2006). Different domains of CD81 mediate distinct stages of hepatitis C virus pseudoparticle entry. *J. Virol.* 80, 4940–4948. doi: 10.1128/JVI.80.10.4940-4948.2006

Bitzegeio, J., Bankwitz, D., Hueging, K., Haid, S., Brohm, C., Zeisel, M. B., et al. (2010). Adaptation of hepatitis C virus to mouse CD81 permits infection of mouse cells in the absence of human entry factors. *PLoS Pathog.* 6:e1000978. doi: 10.1371/journal.ppat.1000978

Burbelo, P. D., Dubovi, E. J., Simmonds, P., Medina, J. L., Henriquez, J. A., Mishra, N., et al. (2012). Serology-enabled discovery of genetically diverse hepaciviruses in a new host. *J. Virol.* 86, 6171–6178. doi: 10.1128/JVI.00250-12

Cannalire, R., Barreca, M. L., Manfroni, G., and Cecchetti, V. (2016). A journey around the medicinal chemistry of hepatitis C virus inhibitors targeting NS4B: from target to preclinical drug candidates. *J. Med. Chem.* 59, 16–41. doi: 10.1021/acs.jmedchem.5b00825

Catanese, M. T., Loureiro, J., Jones, C. T., Dorner, M., von Hahn, T., and Rice, C. M. (2013). Different requirements for scavenger receptor class B type I in hepatitis C virus cell-free versus cell-to-cell transmission. *J. Virol.* 87, 8282–8293. doi: 10.1128/JVI.01102-13

Chan, D. P., Sun, H. Y., Wong, H. T., Lee, S. S., and Hung, C. C. (2016). Sexually acquired hepatitis C virus infection: a review. *Int. J. Infect. Dis.* 49, 47–58. doi: 10.1016/j.ijid.2016.05.030

Chen, Z. W., Li, H., Ren, H., and Hu, P. (2016). Global prevalence of pre-existing HCV variants resistant to direct-acting antiviral agents (DAAs): mining the GenBank HCV genome data. *Sci. Rep.* 6:20310. doi: 10.1038/srep20310

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* 32, W96–W99. doi: 10.1093/nar/gkh354

Corman, V. M., Grundhoff, A., Baechlein, C., Fischer, N., Gmyl, A., Wollny, R., et al. (2015). Highly divergent hepaciviruses from African cattle. *J. Virol.* 89, 5876–5882. doi: 10.1128/JVI.00393-15

de S Leal, E., Holmes, E. C., and Zanotto, P. M. (2004). Distinct patterns of natural selection in the reverse transcriptase gene of HIV-1 in the presence and absence of antiretroviral therapy. *Virology* 325, 181–191. doi: 10.1016/j.virol.2004.04.004

Delport, W., Scheffler, K., Botha, G., Gravenor, M. B., Muse, S. V., and Kosakovsky Pond, S. L. (2010). CodonTest: modeling amino acid substitution preferences in coding sequences. *PLoS Comput. Biol.* 6:e1000885. doi: 10.1371/journal.pcbi.1000885

Ding, Q., von Schaewen, M., and Ploss, A. (2014). The impact of hepatitis C virus entry on viral tropism. *Cell Host Microbe* 16, 562–568. doi: 10.1016/j.chom.2014.10.009

Dobson, A. P., and Carper, E. R. (1996). Infectious diseased and human population history. *Bioscience* 46, 115–126. doi: 10.2307/1312814

Donovan, J., Dufner, M., and Korennykh, A. (2013). Structural basis for cytosolic double-stranded RNA surveillance by human oligoadenylate synthetase 1. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1652–1657. doi: 10.1073/pnas.1218528110

Drexler, J. F., Corman, V. M., Muller, M. A., Lukashev, A. N., Gmyl, A., Coutard, B., et al. (2013). Evidence for novel hepaciviruses in rodents. *PLoS Pathog.* 9:e1003438. doi: 10.1371/journal.ppat.1003438

Drummer, H. E., Boo, I., and Poumbourios, P. (2007). Mutagenesis of a conserved fusion peptide-like motif and membrane-proximal heptad-repeat region of hepatitis C virus glycoprotein E1. *J. Gen. Virol.* 88(Pt 4), 1144–1148. doi: 10.1099/vir.0.82567-0

Drummer, H. E., Wilson, K. A., and Poumbourios, P. (2002). Identification of the hepatitis C virus E2 glycoprotein binding site on the large extracellular loop of CD81. *J. Virol.* 76, 11143–11147. doi: 10.1128/JVI.76.21.11143-11147.2002

Duchene, S., Duchene, D., Holmes, E. C., and Ho, S. Y. (2015). The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* 32, 1895–1906. doi: 10.1093/molbev/msv056

Duchene, S., Holmes, E. C., and Ho, S. Y. (2014). Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B Biol. Sci.* 281:20140732. doi: 10.1098/rspb.2014.0732

Flint, M., von Hahn, T., Zhang, J., Farquhar, M., Jones, C. T., Balfe, P., et al. (2006). Diverse CD81 proteins support hepatitis C virus infection. *J. Virol.* 80, 11331–11342. doi: 10.1128/JVI.00104-06

Forni, D., Cagliani, R., Clerici, M., and Sironi, M. (2017). Molecular evolution of human coronavirus genomes. *Trends Microbiol.* 25, 35–48. doi: 10.1016/j.tim.2016.09.001

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* 49, 6177–6196.

Gale, M. Jr., Blakely, C. M., Kwieciszewski, B., Tan, S. L., Dossett, M., Tang, N. M., et al. (1998). Control of PKR protein kinase by hepatitis C virus nonstructural 5A protein: molecular mechanisms of kinase regulation. *Mol. Cell. Biol.* 18, 5208–5218. doi: 10.1128/MCB.18.9.5208

Gouttenoire, J., Castet, V., Montserret, R., Arora, N., Raussens, V., Ruysschaert, J. M., et al. (2009). Identification of a novel determinant for membrane association in hepatitis C virus nonstructural protein 4B. *J. Virol.* 83, 6257–6268. doi: 10.1128/JVI.02663-08

Guindon, S., Delsuc, F., Dufayard, J. F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* 537, 113–137. doi: 10.1007/978-1-59745-251-9_6

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2016). OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* 12, 281–296. doi: 10.1021/acs.jctc.5b00864

He, J., Sun, E., Bujny, M. V., Kim, D., Davidson, M. W., and Zhuang, X. (2013). Dual function of CD81 in influenza virus uncoating and budding. *PLoS Pathog.* 9:e1003701. doi: 10.1371/journal.ppat.1003701

Ho, S. Y., Duchene, S., Molak, M., and Shapiro, B. (2015). Time-dependent estimates of molecular evolutionary rates: evidence and causes. *Mol. Ecol.* 24, 6007–6012. doi: 10.1111/mec.13450

Hung, H., Nguyen, K. D., Bellwood, P., and Carson, M. T. (2013). Coastal connectivity: long-term trading networks across the South China Sea. *J. Island Coast. Archaeol.* 8, 384–404. doi: 10.1080/15564894.2013.781085

Jackowiak, P., Kuls, K., Budzko, L., Mania, A., Figlerowicz, M., and Figlerowicz, M. (2014). Phylogeny and molecular evolution of the hepatitis C virus. *Infect. Genet. Evol.* 21, 67–82. doi: 10.1016/j.meegid.2013.10.021

Kapoor, A., Simmonds, P., Gerold, G., Qaisar, N., Jain, K., Henriquez, J. A., et al. (2011). Characterization of a canine homolog of hepatitis C virus. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11608–11613. doi: 10.1073/pnas.1101794108

Kapoor, A., Simmonds, P., Scheel, T. K., Hjelle, B., Cullen, J. M., Burbelo, P. D., et al. (2013). Identification of rodent homologs of hepatitis C virus and pegiviruses. *mBio* 4:e00216-13. doi: 10.1128/mBio.00216-13

Kosakovsky Pond, S. L., and Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222. doi: 10.1093/molbev/msi105

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23, 1891–1901. doi: 10.1093/molbev/msl051

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., et al. (2017). The ClusPro web server for protein-protein docking. *Nat. Protoc.* 12, 255–278. doi: 10.1038/nprot.2016.169

Lange, C. M., Bellecave, P., Dao Thi, V. L., Tran, H. T., Penin, F., Moradpour, D., et al. (2014). Determinants for membrane association of the hepatitis C virus NS2 protease domain. *J. Virol.* 88, 6519–6523. doi: 10.1128/JVI.00224-14

Lavie, M., Sarrazin, S., Montserret, R., Descamps, V., Baumert, T. F., Duverlie, G., et al. (2014). Identification of conserved residues in hepatitis C virus envelope glycoprotein E2 that modulate virus dependence on CD81 and SRB1 entry factors. *J. Virol.* 88, 10584–10597. doi: 10.1128/JVI.01402-14

Lawler, A. (2014). Sailing Sinbad's seas. *Science* 344, 1440–1445. doi: 10.1126/science.344.6191.1440

Li, C., Lu, L., Murphy, D. G., Negro, F., and Okamoto, H. (2014). Origin of hepatitis C virus genotype 3 in Africa as estimated through an evolutionary analysis of the full-length genomes of nine subtypes, including the newly sequenced 3d and 3e. *J. Gen. Virol.* 95(Pt 8), 1677–1688. doi: 10.1099/vir.0.065128-0

Li, H. F., Huang, C. H., Ai, L. S., Chuang, C. K., and Chen, S. S. (2009). Mutagenesis of the fusion peptide-like domain of hepatitis C virus E1 glycoprotein: involvement in cell fusion and virus entry. *J. Biomed. Sci.* 16:89. doi: 10.1186/1423-0127-16-89

Lindenbach, B. D., and Rice, C. M. (2013). The ins and outs of hepatitis C virus entry and assembly. *Nat. Rev. Microbiol.* 11, 688–700. doi: 10.1038/nrmicro3098

Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., and Jiggins, F. M. (2014). The evolution and genetics of virus host shifts. *PLoS Pathog.* 10:e1004395. doi: 10.1371/journal.ppat.1004395

Lontok, E., Harrington, P., Howe, A., Kieffer, T., Lennerstrand, J., Lenz, O., et al. (2015). Hepatitis C virus drug resistance-associated substitutions: state of the art summary. *Hepatology* 62, 1623–1632. doi: 10.1002/hep.27934

Loytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10557–10562. doi: 10.1073/pnas.0409137102

Lu, L., Li, C., Xu, Y., and Murphy, D. G. (2014). Full-length genomes of 16 hepatitis C virus genotype 1 isolates representing subtypes 1c, 1d, 1e, 1g, 1h, 1i, 1j and 1k, and two new subtypes 1m and 1n, and four unclassified variants reveal ancestral relationships among subtypes. *J. Gen. Virol.* 95(Pt 7), 1479–1487. doi: 10.1099/vir.0.064980-0

Mark, S. (2002). Alexander the Great, seafaring, and the spread of leprosy. *J. Hist. Med. Allied Sci.* 57, 285–311. doi: 10.1093/jhmas/57.3.285

Martin, D., and Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16, 562–563. doi: 10.1093/bioinformatics/16.6.562

Martin, D. P., Murrell, B., Khoosal, A., and Muhire, B. (2017). Detecting and analyzing genetic recombination using RDP4. *Methods Mol. Biol.* 1525, 433–460. doi: 10.1007/978-1-4939-6622-6_17

Messina, J. P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G. S., Pybus, O. G., et al. (2015). Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* 61, 77–87. doi: 10.1002/hep.27259

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D., and Wiersma, S. T. (2013). Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 57, 1333–1342. doi: 10.1002/hep.26141

Murray, G. G., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., et al. (2016). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* 7, 80–89. doi: 10.1111/2041-210X.12466

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., et al. (2015). Gene-wide identification of episodic selection. *Mol. Biol. Evol.* 32, 1365–1371. doi: 10.1093/molbev/msv035

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764. doi: 10.1371/journal.pgen.1002764

Muse, S. V., and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature* 541, 302–310. doi: 10.1038/nature21347

O'Brien, T. R., Prokunina-Olsson, L., and Donnelly, R. P. (2014). IFN-lambda4: the paradoxical new member of the interferon lambda family. *J. Interferon Cytokine Res.* 34, 829–838. doi: 10.1089/jir.2013.0136

Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., et al. (2009). The earliest horse harnessing and milking. *Science* 323, 1332–1335. doi: 10.1126/science.1168594

Patino-Galindo, J. A., and Gonzalez-Candelas, F. (2017). Comparative analysis of variation and selection in the HCV genome. *Infect. Genet. Evol.* 49, 104–110. doi: 10.1016/j.meegid.2017.01.010

Patino-Galindo, J. A., Salvatierra, K., Gonzalez-Candelas, F., and Lopez-Labrador, F. X. (2016). Comprehensive screening for naturally occurring hepatitis C virus resistance to direct-acting antivirals in the NS3, NS5A, and NS5B genes in worldwide isolates of viral genotypes 1 to 6. *Antimicrob. Agents Chemother.* 60, 2402–2416. doi: 10.1128/AAC.02776-15

Pavio, N., Romano, P. R., Graczyk, T. M., Feinstone, S. M., and Taylor, D. R. (2003). Protein synthesis and endoplasmic reticulum stress can be modulated by the hepatitis C virus envelope protein E2 through the eukaryotic initiation factor 2alpha kinase PERK. *J. Virol.* 77, 3578–3585. doi: 10.1128/JVI.77.6.3578-3585.2003

Pearson, M. (2016). *Trade, Circulation, and Flow in the Indian Ocean World*. Berlin: Springer.

Pedergnana, V., Smith, D., STOP-HCV Consortium Klenerman, P., Barnes, E., and Spencer, C. C. (2016). Interferon lambda 4 variant rs12979860 is not associated with RAV NS5A Y93H in hepatitis C virus genotype 3a. *Hepatology* 64, 1377–1378. doi: 10.1002/hep.28533

Peiffer, K. H., Sommer, L., Susser, S., Vermehren, J., Herrmann, E., Doring, M., et al. (2016). Interferon lambda 4 genotypes and resistance-associated variants in patients infected with hepatitis C virus genotypes 1 and 3. *Hepatology* 63, 63–73. doi: 10.1002/hep.28255

Perales, C., Beach, N. M., Gallego, I., Soria, M. E., Quer, J., Esteban, J. I., et al. (2013). Response of hepatitis C virus to long-term passage in the presence of alpha interferon: multiple mutations and a common phenotype. *J. Virol.* 87, 7593–7607. doi: 10.1128/JVI.02824-12

Pfaender, S., Brown, R. J., Pietschmann, T., and Steinmann, E. (2014). Natural reservoirs for homologs of hepatitis C virus. *Emerg. Microbes Infect.* 3:e21. doi: 10.1038/emi.2014.19

Pflugheber, J., Fredericksen, B., Sumpter, R. Jr., Wang, C., Ware, F., Sodora, D. L., et al. (2002). Regulation of PKR and IRF-1 during hepatitis C virus RNA replication. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4650–4655. doi: 10.1073/pnas.062055699

Posada, D., and Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13757–13762. doi: 10.1073/pnas.241370698

Privman, E., Penn, O., and Pupko, T. (2012). Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.* 29, 1–5. doi: 10.1093/molbev/msr177

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C., and Harvey, P. H. (2001). The epidemic behavior of the hepatitis C virus. *Science* 292, 2323–2325. doi: 10.1126/science.1058321

Pybus, O. G., and Gray, R. R. (2013). Virology: the virus whose family expanded. *Nature* 498, 310–311. doi: 10.1038/498310b

Pybus, O. G., Markov, P. V., Wu, A., and Tatem, A. J. (2007). Investigating the endemic transmission of the hepatitis C virus. *Int. J. Parasitol.* 37, 839–849. doi: 10.1016/j.ijpara.2007.04.009

Pybus, O. G., and Theze, J. (2016). Hepacivirus cross-species transmission and the origins of the hepatitis C virus. *Curr. Opin. Virol.* 16, 1–7. doi: 10.1016/j.coviro.2015.10.002

Quan, P. L., Firth, C., Conte, J. M., Williams, S. H., Zambrana-Torrelio, C. M., Anthony, S. J., et al. (2013). Bats are a major natural reservoir for hepaciviruses and pegiviruses. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8194–8199. doi: 10.1073/pnas.1303037110

Rai, R., and Deval, J. (2011). New opportunities in anti-hepatitis C virus drug discovery: targeting NS4B. *Antiviral Res.* 90, 93–101.

Richens, J. (2006). Donovanosis (granuloma inguinale). *Sex. Transm. Infect.* 82(Suppl. 4), iv21–iv22. doi: 10.1136/sti.2006.023135

Russell, R. S., Kawaguchi, K., Meunier, J. C., Takikawa, S., Faulk, K., Bukh, J., et al. (2009). Mutational analysis of the hepatitis C virus E1 glycoprotein in retroviral pseudoparticles and cell-culture-derived H77/JFH1 chimeric infectious virus particles. *J. Viral Hepat.* 16, 621–632. doi: 10.1111/j.1365-2893.2009.01111.x

Sakamoto, H., Okamoto, K., Aoki, M., Kato, H., Katsume, A., Ohta, A., et al. (2005). Host sphingolipid biosynthesis as a target for hepatitis C virus therapy. *Nat. Chem. Biol.* 1, 333–337. doi: 10.1038/nchembio742

Samper, J. C., and Tibary, A. (2006). Disease transmission in horses. *Theriogenology* 66, 551–559. doi: 10.1016/j.theriogenology.2006.04.019

Sanderson, M. J. (2003). R8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6, 526–538.

Schaer, J., Perkins, S. L., Decher, J., Leendertz, F. H., Fahr, J., Weber, N., et al. (2013). High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17415–17419. doi: 10.1073/pnas.1311016110

Scheel, T. K., Simmonds, P., and Kapoor, A. (2015). Surveying the global virome: identification and characterization of HCV-related animal hepaciviruses. *Antiviral Res.* 115, 83–93. doi: 10.1016/j.antiviral.2014.12.014

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33, W363–W367. doi: 10.1093/nar/gki481

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14. doi: 10.1093/nar/gkv318

Shepard, C. W., Finelli, L., and Alter, M. J. (2005). Global epidemiology of hepatitis C virus infection. *Lancet Infect. Dis.* 5, 558–567. doi: 10.1016/S1473-3099(05)70216-4

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 49, 534–553. doi: 10.1021/jm050540c

Silvie, O., Rubinstein, E., Franetich, J. F., Prenant, M., Belnoue, E., Renia, L., et al. (2003). Hepatocyte CD81 is required for *Plasmodium falciparum* and *Plasmodium yoelii* sporozoite infectivity. *Nat. Med.* 9, 93–96. doi: 10.1038/nm808

Simmonds, P. (2013). The origin of hepatitis C virus. *Curr. Top. Microbiol. Immunol.* 369, 1–15. doi: 10.1007/978-3-642-27340-7_1

Sironi, M., Cagliani, R., Forni, D., and Clerici, M. (2015). Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* 16, 224–236. doi: 10.1038/nrg3905

Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T., et al. (2014). Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology* 59, 318–327. doi: 10.1002/hep.26744

Smith, D. B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P. L., et al. (1997). The origin of hepatitis C virus genotypes. *J. Gen. Virol.* 78(Pt 2), 321–328. doi: 10.1099/0022-1317-78-2-321

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353. doi: 10.1093/molbev/msv022

Sumpter, R. Jr., Wang, C., Foy, E., Loo, Y. M., and Gale, M. Jr. (2004). Viral evolution and interferon resistance of hepatitis C virus RNA replication in a cell culture model. *J. Virol.* 78, 11591–11604. doi: 10.1128/JVI.78.21.11591-11604.2004

Taguchi, T., Nagano-Fujii, M., Akutsu, M., Kadoya, H., Ohgimoto, S., Ishido, S., et al. (2004). Hepatitis C virus NS5A protein interacts with 2′,5′-oligoadenylate synthetase and inhibits antiviral activity of IFN in an IFN sensitivity-determining region-independent manner. *J. Gen. Virol.* 85(Pt 4), 959–969. doi: 10.1099/vir.0.19513-0

Tanaka, T., Kuroda, K., Ikeda, M., Wakita, T., Kato, N., and Makishima, M. (2013). Hepatitis C virus NS4B targets lipid droplets through hydrophobic residues in the amphipathic helices. *J. Lipid Res.* 54, 881–892. doi: 10.1194/jlr.M026443

Taylor, D. R., Shi, S. T., Romano, P. R., Barber, G. N., and Lai, M. M. (1999). Inhibition of the interferon-inducible protein kinase PKR by HCV E2 protein. *Science* 285, 107–110. doi: 10.1126/science.285.5424.107

Tellinghuisen, T. L., Marcotrigiano, J., and Rice, C. M. (2005). Structure of the zinc-binding domain of an essential component of the hepatitis C virus replicase. *Nature* 435, 374–379. doi: 10.1038/nature03580

Van Nguyen, D., Van Nguyen, C., Bonsall, D., Ngo, T. T., Carrique-Mas, J., Pham, A. H., et al. (2018). Detection and characterization of homologues of human hepatitis viruses and pegiviruses in rodents and bats in Vietnam. *Viruses* 10:E102. doi: 10.3390/v10030102

Walter, S., Rasche, A., Moreira-Soto, A., Pfaender, S., Bletsa, M., Corman, V. M., et al. (2016). Differential infection patterns and recent evolutionary origins of equine hepaciviruses in donkeys. *J. Virol.* 91:e01711-16.

Weng, L., Hirata, Y., Arai, M., Kohara, M., Wakita, T., Watashi, K., et al. (2010). Sphingomyelin activates hepatitis C virus RNA polymerase in a genotype-specific manner. *J. Virol.* 84, 11761–11770. doi: 10.1128/JVI.00638-10

Wernersson, R., and Pedersen, A. G. (2003). RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* 31, 3537–3539. doi: 10.1093/nar/gkg609

Wertheim, J. O., Chu, D. K., Peiris, J. S., Kosakovsky Pond, S. L., and Poon, L. L. (2013). A case for the ancient origin of coronaviruses. *J. Virol.* 87, 7039–7045. doi: 10.1128/JVI.03273-12

Wertheim, J. O., and Kosakovsky Pond, S. L. (2011). Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* 28, 3355–3365. doi: 10.1093/molbev/msr170

Wertheim, J. O., Smith, M. D., Smith, D. M., Scheffler, K., and Kosakovsky Pond, S. L. (2014). Evolutionary origins of human herpes simplex viruses 1 and 2. *Mol. Biol. Evol.* 31, 2356–2364. doi: 10.1093/molbev/msu185

White, J. A. (2009). Manifestations and management of lymphogranuloma venereum. *Curr. Opin. Infect. Dis.* 22, 57–66. doi: 10.1097/QCO.0b013e328320a8ae

Xu, Z., Choi, J., Yen, T. S., Lu, W., Strohecker, A., Govindarajan, S., et al. (2001). Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. *EMBO J.* 20, 3840–3848. doi: 10.1093/emboj/20.14.3840

Yalaoui, S., Zougbede, S., Charrin, S., Silvie, O., Arduise, C., Farhati, K., et al. (2008). Hepatocyte permissiveness to *Plasmodium* infection is conveyed by a short and structurally conserved region of the CD81 large extracellular domain. *PLoS Pathog.* 4:e1000010. doi: 10.1371/journal.ppat.1000010

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, Z., Wong, W. S., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107–1118. doi: 10.1093/molbev/msi097

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479. doi: 10.1093/molbev/msi237

Zuiani, A., Chen, K., Schwarz, M. C., White, J. P., Luca, V. C., Fremont, D. H., et al. (2016). A library of infectious hepatitis C viruses with engineered mutations in the E2 gene reveals growth-adaptive mutations that modulate interactions with scavenger receptor class B type I. *J. Virol.* 90, 10499–10512. doi: 10.1128/JVI.01011-16