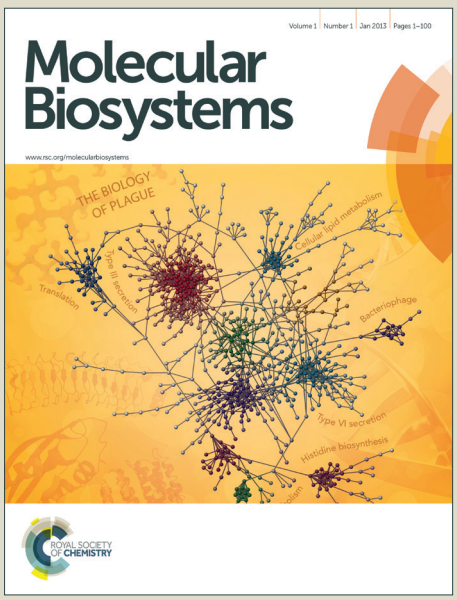
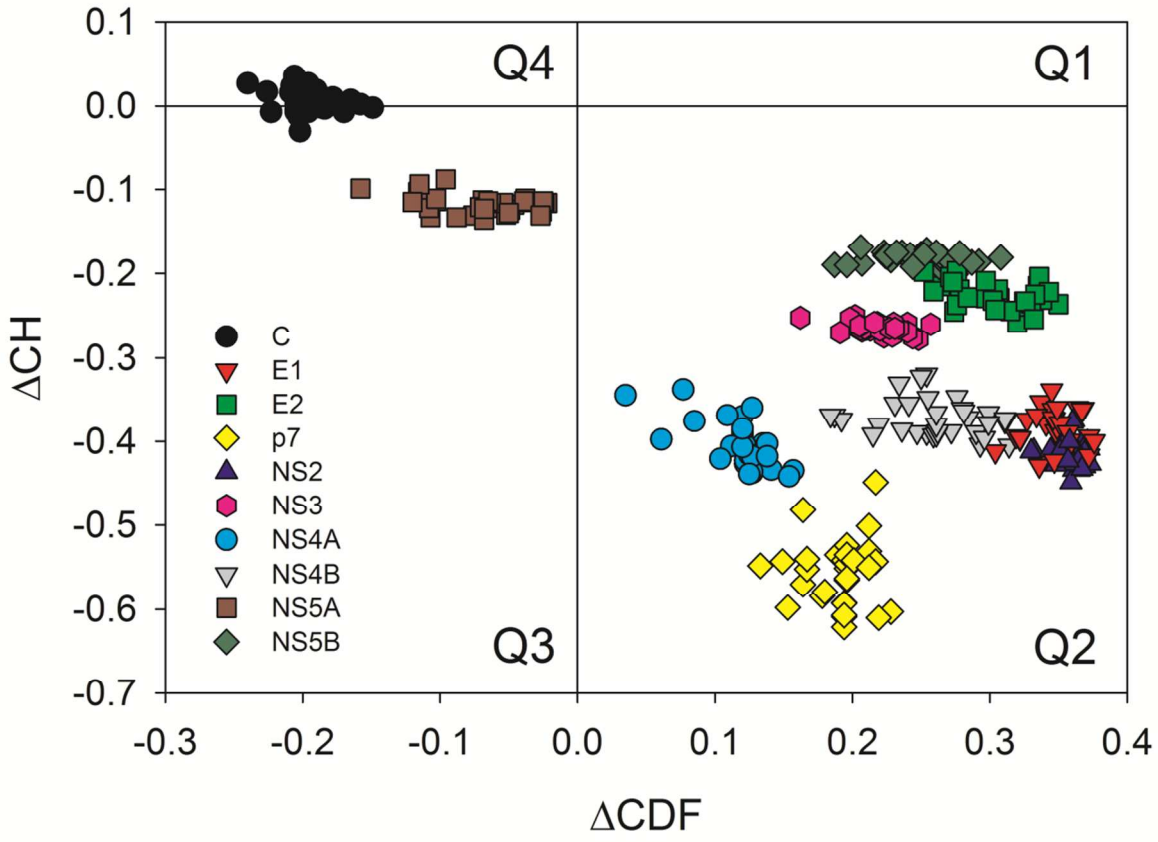
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



We showed that intrinsically disordered regions are common in the human hepatitis C virus proteins and possess numerous important functions.

The intrinsic disorder status of the human hepatitis C virus proteome

Xiao Fan,[1] Bin Xue,[2] Patrick T. Dolan,[3] Douglas J. LaCount,[3] Lukasz Kurgan, [1,]*

and Vladimir N. Uversky[4,5,6]*

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta AB T6G 2V4, Canada;

[2]Department of Cell Biology, Microbiology and Molecular Biology, College of Fine Arts and Sciences, University of South Florida, Tampa, Florida 33612 USA;

[3]Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, Heine Pharmacy Building, 575 Stadium Mall Drive, West Lafayette, Indiana 47906, USA;

[4]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, Florida 33612, USA;

[5]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia;

[6]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

AUTHOR INFORMATION

Corresponding Author

* To whom correspondence should be addressed: LK, Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 2V4, Canada; Phone: (780) 492-5488; Fax: (780) 492-1811; E-mail: lkurgan@ece.ualberta.ca; VNU, Department of Molecular Medicine, University of South Florida, 12901 Bruce B. Downs Blvd. MDC07, Tampa, Florida 33612, USA; Phone: 1-813-0748-5816; Fax: 1-813-974-7357; E-mail: vuversky@health.usf.edu

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript

Funding Sources

This work was supported in part by the Programs of the Russian Academy of Sciences for the "Molecular and Cellular Biology" (to V.N.U) and the Natural Sciences and Engineering Research Council (NSERC) Discovery grant (to L.K.).

ABSTRACT

Many viral proteins or their biologically important regions are disordered as a whole, or contain long disordered regions. These intrinsically disordered proteins/regions do not possess unique structures and possess functions that complement the functional repertoire of "normal" ordered proteins and domains, with many protein functional classes being heavily dependent on intrinsic disorder. Viruses commonly use these highly flexible regions to invade the host organisms, to hijack various host systems, and to help viruses in accommodation to their hostile habitats and to manage their economic usage of genetic material. In this article, we focus on the structural peculiarities of proteins from human hepatitis C virus (HCV) and use a wide spectrum of bioinformatics techniques to evaluate the abundance of intrinsic disorder in the completed proteomes of several human HCV genotypes, to analyze the peculiarities of disorder distribution within the individual HCV proteins, to establish potential roles of the structural disorder in functions of the ten HCV proteins. We show that intrinsic disorder or increased flexibility is not only abundant in these proteins, but is absolutely necessary for their functions, playing a crucial role in the proteolytic processing of the HCV polyprotein, the maturation of the individual HCV proteins, and being related to the posttranslational modifications of these proteins and their interactions with DNA, RNA, and various host proteins.

KEYWORDS: Intrinsically disordered protein; hepatitis C virus; proteome; protein structure; protein function; protein folding; partially folded conformation.

The *Flaviviridae* family includes small spherical enveloped viruses (~40-60 nm in diameter) with single-stranded positive-sense RNA genomes ranging in length from 9.6 to 12.3 kilobases. These viruses are predominantly spread by ticks and mosquitoes. Among four genera of this family are *Flavivirus* (contains about 70 human and animal viruses, such as yellow fever virus, West Nile virus, Dengue Fever), *Hepacivirus* [hepatitis C virus (HCV), GB virus B (GBV-B, also known as hepatitis G virus (HGV), canine hepacivirus, horse hepacivirus, bat hepacivirus, rodent hepacivirus], *Pegivirus* (GB virus A, GB virus C, and GB virus D), and *Pestivirus* (bovine viral diarrhea virus, classical swine fever, dog cholera), members of which are causative agents of various human diseases ranging from Dengue fever to yellow fever, to several forms of encephalitis, and to hepatitis C.

The arguably most famous member of the *Hepacivirus* genus is human HCV, which replicates mainly in the hepatocytes of the liver causing the development of chronic hepatitis C, liver cirrhosis, and hepatocellular carcinoma in humans.[1] Hepatitis C is a contagious liver disease, and HCV is transmitted most efficiently via the inadvertent exposure to infected blood, e.g., through needle sharing at intravenous drug use, healthcare exposure (blood transfusion, organ transplantation, contaminated medical instruments, accidental needle stick injury in health care workers), vertical transmission (passage from infected mother to unborn child), body modification (tattooing and piercing), sexual intercourse, sharing of personal items, etc. The predominant ways of HCV transmission in the developed and developing worlds are intravenous drug use and blood transfusion/unsafe medical procedures, respectively.[2] According to the World Health Organization (WHO, http://www.who.int/mediacentre/factsheets/fs164/en/), hepatitis C is found worldwide, chronically affecting 150-200 million individuals (~3% of world population), with some countries showing chronic HCV infection rates exceeding 5% (e.g., 22% of Egypt population is chronically infected with this virus).[3, 4] HCV infects 3-4 million people every year, with the hepatitis C-related

liver aliments being responsible for more than 350,000 annual deaths worldwide (http://www.who.int/mediacentre/factsheets/fs164/en/), and the annual mortality rate from hepatitis C exceeding 16,000 in USA (http://www.cdc.gov/hepatitis/Statistics/index.htm).

Genetically, HCV is grouped in six major genotypes, which are further subdivided into several subtypes, number of which depends on the genotype.[5] In the infected population, genotypes are distributed disproportionally, and the worldwide distribution and relative prevalence of the HCV genotypes varies from one geographic region to another. For example, in the United States, 70% of the HCV cases are caused by genotype 1, 20% by genotype 2, and ~1% by each of the genotypes 3, 4, 5, and 6.[6] Similarly, HCV genotype 1 is the most common cause of infection in South America and Europe.[1] On the contrary, the recent analysis of the HCV RNA positive patients in the district Swat of Khyber Pakhtoonkhaw (Pakistan) revealed that the most prevalent HCV genotype was 3a (34.1%), followed by 2a (8.1%), 3b (7%) and 1a (5.4%).[7] In North Africa and the Middle East, the most common genotype is HCV-4, whereas genotypes HCV-5 and HCV-6 are commonly found in South Africa and Hong Kong, respectively.[8, 9] Some non-common genotypes are very unique to specific geographical locations, e.g., genotypes 7, 8, and 9 have been reported only in Vietnamese,[10, 11] and genotypes 10 and 11 are widely distributed among Indonesian patients.[12] Curiously, the large disproportionality is also evident for the geographical distribution of the HCV subtypes. For example, the HCV subtypes 1a and 1b are the most common subgenotypes in the USA and Europe,[13-15] whereas the prevalent genotype affecting Japanese patients is 1b.[16] HCV subtypes 2a and 2b are mostly found in North America, Europe, and Japan whereas HCV-2c is exclusively found in northern Italy.[13, 14, 16, 17]

The genome of HCV consists of a single open reading frame that is 9,600 nucleotide bases long[18] that encodes for a single polypeptide, HCV genome polyprotein, containing about 3,000 amino acids.[19] This polyprotein is processed by cellular and viral proteases into the ten smaller proteins

necessary for viral replication within the host cell, or assemble into the mature viral particles.[19] Depending on their roles in the formation of viral particle, these ten proteins are classified as structural proteins, such as core or nucleocapsid protein C (p22), two envelope glycoproteins, E1 (or gp35) and E2 (or gp70), and a viral channel forming protein p7; and non-structural proteins, such as transmembrane protein NS2 (p23), protease/RNA helicase NS3 (p70), cofactor NS4A (p8), cofactor NS4b (p27), interferon resisting protein NS5A (p56/568), and RNA polymerase NS5B (p68).[19]

Viruses represent an interesting example of adaptation to extreme conditions, which include both environmental peculiarities and biological and genetic features of the hosts. They have to survive outside and within the host cell (some viruses are known to infect Archaea isolated from the geothermally heated hot environments[20]) and need to infect the host organism and replicate their genes while avoiding the host's countermeasures.[21] Furthermore, viral genomes are unusually compact and often contain overlapping reading frames. Due to the extremely abbreviated proteomes of viruses, which typically have just a minimal set of specific structural proteins crucial for the viral particle assembly and a set of non-structural proteins that are used by the virus to hijack many functional pathways of the host cell, many viral proteins are known to be multifunctional and usually need to perform numerous interactions with host cell components during the different steps of the virus life cycle from entry to replication to formation and exit of new infectious particles.

The mentioned multifunctionality and robustness can be at least in part explained by the fact that the viral proteins possess unique structural features and characteristics.[22] For example, viral proteins can greatly benefit from the functional and structural flexibility granted by partially folded or unfolded protein domains.[23] In agreement with this hypothesis, a comprehensive bioinformatics analysis revealed that in comparison with proteins from their hosts, viral proteins are less densely

packed, possess a much weaker network of inter-residue interactions (manifested by the lower contact density parameters, the increased fraction of residues not involved in secondary structure elements, and the abundance of short disordered regions), the unusually high occurrence of polar residues, and are characterized by the lower destabilizing effects of mutations.[22] Furthermore, it has been concluded that the adaptive forces that shape viral proteins were different from those responsible for evolution of proteins of their hosts. In fact, the abundance of polar residues, the lower van der Waals contact densities, high resistance to mutations, and the relatively high occurrence of flexible and intrinsically disordered regions suggested that viral proteins are not likely to have evolved for higher thermodynamic stability but rather to be more adaptive for fast change in their biological and physical environment.[22]

Recent years witnessed the rise of an idea that to be biologically active, some many proteins do not necessarily require a unique 3-D structure as a whole or in part.[24-32] These intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs) exist as dynamic conformational ensembles,[24, 26, 30, 33-36] ranging from collapsed (molten globule-like), to partially collapsed (pre-molten globule-like) and even highly extended (coil-like).[27, 29, 36, 37] IDPs/IDPRs are very common in all proteomes analyzed so far,[38] and are often involved in various human diseases.[39, 40] These proteins are typically involved in regulation, signaling, and control pathways,[26, 29, 31, 36, 41, 42] thereby complementing the traditional catalytic or transport functions of ordered proteins.[34, 43-47] Functionally important IDPs ranges from transcription factors in general,[48, 49] to the reprogramming transcription factors responsible for the conversion of terminally differentiated somatic cells to the induced pluripotent stem (iPS) in particular,[50] and from proteins of the extracellular matrix[51] to mitochondrial,[52] ribosomal,[53] and nuclear proteins.[54, 55]

Based on the recent analysis of the abundance of intrinsic disorder in various organisms it has been concluded that viruses have the largest variation range in the content of disordered residues

in their proteomes.[38] For example, in human coronavirus NL63 only as few as 7.3% of residues are predicted to be disordered, whereas this percentage reaches a value as high as 77.3% in the case of *Avian carcinoma virus* proteome. Some viral species are highly enriched in intrinsic disorder, and more than 20 small viruses with 5 or less proteins were shown to have 50% or higher disordered residues in their proteomes.[38] These small viruses have the highest fraction of intrinsic disorder among all species. When the proteome size increases, the fraction of disordered residues in the proteomes of various viruses seem to converge to a range between 20% and 40%. The conclusion on the variability of protein intrinsic disorder within and between viral families has been supported by a recent comprehensive analysis of 2,278 available viral genomes in 41 families.[56] This analysis revealed that the substantial variation in the level of disorder in viral proteins did not follow the established trend among their hosts, where the abundance of disordered proteins and proteins with disordered regions is predicted to increase from eubacteria, to archaebacteria, to protists, and to multicellular eukaryotes. In fact, large variation in the disorder level is seen even for virus infecting similar hosts, e.g. poxviruses and herpesviruses (which are large mammalian viruses) showed markedly differing disorder levels (5.6% and 17.9%, respectively).[56]

This high content of predicted intrinsic disorder in viruses is in agreement with another study which showed that, in comparison with Archaea and bacteria, viral and bacteriophagic proteins are significantly enriched in polar residues and depleted in hydrophobic residues.[57] The high levels of disorder in many proteins can be explained by the fact that protein intrinsic disorder and the functional advantages that it confers may be a way for viral proteins to fulfill their numerous functions based on numerous interactions with host membranes, host nucleic acids and host proteins. In fact, the lack of a rigid 3D structure enables IDPs/IDRs to be highly promiscuous and take part in several interactions with multiple partners. IDRs in particular can act as flexible linkers

between functional domains enabling mechanisms that will facilitate binding and promiscuity. These flexible linkers can also help viral proteins to elude the host cell's immune system thanks to interactions with host proteins that make viral epitopes poorly recognized by the innate immune system. Finally, the lack of structural constraints of IDRs can represent a way to resist the high mutation rates that are characteristic for viruses.[22]

In this study, we used a wide spectrum of bioinformatics techniques to evaluate the abundance of intrinsic disorder in the completed proteomes of several human HCV genotypes (such as 1a, 1b, 1c, 2a, 2b, 2c, 2k, 3a, 3b, 3k, 4a, 5a, 6a, 6b, 6d, 6g, 6h, and 6k), to analyze the peculiarities of disorder distribution within the individual HCV proteins, and to establish a potential conjunction between the structural disorder and functions of the ten HCV proteins.

Materials and Methods

Dataset

All high quality (i.e., complete and reviewed) HCV genotypes were collected from Uniprot release 2013_09.[58] We extracted 32 HCV polyproteins across the 18 genotypes with 8 isolates for genotype 1b, 2 isolates for genotypes 1a, 1c, 2a, 2b, 3a, 5a, 6a, and 1 isolate for genotypes 2c, 2k, 3b, 3k, 4a, 6b, 6d, 6g, 6g, and 6k (Uniprot IDs: P26664, P27958, P26663, Q9WMX2, Q03463, Q00269, Q913V3, P26662, P29846, O92972, Q81754, Q913D4, P26660, Q99IB8, P26661, Q9DHD6, Q68749, Q9QAX1, Q81495, Q81258, Q81487, Q68801, O39929, O39928, O91936, Q5I2N3, O39927, O92529, O92530, Q68798, O92532, O92531). Each of the polyproteins is approximately 3000 amino acids long and includes 10 protein chains, where the sizes of the individual proteins range between 54 and 631 residues; the total number of HCV proteins across all genotypes and isolates is 320. The sequence for genotypes 7, 8, 9, 10, and 11

were not included since they were not reviewed and some of them are not complete. Moreover, some researchers suggest that these genotypes are subtypes of genotypes 3 and 6.[59-61]

Annotation of the HCV proteins

We collected a rich set of annotations for each protein from each genotype and isolate including annotations of transmembrane regions, protein-, DNA- and RNA-binding regions, post-transcriptional modification (PTM) sites, amino acid polymorphisms (AAPs), and putative annotations of IDRs and molecular recognition features (MoRF) regions. MoRFs are short disordered regions that undergo disorder-to-order transition when binding to their protein partners. These regions were found to be involved in signaling and regulatory functions.[62-65]

The main source of annotations was Uniprot.[58] The transmembrane regions were derived using TRANSMEM tag, PTM sites from MOD_RES tag, and protein-, DNA- and RNA-binding regions from REGION tag. The protein-binding annotations include interactions with NS3, NS4B, CD81, APOA2, STAT1, and DDX3X and mitochondrial targeting signal region. The DNA-binding annotations encompass HVR1-, HVR2n, and FKBP8-binding and transcriptional activation region. The RNA-binding annotations involve interactions with PKR and PKR/eIF2-alpha phosphorylation homology domain (PePHD).

The AAPs were retrieved from multiple sequence alignments (MSA) with ClustalW[66] We aligned subtypes and isolates per each of the six major genotype (1, 2, 3, 4, 5, and 6), and across all 32 HCV polyproteins. A given residue is defined as polymorphic if it varies at a given position in the alignment between isolates/genotypes. We considered two levels of polymorphisms: strong where over half of the considered chains have a different amino acid type

at a given position in the alignment, and weak where at least one and no more than half of chains have a different amino acid type.

IDRs were predicted using three disorder prediction methods: PONDR® VLXT,[67] MFDp,[68] and PONDR-FIT.[69] PONDR® VLXT is not the most accurate predictor but has high sensitivity to local sequence peculiarities which are often associated with disorder-based interaction sites.[67] MFDp and PONDR-FIT are meta-predictors, which means that they combine predictions of several other disorder predictors to boost predictive performance, and these sets of predictors are different when compared between MFDp and PONDR-FIT. They also use different underlying prediction models, with neural network and support vector machine used in PONDR-FIT and MFDp, respectively. As a result, these two methods generate complementary predictions that are characterized by high predictive performance.[70] The putative IDRs were obtained by combining the disordered regions predicted by MFDp and PONDR-FIT. MoRFpred method, the leading predictor of MoRFs,[71] was used to annotate MoRF regions. Short IDRs and MoRF segments that have less than 4 and 5 residues, respectively, were removed. This is consistent with their respective definitions.[72]

Two binary disorder classifiers, charge-hydropathy (CH) plot[30, 73] and cumulative distribution function (CDF) plot,[73, 74] as well as their combination known as four quadrants CH-CDF plot,[74-76] were also used in this study.

The putative IDR regions were also annotated functionally following protocol defined in refs.[53, 77] The protocol is based on a local pairwise alignment against functionally annotated IDRs from DisProt version 6.02.[78] Briefly, we aligned each disordered segments extracted from HCV proteins across all genotypes and isolates into the set of 862 disordered segments collected from the DisProt database that have functional annotation. The alignment was computed using the

Smith-Waterman algorithm[79] using the EMBOSS implementation with default parameters (gap_open=10, gap_extend=0.5, and blosum62 matrix). Sequence identity was calculated as the number of identical residues in the local alignment divided by the length of the local alignment or the length of the shorter of the two being aligned segments, whichever is larger. A given functional annotation from the DisProt database was transferred if the identity was above 80%.

Results and Discussion

Intrinsic disorder and proteolytic processing of the HCV polyprotein

The translation of HCV genome consisting of a single open reading frame of about 9600 nucleotides depends on an internal ribosome entry site (IRES) within the 5′-noncoding region that binds 40S ribosomal subunits directly and avidly, bypassing the need for pre-initiation factors, and inducing an mRNA-bound conformation in the 40S subunit.[80] After the recruitment of other necessary components, translationally active 80S complex is formed, and the translation of HCV genome produces a single polypeptide, an HCV polyprotein (~3010 residues), that requires further proteolytic processing to generate 10 active viral proteins, that are classified either as structural or non-structural (see Figure 1). All the structural proteins are located within the N-terminal one-third of the polyprotein. Among these structural proteins are the highly basic core (C) protein, and glycoproteins E1 and E2. A small integral membrane protein, p7, potentially acting as an ion channel,[81, 82] is found next to the structural proteins. The remainder of the genome encodes the nonstructural (NS) proteins NS2, NS3, NS4A, NS4B, NS5A and NS5B, which play crucial roles in controlling, coordinating and regulating the various intracellular processes of the virus life cycle. The maturation of structural proteins is driven by the signal peptidase (sp) cleavages between

C/E1, E1/E2 and E2/p7. The protein C is cleaved from the polyprotein by a host sp into a precursor protein of 191 amino acids (p23) containing a hydrophobic C-terminal tail that is responsible for anchoring this protein to endoplasmic reticulum.[83] The next step in the C maturation is driven by the cleavage of p23 by human signal peptide peptidase (hspp) which separates the C-terminus C from the rest of the C protein and generates the mature form of C, p21.[84, 85] The host cell sp is also responsible for the cleavage of the p7/NS2 junction, whereas maturation of the non-structural proteins is determined by the activity of two viral enzymes, the NS2/3 autoprotease, which cleaves at the NS2/3 junction; and the NS3-4A serine protease, which cleaves at all downstream sites (see Figure 1) http://www.nature.com/nature/journal/v436/n7053/full/nature04077.html - f2.[86]

Figure 1A represents the proteome map of the human HCV-1b (UniProt ID: P26662) where a polyprotein (Figure 1A) and each viral protein (Figure 1C) is presented as a bar whose location corresponds to the location of the corresponding gene within the HCV genome. PONDR® VLXT disorder predictions for the HCV-1b polyprotein and each of the 10 viral proteins are shown as red solid lines inside the corresponding bars. In such representation, a residue is considered to be disordered if its score is above 0.5. In Figure 1C, the viral proteins are color coded according to their intrinsic disorder content evaluated by PONDR® VLXT, were two arbitrary cutoffs for the levels of intrinsic disorder were used to classify proteins as highly ordered (0–10% of the sequence is disordered, cyan rectangles), moderately disordered (11%–30% of the sequence is disordered, yellow rectangles) and highly disordered (31%–100% of the sequence is disordered, light red rectangles).[87] Figure 1C also shows that there is no HCV protein that would possess an average disorder score less than 0.1. In other words, if another type of disorder classification is used, where highly ordered, moderately disordered, and highly disordered proteins are defined as proteins with predicted mean intrinsic disorder (PID) score < 0.1, 0.1 < PID score < 0.3, and PID score > 0.3,

respectively, then HCV does not contain any highly ordered proteins, possessing 7 moderately disordered proteins and 3 highly disordered proteins.

Figure 2 shows the proteome map of the considered HCV polyproteins where data are summarized (based on the multiple sequence alignment) over all 32 polyproteins (dark red lines) and over the polyproteins that correspond to each of the six genotypes (see legend for color codes). Individual viral proteins are named at the top of the figure and the cleavage sites are denoted with vertical gray lines. The putative disorder annotations are shown in the bottom half of the Figure 2. The predictions are provided in two formats: the real-valued, in the [0, 1] range, disorder propensity profiles generated by MFDp and PONDR-FIT that are shown in the middle of the Figure; and binary profiles denoted as IDR that are given immediately below the real-values profiles where only the disordered residues (that combine predictions from MFDp and PONDR-FIT) are shown. Since proteolytic digestion is orders of magnitude faster in unstructured as compared to structured protein regions,[88-93] it is extremely important for the protein cleavage process that the sites of cleavage be located in regions that lack structure or possess high structural flexibility. Based on the real-valued disorder propensities generated by MFDp and PONDR-FIT analysis, disordered regions are regions with the scores above the dotted line, whereas the scores below the line indicate degree of flexibility. Figures 1B and 2 provide some important clues on the structural prerequisites for maturation of the viral proteins and show that the cleavage sites are predominantly located within or in close proximity to the regions with increased flexibility. This can be observed by analyzing the propensity scores of MFDp and/or PONDR-FIT in the vicinity of the vertical gray lines, which spike to relatively high values there. Some discrepancies in the positions of the cleavage sites in polyprotein shown in Figure 1B and Figure 2 are due to the fact that Figure 2 uses the “aligned” polyprotein, which shifts/renumbers positions compared with the single genotype analysis shown in Figure 1B.

Curiously, the majority of cleavage sites are located within polymorphic regions predicted to be disordered (as shown in Figures 1 and 2). On the other hand, some sequence conservation is obvious at these cleavage sites and in their vicinity. In fact, the eukaryotic cell signal peptidase, sp, is an integral membrane protein complex of the endoplasmic reticulum capable of the endoproteolytic cleavage of signal peptides from preproteins during co- and post-translational translocation across the endoplasmic reticulum bilayer.[94-96] In part, the enzymatic specificity of the eukaryotic cell sp is determined by the primary amino acid sequence of its substrates, with the amino acids found at positions -1 and -3 almost always containing small-neutral side chains (e.g., Ala, Gly, Ser, Thr, and Cys), and with position -3 being somewhat less restrictive including hydrophobic residues (e.g., Leu, Val, Ile). This observation constitutes the "(-3,-1) rule"[97-99] and the "A-X-BJ" model.[100] Hspp is a presenilin-related aspartic protease that catalyses intramembrane proteolysis of membrane protein signal sequences with type II orientation (N- to C-terminus from the cytosol to the endoplasmic reticulum lumen).[101] In HCV polyprotein, the cell hssp targets a hydrophobic sequence at the junction between the C protein and the envelope glycoprotein E1.[102] The minimal domain for activity of the NS2/3 autoprotease has been mapped to residues 907–1206 containing the C-terminal portion of NS2, immediately following the hydrophobic region, as well as the N-terminal protease domain of NS3.[103] Although the HCV NS2/3 cysteine protease shows no sequence motifs typical of known proteases, it cleaves at the conserved NS2/3 site (L1026/A1027).[104] The NS3 substrate specificity is determined by the conserved residues at the corresponding cleavage sites of all HCV strains, which are an acidic residue at the P6 position, a Cys/Thr at the P1 and a Ser/Ala residue at the P1' site.[105] Therefore, despite the fact that cleavage sites of sp, spp, NS2/3 and NS3 are located within disordered polymorphic regions, there are some mutational restraints on the residues forming the cleavage sites in order for these sites to be always recognizable by the enzymes.

Abundance of intrinsic disorder in the HCV proteins

Disorder status of 10 proteins in the HCV proteome

Figure 2 provides a comprehensive overview of the abundance of intrinsic disorder in 10 HCV proteins across the considered HCV genotypes. The analysis is summarized (based on the multiple sequence alignment) over all 32 polyproteins (dark red lines) and over the polyproteins that correspond to each of the six genotypes (see legend for color codes). The IDR lines that are shown towards the bottom of the Figure reveal the location of IDRs. Even a superficial analysis of these lines suggests that only four HCV proteins (E1, E2, p7, and NS2) are predicted to be primarily structured, whereas other proteins are moderately (NS3, NS4A, NS4B, and NS5B) or highly disordered (C and NS5A). Based on analysis of all 32 polyproteins (dark red IDR lines in Figure 2) and using a classification scheme where highly ordered, moderately disordered, and highly disordered proteins are defined as proteins with predicted content (fraction) of disordered residues <10%, 10 to 30%, and >30%, respectively, then HCV has five highly ordered protein (E1 with no disordered residues, E2 with 1.4%, p7 with 4.8%, NS4A with 5.6%, and NS2 with 7.4% of disordered residues), three moderately disordered proteins (NS4B with 18.9%, NS5B with 19.1%, and NS3 with 25.2% of disordered residues), and two highly disordered proteins (NS5A with 51.5% and C with 60.2% of disordered residues). The overall disorder content of the entire HCV polyprotein is at 23.3%. These are rather unexpected observations since all the HCV proteins are anchored to the intracellular membranes via specific determinants that are essential for protein function in the cell,[106] since the HCV proteome includes three structural proteins and three enzymes,[21] and since based on the earlier studies, enzymes and transmembrane proteins are expected to be mostly ordered.[45-47] Furthermore, we also observe that the abundance of disorder is

similar for each of the considered genotypes; this conclusion is based on comparison of the dark red and other IDR lines in Figure 2.

Figure 3A summarizes sizes of IDRs for the 10 HCV proteins. We observe that 85% of these IDRs are relatively short, between 4 and 30 residues (shown using blue bars). Five proteins including E1, E2, p7, NS2 and NS4A have only these short IDRs, while the other five proteins also have long (over 30 residues; shown using red bars) disordered regions. NS5A includes the largest number of the long IDRs, with the longest one in genotype 2 that consist of 282 residues. The two proteins with the largest total count of IDRs over the considered 32 polyproteins, NS3 with 300 IDRs and NS5B with 273 IDRs, include both short and long regions.

Disorder status of the HCV proteins from the CH-CDF analysis

Figure 4 represents the results of the CH-CDF analysis of the 10 HCV proteins from 32 isolates and provides further support to the highly disordered nature of C and NS5A proteins. CH-plot,[30, 73] CDF-plot[73, 74] and CH-CDF analysis[74-76] represent computational tools for the binary classification of the disorder status of whole proteins, providing evidence of whether a given protein is expected to be ordered or disordered as a whole. In the CH-CDF plot, each spot correspond to a single protein, and the coordinates of each spot are calculated as a distance of the corresponding protein in the CH-plot (charge-hydropathy plot) from the boundary (Y-coordinate) and an average distance of the respective cumulative distribution function (CDF) curve from the CDF boundary (X-coordinate).[74-76] The quadrants of CDF-CH phase space correspond to the following expectations: Q1, proteins predicted to be disordered by CH-plots, but ordered by CDFs; Q2, ordered proteins; Q3, proteins predicted to be disordered by CDFs, but compact by CH-plots (i.e., putative molten globules or hybrid proteins with alternating ordered and disordered regions); Q4, proteins predicted to be disordered by both methods (i.e., proteins with extended disorder).

Although these classifications could be questionable for large, multidomain proteins, they provide relatively unbiased description of small proteins.

Figure 4 shows that, according to the overall level of intrinsic disorder, HCV proteins can be grouped into three classes related to their localization within the CH-CDF phase space. Here, C proteins from almost all HCV isolates are located predominantly in the quadrant Q4 and therefore are expected to behave as native coils or native pre-molten globules. All NS5A proteins from all HCV isolates are predicted as potential native molten globules or hybrid proteins with alternating ordered and disordered regions. Finally, all E1, E2, p7, Ns2, NS3, NS4A, NS4B, and NS5B from all HCV isolates are predicted to be mostly ordered. Although there is an apparent contradiction between the results of disorder analyses shown in Figure 2 and 4 (where 8 proteins out of 10 are predicted to be ordered by the CH-CDF plot (Figure 4) whereas only 5 proteins are predicted to be ordered based on the analyses summarized in Figure 2), this difference is attributed to the methodological differences in the evaluating of disorder by various techniques. The CH-CDF plot is a per-protein disorder analysis leading to the rather crude classification of a whole target protein to one of four broad groups, whereas Figure 2 is generated based on the fine per-residue disorder analysis.

Functional analysis of intrinsically disordered regions in the HCV proteins

Given the relatively high abundance of disorder in the HCV polyproteins, we investigated whether and what functional roles the corresponding IDRs would carry. Utilizing protocol from ref.,[53, 77] 228 IDRs from among 1102 IDRs found across the considered 32 HCV polyproteins were predicted with 23 functional subclasses, which are defined in DisProt.[78] We discuss 13 functional annotation subclasses that were transferred into at least three IDRs (Figure 5) to exclude

annotations with a higher likelihood of being incorrect. Some of the IDRs were annotated with multiple putative functions, with the total of 509 annotations. Figure 5 shows that IDRs facilitate various binding events including interactions with proteins, RNAs, DNAs, metals, and ligands. The protein-protein interactions are implemented by over half of the functionally annotated IDRs. They also serve as flexible linkers and are present in phosphorylation sites. Some other putative functions were found to be present in a smaller subset of three of four genotypes and they include transactivation, intra-protein interactions, membrane (channel) transport, regulation of apoptosis, and autoregulatory functions. These abovementioned functions are typical to HCV.

Figure 3B shows that sizes of IDRs with the putative functional annotations vary between the 13 functional subclasses. Several functions, such as transactivation, autoregulatory and channel transport are associated with short IDRs that range between 5 and 10 residues in length. Most of the phosphorylation sites (80%) are also found in such short IDRs. On the other hand, we found that long IDRs (over 30 residues in length) are associated with regulation of apoptosis (72.7% of the IDRs found to have this function are long), and with protein-protein and protein-RNA binding. We found a substantial number of IDRs that are both short and long (including regions that are over 85 residues in length) that are annotated with the latter two functions.

Detailed localization of the functionally characterized IDRs per HCV protein and genotype is summarized in Figure 6. We associate a given functional IDR with a given genome if it is present in at least half of its isolates. We observe that eight out of ten HCV proteins are annotated with the putative functions of IDRs. Protein C includes putative annotations of protein-protein and protein-RNA binding across all six genotypes. Similarly universal annotation of flexible linkers/spacers in found in the NS4B protein. The protein-protein binding regions were found in all genotypes and in all but three proteins. Moreover, there are multiple protein binding regions in several protein including C, E2, NS3, and NS5A. Similarly, the protein-RNA binding regions were predicted in

all genotypes and in two proteins, C and NS5A. The largest number of the functionally annotated IDRs is in the NS5A chain. We found all but one (protein-DNA binding) of the annotated function types in the genotype 1.

These functions are likely to be implemented by the disordered regions due to their high, 80% or higher, sequence identity to experimentally annotated disordered regions that is required by our annotation protocol. We also note that some of these functions are specific to certain genotypes, although this finding is weakened by the relatively low coverage of our annotations, i.e., only 20% of the IDRs were annotated. Moreover, our analysis suggests that individual genotypes vary in functions that are facilitated by IDRs, which likely stems from the high mutation rates that are characteristic to viruses.

Enrichment of intrinsic disorder in annotated functional regions of the HCV proteins

The top part of Figure 2 includes annotations of various functional regions in the HCV polyproteins. These include location of transmembrane regions (TMR), protein-, DNA- and RNA-binding regions, and PTM sites. We also include annotations of MoRF regions that are associated with protein-protein interactions (bottom of Figure 2) and amino acid polymorphisms (AAP; top of Figure 2), where a given residue is defined as polymorphic if its amino acid type varies at a given position in the alignment between isolates/genotypes. We link the intrinsic disorder to these functional annotations by investigating its enrichment when compared to the overall amount of disorder in the HCV polyproteins, see Table 1. A score above/below 1 means that disorder amount in a given functional annotation is enriched/depleted by the corresponding ratio, e.g., we observe that intrinsically disordered residues occur 2.11 times more often in protein-binding regions in genotype 1 when compared with their rate of occurrence in the whole polyprotein. The last row in Table 1 that gives the average, over the 6 genotypes, results indicates that disorder is substantially

enriched in protein-, DNA-, and RNA-binding regions as well as in the PTM sites; this is also true for each of the six genotypes. This observation agrees with the analysis performed in Figure 5, providing further support to our claim that disorder carries these functional roles in the HCV. The enrichment of disorder in the MoRF regions is expected, as these regions are disordered and undergo disorder-to-order transition upon binding to protein partners. Table 1 also reveals that intrinsic disorder is strongly depleted in the transmembrane regions across the six genotypes. Moreover, the rate of disordered residues at the polymorphic positions on the HCV polyprotein is similar to the overall rate of disorder, which indicates that disordered regions have similar mutation rates as the whole HCV polyprotein. The only visible exceptions are the disordered regions in the core protein C (see AAP and IDR lines in Figure 2), where the mutation rates are lower.

MoRF regions in the HCV proteins

The putative MoRF regions are shown at the bottom of Figure 2. These regions are disordered when the corresponding chain is isolated from its protein binding partner and become structured upon binding. We observe that several HCV proteins have MoRF regions, including the core protein C, integral membrane protein NS2, NS3, NS4A, NS5A, and NS5B. Some of the MoRF regions are aligned with the known protein-binding regions, including the long MoRF regions in C and NS4A (see protein binding and MoRF lines in Figure 2), which suggests that these interactions occur by coupled binding and folding. The other regions constitute putative protein-binding events.

Functional implications of intrinsic disorder in the HCV proteins

Core protein C

HCV core or capsid protein(C or HCV-C) is the first protein that is cleaved from the large

polyprotein. It is released from the N-terminal region of the viral polyprotein via cleavage by host-encoded proteinases to generate an immature and a mature core protein of 191 and 173 amino acids, respectively. Therefore, core protein exists in two, or alternatively three, forms with the molar mass of 21 kDa (191 amino acids), 19 kDa (173 amino acids) and 16 kDa.[107, 108]The smallest form of core protein (16 kDa) is not generated by post-translational cleavage, but by translation from the second (alternative) reading frame.[109] The 19 and 21 kDa proteins are bound on membranes of ER, whereas the 16 kDa truncated form is localized preferentially to the perinuclear space.[107]

Similar to other proteins C of the *Flaviviridae*, HCV core protein is a highly basic peptide that can interact with membrane and binds RNA with broad sequence specificity, possessing RNA chaperone activities *in vitro* and being responsible for the condensation and packaging of the viral genomic RNA during virion morphogenesis.[110-113] This protein is important for the viral replication cycle and regulates and controls a complex and dynamic network of host cell proteins contributing to the viral persistence and pathogenicity. The ability of HCV-C to be involved in the numerous promiscuous interactions with a plethora of structurally unrelated partners and produce diverse and dynamic protein-protein, protein-RNA, and protein-lipid complexes during the viral replicative cycle is determined by its intrinsically disordered nature[110-113], which is also apparent from our analysis (Figure 2).

Mature HCV-C consists of two domains (referred to as domain 1 and domain 2) and is predicted to contain only few structural elements.[114] The C-terminal domain (or domain 2) is enriched in hydrophobic residues and serves as a membrane-binding module. The domain 1 encompasses residues 1 to 117 and contains three highly basic amino acid clusters that mediate RNA binding and promote RNA-structural rearrangements. This domain is sufficient for the assembly in nucleocapsid-like-particles (NLPs) in the presence of structured RNA. Besides RNA binding, the

majority of mapped protein interaction sites also fall within domain 1, indicating that this domain is the major organizer of the HCV infection network.

The biophysical characteristics of the N-terminal domain 1 of HCV-C and its fragments have been extensively characterized by a variety of methods, demonstrating the lack of any stable secondary or tightly folded tertiary structure.[110] This domain is very sensitive to proteolytic degradation and shows aberrant electrophoretic mobility on SDS gels,[115] features characteristic of IDPs.[116, 117] In agreement with this and with bioinformatics analyses that predicted that domain 1 is mostly unstructured (including our analysis in Figure 2), far-UV circular dichroism (CD) spectra of the N-terminal 124 (C124), 117 (C117) or 82 (C82) amino acid long fragments of HCV-C were shown to be typical for random coil-like polypeptides.[112, 115, 118] NMR spectroscopic analysis further supported the highly disordered nature of C82.[115] Overall, the highly disordered status of this domain provides an explanation for the ability of HCV-C to interact with several unrelated host proteins, such as the C-terminus of p53, the intracellular domain of lymphotoxin β receptor, the DEAD-box protein (DDX3, CAP- Rf), the 14-3-3 protein, and the p21Waf1/Cip1/Sid1.[115] Curiously, the intrinsically disordered N-terminal region of the domain 1 contains immunodominant antigenic sites, and NMR analysis of the 2-45 fragment of HCV-C shows that a helix–loop–helix motif is formed within the 17-37 region (PDB ID: 1CWX) carrying at least one conformational epitope.[119] Furthermore, in a crystal structure of a complex between a hepatitis C virus (HCV) core protein-derived peptide (residues 13-40) and the Ab fragment of a murine monoclonal antibody 19D9D6, only 16 residues of the peptide were resolved of which only 5 (residues 32-36) were involved in the formation of an α-helical structure (PDB ID: 1N64).[120]

Similarly to many known RNA chaperones, which are typically disordered,[43] HCV-C was found to be intrinsically disordered and possess the RNA chaperoning activity in a wide range of environmental conditions (including high temperatures). Furthermore, Fourier transform infrared

(FTIR) spectroscopy analysis revealed that the binding of the intrinsically unstructured domain 1 of the HCV core protein to its specific target (the SLIIId subdomain of the HCV internal ribosome entry site, IRES) leads to the formation of a noticeable ordered secondary structure (~22% of β-sheet), although the majority of the protein remained mostly unstructured.[121]

Figure 7 represents known interactions of HCV and host cell proteins and shows that the HCV proteins with the highest percentage of disordered residues have the greatest numbers of interacting proteins. Surprisingly, the partners of C, NS5A and NS3 also appear to be more interconnected than the partners of HCV proteins with less disorder (see light green lines at Figure 7).

Envelope glycoproteins E1 (gp35) and E2 (gp70)

Structural proteins E1 (30–35 kDa) and E2 (70–72 kDa) are type-I transmembrane glycoproteins which are highly N-glycosylated, containing up to 5 and 11 glycosylation sites, respectively. They comprised of the N-terminal ectodomains of 160 and 334 amino acids, respectively, and a short C-terminal transmembrane domain (TMD) of ~30 amino acids.[119] Both TMDs contain two short stretches of hydrophobic amino acids separated by a short polar segment containing fully conserved charged residues.[122] The E1 and E2 glycoproteins form a functional heterodimer and their TMDs play a major role in the biogenesis of the E1-E2 heterodimer.[123] This E1-E2 heterodimer is involved in virus attachment to the host cell, virion internalization through clathrin-dependent endocytosis and fusion with host membrane.[124] Although E1/E2 heterodimer is known to interact with human LDLR, CD81 and SCARB1/SR-BI receptors, this binding is not sufficient for infection, and some additional liver specific cofactors may be needed. In addition, E2 binds and inhibits human EIF2AK2/PKR, and also binds human CD209/DC-SIGN and CLEC4M/DC-SIGNR. These distinct functions imply that the envelope proteins adopt markedly different

conformations and that these conformations and transitions between them must be controlled tightly to occur at the appropriate phases of the replicative cycle.[119]

In the infected cells, E1 and E2 are localized in the lumen of the endoplasmic reticulum (ER), where they interact with other proteins on ER membranes, such as calnexin, calreticulin, and BiP (heavy immunoglobulin chain binding protein).[125] The fact that HCV-E1 and HCV-E2 glycoproteins are localized solely within the membranes of ER and are undetectable on the plasma membrane of infected cells suggests that, similar to other members of *Flaviviridae* family, HCV is released from host cells by budding from ER and subsequent exocytosis. Absence of the viral proteins on the surface of the infected cells limits the host immune response potential and contributes to the establishment and maintenance of chronic HCV infection.[125]

The E2 envelope glycoprotein contains two hypervariable regions (HVR1, residues 385-411 and HVR2, 475-481 in the polyprotein P29846, or residues 2-27 and 91-97 in the corresponding E2 protein).[126, 127]These amino acid stretches differ by up to 80% among HCV genotypes, and are quite different even among subtypes of the same genotype.[119] Curiously, HVR1 is located within the intrinsically disordered region, whereas HVR2 coincide with the region with increased conformational flexibility (see Figure 2). HVR1 is a globally basic region, which, despite its sequence variability, is characterized by a high conservation of the physicochemical properties of the residues at each position.[119] Since the only established so far neutralizing epitope is located within the HVR1,[128] it is assumed that the HVR1 variability is determined by antibody selection of immune-escape variants.[119] Furthermore, specific distribution of basic residues within this region is likely to be related to the ability of E2 to interact with negatively charged molecules at the cell surface and can be related to the host cell recognition and attachment, as well as to the cellular compartmentalization of the virus.[119] As far as HVR2 is concerned, although this region

shows 100% sequence variability between HCV genotypes, it is reported to act together with HVR1 in regulation of the E2 binding to the receptor.[129, 130]

Intrinsic membrane protein p7

Structurally, the intrinsic membrane protein p7 (63 residues) has a double membrane-spanning topology, with its N- and C-terminal ends facing cytosol and the very short connecting loop facing the ER lumen,[131] and with both transmembrane passages (or transmembrane segments, TMSs) being predicted to form α-helices.[119] In agreement with this predictions, the NMR analyses revealed that p7 possesses a hairpin structure, where the TMS1 consists of two helical parts including residues 6-16 and 17-27, and the TMS2 (residues 41 to 57) is also made of two α-helices.[132] NMR analysis also revealed that the five residues at both the N- and C-termini are disordered.[132] This finding is in agreement with the results of our intrinsic disorder propensity analysis of p7 which also showed that the termini of p7 are expected to be intrinsically disordered (see Figure 2; the disorder propensity profiles generated by both PONDR-FIT and MFDp have high values at both termini of p7).

p7 can form hexamers and mediate membrane ion permeability,[81, 82] thereby belonging to the members of the viroporin family,[133] which is a class of small virus-encoded hydrophobic proteins that oligomerize and form ion channels, modifying membrane permeability.[134] Recent NMR analysis of the HCV viroporin revealed an unusual mode of hexameric assembly, where the individual p7 monomers, *i*, not only interact with their immediate neighbors, but also reach farther to associate with the (*i+2*) and (*i+3*) monomers. This complex set of intermolecular interactions produces a sophisticated, funnel-like architecture.[135]

Functionally, besides being responsible for the formation of an ion channel, p7 is involved in interaction with several host globular and transmembrane proteins,[136] such as Netrin-G1 ligand,

which stimulates the growth of embryonic thalamic axons by binding of netrin-G1,[137]Preadipocyte factor 1, involved in cell differentiation of adipocytes,[138] and Notches 4 and 2 (cell surface receptors) involved in cell signaling.[139]

NS2 integral membrane protein

NS2 is another integral membrane protein with the only known function being participation in the proteolytic cleavage of the polyprotein at the NS2-NS3 junction, which represents the first posttranslational autocatalytic cleavage of the large HCV polyprotein.[140] NS2 is a non-glycosylated hydrophobic membrane protein with a molecular mass of ~23 KDa which is mostly not exposed in the cytosol.[141]NS2 has four potential TMSs and is targeted to the ER membrane by the signal sequence located in its preceding p7 protein and by the two internal signal sequences located within the NS2 itself.[142] NS2 contains a domain that is involved in the interaction with the N-terminus of the adjacent NS3 serine protease.[125] The zinc-dependent NS2-NS3 proteinase function requires most of the NS2 sequence,[143] and two residues, His 143 and Cys 184, were established to be essential for the NS2 catalytic activity.[144] NS2 also interacts with several host proteins (see Figure 7). Looking at the peculiarities of disorder propensity distribution within this protein (see Figure 2) one can assume that the C-terminal region of the protein with some disorder propensity and the increase structural flexibility could be responsible for interaction of NS2 with its binding partners, which is further supported by the existence of the putative MoRF region.

NS3 bifunctional protein

HCV-NS3 is a 631 amino acid residue bifunctional enzyme with two domains and a molar mass of 70 kDa. It has a proteinase domain at the protein N-terminus (189 N-terminal amino acids), and an NTPase–helicase domain that occupies the remaining two-thirds of the sequence at the C-

terminus (the 442 C-terminal amino acids).[145] Structurally, NS3 is one of the most well-studied HCV proteins and the 3D structures are available for full-length NS3, the NS3 serine proteinase domain (free[146] and complexed with NS4A co-factor and/or with inhibitors[147, 148]), and the NS3 helicase domain (free[149, 150] and complexed with single-stranded DNA[151]).

The major role of the HCV-NS3 N-terminal serine proteinase is post-translational cleavage of the HCV polyproteinat NS3/4A, NS4A/4B, NS4B/5A and NS5A/5B sites. This domain is also known to be necessary for the HCV infectivity.[125] All but NS3/4A cleavages catalyzed by the NS3 proteinase are done in *trans*. With the exception for the cleavage at the NS5A/5B site, all the other cleavages require the presence of the NS4A serine proteinase co-factor (54 residues).[125] The structure of the NS3 serine proteinase domain is described as a chymotrypsin like fold, with two six-stranded β-barrel subdomains of identical topology.[119] The proper folding of the HCV-NS3 protease is controlled by interaction with the NS4A, central part of which forms a β-strand inserted into the N-terminal β-barrel of NS3.[119]

The HCV-NS3 RNA helicase consists of three well-defined subdomains (two structurally related subdomains folded with β-α-β topology, and a third C-terminal subdomain containing seven α-helices and three β-strands) which all contribute to the protein's helicase activity.[119] The second subdomain of the HCV helicase is flexibly linked to the remainder of the NS3 protein and could undergo rigid-body movements during the unwinding of double-stranded RNA.[152] Based on the NMR analysis of the solution structure of the subdomain 2 of the HCV-NS3 RNA helicase it has been concluded that this domain was globular and well-structured in solution even in the absence of the remaining parts of the NS3 protein, being characterized by a fold consisting of a six-stranded parallel β-sheet with the β1-β8-β7-β4-β6-β5 topology.[152] This core is sandwiched between two regions characterized by very different modes of interaction with the central β-sheet. Although one side of this sheet (which in the full-length protein is oriented away from the subdomain interfaces)

is involved in extensive hydrophobic interactions with residues Ile347 (β2), Phe349 (β2), Ile354 (β3), Val358 (α1), Ile359 (α1), Leu377 (α2), Leu381 (α2), and Ile386 (α2-β5), the another side of the core interact with residues Phe422, Leu414, and Val397 which all are from rather flexible regions of the structure.[152] In fact, residues on this opposite side are involved in the formation of two long flexible loops (Tyr392-Asp405 and Thr411-Asp423), an α-helix (α3, Ala455-Arg464), and a third loop (Thr465-Pro470). It was proposed that the conformational flexibility of these loops may facilitate conformational changes required for helicase function.[152] Furthermore, Figure 6 shows that the NS3 protein contains several short disordered regions involved in protein-protein interactions. There is also short disordered region related to the transactivation activity of this protein (see Figure 6).

NS4A co-factor of NS3 proteinase

As it was mentioned above, the NS4A protein (8 kDa, 54 residues) serves as a specific co-factor of the NS3 proteinase responsible for the posttranslational cleavage of the primary viral polyprotein.[153, 154] Here, NS4A was shown to stabilize NS3, being incorporated as an integral component into the enzyme core, and assist in localization of NS3in the membrane.[153] It was also shown that the full protease activity of the HCV N-terminal serine protease domain of the NS3 protein, which is crucial for the processing of the non-structural region of the HCV polyprotein at the NS3/4A, NS4A/4B, NS4B/5A and NS5A/5B, is only achieved upon interaction of this proteinase with the NS4A protein.[154] The minimal NS4A domain required for the increase in the cleavage efficiency of NS3 is the NS4A 21-32 region,[154] whereas the N-terminally located hydrophobic domain of NS4A is responsible for interaction with the membranes and with other replicase components.[153]

According to Figure 2, NS4A is a moderately disordered protein, with all the disordered and flexible residues being located in the C-terminal domain. Unfortunately, this disordered C-terminal domain was not functionally annotated as of yet.

Integral membrane protein NS4B

NS4B is a highly hydrophobic, integral membrane protein that cotranslationally associates with the ER membrane, presumably via an internal signal-like sequence, or with some ER-derived modified compartment.[155] NS4B is engaged in virus assembly and release[156] and can induce the formation of the so called membranous web potentially representing the HCV RNA replication complex.[157] Formation of the HCV RNA replication complex requires interactions among the HCV non-structural (NS) proteins and a human cellular vesicle membrane transport protein referred to as hVAP-33.[158] The formation of this HCV replicon is initiated by the precursor of NS4B, which is able to anchor to lipid rafts. Most of the other HCV nonstructural proteins, including NS5A, NS5B, and NS3, are also localized to these lipid raft membranes, suggesting that protein-protein interactions among the various HCV nonstructural proteins and hVAP-33 are important for the formation of the HCV replication complex.[158]

Structural information on NS4B protein is very limited due to its highly hydrophobic nature. This protein is predicted to have four to five transmembrane regions, with the N-terminus being placed in the lumen and the C-terminus located in the cytoplasm.[155, 159, 160] A putative N-terminal amphipathic helix of NS4B was shown to mediate membrane association,[160] whereas mutations in a C-terminally located helix were shown to abolish replication.[156, 161] Based on the analysis of the peptide library derived from the NS4B protein it has been concluded that the different NS4B domains have different propensities to bind, interact, and affect different model membranes.[162] Unexpectedly, despite its high overall hydrophobicity, NS4B is predicted to be a moderately

disordered protein with several disordered regions likely connecting transmembrane regions (see TMR and IDR lines in Figure 2). It is tempting to hypothesize that these disordered regions can be involved in interaction of NS4B with some of its partners. Unfortunately, these disordered regions were not functionally annotated as of yet.

Membrane-associated phosphoprotein NS5A

HCV NS5A is a 49-KDa well-studied protein that has a key role in viral replication and is also involved in particle assembly.[163] Numerous protein-protein interactions have been reported for NS5A, including viral or host cell.[164] NS5A is a membrane-associated protein with an anchor on its N-terminal. Its cytoplasmic portion, which is divided into three domains, encompasses disordered regions. Domain 1 (D1, residues 27-213 in the isolated protein or residues 2008-2194 in Figure 2) of NS5A is highly conserved and its structure has been solved revealing a structural scaffold with a novel zinc-binding motif and a disulfide bond,[165, 166] whereas domains 2 and 3 (D2 and D3, residues 250-342 (residues 2231-2332 in Figure 2) and 356–447 (residues 2335-2428 in Figure 2), respectively) are less conserved and possess highly disordered regions.[167, 168] Both NS5A-D2 and NS5A-D3 are known to establish a complex molecular partnership.[169, 170] The absence of an ordered conformation and the highly dynamic behavior of both NS5A-D2 and NS5-D3 provide an underlying molecular basis enabling interactions with multiple partners and conferring to NS5A a hub-like character.

Domain 2 of HCV NS5A (NS5A-D2) is important for functions of NS5A and is involved in molecular interactions with the RdRp (NS5B) and PKR, a cellular interferon-inducible serine/threonine specific protein kinase. Thus, the interactions established by NS5A-D2 interfere with host regulation processes such as signaling pathways and apoptosis.[171] Liang and co-workers carried out a structural analysis of NS5-D2 using NMR spectroscopy. The analysis of the backbone

^{1}H, ^{13}C, and ^{15}N resonances, $^{3}J_{HN\alpha}$ coupling constants, and 3D NOE data indicates that NS5A-D2 lacks secondary structural elements and reveals characteristics of unfolded proteins. NMR relaxation parameters confirmed the lack of rigid structure in the domain.[172]

Likewise, sequence analysis indicates that NS5A-D3 is mostly unstructured although short structural elements may exist at its N-terminus (see Figure 2). In agreement, gel filtration chromatography, CD and NMR spectroscopy all pointed out the disordered nature of purified recombinant NS5A-D3.[167] However, in a more recent study by the same group, two NS5A-D3s from two HCV strains were found to exhibit propensity to partially fold into an α-helix.[173] NMR analysis identifies two putative α-helices, for which a molecular model could be obtained. The amphipathic nature of the first helix and its conservation in all genotypes suggest that it might correspond to a MoRF, and as such promote the interaction with relevant biological partner(s). One such a partner is Cyclophilin A (CypA).[173] Cyclophilins are host cell factors that are essential for HCV replication. NMR heteronuclear exchange experiments demonstrate that CypA has *in vitro* peptidyl-prolylcis/trans isomerase activity toward some of the peptidyl-prolyl bonds in NS5A-D3.[173] Interestingly, the interaction between HCV NS5A-D3 and CypA is completely abrogated by Cyclosporin A (CsA), a discovery that designates inhibitors of CypA, such as CsA or non-immunosuppressive analogues, as candidates for development of antiviral strategies.[173]

Tail-anchored RNA-dependent RNA polymerase NS5B

It was pointed out that NS5B region is highly heterogeneous among some HCV strains,[9] and this heterogeneity represents foundation of the existing classification of HCV strains into genotypes and subgenotypes.[9] RNA-dependent RNA polymerase NS5B is a key enzyme controlling two important steps in HCV replication, the synthesis of a complementary minus-strand RNA using the genome as template, and subsequent synthesis of genomic plus-strand RNA from this minus-

strand RNA template.[119] The biochemical activity of this viral RNA-dependent RNA polymerase is dependent on divalent cations (Mn^{2+} or Mg^{2+}), pH near to neutral and a very low concentration of salts.[125]

NS5B is a well-characterized tail-anchored membrane protein[174, 175] which is posttranslationally targeted to the membrane *via* a hydrophobic C-terminal insertion sequence (the 21 C-terminal amino acid residues) leading to the integral membrane association, and cytosolic orientation of the functional protein domain.[176] The catalytic domain of NS5B is formed by the 530 N-terminal residues (of the 591 amino acids in the full-length protein) and possesses the classical "fingers," "palm," and "thumb" subdomains typical for other single-chain polymerases. What differentiates NS5B from the other polymerases is the presence of an extension in the fingers subdomain. This extension is involved in the interaction with the thumb subdomain and restricts the mobility of one subdomain with respect to the other, leading to a fully enclosed active site into which NTP molecules readily can bind with no further rearrangement of the domains.[119] Using limited trypsinolysis it has been established that the NS5B structure is noticeably affected by the ligand binding. Here, although in the absence of inhibitors and RNA NS5B did not possess any specific hypersensitivity to trypsin, binding of P495-site inhibitors or the RNA template to NS5B induced specific trypsin hypersensitivity attributed to movement of the Δ1 loop.[177]

As already mentioned, CypA is critical for HCV replication, being involved, together with NS5A and NS5B, in the formation of the membrane-associated multiprotein complex supporting RNA transcription and replication. Recently, NMR spectroscopy was used to characterize the peculiarities of the molecular interactions between a truncated form of NS5B (NS5BΔ21), and NS5A-D2, NS5A-D3, and CypA.[178] This analysis revealed that NS5A-D2 interacts with NS5BΔ21, whereas NS5A-D3 is not involved in this interaction. In addition, both NS5BΔ21 and CypA were found to share a common binding site on NS5A. No direct molecular interaction was

detected between HCV NS5BΔ21 and host CypA.[178] Addition of CsA to a sample containing NS5BΔ21, NS5A-D2, and CypA specifically inhibits the interaction between CypA and NS5A-D2 without altering the interaction between NS5A-D2 and NS5BΔ21.[178]Authors were able to retrieve a high quality heteronuclear NMR spectrum of HCV NS5BΔ21, which allowed characterization of the NS5-D2-binding site on the polymerase.[178]In the ^{1}H,^{15}N-TROSY spectrum of NS5BΔ21, nearly 490 peaks were detected, which is less than the 578 residues (including His tag) in the NS5BΔ21 used for the analysis.[178]

The formation of the membrane-associated machinery containing both HCV non-structural proteins (including NS5B) and human host factors such as vesicle-associated membrane protein-associated protein subtypes A and B (VAP-A and VAP-B) is an inevitable step in the HCV genome replication.[158, 179]Curiously, a splicing variant of VAP-B, the 99-residue protein VAP-C, was shown to serve as an endogenous inhibitor of HCV infection by inhibiting HCV replication via binding to NS5B.[180, 181]Structural characterization of VAP-C by CD and NMR spectroscopies revealed VAPC is entirely unstructured in solution. Despite its highly disordered nature, this protein is capable of specific binding to NS5B leading to the formation of a "fuzzy complex", in which VAP-C remained substantially disordered.[182]

Alternative reading frame protein ARFP or HCV Core+1/S polypeptide

The HCV Core+1/S polypeptide, also known as the alternative reading frame protein (ARFP), provides another example of the intrinsically disordered regulatory protein from HCV.[183] Core+1/S is an alternative reading frame protein that is expressed from the Core coding region of the viral genome. This ORF is responsible for the expression of various ARFPs, also named Core+1 proteins, resulting from mechanisms such as ribosomal frame shifting and internal initiation at alternative AUG or non-AUG codons.[183] Although Core+1 proteins were shown not

to be required for HCV replication, they were found to be expressed during HCV infection and interfere with apoptosis and cell cycle regulation, suggesting a possible role of these proteins in HCV pathogenesis.[183] Core+1/S corresponds to the C-terminal fragment of the Core+1 ORF, and to date is the shortest ARFP form described.[183] This protein originates as the result of the internal initiation at alternative AUG codons (85–87) located downstream of the polyprotein codon initiator.[183] Core+1/S is a highly basic polypeptide that lacks significant secondary structure *in vitro*.[183] The intrinsically disordered nature of this protein was evidenced by the sequence-based disorder predictions, size exclusion chromatography, dynamic light scattering (DLS), fluorescence, CD, and NMR studies.[183] It was proposed that intrinsic disorder is used by Core+1/S for the recognition of diverse molecular partners and/or for the assembly.[183]

ACKNOWLEDGMENT

This work was supported in part by the Programs of the Russian Academy of Sciences for the "Molecular and Cellular Biology" (to V.N.U) and the Natural Sciences and Engineering Research Council (NSERC) Discovery grant (to L.K.).

REFERENCES

1. H. R. Rosen, *N Engl J Med*, 2011, **364**, 2429-2438.
2. A. Maheshwari and P. J. Thuluvath, *Clin Liver Dis*, 2010, **14**, 169-176; x.
3. L. Gravitz, *Nature*, 2011, **474**, S2-4.
4. K. Mohd Hanafiah, J. Groeger, A. D. Flaxman and S. T. Wiersma, *Hepatology*, 2013, **57**, 1333-1342.
5. T. Wilkins, J. K. Malcolm, D. Raina and R. R. Schade, *Am Fam Physician*, 2010, **81**, 1351-1357.
6. T. Nakano, G. M. Lau, G. M. Lau, M. Sugiyama and M. Mizokami, *Liver Int*, 2012, **32**, 339-345.
7. Inamullah, M. Idrees, H. Ahmed, g. Sajid ul, M. Ali, L. Ali and A. Ahmed, *Virol J*, 2011, **8**, 16.
8. T. A. Cha, J. Kolberg, B. Irvine, M. Stempien, E. Beall, M. Yano, Q. L. Choo, M. Houghton, G. Kuo, J. H. Han and et al., *J Clin Microbiol*, 1991, **29**, 2528-2534.
9. P. Simmonds, E. C. Holmes, T. A. Cha, S. W. Chan, F. McOmish, B. Irvine, E. Beall, P. L. Yap, J. Kolberg and M. S. Urdea, *J Gen Virol*, 1993, **74 (Pt 11)**, 2391-2399.
10. H. Tokita, S. M. Shrestha, H. Okamoto, M. Sakamoto, M. Horikita, H. Iizuka, S. Shrestha, Y. Miyakawa and M. Mayumi, *J Gen Virol*, 1994, **75 (Pt 4)**, 931-936.
11. V. Gouvea, N. Snellings, S. J. Cohen, R. L. Warren, K. S. Myint, M. P. Shrestha, D. W. Vaughn, C. H. Hoke, Jr. and B. L. Innis, *Virus Res*, 1997, **52**, 87-96.
12. H. Tokita, H. Okamoto, H. Iizuka, J. Kishimoto, F. Tsuda, L. A. Lesmana, Y. Miyakawa and M. Mayumi, *J Gen Virol*, 1996, **77 (Pt 2)**, 293-301.
13. F. McOmish, P. L. Yap, B. C. Dow, E. A. Follett, C. Seed, A. J. Keller, T. J. Cobain, T. Krusius, E. Kolho, R. Naukkarinen and et al., *J Clin Microbiol*, 1994, **32**, 884-892.
14. J. B. Nousbaum, S. Pol, B. Nalpas, P. Landais, P. Berthelot and C. Brechot, *Ann Intern Med*, 1995, **122**, 161-168.
15. N. N. Zein, J. Rakela, E. L. Krawitt, K. R. Reddy, T. Tominaga and D. H. Persing, *Ann Intern Med*, 1996, **125**, 634-639.
16. N. Takada, S. Takase, A. Takada and T. Date, *J Hepatol*, 1993, **17**, 277-283.
17. G. Dusheiko, H. Schmilovitz-Weiss, D. Brown, F. McOmish, P. L. Yap, S. Sherlock, N. McIntyre and P. Simmonds, *Hepatology*, 1994, **19**, 13-18.
18. N. Kato, *Microb Comp Genomics*, 2000, **5**, 129-151.
19. J. Dubuisson, *World J Gastroenterol*, 2007, **13**, 2406-2415.
20. D. Prangishvili, P. Forterre and R. A. Garrett, *Nat Rev Microbiol*, 2006, **4**, 837-848.
21. D. C. Reanney, *Annu Rev Microbiol*, 1982, **36**, 47-73.
22. N. Tokuriki, C. J. Oldfield, V. N. Uversky, I. N. Berezovsky and D. S. Tawfik, *Trends Biochem Sci*, 2009, **34**, 53-59.
23. B. Xue, R. W. Williams, C. J. Oldfield, G. K. Goh, A. K. Dunker and V. N. Uversky, *Protein Pept Lett*, 2010, **17**, 932-951.
24. A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J Mol Graph Model*, 2001, **19**, 26-59.
25. A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Inform Ser Workshop Genome Inform*, 2000, **11**, 161-171.
26. P. Tompa, *Trends Biochem Sci*, 2002, **27**, 527-533.

27. V. N. Uversky, *Protein Sci*, 2002, **11**, 739-756.
28. V. N. Uversky, *J Biomed Biotechnol*, 2010, **2010**, 568068.
29. V. N. Uversky and A. K. Dunker, *Biochimica et biophysica acta*, 2010, **1804**, 1231-1264.
30. V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins*, 2000, **41**, 415-427.
31. A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *The FEBS journal*, 2005, **272**, 5129-5148.
32. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J Mol Biol*, 2004, **337**, 635-645.
33. A. K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger and J. E. Villafranca, *Pac Symp Biocomput*, 1998, 473-484.
34. P. E. Wright and H. J. Dyson, *J Mol Biol*, 1999, **293**, 321-331.
35. G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese and A. K. Dunker, in *Handbook of Protein Folding*, eds. J. Buchner and T. Kiefhaber, Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany, 2005, pp. 271-353.
36. V. N. Uversky, *Biochim Biophys Acta*, 2013, **1834**, 932-951.
37. A. K. Dunker and Z. Obradovic, *Nat Biotechnol*, 2001, **19**, 805-806.
38. B. Xue, A. K. Dunker and V. N. Uversky, *J Biomol Struct Dyn*, 2012, **30**, 137-149.
39. V. N. Uversky, C. J. Oldfield and A. K. Dunker, *Annu Rev Biophys*, 2008, **37**, 215-246.
40. V. Vacic, P. R. Markwick, C. J. Oldfield, X. Zhao, C. Haynes, V. N. Uversky and L. M. Iakoucheva, *PLoS Comput Biol*, 2012, **8**, e1002709.
41. L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker, *J Mol Biol*, 2002, **323**, 573-584.
42. H. J. Dyson and P. E. Wright, *Nat Rev Mol Cell Biol*, 2005, **6**, 197-208.
43. P. Tompa, *FEBS Lett*, 2005, **579**, 3346-3354.
44. P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky and A. K. Dunker, *Biophys J*, 2007, **92**, 1439-1456.
45. S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J Proteome Res*, 2007, **6**, 1899-1916.
46. H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J Proteome Res*, 2007, **6**, 1917-1932.
47. H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J Proteome Res*, 2007, **6**, 1882-1898.
48. J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2006, **45**, 6873-6888.
49. Y. Minezaki, K. Homma, A. R. Kinjo and K. Nishikawa, *J Mol Biol*, 2006, **359**, 1137-1149.
50. B. Xue, C. J. Oldfield, Y. Y. Van, A. K. Dunker and V. N. Uversky, *Mol Biosyst*, 2012, **8**, 134-150.
51. F. Peysselon, B. Xue, V. N. Uversky and S. Ricard-Blum, *Mol Biosyst*, 2011, **7**, 3353-3365.
52. M. Ito, Y. Tohsato, H. Sugisawa, S. Kohara, S. Fukuchi, I. Nishikawa and K. Nishikawa, *Genes Cells*, 2012, **17**, 817-825.
53. Z. Peng, C. J. Oldfield, B. Xue, M. J. Mizianty, A. K. Dunker, L. Kurgan and V. N. Uversky, *Cell Mol Life Sci*, 2013.
54. K. Homma, S. Fukuchi, K. Nishikawa, S. Sakamoto and H. Sugawara, *Mol Biosyst*, 2012, **8**, 247-255.

55. Z. Peng, M. J. Mizianty, B. Xue, L. Kurgan and V. N. Uversky, *Mol Biosyst*, 2012, **8**, 1886-1901.
56. R. Pushker, C. Mooney, N. E. Davey, J. M. Jacque and D. C. Shields, *PLoS One*, 2013, **8**, e60724.
57. I. N. Berezovsky, *Phys Biol*, 2011, **8**, 035002.
58. C. UniProt, *Nucleic Acids Res*, 2013, **41**, D43-47.
59. P. Simmonds, J. Mellor, T. Sakuldamrongpanich, C. Nuchaprayoon, S. Tanprasert, E. C. Holmes and D. B. Smith, *J Gen Virol*, 1996, **77 (Pt 12)**, 3013-3024.
60. X. de Lamballerie, R. N. Charrel, H. Attoui and P. De Micco, *J Gen Virol*, 1997, **78 (Pt 1)**, 45-51.
61. H. Tokita, H. Okamoto, H. Iizuka, J. Kishimoto, F. Tsuda, Y. Miyakawa and M. Mayumi, *J Gen Virol*, 1998, **79 (Pt 8)**, 1847-1857.
62. C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 12454-12470.
63. A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker and V. N. Uversky, *J Mol Biol*, 2006, **362**, 1043-1059.
64. V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky and A. K. Dunker, *J Proteome Res*, 2007, **6**, 2351-2366.
65. Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2007, **46**, 13468-13477.
66. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins, *Bioinformatics*, 2007, **23**, 2947-2948.
67. P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins*, 2001, **42**, 38-48.
68. M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani and L. Kurgan, *Bioinformatics*, 2010, **26**, i489-496.
69. B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker and V. N. Uversky, *Biochim Biophys Acta*, 2010, **1804**, 996-1010.
70. Z. L. Peng and L. Kurgan, *Curr Protein Pept Sci*, 2012, **13**, 6-18.
71. F. M. Disfani, W. L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker, V. N. Uversky and L. Kurgan, *Bioinformatics*, 2012, **28**, i75-83.
72. B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano and A. Kryshtafovych, *Proteins*, 2011, **79 Suppl 10**, 107-118.
73. C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 1989-2000.
74. B. Xue, C. J. Oldfield, A. K. Dunker and V. N. Uversky, *FEBS Lett*, 2009, **583**, 1469-1474.
75. A. Mohan, W. J. Sullivan, Jr., P. Radivojac, A. K. Dunker and V. N. Uversky, *Mol Biosyst*, 2008, **4**, 328-340.
76. F. Huang, C. Oldfield, J. Meng, W. L. Hsu, B. Xue, V. N. Uversky, P. Romero and A. K. Dunker, *Pac Symp Biocomput*, 2012, 128-139.
77. Z. Peng, B. Xue, L. Kurgan and V. N. Uversky, *Cell death and differentiation*, 2013, **20**, 1257-1267.
78. M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res*, 2007, **35**, D786-793.

79. T. F. Smith and M. S. Waterman, *J Mol Biol*, 1981, **147**, 195-197.
80. C. M. Spahn, J. S. Kieft, R. A. Grassucci, P. A. Penczek, K. Zhou, J. A. Doudna and J. Frank, *Science*, 2001, **291**, 1959-1962.
81. D. Pavlovic, D. C. Neville, O. Argaud, B. Blumberg, R. A. Dwek, W. B. Fischer and N. Zitzmann, *Proc Natl Acad Sci U S A*, 2003, **100**, 6104-6108.
82. S. D. Griffin, L. P. Beales, D. S. Clarke, O. Worsfold, S. D. Evans, J. Jaeger, M. P. Harris and D. J. Rowlands, *FEBS Lett*, 2003, **535**, 34-38.
83. E. Santolini, G. Migliaccio and N. La Monica, *J Virol*, 1994, **68**, 3631-3641.
84. P. Hussy, H. Langen, J. Mous and H. Jacobsen, *Virology*, 1996, **224**, 93-104.
85. T. Ogino, H. Fukuda, S. Imajoh-Ohmi, M. Kohara and A. Nomoto, *J Virol*, 2004, **78**, 11766-11777.
86. B. D. Lindenbach and C. M. Rice, *Nature*, 2005, **436**, 933-938.
87. K. Rajagopalan, S. M. Mooney, N. Parekh, R. H. Getzenberg and P. Kulkarni, *J Cell Biochem*, 2011, **112**, 3256-3267.
88. P. P. de Laureto, L. Tosatto, E. Frare, O. Marin, V. N. Uversky and A. Fontana, *Biochemistry*, 2006, **45**, 11523-11531.
89. A. Fontana, P. P. de Laureto, B. Spolaore, E. Frare, P. Picotti and M. Zambonin, *Acta Biochim Pol*, 2004, **51**, 299-321.
90. A. Fontana, G. Fassina, C. Vita, D. Dalzoppo, M. Zamai and M. Zambonin, *Biochemistry*, 1986, **25**, 1847-1851.
91. A. Fontana, P. Polverino de Laureto, V. De Filippis, E. Scaramella and M. Zambonin, *Fold Des*, 1997, **2**, R17-26.
92. L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, J. E. Bruce, E. C. Garner, C. J. Brown, A. K. Dunker, R. D. Smith and E. J. Ackerman, *Protein Sci*, 2001, **10**, 560-571.
93. P. Polverino de Laureto, V. De Filippis, M. Di Bello, M. Zambonin and A. Fontana, *Biochemistry*, 1995, **34**, 12596-12604.
94. A. W. Strauss, M. Zimmerman, I. Boime, B. Ashe, R. A. Mumford and A. W. Alberts, *Proc Natl Acad Sci U S A*, 1979, **76**, 4225-4229.
95. G. Kreibich, M. Czako-Graham, R. C. Grebenau and D. D. Sabatini, *Ann N Y Acad Sci*, 1980, **343**, 17-33.
96. R. J. Folz, S. F. Nothwehr and J. I. Gordon, *J Biol Chem*, 1988, **263**, 2070-2078.
97. G. von Heijne, *Eur J Biochem*, 1983, **133**, 17-21.
98. G. von Heijne, *EMBO J*, 1984, **3**, 2315-2318.
99. G. von Heijne, *J Mol Biol*, 1984, **173**, 243-251.
100. D. Perlman and H. O. Halvorson, *J Mol Biol*, 1983, **167**, 391-409.
101. T. Sato, A. C. Nyborg, N. Iwata, T. S. Diehl, T. C. Saido, T. E. Golde and M. S. Wolfe, *Biochemistry*, 2006, **45**, 8649-8656.
102. J. McLauchlan, M. K. Lemberg, G. Hope and B. Martoglio, *EMBO J*, 2002, **21**, 3980-3988.
103. M. Pallaoro, A. Lahm, G. Biasiol, M. Brunetti, C. Nardella, L. Orsatti, F. Bonelli, S. Orru, F. Narjes and C. Steinkuhler, *J Virol*, 2001, **75**, 9939-9946.
104. S. Welbourn and A. Pause, *Curr Issues Mol Biol*, 2007, **9**, 63-69.
105. J.-H. Lee, K. Bang, J.-W. Jung, I.-A. Ahn, S. Ro and W. Lee, *Bull. Korean Chem. Soc.*, 1999, **20**, 301-306.
106. D. Moradpour and F. Penin, *Curr Top Microbiol Immunol*, 2013, **369**, 113-142.
107. S. Y. Lo, F. Masiarz, S. B. Hwang, M. M. Lai and J. H. Ou, *Virology*, 1995, **213**, 455-461.

108. K. Yasui, T. Wakita, K. Tsukiyama-Kohara, S. I. Funahashi, M. Ichikawa, T. Kajita, D. Moradpour, J. R. Wands and M. Kohara, *J Virol*, 1998, **72**, 6048-6055.
109. Z. Xu, J. Choi, T. S. Yen, W. Lu, A. Strohecker, S. Govindarajan, D. Chien, M. J. Selby and J. Ou, *EMBO J*, 2001, **20**, 3840-3848.
110. R. Ivanyi-Nagy and J. L. Darlix, *Protein Pept Lett*, 2010, **17**, 1019-1025.
111. R. Ivanyi-Nagy and J. L. Darlix, *Adv Exp Med Biol*, 2012, **725**, 142-158.
112. R. Ivanyi-Nagy, J. P. Lavergne, C. Gabus, D. Ficheux and J. L. Darlix, *Nucleic Acids Res*, 2008, **36**, 712-725.
113. R. Ivanyi-Nagy, E. Pécheur and J. L. Darlix, in *Flexible viruses: structural disorder in viral proteins*, eds. V. N. Uversky and S. Longhi, John Wiley and Sons, Hoboken, New Yersey, 2012, pp. 375-407.
114. M. Kunkel and S. J. Watowich, *FEBS Lett*, 2004, **557**, 174-180.
115. J. B. Duvignaud, C. Savard, R. Fromentin, N. Majeau, D. Leclerc and S. M. Gagne, *Biochem Biophys Res Commun*, 2009, **378**, 27-31.
116. V. Receveur-Brechot, J. M. Bourhis, V. N. Uversky, B. Canard and S. Longhi, *Proteins*, 2006, **62**, 24-45.
117. V. N. Uversky and A. K. Dunker, *Anal Chem*, 2012, **84**, 2096-2104.
118. S. Boulant, C. Vanbelle, C. Ebel, F. Penin and J. P. Lavergne, *J Virol*, 2005, **79**, 11353-11365.
119. F. Penin, J. Dubuisson, F. A. Rey, D. Moradpour and J. M. Pawlotsky, *Hepatology*, 2004, **39**, 5-19.
120. R. Menez, M. Bossus, B. H. Muller, G. Sibai, P. Dalbon, F. Ducancel, C. Jolivet-Reynaud and E. A. Stura, *J Immunol*, 2003, **170**, 1917-1924.
121. P. Carmona and M. Molina, *Biochemistry*, 2010, **49**, 4724-4731.
122. L. Cocquerel, C. Wychowski, F. Minner, F. Penin and J. Dubuisson, *J Virol*, 2000, **74**, 3623-3633.
123. A. Op De Beeck, L. Cocquerel and J. Dubuisson, *J Gen Virol*, 2001, **82**, 2589-2595.
124. M. Flint, J. M. Thomas, C. M. Maidens, C. Shotton, S. Levy, W. S. Barclay and J. A. McKeating, *J Virol*, 1999, **73**, 6782-6790.
125. L. Krekulova, V. Rehak and L. W. Riley, *Folia Microbiol (Praha)*, 2006, **51**, 665-680.
126. A. J. Weiner, C. Christopherson, J. E. Hall, F. Bonino, G. Saracco, M. R. Brunetto, K. Crawford, C. D. Marion, K. A. Crawford, S. Venkatakrishna and et al., *J Hepatol*, 1991, **13 Suppl 4**, S6-14.
127. N. Kato, *Acta Med Okayama*, 2001, **55**, 133-159.
128. P. Farci, A. Shimoda, D. Wong, T. Cabezon, D. De Gioannis, A. Strazzera, Y. Shimizu, M. Shapiro, H. J. Alter and R. H. Purcell, *Proc Natl Acad Sci U S A*, 1996, **93**, 15394-15399.
129. E. Scarselli, H. Ansuini, R. Cerino, R. M. Roccasecca, S. Acali, G. Filocamo, C. Traboni, A. Nicosia, R. Cortese and A. Vitelli, *EMBO J*, 2002, **21**, 5017-5025.
130. R. Roccasecca, H. Ansuini, A. Vitelli, A. Meola, E. Scarselli, S. Acali, M. Pezzanera, B. B. Ercole, J. McKeating, A. Yagnik, A. Lahm, A. Tramontano, R. Cortese and A. Nicosia, *J Virol*, 2003, **77**, 1856-1867.
131. S. Carrere-Kremer, C. Montpellier-Pala, L. Cocquerel, C. Wychowski, F. Penin and J. Dubuisson, *J Virol*, 2002, **76**, 3720-3730.
132. G. A. Cook and S. J. Opella, *Biochim Biophys Acta*, 2011, **1808**, 1448-1453.
133. L. Carrasco, *Adv Virus Res*, 1995, **45**, 61-112.
134. M. E. Gonzalez and L. Carrasco, *FEBS Lett*, 2003, **552**, 28-34.

135. B. OuYang, S. Xie, M. J. Berardi, X. Zhao, J. Dev, W. Yu, B. Sun and J. J. Chou, *Nature*, 2013, **498**, 521-525.
136. Q. Li, A. L. Brass, A. Ng, Z. Hu, R. J. Xavier, T. J. Liang and S. J. Elledge, *Proc Natl Acad Sci U S A*, 2009, **106**, 16410-16415.
137. J. C. Lin, W. H. Ho, A. Gurney and A. Rosenthal, *Nat Neurosci*, 2003, **6**, 1270-1276.
138. Y. Wang, C. Hudak and H. S. Sul, *Clin Lipidol*, 2010, **5**, 109-115.
139. S. J. Bray, *Nat Rev Mol Cell Biol*, 2006, **7**, 678-689.
140. M. Hijikata, H. Mizushima, T. Akagi, S. Mori, N. Kakiuchi, N. Kato, T. Tanaka, K. Kimura and K. Shimotohno, *J Virol*, 1993, **67**, 4665-4675.
141. E. Santolini, L. Pacini, C. Fipaldini, G. Migliaccio and N. Monica, *J Virol*, 1995, **69**, 7461-7471.
142. A. K. Yamaga and J. H. Ou, *J Biol Chem*, 2002, **277**, 33228-33234.
143. K. E. Reed and C. M. Rice, *Curr Top Microbiol Immunol*, 2000, **242**, 55-84.
144. A. Grakoui, D. W. McCourt, C. Wychowski, S. M. Feinstone and C. M. Rice, *Proc Natl Acad Sci U S A*, 1993, **90**, 10583-10587.
145. D. W. Kim, Y. Gwack, J. H. Han and J. Choe, *Biochem Biophys Res Commun*, 1995, **215**, 160-166.
146. R. A. Love, H. E. Parge, J. A. Wickersham, Z. Hostomsky, N. Habuka, E. W. Moomaw, T. Adachi and Z. Hostomska, *Cell*, 1996, **87**, 331-342.
147. J. L. Kim, K. A. Morgenstern, C. Lin, T. Fox, M. D. Dwyer, J. A. Landro, S. P. Chambers, W. Markland, C. A. Lepre, E. T. O'Malley, S. L. Harbeson, C. M. Rice, M. A. Murcko, P. R. Caron and J. A. Thomson, *Cell*, 1996, **87**, 343-355.
148. Y. Yan, Y. Li, S. Munshi, V. Sardana, J. L. Cole, M. Sardana, C. Steinkuehler, L. Tomei, R. De Francesco, L. C. Kuo and Z. Chen, *Protein Sci*, 1998, **7**, 837-847.
149. N. Yao, T. Hesson, M. Cable, Z. Hong, A. D. Kwong, H. V. Le and P. C. Weber, *Nat Struct Biol*, 1997, **4**, 463-467.
150. H. S. Cho, N. C. Ha, L. W. Kang, K. M. Chung, S. H. Back, S. K. Jang and B. H. Oh, *J Biol Chem*, 1998, **273**, 15045-15052.
151. J. L. Kim, K. A. Morgenstern, J. P. Griffith, M. D. Dwyer, J. A. Thomson, M. A. Murcko, C. Lin and P. R. Caron, *Structure*, 1998, **6**, 89-100.
152. D. Liu, Y. S. Wang, J. J. Gesell and D. F. Wyss, *J Mol Biol*, 2001, **314**, 543-561.
153. Y. Tanji, M. Hijikata, S. Satoh, T. Kaneko and K. Shimotohno, *J Virol*, 1995, **69**, 1575-1581.
154. L. Tomei, C. Failla, R. L. Vitale, E. Bianchi and R. De Francesco, *J Gen Virol*, 1996, **77 (Pt 5)**, 1065-1070.
155. T. Hugle, F. Fehrmann, E. Bieck, M. Kohara, H. G. Krausslich, C. M. Rice, H. E. Blum and D. Moradpour, *Virology*, 2001, **284**, 70-81.
156. D. M. Jones, A. H. Patel, P. Targett-Adams and J. McLauchlan, *J Virol*, 2009, **83**, 2163-2177.
157. D. Egger, B. Wolk, R. Gosert, L. Bianchi, H. E. Blum, D. Moradpour and K. Bienz, *J Virol*, 2002, **76**, 5974-5984.
158. L. Gao, H. Aizaki, J. W. He and M. M. Lai, *J Virol*, 2004, **78**, 3480-3488.
159. M. Lundin, M. Monne, A. Widell, G. Von Heijne and M. A. Persson, *J Virol*, 2003, **77**, 5428-5438.
160. M. Elazar, P. Liu, C. M. Rice and J. S. Glenn, *J Virol*, 2004, **78**, 11393-11400.
161. H. Lindstrom, M. Lundin, S. Haggstrom and M. A. Persson, *Virus Res*, 2006, **121**, 169-178.

162. J. Guillen, A. Gonzalez-Alvarez and J. Villalain, *Biochim Biophys Acta*, 2009, **1798**, 327-337.
163. T. L. Foster, T. Belyaeva, N. J. Stonehouse, A. R. Pearson and M. Harris, *J Virol*, 2010, **84**, 9267-9277.
164. B. de Chassey, V. Navratil, L. Tafforeau, M. S. Hiet, A. Aublin-Gex, S. Agaugue, G. Meiffren, F. Pradezynski, B. F. Faria, T. Chantier, M. Le Breton, J. Pellet, N. Davoust, P. E. Mangeot, A. Chaboud, F. Penin, Y. Jacob, P. O. Vidalain, M. Vidal, P. Andre, C. Rabourdin-Combe and V. Lotteau, *Mol Syst Biol*, 2008, **4**, 230.
165. T. L. Tellinghuisen, J. Marcotrigiano and C. M. Rice, *Nature*, 2005, **435**, 374-379.
166. R. A. Love, O. Brodsky, M. J. Hickey, P. A. Wells and C. N. Cronin, *J Virol*, 2009, **83**, 4395-4403.
167. X. Hanoulle, D. Verdegem, A. Badillo, J. M. Wieruszeski, F. Penin and G. Lippens, *Biochem Biophys Res Commun*, 2009, **381**, 634-638.
168. X. Hanoulle, A. Badillo, D. Verdegem, F. Penin and G. Lippens, *Protein Pept Lett*, 2010, **17**, 1012-1018.
169. A. Macdonald, K. Crowder, A. Street, C. McCormick and M. Harris, *J Gen Virol*, 2004, **85**, 721-729.
170. A. Macdonald and M. Harris, *J Gen Virol*, 2004, **85**, 2485-2502.
171. S. Feuerstein, Z. Solyom, A. Aladag, A. Favier, M. Schwarten, S. Hoffmann, D. Willbold and B. Brutscher, *J Mol Biol*, 2012, **420**, 310-323.
172. Y. Liang, H. Ye, C. B. Kang and H. S. Yoon, *Biochemistry*, 2007, **46**, 11550-11558.
173. D. Verdegem, A. Badillo, J. M. Wieruszeski, I. Landrieu, A. Leroy, R. Bartenschlager, F. Penin, G. Lippens and X. Hanoulle, *J Biol Chem*, 2011, **286**, 20441-20454.
174. J. Schmidt-Mende, E. Bieck, T. Hugle, F. Penin, C. M. Rice, H. E. Blum and D. Moradpour, *J Biol Chem*, 2001, **276**, 44052-44063.
175. U. Kutay, E. Hartmann and T. A. Rapoport, *Trends Cell Biol*, 1993, **3**, 72-75.
176. B. Wattenberg and T. Lithgow, *Traffic*, 2001, **2**, 66-71.
177. K. Rigat, Y. Wang, T. W. Hudyma, M. Ding, X. Zheng, R. G. Gentles, B. R. Beno, M. Gao and S. B. Roberts, *Antiviral Res*, 2010, **88**, 197-206.
178. C. Rosnoblet, B. Fritzinger, D. Legrand, H. Launay, J. M. Wieruszeski, G. Lippens and X. Hanoulle, *J Biol Chem*, 2012, **287**, 44249-44260.
179. I. Hamamoto, Y. Nishimura, T. Okamoto, H. Aizaki, M. Liu, Y. Mori, T. Abe, T. Suzuki, M. M. Lai, T. Miyamura, K. Moriishi and Y. Matsuura, *J Virol*, 2005, **79**, 13473-13482.
180. H. Kukihara, K. Moriishi, S. Taguwa, H. Tani, T. Abe, Y. Mori, T. Suzuki, T. Fukuhara, A. Taketomi, Y. Maehara and Y. Matsuura, *J Virol*, 2009, **83**, 7959-7969.
181. X. Wen, T. Abe, H. Kukihara, S. Taguwa, Y. Mori, H. Tani, N. Kato, T. Suzuki, M. Tatsumi, K. Moriishi and Y. Matsuura, *PLoS One*, 2011, **6**, e15967.
182. S. Goyal, G. Gupta, H. Qin, M. H. Upadya, Y. J. Tan, V. T. Chow and J. Song, *PLoS One*, 2012, **7**, e40341.
183. A. Boumlic, Y. Nomine, S. Charbonnier, G. Dalagiorgou, N. Vassilaki, B. Kieffer, G. Trave, P. Mavromara and G. Orfanoudakis, *FEBS J*, 2010, **277**, 774-789.
184. S. K. Kwofie, U. Schaefer, V. S. Sundararajan, V. B. Bajic and A. Christoffels, *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 2011, **11**, 1971-1977.
185. P. T. Dolan, C.-y. Zhang, S. Khadka, V. Arumugaswami, A. D. Vangeloff, N. S. Heaton, S. Sahasrabudhe, G. Randall, R. Sun and D. LaCount, *Molecular BioSystems*, 2013, **9**, 3199-3209.

186. M. A. Germain, L. Chatel-Chaix, B. Gagne, E. Bonneil, P. Thibault, F. Pradezynski, B. de Chassey, L. Meyniel-Shicklin, V. Lotteau, M. Baril and D. Lamarre, *Mol Cell Proteomics*, 2014, **13**, 184-203.

FIGURE CAPTIONS

Figure 1. The proteome map of the human HCV-1b (UniProt ID: P26662). Similar to other positive-strand RNA viruses, upon infection of a hepatic cell the HCV genomic RNA serves as messenger RNA for the translation of viral proteins. The linear molecule contains a single open reading frame coding for a precursor polyprotein (~3000 aminoacid residues) consisting of 10 proteins that must be cleaved in order to be functional. A polyprotein (plot **A**) and each viral protein (plot **C**) are presented as a bar whose location corresponds to the location of the corresponding gene within the HCV genome. PONDR® VLXT disorder predictions for polyprotein and each of the 10 viral proteins are shown as red solid lines inside the corresponding rectangle. Plot **B** represents disorder propensities of regions containing cleavage sites. Disorder propensities were evaluated by PONDR® VLXT and PONDR FIT (red and blue lines respectively). Cleavage sites are shown as gray bars. In plot **A**, structural and non-structural proteins are shown by cyan and yellow rectangles, respectively. In plot **C**, the viral proteins are color coded according to their intrinsic disorder content evaluated by PONDR® VLXT, were two arbitrary cutoffs for the levels of intrinsic disorder were used to classify proteins as highly ordered (0–10% of the sequence is disordered, cyan rectangles), moderately disordered (11%–30% of the sequence is disordered, yellow rectangles) and highly disordered (31%–100% of the sequence is disordered, light red rectangles). Percentages and numbers in brackets at the bottom of figure correspond to the percentage of residues predicted to be disordered and the corresponding mean disorder score, respectively, evaluated by PONDR® VLXT for a given protein. Black horizontal line at the middle of each plot represents the threshold line at 0.5 above which residue/region is predicted to be disordered.

Figure 2. Functional and structural annotations of the 32 HCV polyproteins. The annotations were summarized over all polyproteins (horizontal line C) and over the polyproteins that belong to each of the 6 genotypes (horizontal lines 1, 2, 3, 4, 5, and 6); each functional/structural characteristic is shown for these seven protein sets. From the top of the figure we include: the protein names; positions in the polyprotein sequence where individual chains are cleaved; the sequences and amino acid polymorphisms (AAP) based on the multiple sequence alignment with ClustalW where strongly/weakly polymorphic residue is defined as having over half/at least one and no more than half of the considered chains with different amino acid type at a given position in the alignment; annotations of transmembrane regions (TMR), protein-binding, DNA-binding, and RNA-binding regions, and post-transcriptional modification (PTM) sites where black/gray lines denotes annotations that are true across all/most of the corresponding isolates and genotypes; disorder predictions with MFDp and PONDR-FIT and MoRF predictions with MoRFpred including profiles of probabilities and binary predictions (disordered vs. ordered) across all polyproteins (dark red) and for polyproteins from each of the six genotypes (color coded as given in the legend). The gray vertical lines demarcate the termini of individual proteins.

Figure 3. The size of intrinsically disordered regions (IDRs) in each HCV protein (plot A) and in each functional subclass (plot B) over the considered 32 polyproteins. Proteins are sorted according to their order in the genotype. Functions are sorted by their count of IDRs, from high to low value. Bars are color-coded to denote size of IDRs, from blue for short IDRs to red for long IDRs.

Figure 4. Evaluating intrinsic disorder in HCV proteins by CH-CDF plot. Here, the individual proteins are presented as points, coordinates of which are calculated as a distance of the corresponding protein in the CH-plot from the boundary (Y-coordinate) and an average distance of the respective CDF curve from the CDF boundary (X-coordinate). The four quadrants correspond to the following predictions: Q1, proteins predicted to be disordered by CH-plots, but ordered by CDFs; Q2, ordered proteins (N); Q3, proteins predicted to be disordered by CDFs, but compact by CH-plots (i.e., putative molten globules or hybrid proteins); Q4, proteins predicted to be disordered by both methods.

Figure 5. Functional annotations of the putative IDR regions found in the HCV polyproteins. Each function is characterized by the corresponding fraction of IDRs (black bars) and genotypes (black lines) that were annotated with this function. The annotations were generated based on protocol from refs.[53, 77]

Figure 6. Localization of the 12 putative functions of IDRs in the ten HCV proteins over the six genotypes. The red line at the top of the figure denotes putative disordered regions derived as a union of the disordered regions across the six genotypes. Annotations of the 12 functions discussed in Figure 5 for individual IDRs are shown horizontally below the red line. Each functionally annotated IDR is assigned to the corresponding genotypes: g1, g2, g3, g4, g5, and g6, or to all genotypes. The lines that denote the functionally annotated IDRs are colored in black if the annotation was found across multiple genotypes, and gray otherwise. The dashed gray vertical lines demarcate the termini of each protein.

Figure 7. Interactions between HCV and cellular proteins. HCV-human protein-protein interactions (grey lines) were downloaded from the HCVPro database[184] and deduced from references.[185, 186] Black spheres represent HCV proteins, white spheres indicate cellular proteins that bind to HCV proteins and red spheres signify human proteins that bind to HCV proteins and that were implicated in HCV replication by at least one large-scale siRNA screen. Green lines represent interactions between cellular proteins that bind to HCV proteins. The total number of cellular proteins that bind to each HCV protein is shown in parentheses beneath the protein name. For simplicity, the names of cellular proteins that were not siRNA hits and interactions between cellular proteins that bound to different HCV proteins were excluded.

Table 1. Enrichment of intrinsic disorder in functionally annotated regions of HCV polyproteins. We report ratios of the rates of occurrence of intrinsically disordered residues between a given type of functional region and overall in the entire polyprotein. Ratios above/below 1 denote enrichment/depletion of intrinsically disordered residues a given type of functional regions, e.g. ratio of 2.11 for DNA binding regions in genotype 1 means that disordered residues occur 2.11 times more often in the DNA binding regions compared to their rate in the whole HCV polyprotein. We include the following functional annotations: amino acid polymorphisms (AAP) sites, transmembrane regions, protein-binding regions, DNA-binding regions, RNA-binding regions, post-transcriptional modification (PTM) sites, and MoRF regions. The ratios shown in bold font identify values > 1.3 and in italics values < 0.7. The results are reported per each genotype and average over these genotypes.

HCV genotype	Enrichment in intrinsic disorder						
	AAP sites	Transmembrane regions	Protein-binding regions	DNA-binding regions	RNA-binding regions	PTM sites	MoRF regions
1	0.94	*0.41*	**2.11**	**2.67**	**2.77**	**4.02**	**2.56**
2	1.05	*0.28*	**1.99**	**2.46**	**2.24**	**3.77**	**2.44**
3	0.94	*0.17*	**2.07**	**2.90**	**3.48**	**3.49**	**2.47**
4	1.00	*0.15*	**2.02**	**2.45**	**1.91**	**3.48**	**2.55**
5	0.89	*0.27*	**1.84**	**2.14**	**1.58**	**4.19**	**2.52**
6	1.05	*0.36*	**2.19**	**2.60**	**2.51**	**3.54**	**3.22**
Average over 6 genotypes	0.98	*0.27*	**2.04**	**2.54**	**2.42**	**3.75**	**2.63**

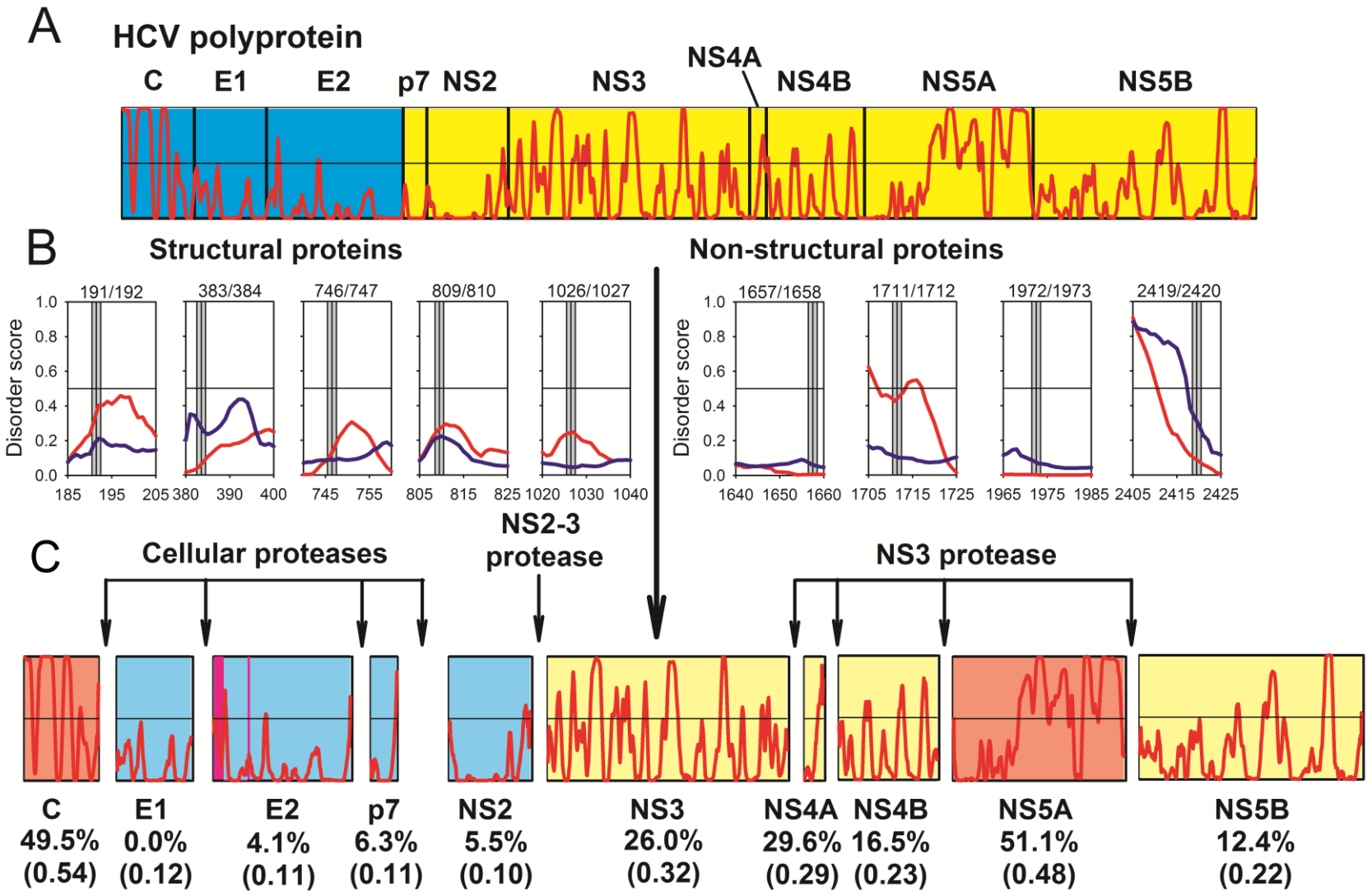


Figure 1

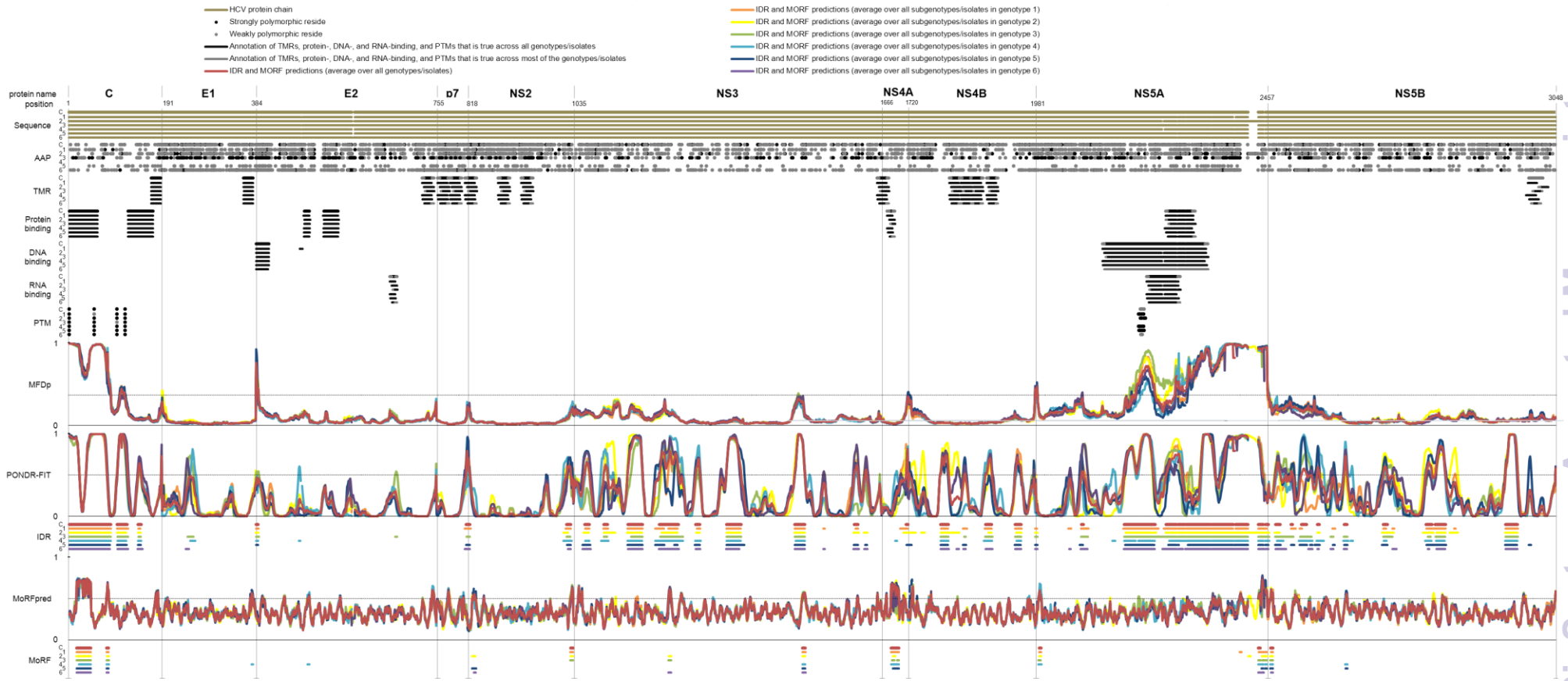


Figure 2

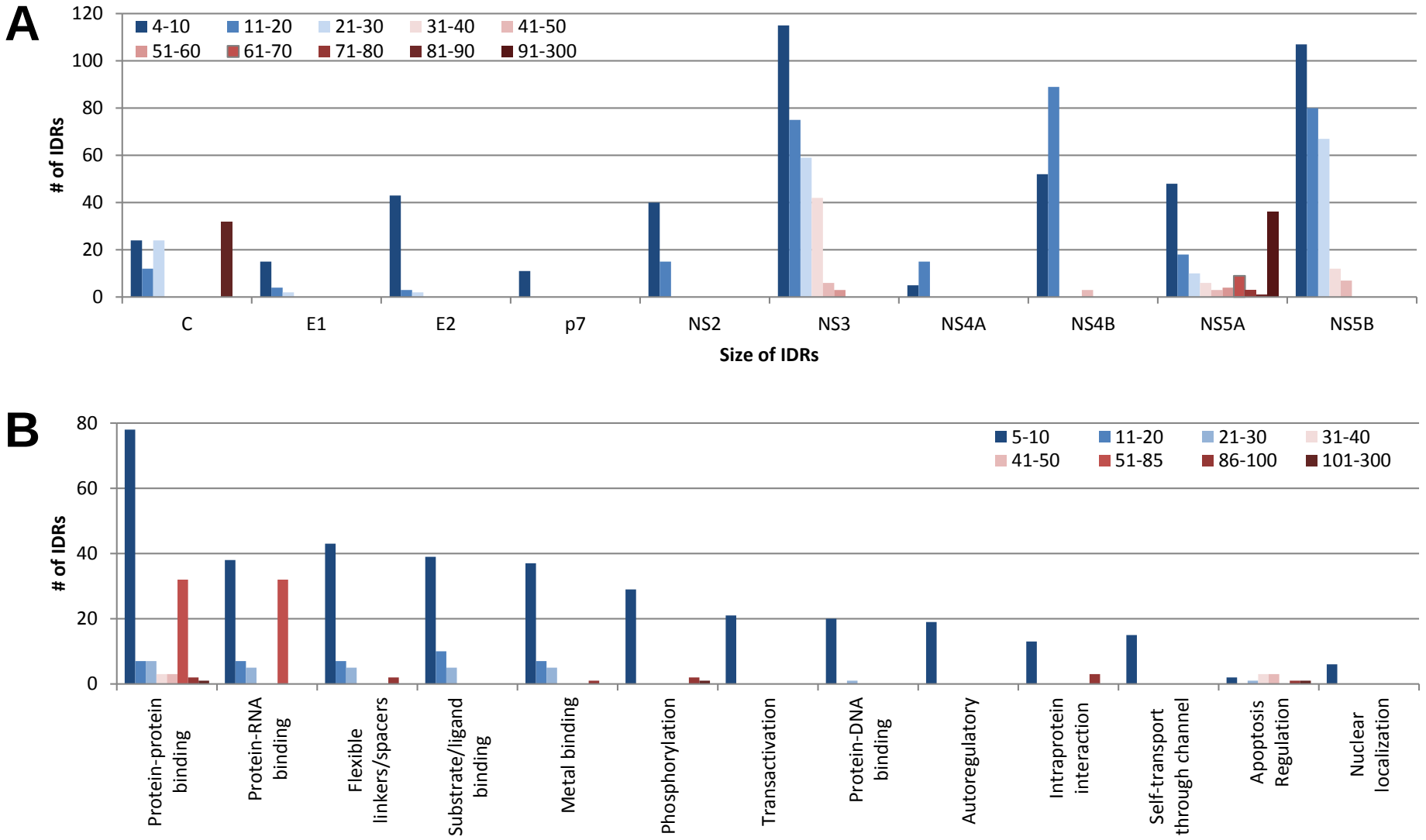


Figure 3

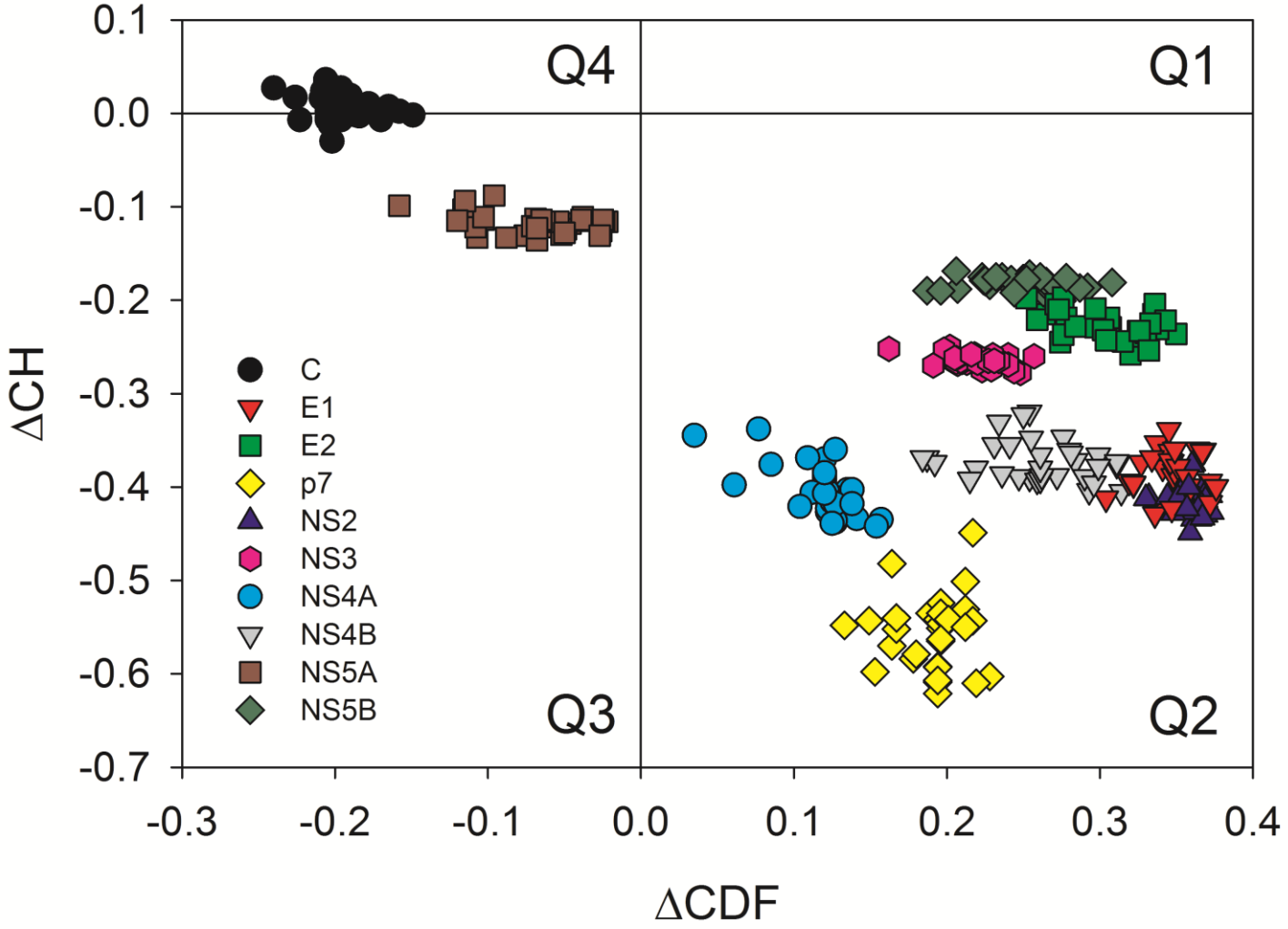


Figure 4

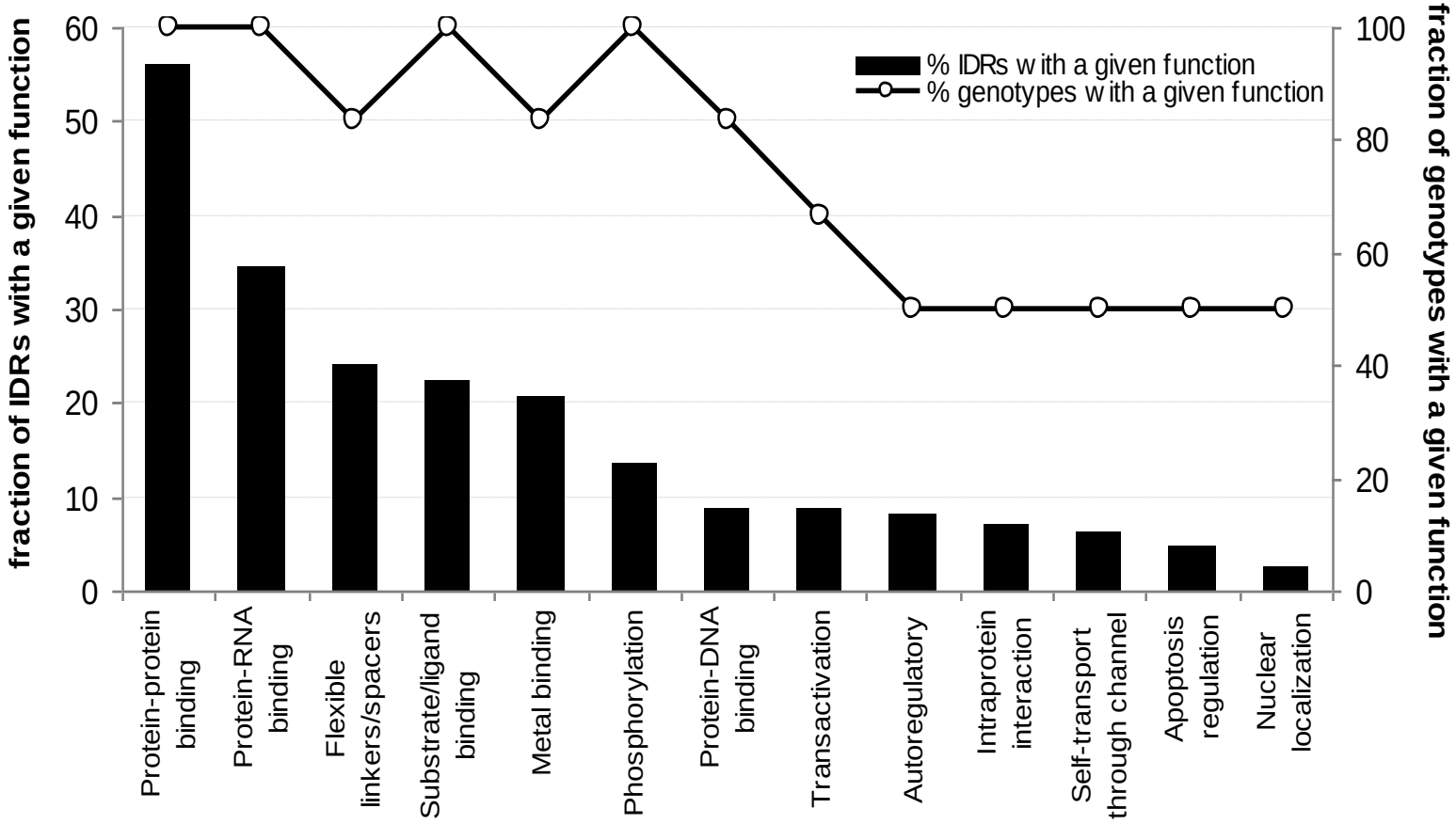


Figure 5

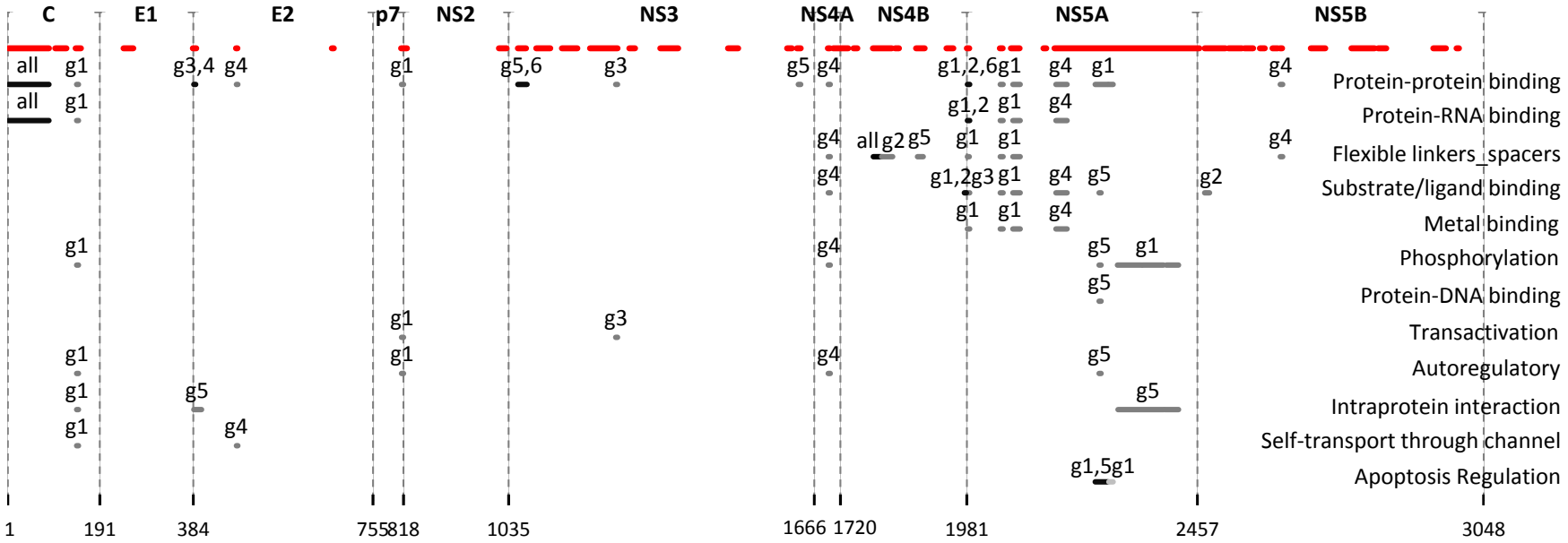


Figure 6

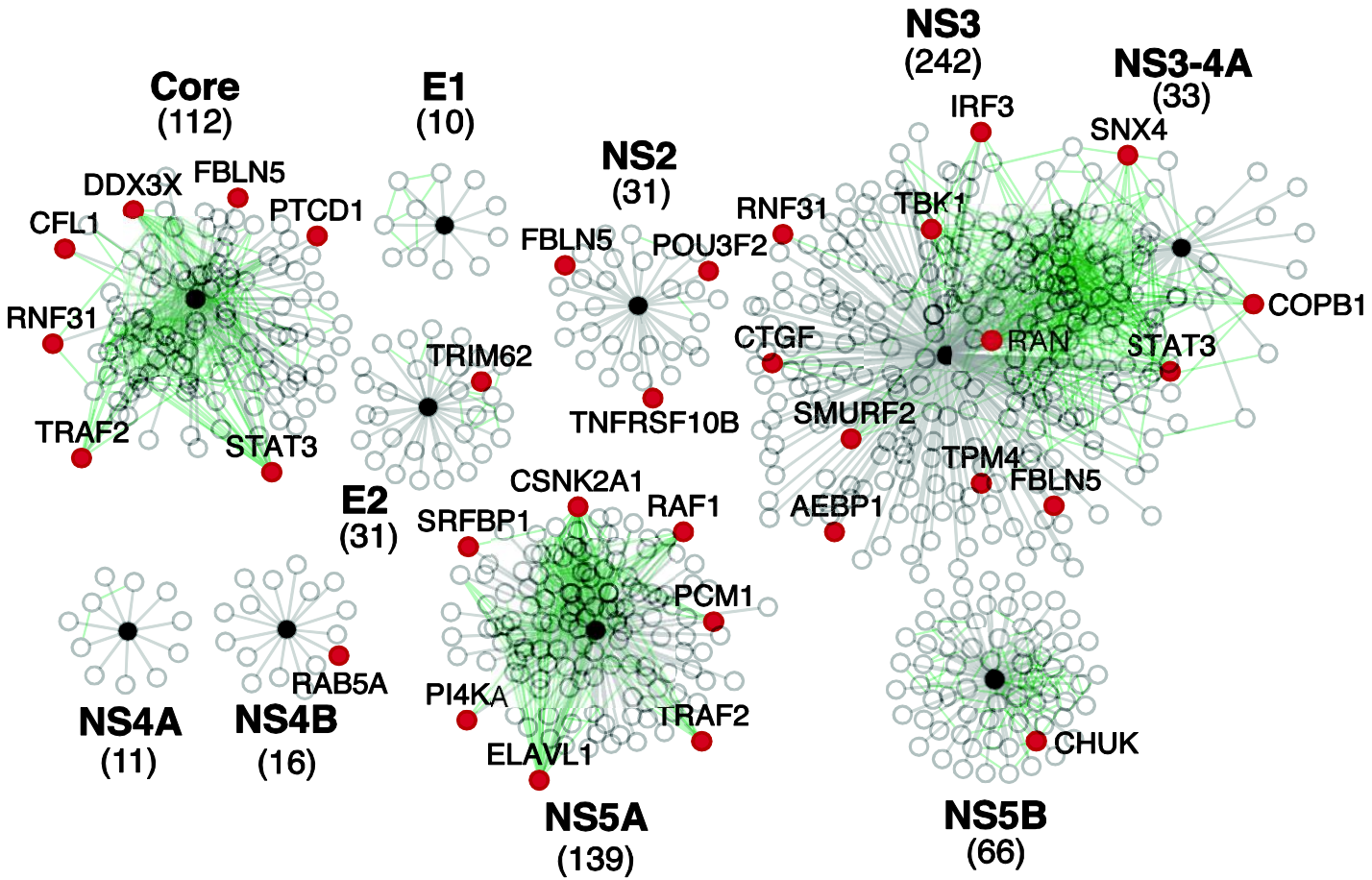


Figure 7