Infection, Genetics and Evolution 69 (2019) 48–60

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Differential interaction strategies of hepatitis c virus genotypes during entry – An in silico investigation of envelope glycoprotein E2 - CD81 interaction☆

Chayan Bhattacharjee[a,1], Suman K. Nandy[b,1,2], Pratik Das[a], Aparna Mukhopadhyay[a,*]

[a] *Department of Life Sciences, Presidency University, 86/1 College Street, Kolkata, 700073, India*
[b] *Department of Biochemistry & Biophysics, University of Kalyani, Kalyani 741235, India*

ARTICLE INFO

Keywords:
HCV
Envelope glycoprotein E2
CD81
Protein docking

ABSTRACT

Hepatitis C Virus is a blood borne pathogen responsible for chronic hepatitis in more than 71 million people. Wide variations across strains and genotypes are one of the major hurdles in therapeutic development. While genotype 1 remains the most extensively studied and abundant strain, genotype 3 is more virulent and second most prevalent. This study aimed to compare differences in the glycoprotein E2 across HCV genotypes at nucleotide, protein and structural levels. Nucleotide sequences of E2 from 29 strains across genotypes 1a, 1b, 3a and 3b revealed a stark preference for C-richness which was attributed to a distinct bias for C-rich codons in genotype 1. Genotype 3 exhibited a similar preference to a lesser extent. Amino acid level comparison revealed majority of the changes at the C-terminal half of the proteins leaving the N-terminal region conspicuously conserved apart from the two hyper variable regions. Amino acid changes across genotypes were mostly polar-nonpolar alterations. In silico models of E2 glycoproteins and docking analysis with the energy minimized PDB-CD81 model revealed unique interacting residues in both E2 and CD81. While several CD81 binding residues were common for all four genotypes, number and composition of interacting residues varied. The interacting residues of E2 were however unique for each genotype. E2 of genotype 3a and CD81 had the strongest interaction. In conclusion this is the first comprehensive study comparing E2 sequences across genotypes 1a, 1b, 3a and 3b revealing stark genotype-specific differences which requires more extensive investigation.

1. Introduction

Hepatitis C virus (HCV), a member of the Flaviviridae family, is one of the leading causes of chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma worldwide. HCV, an enveloped virus with an approximately 9.6 kb long, positive stranded RNA genome, encodes a single polyprotein of about 3010 amino acids. Post-translationally, this polyprotein is cleaved (Fields et al., 2013) to generate at least 11 structural and non-structural proteins. Structural proteins include two viral glycoproteins, E1 and E2 (Yagnik et al., 2000) located at the N terminus of the polyprotein (Moradpour and Penin, 2013). HCV heterodimer E1/E2, the complex responsible for viral entry and tropism, are type I transmembrane proteins with an N terminal ectodomain and C- terminal transmembrane domain (TMD). The exact role of E1 in viral entry is poorly understood, but is believed to be related to the fusion process (Boo et al., 2012; Callens et al., 2005). E2, however, is currently the better characterized subunit where it plays pivotal roles in HCV entry, i.e. interaction with entry factors like scavenger receptor BI (SR-B1) (Scarselli et al., 2002), tetraspanin CD81 (Pileri et al., 1998) along with several tight junction proteins, Claudin-1 (Evans et al., 2007) and occludin receptors (Ciesek et al., 2011). E2 is also one of the most diverse HCV proteins characterized by two hypervariable regions (HVR), HVR1 and HVR2. Previous reports have shown that both HVR1 and HVR2 are important for the interaction with the cellular receptor

Abbreviations: HCV, Hepatitis C Virus; GP, Glycoprotein; CD81, Cluster of Differentiation 81, Tetraspanin protein; E1 /E2, Envelope glycoprotein on HCV; HCVpp, HCV pseudoparticle; HCVcc, Cell culture adaptable HCV; TMD, Transmembrane domain; CUB, codon usage bias; RMS, Root Mean Square; PROVE, PROtein Volume Evaluation; SAVES, Structural Analysis and Verification Server; BLAST, Basic Local alignment Search Tool; PDB, Protein data bank; SASA, solvent accessible surface area; IE, Interaction energy; HB, hydrogen bond; VDW, Van der Waals

☆ Data in this paper is part of a doctoral thesis to be submitted in partial fulfilment of the requirements of a degree of Doctor of Philosophy from Presidency University, Kolkata.

* Corresponding author.

E-mail address: aparna.dbs@presiuniv.ac.in (A. Mukhopadhyay).

[1] Both authors contributed equally to this work.

[2] Present Address: Bioinformatics Infrastructure Facility, North-Eastern Hill University, Tura Campus, Tura, Meghalaya-794002, India.

https://doi.org/10.1016/j.meegid.2019.01.008
Received 3 August 2018; Received in revised form 12 December 2018; Accepted 8 January 2019
Available online 11 January 2019

protein CD81 (Callens et al., 2005; McCaffrey et al., 2011; Roccasecca et al., 2003), a 26 kDa integral membrane protein (Pileri et al., 1998). The receptor is essential for HCV glycoprotein mediated entry for all genotypes (Roccasecca et al., 2003). The CD81 binding site in E2 has been investigated in many studies primarily using genotype 1a (strain H77) and 2a (using JFH-1) (Callens et al., 2005; Roccasecca et al., 2003).

Like other RNA viruses HCV also exhibits a high degree of genetic diversity, creating a major challenge for the development of both HCV vaccines and pan-genotypic therapeutics (Timm and Roggendorf, 2007). The virus has a total of seven genotypes, with > 50 subtypes and millions of quasispecies (Cuypers et al., 2015). Among the seven genotypes of HCV, genotype 1 is the most prevalent worldwide, mainly present in Europe and America followed by genotype 3 (Messina et al., 2015).

The seven genotypes of HCV have been reported to differ by as much as 33%, with variations distributed throughout the genome (Okamoto et al., 1992). Most diversity however lies in the HVRs of E2 (Okamoto et al., 1991). Majority of our knowledge on HCV biology is based on genotypes 1 and 2 as reagents such as the subgenomic replicon system, HCVcc (cell culture adaptable HCV) and HCVpp (HCV pseudoparticles) have been prepared from these strains (Catanese and Dorner, 2015). However, clinical information on other strains are quite well documented and it is widely accepted that the pathogenesis and prognosis of infection by various genotypes vary widely (Alonso et al., 2016; Bochud et al., 2009; Zein, 2000), with genotype 3 being more virulent and pathogenic compared to the others (Chayama and Hayes, 2011). Genotype 3 has therefore been the focus of this study in comparison to genotype 1.

In this study, we compared E2 sequences between genotypes 1 and 3 at the nucleotide level and then extended our quest to the amino acid and structural level. As nucleotide differences must be translated into differences at the amino acid as well, codon usage bias (CUB) where preference is given to one or two codons over the other synonymous ones was investigated. CUB stems from translational pressure governed by the conditions present in the host as well as mutational pressure (Chen, 2013). Such CUB has also been reported in HCV, this present study aimed to explore how such codon bias is maintained in spite of the wide differences at nucleotide and protein level. Furthermore using the protein modelling tools, E2 models were generated for genotypes 1 and 3 and their binding affinity to CD81 was calculated. This study brings to light an important aspect of E2-CD81 interaction; differences between genotypes 1 and 3 which once validated experimentally could be a major step towards design of better genotype specific therapeutics.

2. Method

2.1. HCV sequence selection and analyses

Sequences and annotations of CD81 and envelope glycoprotein E2 of Hepatitis C Virus (HCV) strains 1a, 1b, 3a and 3b were downloaded from UniProt (http://www.uniprot.org) (UniProt ID P60033) and NCBI (http://www.ncbi.nlm.nih.gov/) Protein databases. The corresponding nucleotide sequences were retrieved by tblastn sequence similarity search. The serial number, GenBank accession numbers and other information are listed in Supplementary Table S1. For analysis of the C content of the entire HCV genome, glycoprotein E1 and RNA dependent RNA polymerase (RdRp), the same sequences were used which are also listed. However as the sequences used for analysis of glycoprotein E1 from genotype 3a were partial coding sequences, a few of the sequences used for other analyses were from other strains (also listed).

Nucleotide composition of the 29 HCV E2 sequences and 27 sequences of HCV whole genome, E1 and RdRp were determined by using CAIcal server (http://genomes.urv.es/CAIcal). Compositional properties such as frequency of Adenine, Thymine, Guanine and Cytosine at the 1st, 2nd and 3rd position of a codon of HCV E2 were analysed using mega7 software (version MEGA 7.0.26) (Kumar et al., 2016). Analysis of the whole genome of HCV, glycoprotein E1 and the RNA dependent RNA polymerase (RdRp) were conducted similarly. Percentage occurrence was calculated manually.

2.2. Codon usage and cluster codon analysis

Frequency, number, and fraction of each amino acid in the 29 sequences were evaluated by web server *Gene Infinity* (2016 Gene Infinity LLC) (http://www.geneinfinity.org/sms/sms_codonusage.html). Amino acid changes, Intragenic, intergenic changes within the same nucleotide position were analysed manually.

2.3. In-silico characterisation of proteins

HCV and CD81 sequences were characterized by Expasy-ProtParam tool (Gasteiger et al., 2005), XtalPred server (Slabinski et al., 2007), Discovery Studio (DS) 2.5 and predicted for intrinsic disordered regions through top CASP performers MetaDisorder (Kozlowski and Bujnicki, 2012) and DISOPRED3 (Jones and Cozzetto, 2015). MetaDisorder combines weighted consensus of thirteen different primary disorder prediction server results, with fold recognition based disorder prediction components, through genetic algorithm optimization. DISOPRED3 on the other hand, identifies the long, intrinsic, disordered regions and different sequence patterns associated with terminal and internal disordered regions. It also predicts the protein binding sites within disordered regions using a neural network and a nearest neighbour classifier. Furthermore, the transmembrane regions of HCV E2s were predicted using TMHMM 2.0 (Sonnichsen et al., 2000) and Phobius (Käll et al., 2007) servers, subjected to at least ten residue overlap between the server results.

2.4. Protein modelling

Due to unavailability of template structures with suitable identity and query coverage in Protein Data Bank (PDB) at www.rcsb.org (Berman et al., 2000), the three-dimensional structure of receptor protein CD81 and E2 from all four HCV strains were modelled by template-based modelling, using remote template, fold recognition, and ab initio approaches through I-TASSER (Yang et al., 2015), Phyre^2 (Kelley et al., 2015) and RaptorX (Källberg et al., 2012) modelling servers. The best models of each server were protonated at physiological pH and ionic strength, subjected to loop refinement and minimization in smart minimizer of DS 2.5 to satisfy a RMS (Root Mean Square) gradient of 0.1 kcal/mol Å. The secondary structure assignment of the best models of each server and the corresponding minimized ones were compared with existing truncated E2 structures (PDB ID 4MWF, chain C, D & 4WEB, chain E) by structural superposition and all the models were further validated for stereochemistry, three dimensional structural propensity, non-bonded interactions, atomic volumes check against curated databases of refined structures through RAMPAGE (Lovell et al., 2003), WHAT_CHECK (Hooft et al., 1996), Verify3D (Bowie et al., 1991; Lüthy et al., 1992), ERRAT (Colovos and Yeates, 1993) and PROtein Volume Evaluation (PROVE) (Pontius et al., 1996) in Structural Analysis and Verification Server v5.0 (SAVES) at http://services.mbi.ucla.edu/SAVES/. To compare our models with published literature, relative solvent accessibility (RSA) which is percentage of residue accessible to solvent was measured. This was achieved using a scale of pre-calculated fully exposed solvent accessibility of a particular residue using Discovery Studio 4.0 with 960 grid points per atom and probe radius of 1.4 Å. Residues with RSA ≥16% was taken as exposed residues as per (Rost and Sander, 1994).

2.5. Protein-protein interaction

In order to determine the binding site of HCV envelope glycoproteins, structure and sequence based prediction servers such as BSpred (Mukherjee and Zhang, 2011), ConSurf 3.0 (Ashkenazy et al., 2010), meta-PPISP (Qin and Zhou, 2007), PRISM 2.0, (Baspinar et al., 2014) were utilized. Moreover, considering the wide variance in binding site prediction of glycoproteins, ZDOCK rigid body docking algorithm in DS 2.5 was employed to reassert the binding site residues of all four E2, from different HCV strains, by docking them against the known binding residues of CD81 (Leu162, Thr163, Ile181, Ile182, Ser183, Asn184, Leu185, Phe186, Lys187) (Kitadokoro et al., 2001; Pierce et al., 2011; Rajesh et al., 2012) without defining the binding site of E2 glycoprotein. The possible transmembrane part at the C-terminal segment was excluded. ZDOCK performed a six-dimensional grid based search in translational and rotational orientational space of the two molecules by rotating the ligand through uniformly distributed Euler angles (6° or 15°) while keeping the receptor fixed. For each rotation, the translational space was scanned with a step size of 1.2 Å and only the best geometric match between the two proteins was retained. All predicted poses were scored by shape complementarity, electrostatics, desolvation energy in ZDOCK and subsequently reranked by ZRANK using a more detailed weighted energy function involving van der Waal (VDW), electrostatics and desolvation terms (Pierce and Weng, 2007). The ranked poses were further refined with RDOCK through a three step optimization procedure by keeping the ionic residues at neutral state in the first two, i.e., while removing the steric clashes by minimizing the VDW and internal (bond & angle) energies and optimizing the polar interaction. The final stage dealt with charge interaction minimization without any constraints (Li et al., 2003; Pierce and Weng, 2008). The binding interface residues of top 100 RDOCK poses were considered to select the binding site residues of E2 according to their frequency of appearance in binding interface. Afterwards in order to determine the final binding conformations of all four E2-CD81 complexes both the proteins were docked through ZDOCK with defined binding site residues, reranked in ZRANK and refined by RDOCK as detailed above. The best poses of each complex were chosen and analysed by RDOCK score, interaction energy (IE), solvent accessible surface area (SASA), hydrogen bond (HB) and binding free energy in DS 2.5.

2.6. Statistical analyses

All statistical analyses were performed using Microsoft Excel and SPSS 20.0.

3. Results

3.1. Nucleotide base composition analyses of HCV strains reveal a Cytosine-bias

To determine nucleotide base composition variation among different E2 sequences of genotypes 1 and 3, protein sequences were downloaded from NCBI (Supplementary Table S1), translated using tblastn and corresponding nucleotide sequences obtained. Only the available complete sequences were taken and hence the numbers of sequences compared varied among the four different genotypes. The nucleotide positions of E2 were marked as indicated.

Nucleotide composition and diversity compared using CAIcal server are tabulated in Supplementary Table S2 and summarized in Table 1. One way analysis of variance (ANOVA) was applied and comparison was made through Bonferroni's test with least significant difference of $p \leq .05$.

The results clearly indicate a marked abundance of cytosine (C) across all genotypes and an almost similar, but lower abundance of Guanosine. Genotype 1a exhibited the highest average base percentage of Cytosine followed by Genotype 1b, 3b and 3a (Table 1). For each genotype, C content was significantly more ($p < 0.001$) than the other nucleotides. Furthermore, although C content within a genotype didn't vary significantly ($p > 0.001$), but genotypes 1 and 3, it varied significantly ($p < 0.001$). The least frequent nucleotide base was Adenine (A) among all genotypes. To test whether the preference for cytosine exhibited by glycoprotein E2 extended to the entire genome and other proteins such as the glycoprotein E1 and the more conserved RNA dependent RNA polymerase (RdRp), similar analyses were conducted using the specific sequences from each of the regions (Supplementary Tables S3-S5). The accession numbers and regions analysed are all detailed in Supplementary Table S1. It was observed that the entire genome (Supplementary Table S3) and RdRp (Supplementary table S5) from all strains revealed a distinct preference (significant difference from other nucleotides in a genotype, $p < 0.05$) for Cytosine, similar to that seen in glycoprotein E2, but Glycoprotein E1 differed slightly. Glycoprotein E1 (Supplementary Table S4) however, exhibited almost similar preference for both Cytosine and Guanosine in all strains. Avoidance of Adenine was however consistent among all proteins and strains analysed with significant differences between cytosine/ guanosine content ($p < .05$). The variations in the nucleotide content in the entire genome varied significantly ($p < .05$) across genotypes but remained more conserved within a genotype. These results are in agreement with previous reports on GC-richness among RNA viruses (van Hemert et al., 2016).

3.2. Compositional property analysis reveals a preference for C at the third position of a codon

Our earlier results led us to further investigate whether the conserved C-richness of E2 is translated into C-preference in the 3rd nucleotide position. In this study, summarized in Table 2, the nucleotide sequences (detailed in supplementary table S6) were analysed by mega7 software (Kumar et al., 2016) to assess the percentage of each nucleotide in the first, second and third position of a codon. Genotype 1 had a greater preference for C (~40% occurrence) in the 3rd position, compared to genotype 3 (~30% occurrence). One-way ANOVA followed by Bonferonni post hoc analysis showed that in all genotypes except for 1b, the C content both in the 1st and 2nd position, didn't vary significantly ($p > 0.001$, SD = 1.65–2.88, Table 2) from the other nucleotides. In contrast, C content for all genotypes varied significantly ($p < 0.001$, standard deviation of percentage distributions of nucleotide ranged from a minimum of 4.14 in 3a to a maximum of 12.55 in 1a, Table 2) at the 3rd position. The least preferred terminal nucleotide for all the amino acids however were A and T. Similar analyses were conducted using the entire genome of the virus, glycoprotein E1 and the RdRp. While the trend resembled that observed in glycoprotein E1 (Supplementary Table S8) and RdRp (Supplementary Table S9), a different result was observed when the entire genome was used (Supplementary Table S7). Sequences of the entire genome revealed a preference for Guanosine in the 3rd position instead of Cytosine in genotype 1a, although the other genotypes (1b, 3a and 3b) reflected a significantly ($p < 0.01$) greater preference for Cytosine, similar to that observed in the individual proteins. In spite of lower preference for Cytosine in the 3rd position, the entire C content of genotype 1a is high due to preference for C in the 2nd and 1st position of a codon. This highlights and subsequently proves that maintaining C and sometimes G in the third position of a codon has a biological significance in the viral life cycle. This analysis also revealed that major differences occur also within a genotype.

3.3. Cluster codon analyses reveal differential preference for C-rich codons across genotypes

In order to further investigate the apparent preference for C or G by HCV, cluster codon analysis was conducted. In this study, we assessed the percentage usage of each codon by calculating each time a particular codon has been used in the coding sequence of E2. This exercise was repeated for each strain and an average usage with standard deviation is presented in Table 3. Consistent with our earlier result

Table 1
Nucleotide composition analysis (Mean ± SD) of HCV glycoprotein E2 from genotypes 1 and 3.

Genotype	%A	%C	%T	%G	SD within a genotype
1a	20.2 ± 0.71	30.5 ± 0.39[a,b,c]	21.8 ± 0.84	27.5 ± 0.18	4.28
1b	20.3 ± 0.59	30.0 ± 0.83[a,b,c]	21.5 ± 0.53	28.2 ± 0.37	4.28
3a	21.3 ± 0.31	27.9 ± 0.68[a,b,d]	24. 6 ± 0.85	26.3 ± 0.0.54	2.55
3b	21.8 ± 1.01	27.9 ± 0.65[a,b,d]	24.2 ± 1.31	26.1 ± 1.42	2.52

[a] Significant difference between Cytosine and other nucleotides within a strain, $p < 0.01$.
[b] Significant differences between Cytosine content of genotype 1 and 3, $p < 0.01$.
[c] Insignificant differences between Cytosine content within genotype 1, $p > 0.01$.
[d] Insignificant differences between Cytosine content within genotype 3, $p > 0.01$.

(Table 2), it was observed that there was a preference for those codons that had a C in the 3rd position (eg. out of two codons of Cys, TGT and TGC, all genotypes had a marked preference for TGC). Similar preference was observed for Phe, Asn, Pro, Thr and Tyr. Exceptions to this were seen in Gln, Ser and Asp where genotype 1 preferred a C-rich codon, but genotype 3 didn't. This explains the lower C preference in genotype 3 compared to 1 (Table 2). Amino acids with four or more codons were observed to preferably use those with either a Guanosine or a Cytosine in the 3rd position of nucleotide with a remarkable avoidance for those with an Adenine or a Thymine. Such a pattern was seen in the codon usage of Gly and Leu. Arg and Val codons ending with Guanosine were preferred over those ending with cytosine in all genotypes. Few amino acids, Ala, Asp, Glu, His, Ile, Lys, Gln and Ser, deviated from any clear preference. Met (ATG) and Trp (TGG) had only one codon. Our results also showed that most GC rich codon among all amino acids was Gly (GGC), Pro (CCC), Arg (CGG) and Val (GTG). As a whole, this study clearly showed, the most preferred terminal nucleotides in codon usage of E2 for all the amino acids were G and C whereas the least preferred terminal nucleotides were T and A. While genotype-specific differences became apparent, few, such as Val (GTG), Cys (TGC), Pro (CCC), Asn (AAC), Tyr (TAC), had similar frequencies in both 1 and 3 genotypes.

In our analysis so far, it was evident that while the basic theme of nucleotide usage remain the same, there were strain specific differences. Hence we decided to extend our analysis to the amino acid level as nucleotide differences don't always translate to functional differences. In this analysis, the amino acid sequence was paired between strains 1a, 1b, 3a and 3b. As it was tedious to compare all 29 sequences with which we were working so far, we decided to take a representative strain, randomly chosen, from each genotype. Using *t*-test we analysed whether there were any significant differences at the nucleotide content between the chosen strain and all other strains that were being analysed so far. t-test results showed that all differences were insignificant ($p > 0.05$) except for T and G content with the strain used of genotype 1b (CH/BID-V294/2002, Accession No. EU155366.2). However BLAST alignment revealed that all strains shared 93% identity with the chosen strain and hence we decided to choose CH/BID-V294/2002 for further studies as it has been shown earlier (Rost, 1999) that despite of sequence dissimilarity by about 30%, structural conservation is still maintained. Hence for structural analysis, random choice of strain was justified. For further analysis, the four strains that were used are detailed in supplementary table S10.

3.4. Genotype 1 is more conserved than genotype 3 at the individual amino acid level

Simple BLAST analysis of E2 amino acids for strains 1a, 1b, 3a and 3b yield 81% identity between 1a and 1b; 71% between 1a and 3b; 80%, 3a and 3b and between 3b and 1b, 70%. These results suggested the need for a more detailed comprehensive study to determine the exact nature of differences in E2 between each of the genotypes. Individual amino acid comparison of the above mentioned strains is detailed in supplementary Table S11. In this analysis, the amino acid sequences for all strains were paired. Individually any change in amino acid between strains and genotypes, were given a colour code as indicated in Fig. 1. In this way categorization was made which is summarized in Fig. 1B and C. As observed in Fig. 1A, it was evident that E2 pairing between genotypes 1 and 3 could be broken into three portions. The N-terminal region of the protein, exhibited comparatively little variation between genotypes 1 and 3 (primarily red-13%, Fig. 1A and B), the middle portion of the protein had conserved amino acids within a genotype but differed between genotypes (primarily green-19%, Fig. 1A and B) while the C-terminal region of the protein had conserved amino acids in genotype 1 only while the amino acids differed between strains of genotype 3 (primarily blue- 48%, Fig. 1A and B). Most profound changes (48%) were those amino acids that were conserved in genotype 1 but differed between strains in genotype 3 (indicated as blue, Fig. 1A and B). This indicated a greater variation in genotype 3. There were only a few that were conserved in genotype 3 but differed in genotype 1 (peach- 10%, Fig. 1A and B).

This marked distinctive partitioning of the protein called for a more comprehensive study to understand how the variations of amino acids affect the biology of the protein, especially in genotype 3 compared to genotype 1. Such differences between genotypes were also apparent when individual strain comparison was done, (Supplementary Fig. S2, Fig1). Biologically significant ones were scored which included 29% change from a polar to a non-polar amino acid and vice-versa (indicated as blue, Fig. 1A and C). Few (10%) changes were from a polar-aromatic to a non-polar aliphatic amino acid (indicated as yellow, Fig. 1A and C) and 8% change from an aromatic to an aliphatic amino acid (indicated as grey, Fig. 1A and C). These changes spanned almost the entire length of the protein but a region spanning from amino acid 406 to 441 remained conspicuously conserved among all strains. The biological significance of such conservation however, still remains to be determined although it has been reported that residues 412–423 represent the most conserved antigenic site (Barone et al., 2016). Between genotypes 1 and 3 however, 29% changes were seen to be from a polar to a non-polar amino acid or vice versa. Only a few were seen to be acidic-basic changes. Purple indicated either unchanged amino acids or those changes that are not typically considered to be biologically significant. The genotype 1 strains remained primarily purple (93%) indicating few biologically significant changes. Most of the changes were concentrated in the C-terminal half of the protein rather than the N-terminus.

Corresponding with amino acid changes, we analysed the position of nucleotide alterations that differed between strains and genotypes for each codon (Supplementary table S12). This analysis revealed that maximum changes (12% in genotype 1 and 7% in genotype 3) were tolerated in the 3rd nucleotide position. While genotype 1 allowed maximum changes in the 3rd nucleotide position, it tolerated less changes in other positions compared to genotype 3. This was in agreement with our amino acid analysis where genotype 1 maintained conserved amino acid within its genotype which was less in genotype 3.

3.5. Computational modelling of glycoprotein E2 and docking with CD81

In our analysis so far, it was evident that between genotypes and

Table 2
Nucleotide content at different codon positions in HCV E2.

	Percentage occurrence of nucleotide at indicated positions in codons												
		1st position (%)				2nd position (%)				3rd position (%)			
Genotype	Number of sequences analysed	T	C	A	G	T	C	A	G	T	C	A	G
1a Average	10	22.8 ± 1.48	22.2 ± 0.46*	26.1 ± 1.48	28.8 ± 0.2	24.4 ± 1.42	25.3 ± 0.43*	23.2 ± 1.06	26.9 ± 0.71	18 ± 0.88	44.1 ± 0.94[a,b]	11.1 ± 0.63	26.7 ± 0.82
SD		2.88				1.65				12.55			
1b Average	7	22.8 ± 0.90	22.6 ± 0.72*	26.6 ± 0.85	27.9 ± 0.49	24 ± 0.00	27.1 ± 0.67 [a]	24 ± 0.33	24.7 ± 0.47	18 ± 1.62	40.3 ± 2.31 [a, b]	10.2 ± 1.31	31.9 ± 0.91
SD		2.44				1.38				12.11			
3a Average	8	23.1 ± 1.25	27 ± 1.05*, b	21.8 ± 0.68	28.2 ± 0.84	23.12 ± 1.24	27 ± 1.06*	21.8 ± 0.68	21.8 ± 0.68	27 ± 0.93	29.8 ± 0.92 [a, b]	19 ± 0.98	24.2 ± 0.83
SD		2.84				2.84				4.14			
3b Average	4	23.8 ± 0.96	22.4 ± 0.51*	25.4 ± 1.07	28.3 ± 1.66	23.8 ± 0.96	22.4 ± 0.5*.b	25.4 ± 1.06	28.3 ± 1.66	20.2 ± 3.0	36.3 ± 1.98 [a, b]	17 ± 1.25	26.7 ± 2.0
SD		2.51				2.5				7.86			

[a] : Significant difference between Cytosine and other nucleotides within a strain, $p < 0.01$.
[b] Significant differences between Cytosine content of other genotypes, $p < 0.01$.
* Not significant against all other nucleotides within a genotype.

even within genotype 3, large variations are present at the amino acid level. Such variation might consequently lead to alterations in the structure and modify their function. Due to unavailability of complete structures for E2, we decided to use the tools of bioinformatics to model E2 from each strain. For this work we used the four representative strains from each genotype which had been subjected to amino acid analysis earlier. To analyse any functional changes in E2, we decided to look at the interaction between E2 and the receptor CD81, as this is one of the major interactions between the virus and host for entry via clathrin-mediated endocytosis (Blanchard et al., 2006).

Initially, all the proteins were characterized using Expasy-ProtParam and XtalPred servers and found to be stable and hydrophilic along with two exceptions, E2 of 3a was recognised as unstable and the receptor protein CD81 was indicated to be a hydrophobic protein. This was expected given CD81's sub-cellular localization on the cell membrane. All disorder prediction servers unanimously identified the N-terminal part of the envelope glycoprotein E2 as 'disordered region' along with random coil regions at residues 660–670 of HCV 3a and 3b. Both terminal regions of the receptor protein CD81 also appeared to be regions of high flexibility. Modelling programs like Phyre^2 and I-TASSER produced complete, globular models for envelope glycoproteins, whereas CD81 was modelled successfully by three servers (including Raptor X which however failed to produce a complete model of E2). Each model was further subjected to energy minimisation and the best model was chosen. As detailed in the methods section, each model was further validated via numerous software such as RAMPAGE and WHAT_CHECK. The energy minimized models with better Ramachandran plot appearances and backbone conformations, were chosen as representative models of the glycoproteins and receptor as detailed in Supplementary Table S13. Phyre^2 produced the best model for 1a, 1b and 3b, whereas the optimum model of 3a was produced by I-TASSER server.

The HCV E2 models exhibited a globular compact core region flanked between the N-terminal disordered region, including HVR1 & HVR2, and C-terminal transmembrane region (Fig. 2). The core region of E2 consisted of a immunoglobulin β-sandwich fold (residues 497–571, 492–566, 493–567, 493–567 for HCV 1a, 1b, 3a, 3b respectively), similar to that reported by Kong et al., 2013 (Kong et al., 2013), along with two helices and random coil regions. Both terminal regions were characterized by multiple short helices and coil regions. While the first ten residues of amino terminal were detected as disordered region, the last twenty-five amino acids of carboxyl terminus were predicted to be anchored in membrane, by Phobius server, although confirmation from TMHMM was not obtained. In all models, despite presence of eight disulphide bonds, 53–60% of the residues were in random coil conformation, similar to that reported by several studies (Barone et al., 2016; Khan et al., 2014; Kong et al., 2013). Further, the modelled structures showed good threading with the known structures, demonstrating more resemblance with HCV E2 of genotype 1a (PDB ID 4MWF (Kong et al., 2013) (RMSD, root mean square deviation varied between 0.65 and 2.00 Å, for C-α atoms) than that of genotype 2a (PDB ID 4WEBKhan et al., 2014), RMSD ranged between 3.37 and 6.08 Å). Greater variation of the terminal and hypervariable regions and lower at the core region of E2 among the different genotypes of HCV (Okamoto et al., 1991), were also reflected in the corresponding structures, as illustrated from the differences in secondary structure assignment and the spatial arrangement of these regions in all four models (Fig. 2). To further compare our models with structural information and presence of antigenic sites as revealed by binding sites of monoclonal antibodies, comparisons were performed as detailed in Table 4. Two regions, spanning amino acids 412–423 and again 427–446 have been reported to be antigenic sites to generate neutralizing antibodies by several studies (Barone et al., 2016; Deng et al., 2014; Deng et al., 2013; Kong et al., 2012a; Kong et al., 2012a; Li et al., 2015; Meola et al., 2015; Pantua et al., 2013). It was observed that all residues known to be serving as antigenic sites were exposed in the E2

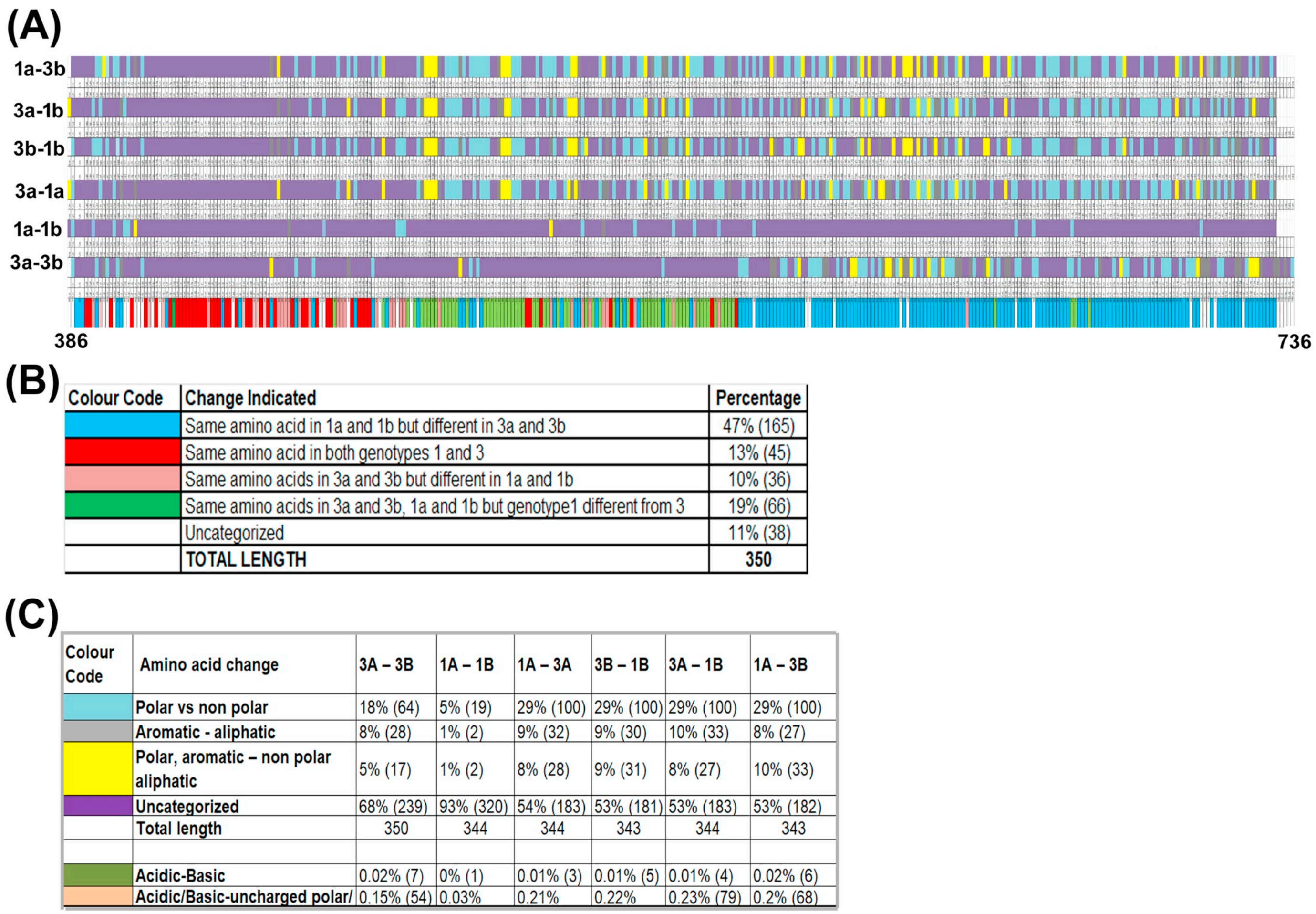
Table 3
Codon usage analysis of E2 across HCV Genotypes. Values in bold indicate the most abundant codon, those in red are abundant codons that are not GC rich.

		1a	1b	3a	3b
Amino acid	**Codon**	**Average ± SD**	**Average ± SD**	**Average ± SD**	**Average ± SD**
Ala	GCG	24.9±5.94	**34.9±9.87**	16.8±5.54	17.1±3.76
Ala	GCA	20.2±3.91	18.3±9.27	27.9±7.51	14.6±2.42
Ala	GCT	7.2±3.83	15.1±6.19	**34.8±3.04**	26.4±7.60
Ala	GCC	**47.7±7.07**	31.8±6.74	20.5±6.46	**41.9±6.39**
Cys	TGT	44.9±3.43	31.0±7.07	33.8±3.47	32.1±9.38
Cys	TGC	**55.1±3.43**	**69.0±7.07**	**66.2±3.47**	**67.9±9.38**
Asp	GAT	36.6±6.94	27.5±12.20	**51.7±8.44**	**57.5±6.38**
Asp	GAC	**63.4±6.94**	**72.5±12.20**	48.3±8.44	42.5±6.38
Glu	GAG	46.9±7.11	**96.0±5.03**	**86.9±7.76**	**70.5±2.88**
Glu	GAA	**53.1±7.11**	4.0±5.03	13.1±7.76	29.5±2.88
Phe	TTT	12.9±5.16	24.0±3.53	35.8±8.19	37.8±6.93
Phe	TTC	**87.1±5.16**	**76.0±3.53**	**64.2±8.19**	**62.2±6.93**
Gly	GGG	19.3±2.58	**43.0±1.62**	**31.1±5.21**	**39.0±6.55**
Gly	GGA	22.6±3.40	8.9±1.60	13.8±4.36	18.6±4.93
Gly	GGT	14.7±1.66	19.6±4.95	24.7±5.97	15.7±3.14
Gly	GGC	**43.4±1.33**	28.5±3.88	30.4±4.89	26.8±1.98
His	CAT	16.5±7.85	25.0±13.65	**50.9±11.44**	34.4±3.59
His	CAC	**83.5±7.85**	**75.0±13.65**	49.1±11.44	**65.6±3.59**
Ile	ATA	10.4±8.05	21.2±6.99	32.8±5.47	**47.3±3.45**
Ile	ATT	28.5±6.30	12.1±7.27	7.4±8.62	13.7±10.30
Ile	ATC	**61.1±12.67**	**66.7±6.52**	**59.8±7.76**	39.1±10.52
Lys	AAG	43.3±7.56	**68.1±13.13**	**60.3±5.60**	**72.5±10.02**
Lys	AAA	**56.7±7.56**	31.9±13.13	39.7±5.60	27.5±10.02
Leu	TTG	18.2±1.10	20.6±3.80	15.9±5.08	21.1±7.53
Leu	TTA	2.9±1.81	3.0±1.49	4.1±2.47	3.0±2.17
Leu	CTG	**29.0±4.60**	26.0±10.16	**31.2±7.37**	19.8±3.59
Leu	CTA	0.3±1.05	7.5±4.18	8.2±5.26	14.7±6.45
Leu	CTT	25.2±2.47	16.4±4.10	17.0±3.58	16.4±2.96
Leu	CTC	24.3±4.05	**26.4±6.39**	23.6±3.84	**24.9±2.64**
Met	ATG	100.0±0.00	100.0±0.00	100.0±0.00	100.0±0.00
Asn	AAT	27.3±5.15	25.6±5.85	36.8±8.31	30.1±3.87
Asn	AAC	**72.7±5.15**	**74.4±5.85**	**63.2±8.31**	**69.9±3.87**
Pro	CCG	22.6±2.08	27.9±4.11	8.9±3.92	14.0±2.10
Pro	CCA	15.8±2.63	15.1±4.32	20.1±1.77	29.4±3.70
Pro	CCT	26.2±3.32	24.0±4.69	32.4±5.82	14.9±3.23
Pro	CCC	**35.5±4.11**	**33.0±7.85**	**38.6±7.04**	**41.7±4.50**
Gln	CAG	**82.0±8.44**	**78.6±10.39**	41.8±7.58	38.1±20.14
Gln	CAA	18.0±8.44	21.4±10.39	**58.2±7.58**	**61.9±20.14**
Arg	AGG	**50.0±1.91**	**39.1±5.27**	**27.8±4.93**	**25.9±4.35**
Arg	AGA	4.4±2.35	9.5±4.07	16.9±6.54	16.3±7.24
Arg	CGG	12.4±3.92	13.4±3.78	23.4±3.69	13.8±6.21
Arg	CGA	5.6±0.20	11.1±4.49	4.7±3.59	11.5±8.68
Arg	CGT	2.9±5.00	9.6±4.31	15.2±4.16	7.2±3.90
Arg	CGC	24.5±6.42	17.3±6.46	12.0±6.05	25.3±3.31
Ser	AGT	10.9±5.33	6.5±5.62	**19.8±8.12**	10.3±6.90
Ser	AGC	**34.3±4.97**	**28.0±7.98**	15.9±4.51	16.8±10.23
Ser	TCG	4.1±1.77	17.1±6.53	11.1±3.81	11.9±6.12
Ser	TCA	13.2±1.90	11.0±5.51	18.3±4.72	**25.5±7.36**
Ser	TCT	12.3±2.80	7.5±2.63	17.5±4.60	14.7±3.37
Ser	TCC	25.2±5.45	**29.9±4.52**	17.5±3.49	20.8±7.79

(*continued on next page*)

Table 3 (*continued*)

Thr	ACG	22.2±2.92	21.2±6.26	15.6±2.90	23.7±7.32
Thr	ACA	11.4±3.45	18.0±3.18	26.3±4.81	22.1±6.97
Thr	ACT	12.2±3.10	21.5±4.17	24.8±6.53	20.4±7.95
Thr	ACC	**54.1±2.37**	**39.3±7.48**	**33.3±5.95**	**33.9±7.51**
Val	GTG	**46.8±3.32**	**48.6±4.43**	**38.9±3.92**	**49.8±6.05**
Val	GTA	8.0±2.87	9.9±3.84	7.6±3.62	12.6±4.53
Val	GTT	6.8±2.22	18.5±4.61	15.6±7.48	8.5±5.61
Val	GTC	38.4±1.67	22.9±5.17	37.9±8.23	29.2±6.00
Trp	TGG	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**	**100.0±0.00**
Tyr	TAT	38.3±16.12	24.1±10.90	34.9±7.36	36.1±19.42
Tyr	TAC	**61.7±16.12**	**75.9±10.90**	**65.1±7.36**	**63.9±19.42**



Colour Code	Change Indicated	Percentage
🟦	Same amino acid in 1a and 1b but different in 3a and 3b	47% (165)
🟥	Same amino acid in both genotypes 1 and 3	13% (45)
🟧	Same amino acids in 3a and 3b but different in 1a and 1b	10% (36)
🟩	Same amino acids in 3a and 3b, 1a and 1b but genotype1 different from 3	19% (66)
	Uncategorized	11% (38)
	TOTAL LENGTH	**350**

Colour Code	Amino acid change	3A – 3B	1A – 1B	1A – 3A	3B – 1B	3A – 1B	1A – 3B
🟦	Polar vs non polar	18% (64)	5% (19)	29% (100)	29% (100)	29% (100)	29% (100)
▨	Aromatic - aliphatic	8% (28)	1% (2)	9% (32)	9% (30)	10% (33)	8% (27)
🟨	Polar, aromatic – non polar aliphatic	5% (17)	1% (2)	8% (28)	9% (31)	8% (27)	10% (33)
🟪	Uncategorized	68% (239)	93% (320)	54% (183)	53% (181)	53% (183)	53% (182)
	Total length	350	344	344	343	344	343
🟩	Acidic-Basic	0.02% (7)	0% (1)	0.01% (3)	0.01% (5)	0.01% (4)	0.02% (6)
🟧	Acidic/Basic-uncharged polar/	0.15% (54)	0.03%	0.21%	0.22%	0.23% (79)	0.2% (68)

Fig. 1. Amino acid changes between genotypes 1a, 1b, 3a and 3b.
Amino acid sequences along with their codons were aligned for comparison as indicated (Supplementary Fig. S3). The individual amino acid changes were compared and colour coded as indicated. (A) Compressed pictorial representation of Table S3 to indicate the overall changes. (B) Colour codes along with percentages of changes scored between genotypes. (C) Percentage changes and colour codes of strain-wise amino acid changes.

model of genotype 1a, further validating our model. However when similar comparisons were made with the models generated from E2 of other genotypes, it was observed that while majority of the antigenic sites were exposed, genotype-specific differences were apparent (Table 4).

In spite of the modelling results, the available theoretical model of CD81 (PDB ID: 2AVZ) gave the best validation server result. As the putative interaction site of CD81 is already known, this sequence was used for docking studies with the four selected E2 models as detailed in materials and methods. From the numerous ranked poses and further analysis, the best poses were chosen and the predicted binding sites of the envelope glycoproteins with CD81 were determined which is summarized in supplementary table S14.

It was observed that the binding region of each E2 from the four strains with CD81 were unique. Although the binding region spanned from 420 to 700 in general, only few residues was common among the different strains. Despite 1a and 1b remaining fairly conserved upon amino acid alignment, only three amino acids (Pro484, Trp487, and

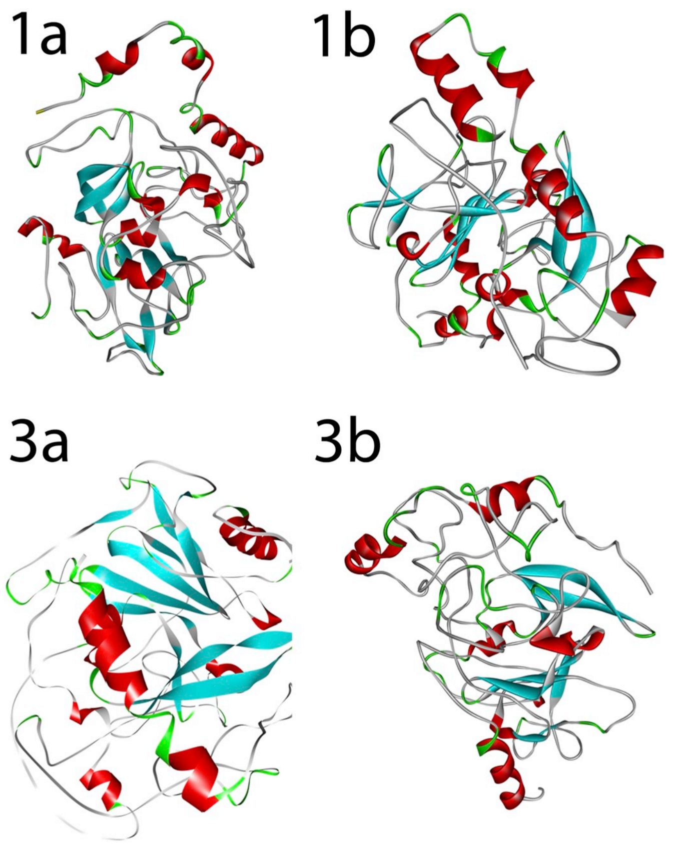


Fig. 2. Predicted models of E2 from genotypes 1a, 1b, 3a and 3b as indicated.

Table 4
Relative Solvent Accessibility of E2 residues of genotypes 1a, 1b, 3a and 3b known to be antigenic determinants. Amino acids are denoted by their standard single letter code.

Position	1a		1b		3a		3b	
412	Q	exposed	Q	exposed	Q	exposed	H	exposed
413	L	exposed	L	buried	L	exposed	L	buried
414	I	exposed	I	buried	V	buried	V	buried
415	N	exposed	N	buried	N	exposed	N	buried
416	T	exposed	T	buried	T	exposed	T	buried
417	N	exposed	N	exposed	N	exposed	N	exposed
418	G	exposed	G	buried	G	exposed	G	exposed
419	S	exposed	S	buried	S	buried	S	exposed
420	W	exposed	W	exposed	W	buried	W	exposed
421	H	exposed	H	exposed	H	exposed	H	exposed
422	I	exposed	I	exposed	I	exposed	I	exposed
423	N	exposed	N	exposed	N	exposed	N	exposed
427	L	exposed	L	exposed	L	exposed	L	exposed
428	N	exposed	N	exposed	N	exposed	S	exposed
429	C	exposed	C	exposed	C	exposed	C	exposed
430	N	exposed	N	exposed	N	exposed	N	exposed
431	E	exposed	D	exposed	E	exposed	D	exposed
432	S	exposed	S	exposed	S	exposed	S	exposed
433	L	exposed	L	exposed	I	exposed	L	exposed
434	N	exposed	Q	exposed	N	exposed	N	exposed
435	T	exposed	T	buried	T	exposed	T	exposed
436	G	exposed	G	buried	G	exposed	G	exposed
437	W	buried	F	buried	F	buried	F	buried
438	L	exposed	I	exposed	I	buried	I	exposed
439	A	exposed	A	buried	A	exposed	A	exposed
440	G	buried	A	exposed	G	buried	G	buried
441	L	exposed	L	exposed	L	buried	L	buried
442	F	exposed	F	exposed	F	exposed	I	exposed
443	Y	exposed	Y	exposed	Y	exposed	Y	exposed
444	Q	exposed	V	exposed	Y	exposed	Y	exposed
445	H	exposed	R	exposed	H	buried	H	buried
446	K	exposed	N	exposed	K	exposed	K	exposed

Tyr489) remained conserved between them. Genotype 3 strains did not have any common binding residues. With the putative interacting residues of E2, a further round of docking and analysis was performed with the entire model of CD81. Fig. 3 exhibits a representative image of the CD81- E2 interaction region from each of the genotypes. Inset is the representation of each of the individual complexes. The residues shown are restricted to the viewing angle and space available for labelling. In each interaction, the type of bond involved is depicted via specific colour codes.

The fresh docking analysis revealed a revised set of interacting residues of both CD81 and E2 (Table 5). This list included some of the earlier identified residues along with some new ones. Furthermore, we observed that even the residues of the host protein CD81 did not remain conserved while interacting with E2 from different strains. Residues such as Thr163 and Asp196 were the only residues that were common among all four strains while Asp117, Ser160, Ala164, Leu165, Thr167, Glu188, Asp189 and Asp195 remained conserved across three strains. E2 residues predicted to be participating in CD81 interaction was found to vary widely between genotypes. Apart from 1b, none of the residues were believed to be part of the HVR-1 although three genotypes, 1a, 1b and 3a revealed residues that were from HVR-2. None of the CD81 binding region was contiguous which was expected given the predicted structure of E2 and the predicted CD81 binding sites from other studies involving genotype 1a glycoprotein E2. The length of the CD81 region varied with 3a exhibiting the largest and 1b the shortest binding region. When genotype comparisons were made, only three residues (Arg483, Trp487 and Tyr489) were found to be common between genotypes 1a and 1b. There weren't any common residues between the E2 glycoproteins of genotypes 3a and 3b between each other or between the other genotypes.

3.6. *Interaction profile of CD81 – HCV E2 Complexes is unique for each genotype*

Binding affinity was determined between E2 of the four genotypes with the corresponding CD81 binding residues. It was observed that the receptor-glycoprotein interaction interface was mostly dominated by electrostatic interactions, numerous hydrogen bonds, hydrophobic interactions, and solvent accessible surface area, SASA > 2000 Å^2 for each complex with considerable binding free energy (Table 6, Fig. 3) - typical for stable non-obligatory complexes (Nooren and Thornton, 2003). E2 of HCV subtypes 3a and 3b illustrated distinctly higher binding affinity towards the receptor compared to others. The greater number of interactive residues with high interaction energy (IE) in the binding interface of subtypes 3a & 3b (Tables 5 and 6), larger SASA and higher number of non-bonded interactions (Tables 5 and 6) compared to other two complexes further explained the variation in binding affinity.

The receptor protein CD81 depicted four distinct interaction patches that contributed heavily in the interaction, namely: residues Asp117; Ser160, Thr163, Ala164, Leu165, Thr167; Glu188, Asp189 and Asp195, Asp196. All these residues were mostly involved in electrostatic interaction. Ala164 actively made both electrostatic and hydrophobic contacts. Thr163 appeared to be a hydrogen bond (HB) centre while Asp196 reported high IE (Table 6).

Electrostatic interaction held the major share (63–72%) of IE in all four complexes, while Van der Waals (VDW) contacts contributed most in E2-CD81 complex for genotype 3a. In 1a_E2-CD81 complex, receptor residues Asp155, Glu188, Asp189, Asp195, Asp196 and E2 residues Arg483, Arg587, Lys588 reported high electrostatic IE (> 50 kcal/mol); while amino acids Phe186 of CD81 & E2 residues Lys588, Ser686 came up as multiple non-bonded interaction centre (Tables 5 and 6, Fig. 3A). Residues Asp189, Asp195 of CD81 & Lys410 of 1b_E2 recorded markedly high electrostatic IE (> 50 kcal/mol) and amino acids Lys410, Arg483 and Tyr485 of 1b_E2 & Ser160, Ala164, Thr167 of CD81 acted as a hub for non-bonded interaction in 1b_E2-CD81 complex (Tables 5

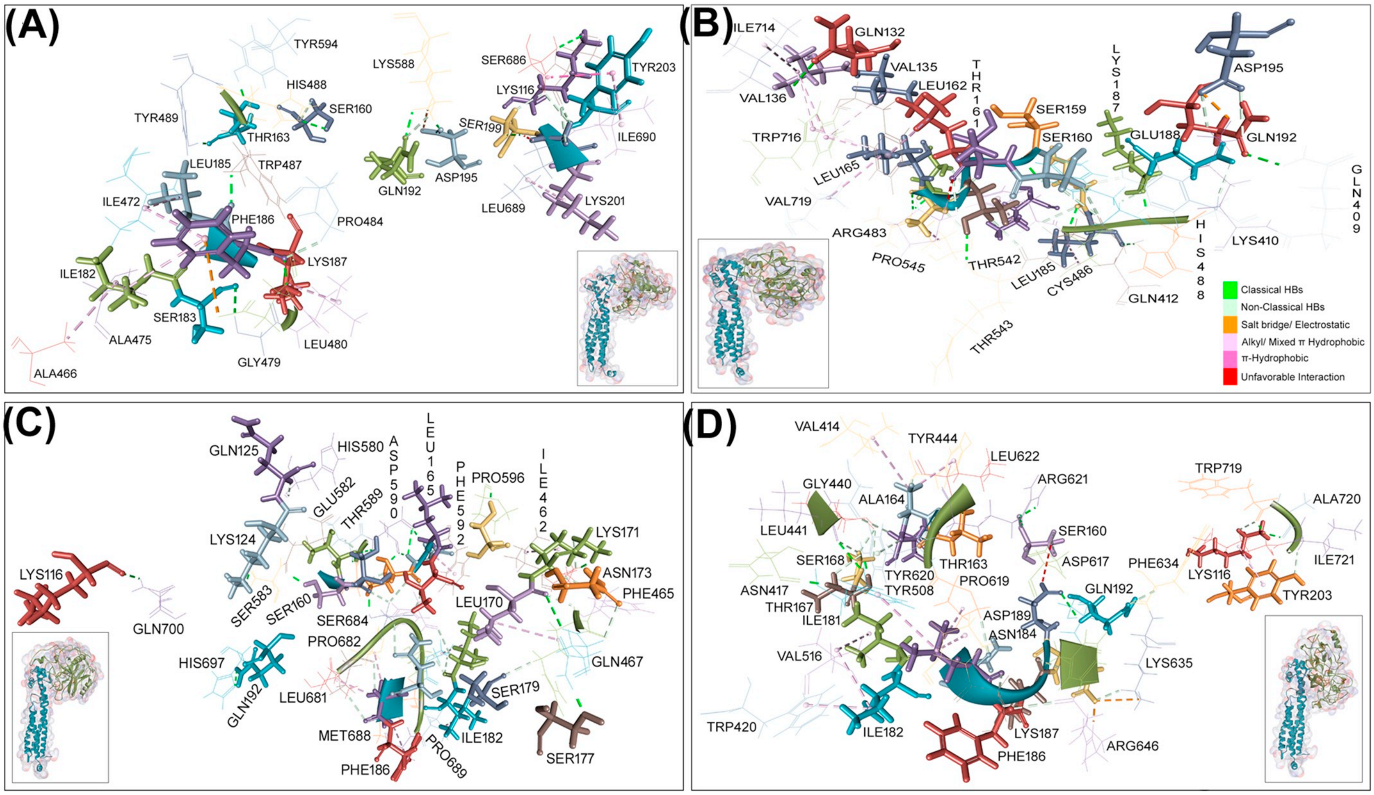


Fig. 3. CD81 - HCV E2 interaction profile.
CD81 interaction with HCV E2 protein of (A) genotype 1a, (B) 1b (C) 3a and (D) 3b. Only the amino acids that participated in non-bonded interaction are shown. Inset: Whole complex. CD81 is represented in cyan, E2 in green, receptor protein amino acids in stick and E2 in solid line. Non-bonded interactions are shown by broken lines, classical hydrogen bonds in green, ionized carbon & π-donor hydrogen bonds in sky blue, salt bridge & electrostatic contacts in orange and hydrophobic interactions in pink. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and 6, Fig. 3B). In 3a_E2-CD81 complex, IE is rather evenly distributed among the residues in binding interface as only Lys124 of CD81 portrayed high electrostatic IE (> 50 kcal/mol) but multiple amino acids such as Gln467, Asp590, Ser684 of E2 & Thr161, Thr163, Ala164, Lys171, Ile181 of CD81 emerged as non-bonded interaction site (Tables 5 and 6, Fig. 3C). The 3b_E2-CD81 complex was again populated with residues of high electrostatic IE (> 50 kcal/mol), viz. Asp117, Glu188, Asp195, Asp196 of CD81 & Arg621, Lys635, Arg646 of envelope glycoprotein 3b_E2 (Tables 5 and 6, Fig. 3D) and multiple non-bonded interaction contacts were observed in amino acids Ser160, Ala164, Thr167, Ser168, Glu188 Of CD81 & Tyr620 of E2.

4. Conclusion and discussion

In this study, we compared E2 glycoprotein sequences of 29 HCV strains belonging to genotypes 1a, 1b, 3a and 3b in an attempt to characterise and thereby predict how the variation among these strains could be affecting biology of the virus. Nucleotide content of each of the strains revealed a preference of the strains to be C rich which was more profound in genotype 1 compared to genotype 3. Codon usage analysis revealed a preference of C-rich codons in the genotype 1 strains, specifically those with a C in the 3rd position. Genotype 3 pattern was similar although the strength of preference was less. When comparison was made at the amino acid level, fewer changes (polar-non polar, aromatic-aliphatic, acidic-basic) were observed within the genotype 1 strains compared to genotype 3. Alignment also revealed that the variations were abundant at the C-terminus of the protein while an N-terminal region (from amino acid 415–427) was conspicuously conserved in all four strains. This region thus represents a conserved antigenic site for all genotypes. In silico modelling analysis of E2 revealed globular shaped models for each of the genotypes. Docking analysis of E2 models with CD81 revealed several conserved amino acids in the receptor across genotypes along with many that were unique to every genotype. The binding sequence prediction of E2 from all four strains were however distinct. Binding energy analysis revealed that the strength of interaction of genotype 3a was the highest among the four strains studied.

Previous reports analysing HCV whole genome has also revealed a bias to be GC rich (van Hemert et al., 2016) in accordance with our results. That particular study compared 29 animal RNA viruses of which one was HCV. We have confirmed these results using 34 sequences of HCV across four genotypes. We have reported a distinct GC bias in the entire genome as well as in the glycoproteins E1, E2 and RNA dependent RNA polymerase (RdRp). While the exact reason for a preference for G and C remains speculative, few reports suggest that GC richness helps in immune evasion as AU rich regions elicit strong innate immune response (Vabret et al., 2012). Codon usage of HCV has been reported to be in the early 50s in a study investigating many RNA viruses (Belalov and Lukashev, 2013). Unicellular organisms such as *E. coli* (Botzman and Margalit, 2011; Pan et al., 1998) also exhibit such bias as use of rare codons was hypothesized to slow down the translation rate to ensure optimum protein expression and folding (Gouy and Gautier, 1982; Thanaraj and Argos, 1996). Synonymous substitutions may result in an altered structure and thereby often helps in innate immune response evasion (Kondili et al., 2016; Simmonds et al., 2004). As RNA viruses create random mutations driven by the poor proof reading activity of most RdRp, translational pressure that stems from the availability of tRNAs in the host required for the viral protein translation becomes important. The liver however has been reported to be quite abundant in its content of tRNAs compared to other tissues such as the brain (Dittmar et al., 2006). Interestingly, tRNAs for hydrophobic amino acids are more pronounced in the liver (Dittmar et al., 2006) and

Table 5
Key E2-CD81 residues of Interaction.

1a - CD 81 Complex		1b-CD81 Complex		3a-CD81 Complex		3b-CD81 Complex	
CD81	**1a**	**CD81**	**1b**	**CD81**	**3a**	**CD81**	**3b**
ASP117	ARG455	ASP128	**GLN409**	ASP117	PRO461	**ASP117**	SER419
ASP155	*ILE472*	GLN132	**LYS410**	**LYS121**	*ILE462*	CYS157	LEU441
THR163	*LEU480*	**ASP137**	*ARG483*	**LYS124**	PHE465	SER160	TYR443
ILE182	**ARG483**	SER160	PRO484	GLN132	GLN467	THR163	SER513
LEU185	TRP487	THR161	*TYR485*	SER159	GLY468	*ALA164*	*VAL516*
PHE186	HIS488	THR163	TRP487	SER160	GLY574	LEU165	VAL616
GLU188	TYR489	*ALA164*	TYR489	**THR161**	**HIS580**	**THR167**	TYR618
ASP189	**ARG587**	*LEU165*	PRO545	LEU162	SER583	SER168	*PRO619*
GLN192	**LYS588**	**THR167**		THR163	PRO588	*ILE181*	TYR620
ASP195	TYR594	ASN184		*ALA164*	ASP590	ASN184	**ARG621**
ASP196	*SER686*	*LEU185*		LEU165	PRO596	PHE186	ILE633
TYR203	GLY688	**GLU188**		THR166	GLU597	**GLU188**	**LYS635**
		ASP189		THR167	PRO682	**ASP189**	**ARG646**
		ASP195		SER168	*CYS683*	**ASP195**	
		ASP196		***LYS171***	SER684	**ASP196**	
				ASN173	PHE685	*TYR203*	
				ILE181	THR686		
				ASP196	HIS699		
					GLN700		

Note: Residues with Interaction Energy (IE) ≥ 10 kcal/mol were reported here, residues with **IE ≥ 25** kcal/mol are in **Bold**, a. a. involved in HB underlined, residues participating in *hydrophobic contacts italicized*, residues common in all four complexes are shown in red and in three in blue.

Table 6
Binding affinity of HCV protein E2-CD81 Complexes.

CD81 Complex with	Non-bonded interactions				ΔSASA (Å^2)	Binding Free Energy (kcal/mol)
	E	SB	HB	H		
1a_E2	2	–	17	9	2247.0	−92.97
1b_E2	–	1	24	11	2135.2	−106.53
3a_E2	–	–	34	10	3521.1	−238.96
3b_E2	1	1	27	9	2776.3	−153.79

Note: E – Electrostatic, SB - Salt Bridge, HB - Hydrogen Bond, H – Hydrophobic.

almost all strains of HCV maintain a similar preference for hydrophobic amino acids (data not shown). Few codons which have been shown to be preferred in the liver (Dittmar et al., 2006) were also seen to be preferred by all genotypes of HCV. These included Arg, Gly, Leu, Ser, Thr and Val. Transfer RNAs (tRNA) with a C or G in their 1st nucleotide for anticodon, obviously undergoes a stronger codon-anti codon interaction resulting in a more efficient translation. In fact binding energy of mRNA with tRNA in HCV is stronger than that in the human host (Allnér and Nilsson, 2011). Furthermore, viral infection may also regulate the activity of RNA polymerase III, thereby affecting tRNA synthesis (Fradkin et al., 1987; Hoeffler and Roeder, 1985). However in spite of many advantages of GC richness, it was intriguing to note that genotype 3 strains had a less preference for C richness compared to genotype1. Furthermore as it is well established that genotype 3 is more pathogenic than genotype1 (Chayama and Hayes, 2011), the implication of such reduced preference remains to be investigated. This codon usage can also contribute to understanding the interaction between virus and the host immune system (Kondili et al., 2016).

Crystal structures from soluble E2 constructs (Khan et al., 2014; Kong et al., 2013) have revealed globular protein with no regular structure in spite of having eight disulphide bonds (Kong et al., 2013). Although 62% of the protein exists in loops or disordered structures, overall it was well-defined with a central β-sheet flanked by loops, short helices and β-sheets at the front and back. Our predicted models of all four strains conform to similar structural organisation as reported in these studies (Balasco et al., 2017; Khan et al., 2014; Kong et al., 2013). Our model is also supported by E2 models as proposed by computer algorithms using class II fusion protein folds supported by FTIR spectroscopy and circular dichroism (Krey et al., 2010). Furthermore, analysis of antigenic residues as predicted to be exposed by several studies (Barone et al., 2016; Deng et al., 2014; Kong et al., 2012b; Leopold Kong et al., 2012b; Li et al., 2015; Meola et al., 2015; Pantua et al., 2013) reporting neutralizing antibodies against the conserved region (412–423 and 427–446) revealed that our model generated from genotype 1a also has the same residues exposed. However it was interesting to note that while the majority of these residues were also exposed in the models from the other genotypes, strain specific differences remained. This explains the lack of pan-genotypic effect of most antibodies. It is also interesting to note that although this region remains mostly conserved across genotypes, subtle structural differences do occur.

The E2 binding sites for CD81 interaction have been extensively studied and have been neatly summarized in a review article by Feneant et al. (Fénéant et al., 2014). Researchers have used multiple tools to identify the E2 binding site which include the use of blocking monoclonal antibodies, E2 deletion mutants, as well as in silico modelling similar to what we have done (Callens et al., 2005; Drummer et al., 2006; Roccasecca et al., 2003; Rothwangl et al., 2008). Surprisingly all the papers report different E2 binding sites. This could be due to a predicted flexible conformation of E2 that varies during interactions with receptors (Owsianka et al., 2001). It has also been suggested that interactions and conformations may vary from cell type to cell type (Roccasecca et al., 2003). Our predicted sites have been supported by quite a few of these studies. Our prediction of the CD81 binding sites on E2 follows the same theme as suggested by Krey et al. (Krey et al., 2010) who predicted CD81 binding residues far apart in primary structure brought into close approximation in secondary structure. Similarly region spanning 407–524 suggested by Patel et al. using chimeras made from genotype 1a strains (McKeating et al., 2000) was also observed in our prediction. HVR2 residues reported to be part of the CD81 binding region using deletion mutants as well as blocking antibodies (McCaffrey et al., 2011; Roccasecca et al., 2003) support our prediction. Several residues identified from studies involving blocking antibodies such as mAb 6/41a and mAb 6/53 (Yagnik et al., 2000) have also been validated to be part of the CD81-E2 binding interface in genotype 1a and 1b. Interestingly these residues were not part of the binding interface for E2 from genotypes 3a and 3b, further highlighting strain specific differences. Although most of the published reports are based on studies on strain H77 of genotype 1a, few are based on the JFH-1 strain (genotype 2a). To the best of our knowledge, there are no reports on the CD81 binding regions in genotype 3. However several residues that we predict to be important in genotypes 3a and 3b have been reported in other studies involving genotype 1a (Drummer et al., 2006; Owsianka et al., 2001). From mutagenic studies on isolates from genotypes 1a and 1b, CD81 binding differs among strains (Roccasecca et al., 2003). While HVR1 region was speculated to inhibit CD81 interaction by steric hindrance, HVR2 (613–620) is believed to be important in CD81 interaction. While we have reported residues from HVR2, no residues of HVR1 are part of our predicted interaction region. Such was also reported in a study using HVR deletions where deletion of HVR1 had no effect whereas deletion of HVR2 abrogated CD81 interaction (McCaffrey et al., 2011). Importance of disulphide bridges between cysteines were shown to be important for E2 interaction (Petracca et al., 2000). However none of these cysteines showed up as residues in the interaction site possibly because cysteines perhaps contribute to the final structure of the CD81 but may not necessarily be part of the interacting residues.

Crystal structure of CD81-LEL have revealed the head sub-domain to be implicated in binding to HCV (Kitadokoro et al., 2001). In a study using African green monkey CD81, T163A, F186 L, E188K and D196E mutations disrupted HCV E2 binding (Higginbottom et al., 2000). In our docking studies, T163 and D196 were present in the interaction sites of all four genotypes whereas F186 was found in 1a, 3b and E188 in 1a, 1b and 3b binding sites. Interestingly, while our predictions are supported by studies using 1a, we report results of other genotypes which have not been revealed earlier.

Binding energy calculation revealed that E2 from genotype 3a had the highest binding energy compared to all other genotypes followed by 3b, 1b and 1a. Stronger binding energy is believed to result in tighter binding. This may be a reason of higher virulence observed in genotype 3 compared to genotype 1 (Chayama and Hayes, 2011; Shaw et al., 2003) although published literature differ regarding the strength of interaction between CD81 and E2 among genotypes. E2 constructs of HCV 3a generated from patient serum in Glasgow failed to bind CD81 (Shaw et al., 2003). However E2 isolated from other patient serum of genotype 3a exhibited similar binding to CD81 compared to genotypes 1a and 1b (Lavillette et al., 2005). There is however no data available on genotype 3b to the best of our knowledge. Hence our results need to be validated experimentally to conclusively prove the strength of interaction between E2 and CD81.

In conclusion, this study reports for the first time a comprehensive analyses on the changes at both nucleotide and amino acid levels of glycoprotein E2 across genotypes 1a, 1b, 3a and 3b. Our results have revealed wide variation in E2 from genotype 1 to 3 where the marked C preference, amino acid changes, codon preferences differ markedly across genotypes. This analysis brings to light a very important question. So far the bulk of our knowledge on HCV biology is centred on the studies conducted using the available reagents from genotype 1a (strain H77) and 2a (JFH-1). But when such variation can be observed between genotypes 1 and 3, its time to rethink and evaluate how the biology of the virus changes across genotypes. In fact such variations explain the different efficacies of available therapeutics and hence further research is needed to evaluate genotype specific virus biology.

Acknowledgements

The authors acknowledge funding from Department of Biotechnology-BioCARe project (BT/Bio-CARe/07/10139/2013-14) awarded to AM and DBT-BUILDER awarded to Presidency University. The authors would also like to acknowledge funding from Presidency University- Faculty Research Professional Development Fund. Thanks are also due to the Bioinformatics Infrastructure Facilities (BIF) at University of Kalyani & North-Eastern Hill University, Tura Campus, where structural analyses were performed.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

Human subjects nor animals were used in this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2019.01.008.

References

Allnér, O., Nilsson, L., 2011. Nucleotide modifications and tRNA anticodon-mRNA codon interactions on the ribosome. RNA 17, 2177–2188. https://doi.org/10.1261/rna.029231.111.

Alonso, S., Riveiro-Barciela, M., Fernandez, I., Rincon, D., Real, Y., Llerena, S., Gea, F., Olveira, A., Fernandez-Carrillo, C., Polo, B., Carrion, J.A., Gomez, A., Devesa, M.J., Baliellas, C., Castro, A., Ampuero, J., Granados, R., Pascasio, J.M., Rubin, A., Salmeron, J., Badia, E., Planas, J.M., Lens, S., Turnes, J., Montero, J.L., Buti, M., Esteban, R., Fernandez-Rodriguez, C.M., 2016. Effectiveness and safety of sofosbuvir-based regimens plus an NS5A inhibitor for patients with HCV genotype 3 infection and cirrhosis. Results of a multicenter real-life cohort. J. Viral Hepat. https://doi.org/10.1111/jvh.12648.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N., 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 38, W529–W533. https://doi.org/10.1093/nar/gkq399.

Balasco, N., Barone, D., Sandomenico, A., Ruggiero, A., Doti, N., Berisio, R., Ruvo, M., Vitagliano, L., 2017. Structural versatility of hepatitis C virus proteins: implications for the design of novel anti-HCV intervention strategies. Curr. Med. Chem. 24, 4081–4101. https://doi.org/10.2174/0929867324666170508105544.

Barone, D., Balasco, N., Autiero, I., Vitagliano, L., 2016. The dynamic properties of the hepatitis C Virus E2 envelope protein unraveled by molecular dynamics. J. Biomol. Struct. Dyn. 35, 1–12. https://doi.org/10.1080/07391102.2016.1162198.

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., Gursoy, A., 2014. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. Nucleic Acids Res. 42, W285–W289. https://doi.org/10.1093/nar/gku397.

Belalov, I.S., Lukashev, A.N., 2013. Causes and implications of codon usage bias in RNA viruses. PLoS ONE 8, e56642. https://doi.org/10.1371/journal.pone.0056642.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Blanchard, E., Belouzard, S., Goueslain, L., Wakita, T., Dubuisson, J., Wychowski, C., Rouillé, Y., 2006. Hepatitis C virus entry depends on clathrin-mediated endocytosis. J. Virol. 80, 6964–6972. https://doi.org/10.1128/JVI.00024-06.

Bochud, P.Y., Cai, T., Overbeck, K., Bochud, M., Dufour, J.F., Mullhaupt, B., Borovicka, J., Heim, M., Moradpour, D., Cerny, A., Malinverni, R., Francioli, P., Negro, F., 2009. Genotype 3 is associated with accelerated fibrosis progression in chronic hepatitis C. J. Hepatol. 51, 655–666. https://doi.org/10.1016/j.jhep.2009.05.016.

Boo, I., teWierik, K., Douam, F., Lavillette, D., Poumbourios, P., Drummer, H.E., 2012. Distinct roles in folding, CD81 receptor binding and viral entry for conserved histidine residues of hepatitis C virus glycoprotein E1 and E2. Biochem. J. 443, 85–94. https://doi.org/10.1042/BJ20110868.

Botzman, M., Margalit, H., 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biol. 12, R109. https://doi.org/10.1186/gb-2011-12-10-r109.

Bowie, J.U., Lüthy, R., Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.

Callens, N., Ciczora, Y., Bartosch, B., Vu-Dac, N., Cosset, F.-L., Pawlotsky, J.-M., Penin, F., Dubuisson, J., 2005. Basic residues in hypervariable region 1 of hepatitis C virus envelope glycoprotein e2 contribute to virus entry. J. Virol. 79, 15331–15341. https://doi.org/10.1128/JVI.79.24.15331-15341.2005.

Catanese, M.T., Dorner, M., 2015. Advances in experimental systems to study hepatitis C virus in vitro and in vivo. Virology 479–480, 221–233. https://doi.org/10.1016/j.virol.2015.03.014.

Chayama, K., Hayes, C.N., 2011. Hepatitis C virus: how genetic variability affects pathobiology of disease. J. Gastroenterol. Hepatol. 26, 83–95. https://doi.org/10.1111/j.1440-1746.2010.06550.x.

Chen, Y., 2013. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. Biomed. Res. Int. 2013, 1–10. https://doi.org/10.1155/2013/406342.

Ciesek, S., Westhaus, S., Wicht, M., Wappler, I., Henschen, S., Sarrazin, C., Hamdi, N., Abdelaziz, A.I., Strassburg, C.P., Wedemeyer, H., Manns, M.P., Pietschmann, T., von Hahn, T., 2011. Impact of intra- and interspecies variation of occludin on its function as coreceptor for authentic hepatitis C virus particles. J. Virol. 85, 7613–7621. https://doi.org/10.1128/JVI.00212-11.

Colovos, C., Yeates, T.O., 1993. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2, 1511–1519. https://doi.org/10.1002/pro.5560020916.

Cuypers, L., Li, G., Libin, P., Piampongsant, S., Vandamme, A.-M., Theys, K., 2015. Genetic Diversity and selective pressure in hepatitis C Virus Genotypes 1-6: significance for Direct-Acting Antiviral Treatment and drug resistance. Viruses 7, 5018–5039. https://doi.org/10.3390/v7092857.

Deng, L., Zhong, L., Struble, E., Duan, H., Ma, L., Harman, C., Yan, H., Virata-Theimer, M.L., Zhao, Z., Feinstone, S., Alter, H., Zhang, P., 2013. Structural evidence for a bifurcated mode of action in the antibody-mediated neutralization of hepatitis C virus. Proc. Natl. Acad. Sci. U. S. A. 110, 7418–7422. https://doi.org/10.1073/pnas.1305306110.

Deng, L., Ma, L., Virata-Theimer, M.L., Zhong, L., Yan, H., Zhao, Z., Struble, E., Feinstone, S., Alter, H., Zhang, P., 2014. Discrete conformations of epitope II on the hepatitis C virus E2 protein for antibody-mediated neutralization and nonneutralization. Proc. Natl. Acad. Sci. U. S. A. 111, 10690–10695. https://doi.org/10.1073/pnas.1411317111.

Dittmar, K.A., Goodenbour, J.M., Pan, T., 2006. Tissue-specific differences in human transfer RNA expression. PLoS Genet. 2, e221. https://doi.org/10.1371/journal.pgen.0020221.

Drummer, H.E., Boo, I., Maerz, A.L., Poumbourios, P., 2006. A conserved Gly436-Trp-Leu-Ala-Gly-Leu-Phe-Tyr motif in hepatitis C Virus glycoprotein E2 is a determinant of CD81 binding and viral entry. J. Virol. 80, 7844–7853. https://doi.org/10.1128/JVI.00029-06.

Evans, M.J., von Hahn, T., Tscherne, D.M., Syder, A.J., Panis, M., Wölk, B., Hatziioannou, T., McKeating, J.A., Bieniasz, P.D., Rice, C.M., 2007. Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry. Nature 446, 801. https://doi.org/10.1038/nature05654https://www.nature.com/articles/nature05654#supplementary-information.

Fénéant, L., Levy, S., Cocquerel, L., 2014. CD81 and hepatitis C virus (HCV) infection. Viruses 6, 535–572. https://doi.org/10.3390/v6020535.

Fields, B.N., Knipe, D.M., David, M., Howley, P.M., 2013. Fields virology. (Wolters Kluwer Health/Lippincott Williams & Wilkins).

Fradkin, L.G., Yoshinaga, S.K., Berk, A.J., Dasgupta, A., 1987. Inhibition of host cell RNA polymerase III-mediated transcription by poliovirus: inactivation of specific transcription factors. Mol. Cell. Biol. 7, 3880–3887.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein Identification and Analysis Tools on the ExPASy Server. In: The Proteomics Protocols Handbook. Humana Press, pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571. (Totowa, NJ).

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

van Hemert, F., van der Kuyl, A.C., Berkhout, B., 2016. Impact of the biased nucleotide composition of viral RNA genomes on RNA structure and codon usage. J. Gen. Virol. 97, 2608–2619. https://doi.org/10.1099/jgv.0.000579.

Higginbottom, A., Quinn, E.R., Kuo, C.C., Flint, M., Wilson, L.H., Bianchi, E., Nicosia, A., Monk, P.N., McKeating, J.A., Levy, S., 2000. Identification of amino acid residues in CD81 critical for interaction with hepatitis C virus envelope glycoprotein E2. J. Virol. 74, 3642–3649.

Hoeffler, W.K., Roeder, R.G., 1985. Enhancement of RNA polymerase III transcription by the E1A gene product of adenovirus. Cell 41, 955–963.

Hooft, R.W.W., Vriend, G., Sander, C., Abola, E.E., 1996. Errors in protein structures. Nature 381, 272. https://doi.org/10.1038/381272a0.

Jones, D.T., Cozzetto, D., 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31, 857–863. https://doi.org/10.1093/bioinformatics/btu744.

Käll, L., Krogh, A., Sonnhammer, E.L.L., 2007. Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Res. 35, W429–W432. https://doi.org/10.1093/nar/gkm256.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., Xu, J., 2012. Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. 7, 1511–1522. https://doi.org/10.1038/nprot.2012.085.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.E., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10, 845–858. https://doi.org/10.1038/nprot.2015.053.

Khan, A.G., Whidby, J., Miller, M.T., Scarborough, H., Zatorski, A.V., Cygan, A., Price, A.A., Yost, S.A., Bohannon, C.D., Jacob, J., Grakoui, A., Marcotrigiano, J., 2014. Structure of the core ectodomain of the hepatitis C virus envelope glycoprotein 2. Nature 509, 381–384. https://doi.org/10.1038/nature13117.

Kitadokoro, K., Bordo, D., Galli, G., Petracca, R., Falugi, F., Abrignani, S., Grandi, G., Bolognesi, M., 2001. CD81 extracellular domain 3D structure: insight into the tetraspanin superfamily structural motifs. EMBO J. 20, 12–18. https://doi.org/10.1093/emboj/20.1.12.

Kondili, M., Roux, M., Vabret, N., Bailly-Bechet, M., 2016. Innate immune system activation by viral RNA: how to predict it? Virology 488, 169–178. https://doi.org/10.1016/J.VIROL.2015.11.007.

Kong, L., Giang, E., Nieusma, T., Robbins, J.B., Deller, M.C., Stanfield, R.L., Wilson, I.A., Law, M., 2012a. Structure of hepatitis C virus envelope glycoprotein E2 antigenic site 412 to 423 in complex with antibody AP33. J. Virol. 86, 13085–13088. https://doi.org/10.1128/JVI.01939-12.

Kong, L., Giang, E., Robbins, J.B., Stanfield, R.L., Burton, D.R., Wilson, I.A., Law, M., 2012b. Structural basis of hepatitis C virus neutralization by broadly neutralizing antibody HCV1. Proc. Natl. Acad. Sci. 109, 9499–9504. https://doi.org/10.1073/pnas.1202924109.

Kong, L., Giang, E., Nieusma, T., Kadam, R.U., Cogburn, K.E., Hua, Y., Dai, X., Stanfield, R.L., Burton, D.R., Ward, A.B., Wilson, I.A., Law, M., 2013. Hepatitis C virus E2 envelope glycoprotein core structure. Science 342, 1090–1094. https://doi.org/10.1126/science.1243876.

Kozlowski, L.P., Bujnicki, J.M., 2012. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinforma. 13, 111. https://doi.org/10.1186/1471-2105-13-111.

Krey, T., D'Alayer, J., Kikuti, C.M., Saulnier, A., Damier-Piolle, L., Petitpas, I., Johansson, D.X., Tawar, R.G., Baron, B., Robert, B., England, P., Persson, M.A.A., Martin, A., Rey, F.A., 2010. The disulfide bonds in glycoprotein E2 of hepatitis C virus reveal the tertiary organization of the molecule. PLoS Pathog. 6, e1000762. https://doi.org/10.1371/journal.ppat.1000762.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger Datasets. Mol. Biol. Evol. 33, 1870–1874. https://doi.org/10.1093/molbev/msw054.

Lavillette, D., Tarr, A.W., Voisset, C., Donot, P., Bartosch, B., Bain, C., Patel, A.H., Dubuisson, J., Ball, J.K., Cosset, F.L., 2005. Characterization of host-range and cell entry properties of the major genotypes and subtypes of hepatitis C virus. Hepatology 41, 265–274. https://doi.org/10.1002/hep.20542.

Li, L., Chen, R., Weng, Z., 2003. RDOCK: refinement of rigid-body protein docking predictions. Proteins 53, 693–707. https://doi.org/10.1002/prot.10460.

Li, Y., Pierce, B.G., Wang, Q., Keck, Z.-Y., Fuerst, T.R., Foung, S.K.H., Mariuzza, R.A., 2015. Structural Basis for Penetration of the Glycan Shield of Hepatitis C Virus E2 Glycoprotein by a Broadly Neutralizing Human Antibody. J. Biol. Chem. 290, 10117–10125. https://doi.org/10.1074/jbc.M115.643528.

Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C., 2003. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. Proteins Struct. Funct. Bioinforma. 50, 437–450. https://doi.org/10.1002/prot.10286.

Lüthy, R., Bowie, J.U., Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. Nature 356, 83–85. https://doi.org/10.1038/356083a0.

McCaffrey, K., Gouklani, H., Boo, I., Poumbourios, P., Drummer, H.E., 2011. The variable regions of hepatitis C virus glycoprotein E2 have an essential structural role in glycoprotein assembly and virion infectivity. J. Gen. Virol. 92, 112–121. https://doi.org/10.1099/vir.0.026385-0.

McKeating, J.A., Patel, A.H., Dubuisson, J., Penin, F., Wood, J., 2000. Construction and characterization of chimeric hepatitis C virus E2 glycoproteins: analysis of regions critical for glycoprotein aggregation and CD81 binding. J. Gen. Virol. 81, 2873–2883. https://doi.org/10.1099/0022-1317-81-12-2873.

Meola, A., Tarr, A.W., England, P., Meredith, L.W., McClure, C.P., Foung, S.K.H., McKeating, J.A., Ball, J.K., Rey, F.A., Krey, T., 2015. Structural flexibility of a conserved antigenic region in hepatitis C virus glycoprotein E2 recognized by broadly neutralizing antibodies. J. Virol. 89, 2170–2181. https://doi.org/10.1128/JVI.02190-14.

Messina, J.P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G.S., Pybus, O.G., Barnes, E., 2015. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology 61. https://doi.org/10.1002/hep.27259.

Moradpour, D., Penin, F., 2013. Hepatitis C Virus Proteins: From Structure to Function. pp. 113–142. https://doi.org/10.1007/978-3-642-27340-7_5.

Mukherjee, S., Zhang, Y., 2011. Protein-Protein complex Structure predictions by Multimeric Threading and Template Recombination. Structure 19, 955–966. https://doi.org/10.1016/j.str.2011.04.006.

Nooren, I.M.A., Thornton, J.M., 2003. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. EMBO J. 22, 3486–3492. https://doi.org/10.1093/emboj/cdg359.

Okamoto, H., Okada, S., Sugiyama, Y., Kurai, K., Iizuka, H., Machida, A., Miyakawa, Y., Mayumi, M., 1991. Nucleotide sequence of the genomic RNA of hepatitis C virus isolated from a human carrier: comparison with reported isolates for conserved and divergent regions. J. Gen. Virol. 72, 2697–2704. https://doi.org/10.1099/0022-1317-72-11-2697.

Okamoto, H., Kurai, K., Okada, S.-I., Yamamoto, K., Lizuka, H., Tanaka, T., Fukuda, S., Tsuda, F., Mishiro, S., 1992. Full-length sequence of a hepatitis C virus genome having poor homology to reported isolates: Comparative study of four distinct genotypes. Virology 188, 331–341. https://doi.org/10.1016/0042-6822(92)90762-E.

Owsianka, A., Clayton, R.F., Loomis-Price, L.D., McKeating, J.A., Patel, A.H., 2001. Functional analysis of hepatitis C virus E2 glycoproteins and virus-like particles reveals structural dissimilarities between different forms of E2. J. Gen. Virol. 82, 1877–1883. https://doi.org/10.1099/0022-1317-82-8-1877.

Pan, A., Dutta, C., Das, J., 1998. Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: Translational selection versus mutational bias. Gene. https://doi.org/10.1016/S0378-1119(98)00257-1.

Pantua, H., Diao, J., Ultsch, M., Hazen, M., Mathieu, M., McCutcheon, K., Takeda, K., Date, S., Cheung, T.K., Phung, Q., Hass, P., Arnott, D., Hongo, J.-A., Matthews, D.J., Brown, A., Patel, A.H., Kelley, R.F., Eigenbrot, C., Kapadia, S.B., 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. https://doi.org/10.1016/j.jmb.2013.02.025.

Petracca, R., Falugi, F., Galli, G., Norais, N., Rosa, D., Campagnoli, S., Burgio, V., Di Stasio, E., Giardina, B., Houghton, M., Abrignani, S., Grandi, G., 2000. Structure-function analysis of hepatitis C virus envelope-CD81 binding. J. Virol. 74, 4824–4830.

Pierce, B., Weng, Z., 2007. ZRANK: Reranking protein docking predictions with an optimized energy function. Proteins Struct. Funct. Bioinforma. 67, 1078–1086. https://doi.org/10.1002/prot.21373.

Pierce, B., Weng, Z., 2008. A combination of rescoring and refinement significantly improves protein docking performance. Proteins 72, 270–279. https://doi.org/10.1002/prot.21920.

Pierce, B.G., Hourai, Y., Weng, Z., 2011. Accelerating Protein Docking in ZDOCK using an Advanced 3D Convolution Library. PLoS ONE 6, e24657. https://doi.org/10.1371/journal.pone.0024657.

Pileri, P., Uematsu, Y., Campagnoli, S., Galli, G., Falugi, F., Petracca, R., Weiner, A.J., Houghton, M., Rosa, D., Grandi, G., Abrignani, S., 1998. Binding of hepatitis C virus to CD81. Science 282 (80), 938–941.

Pontius, J., Richelle, J., Wodak, S.J., 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol. 264, 121–136. https://doi.org/10.1006/jmbi.1996.0628.

Qin, S., Zhou, H.-X., 2007. Meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics 23, 3386–3387. https://doi.org/10.1093/bioinformatics/btm434.

Rajesh, S., Sridhar, P., Tews, B.A., Fénéant, L., Cocquerel, L., Ward, D.G., Berditchevski, F., Overduin, M., 2012. Structural basis of ligand interactions of the large extracellular domain of tetraspanin CD81. J. Virol. 86, 9606–9616. https://doi.org/10.1128/JVI.00559-12.

Roccasecca, R., Ansuini, H., Vitelli, A., Meola, A., Scarselli, E., Acali, S., Pezzanera, M., Ercole, B.B., McKeating, J., Yagnik, A., Lahm, A., Tramontano, A., Cortese, R., Nicosia, A., 2003. Binding of the hepatitis C virus E2 glycoprotein to CD81 is strain specific and is modulated by a complex interplay between hypervariable regions 1 and 2. J. Virol. 77, 1856–1867.

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng. 12, 85–94.

Rost, B., Sander, C., 1994. Conservation and prediction of solvent accessibility in protein families. Proteins Struct. Funct. Genet. 20, 216–226. https://doi.org/10.1002/prot.340200303.

Rothwangl, K.B., Manicassamy, B., Uprichard, S.L., Rong, L., 2008. Dissecting the role of putative CD81 binding regions of E2 in mediating HCV entry: putative CD81 binding region 1 is not involved in CD81 binding. Virol. J. 5, 46. https://doi.org/10.1186/1743-422X-5-46.

Scarselli, E., Ansuini, H., Cerino, R., Roccasecca, R.M., Acali, S., Filocamo, G., Traboni, C., Nicosia, A., Cortese, R., Vitelli, A., 2002. The human scavenger receptor class B type I is a novel candidate receptor for the hepatitis C virus. EMBO J. 21, 5017–5025.

Shaw, M.L., McLauchlan, J., Mills, P.R., Patel, A.H., McCruden, E.A.B., 2003. Characterisation of the differences between hepatitis C virus genotype 3 and 1 glycoproteins. J. Med. Virol. 70, 361–372. https://doi.org/10.1002/jmv.10404.

Simmonds, P., Tuplin, A., Evans, D.J., 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA 10, 1337–1351. https://doi.org/10.1261/rna.7640104.

Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I.A., Lesley, S.A., Godzik, A., 2007. XtalPred: a web server for prediction of protein crystallizability. Bioinformatics 23, 3403–3405. https://doi.org/10.1093/bioinformatics/btm477.

Sonnichsen, B., De Renzis, S., Nielsen, E., Rietdorf, J., Zerial, M., 2000. Distinct membrane domains on endosomes in the recycling pathway visualized by multicolor imaging of Rab4, Rab5, and Rab11. J. Cell Biol. 149, 901–914.

Thanaraj, T.A., Argos, P., 1996. Ribosome-mediated translational pause and protein domain organization. Protein Sci. https://doi.org/10.1002/pro.5560050814.

Timm, J., Roggendorf, M., 2007. Sequence diversity of hepatitis C virus: implications for immune control and therapy. World J. Gastroenterol. 13, 4808–4817.

Vabret, N., Bailly-Bechet, M., Najburg, V., Müller-Trutwin, M., Verrier, B., Tangy, F., 2012. The Biased Nucleotide Composition of HIV-1 Triggers Type I Interferon Response and Correlates with Subtype D increased Pathogenicity. PLoS ONE 7, e33502. https://doi.org/10.1371/journal.pone.0033502.

Yagnik, A.T., Lahm, A., Meola, A., Roccasecca, R.M., Ercole, B.B., Nicosia, A., Tramontano, A., 2000. A model for the hepatitis C virus envelope glycoprotein E2. Proteins 40, 355–366 (10.1002/1097-0134(20000815)40:3 < 355::AID-PROT20 > 3.0.CO;2-K pii).

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y., 2015. The I-TASSER Suite: protein structure and function prediction. Nat. Methods 12, 7–8. https://doi.org/10.1038/nmeth.3213.

Zein, N.N., 2000. Clinical significance of hepatitis C virus genotypes. Clin. Microbiol. Rev. 13, 223–235.