JCI The Journal of Clinical Investigation

Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans

Rajeev Aurora, … , John E. Tavis, the Virahep-C Study Group

J Clin Invest. 2009;119(1):225-236. https://doi.org/10.1172/JCI37085.

Technical Advance

Hepatitis C virus (HCV) is a common RNA virus that causes hepatitis and liver cancer. Infection is treated with IFN-α and ribavirin, but this expensive and physically demanding therapy fails in half of patients. The genomic sequences of independent HCV isolates differ by approximately 10%, but the effects of this variation on the response to therapy are unknown. To address this question, we analyzed amino acid covariance within the full viral coding region of pretherapy HCV sequences from 94 participants in the Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C) clinical study. Covarying positions were common and linked together into networks that differed by response to therapy. There were 3-fold more hydrophobic amino acid pairs in HCV from nonresponding patients, and these hydrophobic interactions were predicted to contribute to failure of therapy by stabilizing viral protein complexes. Using our analysis to detect patterns within the networks, we could predict the outcome of therapy with greater than 95% coverage and 100% accuracy, raising the possibility of a prognostic test to reduce therapeutic failures. Furthermore, the hub positions in the networks are attractive antiviral targets because of their genetic linkage with many other positions that we predict would suppress evolution of resistant variants. Finally, covariance network analysis could be applicable to any virus with sufficient genetic variation, including most human […]

Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans

Rajeev Aurora,[1] Maureen J. Donlin,[1,2] Nathan A. Cannon,[1] and John E. Tavis[1,3] for the Virahep-C Study Group

[1]Department of Molecular Microbiology and Immunology, [2]Department of Biochemistry and Molecular Biology, and [3]Saint Louis University Liver Center, Saint Louis University School of Medicine, St. Louis, Missouri, USA.

Hepatitis C virus (HCV) is a common RNA virus that causes hepatitis and liver cancer. Infection is treated with IFN-α and ribavirin, but this expensive and physically demanding therapy fails in half of patients. The genomic sequences of independent HCV isolates differ by approximately 10%, but the effects of this variation on the response to therapy are unknown. To address this question, we analyzed amino acid covariance within the full viral coding region of pretherapy HCV sequences from 94 participants in the Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C) clinical study. Covarying positions were common and linked together into networks that differed by response to therapy. There were 3-fold more hydrophobic amino acid pairs in HCV from nonresponding patients, and these hydrophobic interactions were predicted to contribute to failure of therapy by stabilizing viral protein complexes. Using our analysis to detect patterns within the networks, we could predict the outcome of therapy with greater than 95% coverage and 100% accuracy, raising the possibility of a prognostic test to reduce therapeutic failures. Furthermore, the hub positions in the networks are attractive antiviral targets because of their genetic linkage with many other positions that we predict would suppress evolution of resistant variants. Finally, covariance network analysis could be applicable to any virus with sufficient genetic variation, including most human RNA viruses.

Introduction

HCV chronically infects about 3.8 million Americans and causes 8,000–10,000 deaths each year in the United States by inducing liver failure or hepatocellular carcinoma (1). HCV infection is treated with a combination of pegylated IFN-α and ribavirin. Treatment for 24–48 weeks clears the virus — referred to herein as sustained viral response (SVR) — in 50%–60% of genotype 1 patients (2, 3). IFN-α provides the primary antiviral effect and can clear HCV when used alone (4, 5). When ribavirin is taken with IFN-α, it roughly doubles the clearance rate (4–6). There are no effective therapies for patients who fail to clear virus following IFN-α plus ribavirin therapy, and the reasons for the high rate of therapeutic failures are unknown.

HCV is a hepatotropic Flavivirus that persistently infects hepatocytes and some lymphocytes (reviewed in ref. 7). The virus is composed of a lipid envelope derived from host cell membranes in which the viral glycoproteins E1 and E2 are embedded. Within the envelope is a capsid formed by the viral core protein that surrounds the viral RNA genome. The approximately 9,600-nt positive-polarity RNA genome is translated to produce a polyprotein of about 3,010 amino acids (Figure 1). The polyprotein is cleaved to produce 10 viral proteins. The nonstructural proteins, P7 through NS5B, replicate the viral RNA on modified host membranes. After genomic replication, virions assemble and are secreted from the cell noncytolytically. It is believed that all HCV proteins form multiprotein complexes during viral replication and/or during assembly of the virion (8, 9). Furthermore, many HCV proteins interact with cellular proteins to modify host cell responses, particularly the type 1 IFN response that is induced by viral replication and whose role is to suppress viral persistence and spread (10).

The HCV genome is highly variable, with 6 genotypes that are less than 72% identical at the nucleotide level (11–14). Within the genotypes, subtypes with nucleotide identities of 75%–86% may occur. Individual isolates of a given subtype typically differ by about 8%–10%, and, as HCV replicates as quasispecies, multiple variants differing by up to a few percent exist even within individual patients. Viral genetic variation at the genotype level clearly affects the outcome of antiviral therapy, because genotype 1 is cleared by IFN-α and ribavirin only about half of the time, whereas response rates for genotypes 2 and 3 are typically greater than 80% (3, 15). However, the effect of HCV sequence variation within the major genotypes on response to therapy is not understood.

The Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C) clinical study previously investigated the efficacy of pegylated IFN-α plus ribavirin for treating genotype 1 HCV (16). As part of the Virahep-C study, we analyzed viral genetic patterns associated with response or failure of therapy (17, 18). We sequenced the complete pretreatment HCV open reading frame (ORF) from 94 patients and found that HCV genetic variability among sequences from patients in whom therapy efficiently suppressed the virus was significantly higher than among sequences from patients in whom suppression was minimal. We interpreted the association of higher interpatient HCV genetic diversity with response to therapy to imply that HCV survived in the nonresponders because there were only a few ways to optimize activity of the viral proteins, but many ways to interfere with their function.

Conflict of interest: The authors have declared that no conflict of interest exists.

Nonstandard abbreviations used: OMES, observed minus expected squared; ORF, open reading frame; SVR, sustained viral response; Virahep-C, Viral Resistance to Antiviral Therapy of Chronic Hepatitis C [study].

Citation for this article: *J. Clin. Invest.* **119**:225–236 (2009). doi:10.1172/JCI37085.

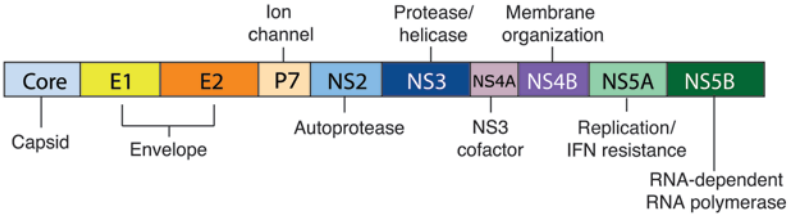

Figure 1
The HCV ORF. The approximately 9,600-nt positive-polarity HCV RNA genome encodes a long ORF. The ORF is translated into an approximately 3,010–amino acid polyprotein that is cleaved to 10 mature proteins. The known or predicted functions of the proteins are indicated.

Our previous analysis of the Virahep-C sequences treated variation at each amino acid position as being independent from all others. However, this is an oversimplification because amino acids interact with other residues, both within a given protein and in other proteins. Hence, we hypothesized that a global analysis of the viral genome in the context of response to therapy would provide new insights on how the virus may evade the selective pressures applied by the drugs.

To test this hypothesis, we analyzed the complete protein coding region of the pretherapy Virahep-C HCV sequences for pairs of amino acid positions that varied in concert among independent viral isolates (i.e., covariance) and then asked whether there were differences in the patterns of these covariances in responders and nonresponders to therapy. Covariance analysis has previously been used to identify residue pairs directly interacting in 3-dimensional structures (19–24), to infer protein/protein interactions (25), to examine allosteric interactions within proteins (26, 27), and to evaluate the relative functional importance of residues in proteins (28). Our genome-wide analysis revealed the existence of genetic interactions, which we believe to be novel, that were interwoven through the HCV genome. Importantly, these interactions were very different in responders and nonresponders to IFN-based therapy; hence, they may permit prediction of the outcome of therapy.

Results

Virahep-C cohort and HCV sequences. The consensus pretreatment sequences for the full-length HCV ORF were previously determined from 94 participants in the Virahep-C study (17). The characteristics of these patients are shown in Table 1. Most analyses were performed on all 94 sequences, stratified either by genotype (1a versus 1b) or by genotype plus outcome of therapy. In most cases, we also stratified the samples by genotype plus the extremes of early (day 28) response to therapy, in order to eliminate confounding nonbiological effects on changes in viral titre (such as insufficient drug intake). The day-28 response categories were Marked, Intermediate, and Poor (see Methods). The day-28 stratifications used the 63 Marked and Poor sequences. Table 2 shows the number of samples in the day-28 and treatment outcome classes for genotypes 1a and 1b.

Identification of covarying amino acid positions. To identify covarying amino acid positions in the HCV ORF, we first created 10 multiple sequence alignments of the Virahep-C sequences: 1a All, 1a Marked, 1a Poor, 1a SVR, 1a Non-SVR, 1b All, 1b Marked, 1b Poor, 1b SVR, and 1b Non-SVR. A covariance score for every possible pair of the 2,955 positions in each alignment was calculated by squaring the difference between the number of observed and expected amino acid pairs and normalizing this difference by the number of entries (excluding gaps) in each column, the observed minus expected squared (OMES) method (29). The null model in this method is the expected number of covarying pairs, based on the independent count of the amino acids at each of the 2 positions. Covarying positions were defined as those pairs with scores of at least 0.5, corresponding to a difference of at least 3 covarying pairs between the observed and expected in an alignment of 16 samples. This cutoff was chosen because it was the lowest value that was not greatly influenced by noise, and it allowed us to retain the maximum amount of potentially informative data (see Methods). The covarying positions and amino acid pairs are shown in Supplemental Table 1 (supplemental material available online with this article; doi:10.1172/JCI37085DS1).

There were 246 covarying positions in genotype 1a and 280 in genotype 1b (Table 2), representing about 10% of the 2,955 columns in the alignment for each genotype. The covarying positions were spread throughout the genome, with positions in each of the 10 viral genes covarying with positions in each of the other 9 genes. These covarying positions, with the exception of position 1,583, are different than the adaptive mutations required for efficient HCV RNA replication in the replicon culture system (30–32). Furthermore, there was a pronounced underrepresentation among the covariances of residues that have been demonstrated through molecular or biochemical analyses to be essential for protein function. This is because covariance by definition requires genetic variation; thus, very highly conserved positions were excluded from this analysis.

The difference between the number of covarying positions in the Marked and Poor response classes (202 vs. 172 for 1a and 265 vs. 195 for 1b; Table 2) was not significant for either genotype by Fisher's exact test ($P > 0.1$). Similarly, no significant differences in the number of covarying positions were found when the SVR and Non-SVR

Table 1
Baseline characteristics of the Virahep-C cohort

Characteristic	Marked ($n = 31$)	Intermediate ($n = 31$)	Poor ($n = 32$)	*P*
AA, no. (%)	16 (51.6%)	15 (48.4%)	16 (50.0%)	0.97[A]
Male, no. (%)	22 (71.0%)	25 (80.7%)	21 (65.6%)	0.40[A]
Age, yr	46.5 (6.2)	48.1 (6.7)	49.4 (8.8)	0.30[B]
Body wt, kg	84.9 (16.8)	88.6 (14.6)	91.0 (13.5)	0.28[B]
HCV RNA, $\log_{10}$ IU/ml	5.9 (0.9)	6.5 (0.6)	6.4 (0.5)	0.003[C]
ALT, U/l	79.0 (52.3)	70.6 (36.3)	86.3 (43.6)	0.17[C]
Albumin, g/dl	4.1 (0.4)	4.2 (0.3)	4.2 (0.3)	0.29[C]
Ishak necroinflammatory[D]	7.2 (2.5)	7.4 (3.0)	7.9 (2.7)	0.52[C]
Ishak fibrosis[E]	1.8 (1.3)	1.8 (1.3)	2.3 (1.4)	0.15[C]

Data adapted from ref. 17. Unless otherwise indicated, data are reported as mean (SD). AA, African American. [A]χ^2 test. [B]ANOVA. [C]Kruskal-Wallis; missing = 1. [D]Necroinflammatory score, graded on a 0–18 scale. [E]Fibrosis score, graded on a 0–6 scale.

Table 2
Network characteristics

Genotype/class	No. samples	Nodes[A]	Edges[B]	Density[C]
1a All	47	246	3,080	0.1022
1a Marked	16	202	1,850	0.0910
1a Poor	16	172	2,119	0.1441
1a SVR	22	223	2,136	0.0863
1a Non-SVR	25	217	1,796	0.0766
1b All	47	280	2,189	0.0560
1b Marked	15	265	1,490	0.0425
1b Poor	16	195	1,237	0.0654
1b SVR	26	262	1,512	0.0442
1b Non-SVR	21	221	1,740	0.0716

[A]Positions in the sequence alignments that covary. [B]Amino acid pairs that covary; each node may have multiple edges. [C]Number of edges per node.

sequences were compared (223 vs. 217 for 1a, 262 vs. 221 for 1b). Therefore, covarying positions were common and widespread in the HCV genome, but there were no significant differences in the number of covarying positions between the response classes.

Covarying positions form networks. Inspection of the covarying positions revealed that one member of a covarying pair often covaried with one or more other positions. This led us to hypothesize that the covarying positions may be linked into a network. To test this hypothesis, we performed a complete clustering analysis on an alignments of all 47 sequences for each genotype, 1a and 1b. The results were presented as graphs, in which the nodes represent positions in the sequence alignment and pairs of nodes were connected by a line (referred to herein as an *edge*) if they had a covariance score of at least 0.5. We found that the covarying positions indeed formed networks (Figure 2, A and C, and Table 2). Both the networks followed the inverse power law distribution, in which the probability that any node has k edges is given by the equation $\Pr(k) = k^{-\gamma}$ (33, 34). For 1a, γ equaled 0.98 (Figure 2B), and for 1b, γ equaled 1.1 (Figure 2D). This finding indicates that the networks had hub-and-spoke architectures, in which a few nodes covaried with many others, but most nodes covaried with only few others. The covariances (edges) between amino acid positions that were most highly connected (hubs) had scores at or near the maximal value possible for alignments of this size. This indicates that there were very strong selective pressures for certain amino acid combinations and against other combinations at these pairs of positions.

To evaluate the generality of these networks, we sought to determine whether the covariance networks found in the Virahep-C sequences were representative of networks derived from HCV sequences in general circulation. A set of 118 full-length genotype 1a sequences from non–Virahep-C patients was collected, and their polyprotein amino acid sequences were deduced. We randomly chose 10 sets of 47 amino acid sequences from the set of 118, and each set was aligned and independently subjected to covariance analysis. The distribution of the covariance scores in the 10 random sets of 47 sequences were compared in a pairwise manner by the Kolmogorov-Smirnov test, and the P values ranged from 0.23 to 0.38 (P = NS). This indicates that the score distributions were statistically indistinguishable among the 10 random sequence sets. The distribution of covariance scores in the Virahep-C 1a All network (47 sequences) was then compared with the distribution in each of the 10 random permutations. These P values ranged from 0.23 to 0.32. Therefore, the distribution of covariance scores within the Virahep-C data set was very similar to that in non–Virahep-C data sets of equivalent size. In addition, the number of nodes, the number of edges, and the edge density were all similar among the 11 sequence sets, and the top 4 hub nodes were the same for all 11 sets. Therefore, the covariances and covariance networks found in the Virahep-C sequences were representative of a randomly chosen set of HCV sequences.

Connectivity of the covariance networks differs by response class. We next examined the characteristics of the networks that were generated from the sequences when they were stratified by response of the patient to antiviral therapy (Marked, Poor, SVR, or Non-SVR). The characteristics of the 8 response-specific covariance networks are shown in Table 2, and the 1a Marked and 1a Poor networks are shown in Figure 3, A and C. The networks all had a hub-and-spoke architecture, with γ ranging about 1.0 to 1.2, and the numbers of nodes and edges were similar between the contrasting response classes (Marked versus Poor and SVR versus Non-SVR; Table 2). Similarly, more than half of the amino acid positions that formed the nodes in the networks were shared between the contrasting response classes (Figure 4, A and B). However, the pairs of positions forming the networks from the contrasting phenotypes were very different, with relatively few edges shared between them (Table 3 and Figure 4, C and D). This difference in connectivity among the nodes led to differences in the identity of the most highly connected hubs between the networks for the various response classes (Table 4). For example, the positions that were most highly interconnected for 1a were within NS2, NS4A, and E2 in the Marked responders, but in P7, E2, and NS5A in the Poor responders. Furthermore, most nodes within the Poor and Non-SVR nonresponder networks were more highly connected than were most nodes in the Marked and SVR responder networks. An example is in Figure 3, B and D, which compare first-neighbor networks for residue 463 in the 1a Marked and 1a Poor networks. Therefore, networks with similar numbers of covarying amino acid positions were found in all response classes, but the patterns of connections among the covarying residues were very different in the contrasting response classes.

To test the possibility that the differences between the Marked and Poor networks may have occurred by chance, we generated alignments in which we randomly exchanged sequences between the Marked and Poor responders. The covariance analysis was repeated for 10 iterations at 3 levels of randomly exchanged sequences: 12.5% (2 of 16 genomes), 25% (4 of 16 genomes), and 37.5% (6 of 16 genomes). An example of 1a Marked responders shuffled with 1a Poor responders is shown in Figure 5. Exchanging 2 sequences in the shuffled networks led to the loss of approximately 15% of the nodes and edges (e.g., the residue positions and their interactions) found in the unshuffled alignments. At 4 sequences shuffled, the number of edges and nodes decreased further, and when 6 sequences were shuffled, there was a large loss of the original nodes and edges. In every case, the covarying positions that were present in the shuffled alignments were also found in the All alignment. The results of these control analyses indicate that (a) the response-specific networks are not so sensitive to replacement of sequences that they exist on the edge of chaos; and (b) although the networks are not hypersensitive to mixing of the phenotypic classes, they do depend on the phenotypic clustering of the sequences for their integrity. Together, these observations indicate that the covarying

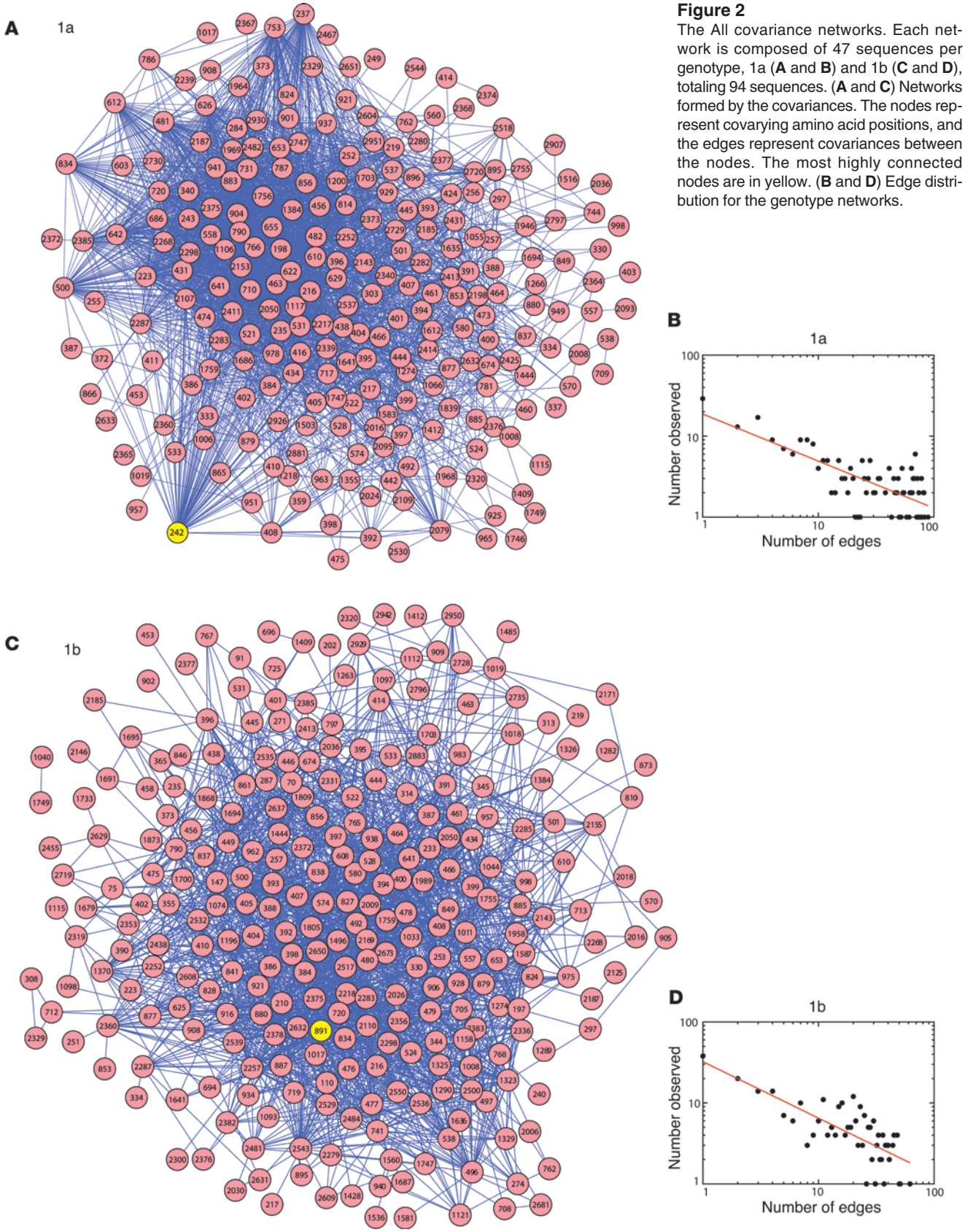

Figure 2
The All covariance networks. Each network is composed of 47 sequences per genotype, 1a (**A** and **B**) and 1b (**C** and **D**), totaling 94 sequences. (**A** and **C**) Networks formed by the covariances. The nodes represent covarying amino acid positions, and the edges represent covariances between the nodes. The most highly connected nodes are in yellow. (**B** and **D**) Edge distribution for the genotype networks.

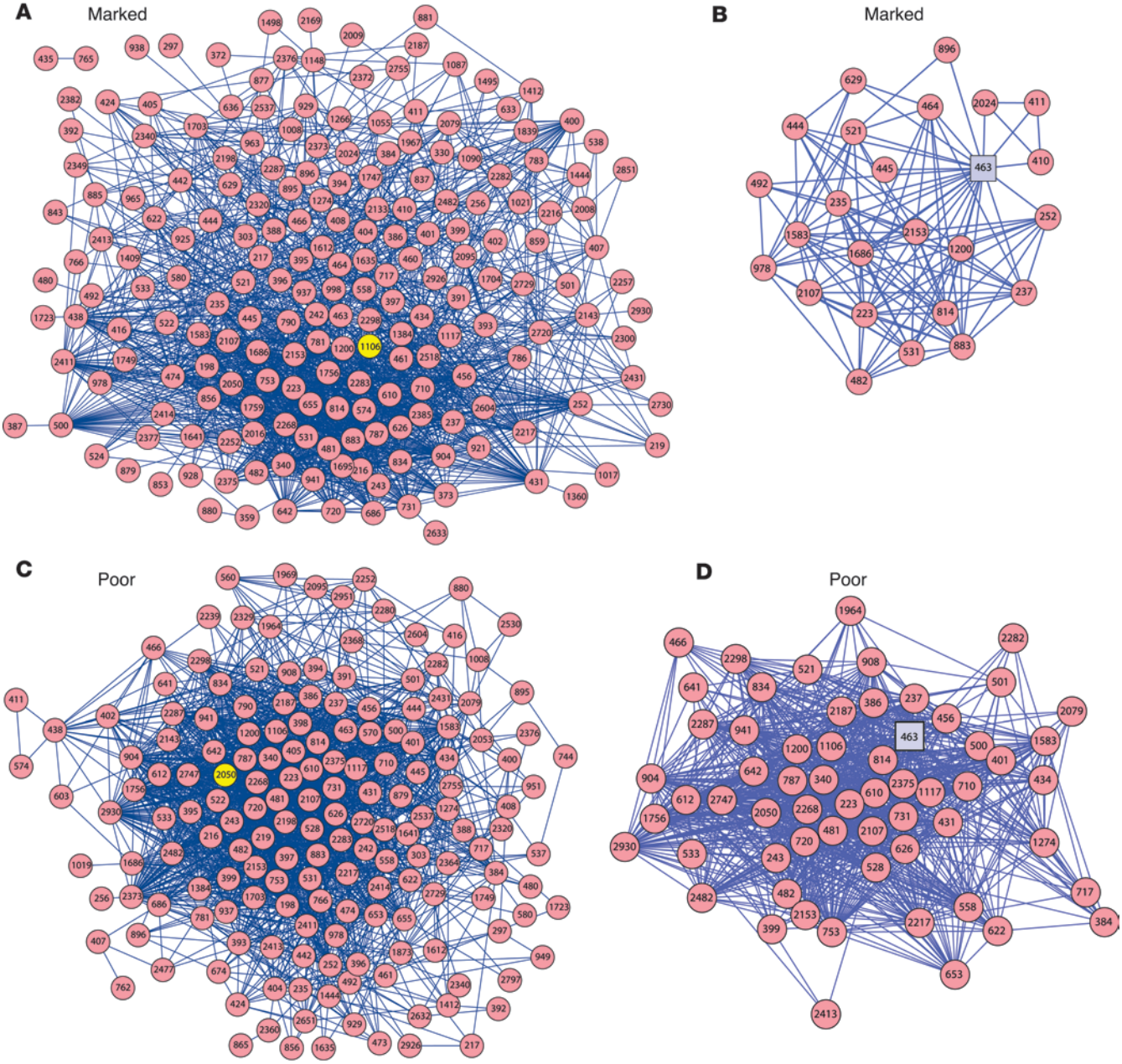

Figure 3
The connectivity of the response-specific covariance networks is different. Shown are 1a Marked (**A** and **B**) and 1a Poor (**C** and **D**) response classes. (**A** and **C**) Networks formed by the covariant pairs of residue positions. The most highly connected nodes are in yellow. (**B** and **D**) Positions that directly covary with position 463 (square node) are shown to highlight the differences between the networks.

pairs identified for the response classes were not generated by random chance and therefore reflect a feature of the viruses infecting the patients in the various response classes.

HCV replicates as a quasispecies, but our analyses were performed with the consensus sequences (the most common residue at each position) from each individual; consequently, the quasispecies variation within an individual could affect the network. The shuffling experiment in Figure 5 revealed that the networks could withstand replacement of at least 25% of the sequences with others that vary by about 10%. The quasispecies variation within an individual is typically approximately 1%, but can reach up to about 4% in some patients; hence, the networks were tolerant of variation that was considerably larger than the quasispecies variability typical for an individual. In addition, analysis of the frequency of each degenerate nucleotide position in the codons for the top 5 hub positions for the Marked and Poor sequences (Table 4) revealed that the probability that a given viral RNA molecule in the quasispecies spectrum encoded all 5 of the hub consensus sequence residues was 93% for the Marked sequences and 82% for the Poor sequences. Finally, near–full-length HCV quasispecies

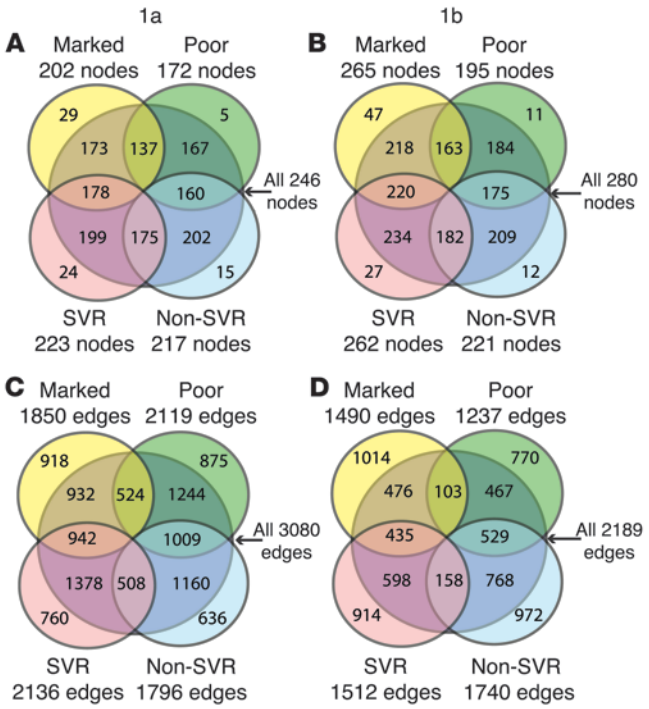

Figure 4
Segregation of the edges and nodes by phenotype. The overlap and segregation of the covarying nodes (**A** and **B**) and edges (**C** and **D**) by response class is shown for genotype 1a (**A** and **C**) and 1b (**B** and **D**).

variants have previously been analyzed in 2 genotype 1a patients, with 6 variants sequenced per patient (35, 36). In one patient, the residues at the top 5 hubs from our All network were identical in the consensus sequence derived from the quasispecies variants and in all 6 of the variants, and in the other patient there was a single substitution at one hub position in one of the 6 quasispecies variants. Therefore, the central portion of the All network was perfectly conserved in 11 of the 12 (92%) quasispecies variants for which the integrity of the networks can be assessed.

The spectrum of HCV isolates circulating in the human population appears to range from relatively resistant to relatively sensitive sequences in their responses to IFN-based therapy (17). Therefore, if the covariance networks accurately reflect the HCV population as a whole, essentially every covariance in the All networks — composed of the full set of 47 sequences for each genotype — should be present in at least one of the response-specific networks (i.e., Marked, Poor, SVR, or Non-SVR). To test this prediction, the All networks, generated from alignments of all 47 1a or 1b sequences, were compared with the response-specific networks. As predicted, every node in the All networks was in at least one of the response-specific networks (Figure 4, A and B). Similarly, every edge in the All networks was also found in one or more of the response-specific networks (Figure 4, C and D). This indicates that no new interactions appeared by chance in the full set of sequences and that the information in the All network was also found within the subsets segregated by treatment response.

Topological assessment of the covarying positions. Genetic covariance is indicative of a functional interaction between the covarying residues. Covariance interactions are often caused by direct binding between the residues, but other interactions, such as compensatory allosteric changes, are also common. HCV replication and assembly involve multicomponent complexes composed of viral and possibly cellular proteins. Furthermore, these processes take place in association with cytoplasmic membranes (8, 9); hence, all or nearly all HCV proteins are membrane associated. Therefore, to evaluate whether the covarying residues could interact directly with each other, we evaluated their topological orientation relative to cellular membranes using the experimentally known orientations or inferred orientations for those residues lacking experimental data (8). In 1a, we found that for the 714 covarying amino acid pairs in which both residues were within the same protein, 672 were in the same compartment, 41 were in adjacent compartments (e.g., cytosol and transmembrane), and 1 was in nonadjacent compartments (e.g., lumen and cytosol). For 1b, there were 460 pairs in the same compartment, 38 in adjacent compartments, and 10 in nonadjacent compartments among the 508 covarying pairs in the same protein. When this analysis was expanded to include all interactions in the networks, we found that nearly 75% of the pairs occurred in the same or adjacent compartments for both subtypes. This indicates that many of the covarying pairs could be in direct contact with each other and/or interact with a common partner.

Structural assessment of covarying positions. The possibility that some of the covarying residues may contact each other can be tested for all or part of NS2, NS3, NS5A, and NS5B because crystal structures are available for these proteins. To this end, we mapped the covariant amino acid pairs within these proteins onto the structures. For 1a, there were 39 covarying pairs within the NS2 crystal structure. The covarying residues were on the solvent-accessible surface of the protein for 29 of the pairs, but none of the paired residues were close enough to bind to each other (≤7.5 Å). For 1b, there were 88 pairs in the structure. Of these, 80 were solvent exposed, but the residues in only 1 pair were within 7.5 Å of each other. For NS3, there were 26 covariant pairs for 1a and 32 for 1b. Although all of these pairs were on solvent-exposed surfaces, none of the paired residues were within 7.5 Å of each other. No intraprotein pairs were found within the N-terminal third of NS5A that is in the crystal structure. There were 21 covarying pairs within 1a NS5B and 67 pairs in 1b, and all of these were on the surface of the protein, but again, none of the covarying residues were within 7.5 Å of each other. Therefore, the large majority of the covarying residues in the available protein structures cannot bind directly to each other. However, most of the covarying residues are on the solvent-exposed surfaces of their respective proteins, where they may be involved in intermolecular interactions with other cellular or viral components.

X-ray cocrystal structures are needed to provide sufficient resolution to evaluate the possibility of direct intermolecular binding between covarying residues, and the only such data currently avail-

Table 3
Segregation of edges by phenotype

Genotype/response	Responders[A]	Intersection	Nonresponders[B]
1a Day 28	1,850	524	2,119
1a Sustained	2,136	508	1,796
1b Day 28	1,490	103	1,237
1b Sustained	1,512	158	1,740

[A]Marked or SVR class. [B]Poor or Non-SVR class.

Table 4
Top 5 most-connected nodes by response class

Rank	All			Marked			Poor			SVR			Non-SVR		
	Node	Edges	Protein	Node	Edges	Protein	Node	Edges	Protein	Node	Edges	Protein	Node	Edges	Protein
1a genotype															
1	242	89	E1	1,686	57	NS4A	2,050	72	NS5A	216	66	E1	610	61	E2
2	482	83	E2	883	57	NS2	610	69	E2	481	65	E2	242	57	E1
3	753	82	P7	904	56	NS2	753	69	P7	1,756	65	NS4B	1,200	57	NS3
4	610	82	E2	655	56	E2	626	68	E2	814	64	NS2	463	56	E2
5	710	81	E2	481	56	E2	482	68	E2	753	63	P7	710	52	E2
1b genotype															
1	891	62	NS2	2,257	38	NS5A	524	31	E2	880	35	NS2	720	45	E2
2	887	51	NS2	856	37	NS2	891	29	NS2	434	35	E2	741	44	E2
3	1,496	50	NS3	2,143	36	NS5A	768	29	P7	608	33	E2	2,543	42	NS5B
4	720	48	E2	887	34	NS2	407	29	E2	2,009	33	NS5A	480	41	E2
5	479	48	E2	916	32	NS2	1,011	29	NS2	2,169	32	NS5A	466	40	E2

able are for the NS3/NS4A complex (NS4A is a peptide cofactor for NS3 that folds into the NS3 structure). There were 2 covariances in our data set between 1a NS3 and NS4A in which both residues were within the structure, residue pairs 1,106/1,686 and 1,117/1,686. Residues 1,106 (NS3) and 1,686 (NS4A) are separated only by an intervening protein chain; therefore, variations at these sites could easily communicate with each other through minor alterations to the local protein structure. Covarying residues 1,117 (NS3) and 1,686 (NS4A) are located far from each other in 3-dimensional space, and consequently a cascade of allosteric interactions through the protein would be needed for them to communicate with each other. Therefore, the limited data available indicate that the covariance interactions may involve both local and long-distance allosteric interactions.

Nonresponder networks have more hydrophobic pairs. To analyze the nature of the interactions between the covarying residues, we asked whether differences exist in the chemical nature of the covarying amino acid pairs within the networks. In both subtypes, hydrophobic pairs (i.e., Ile-Leu or Phe-Leu) were significantly more common in the Poor responders than in the Marked responders (3.2-fold in 1a and 3.5-fold in 1b; Table 5). Similar excesses of hydrophobic pairs were found in the Non-SVR networks relative to the SVR networks. The high frequency of hydrophobic amino acid pairs was a property of the covarying residue pairs in the networks, and not a feature of the sequences as a whole because there were no significant differences in the average number of hydrophobic amino acids between the Marked and Poor or SVR and Non-SVR sequences in either genotype (Table 5).

Identification of potential biomarkers for prediction of therapy outcome. The differences between the SVR and Non-SVR covariance networks imply that the patterns of covariances in the pretreatment sequences could be used to predict the outcome of therapy. Covariance by definition requires that the amino acid positions being compared are variable in a sequence set; consequently, only a fraction of the sequences will contain any given covarying pair. In our analysis, a given covarying pair was found in at most 60% of the sequences. Therefore, we asked whether subnetworks with a limited number of nodes could be identified that could be used as biomarkers to predict treatment outcome in a majority of patients.

To identify potentially predictive subnetworks, we began with each covariance in the SVR or Non-SVR networks and then added the nearest-neighbor covariances in a stepwise fashion, keeping covariances in the growing subnetworks if they increased the proportion of patients for which the subnetwork was accurately associated with outcome of therapy (i.e., coverage). This process was repeated exhaustively until the coverage for each growing subnetwork could no longer be increased. More than 64,000 subnetworks were created by this process. We then selected the subnetworks that were associated with outcome at 100% accuracy with 0% false positive associations (i.e., the covariances were found in the only the desired outcome class and never in the contrasting class), and then ranked them by their coverage to identify the sub-

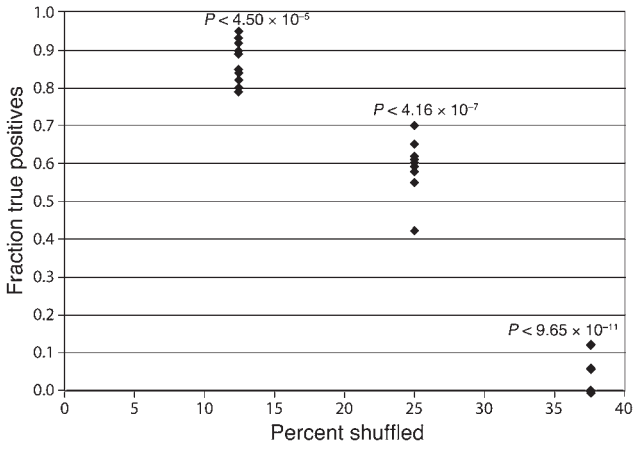

Figure 5
Random shuffling of sequences causes loss of information in the response-specific networks. We generated 10 independent alignments, in which 1a Marked sequences were randomly replaced with 1a Poor sequences at 3 levels of replacement: 2, 4, or 6 sequences, giving 12.5%, 25%, and 37.5% sequences shuffled, respectively. Shown are proportions of true positive covarying pairs relative to the unshuffled Marked sequences. *P* values showing significance of the differences in conserved edges relative to the unshuffled network were determined by Student's *t* test.

Table 5
Number of hydrophobic amino acids in the genome and hydrophobic residue pairs in the covariance networks

Class	Network pairs		Residues per genome	
	1a	1b	1a	1b
Poor	1,135	773	1,605 ± 3.6	1,599 ± 5.1
Marked	354	223	1,604 ± 6.4	1,603 ± 6.4
Total	3,445	2,624		
P	2×10^{-10}	2×10^{-16}		
Non-SVR	685	1,428	1,605 ± 3.4	1,599 ± 6.1
SVR	214	408	1,604 ± 6.0	1,602 ± 5.8
Total	3,424	3,094		
P	2×10^{-7}	4×10^{-16}		

Hydrophobic residues per genome are shown as mean ± SD. *P* values were calculated by Fisher's exact test.

networks that covered the maximal number of patients. Several hundred subnetworks were found that perfectly associated with outcome and had maximal coverage (95.5%–100%, depending on outcome class). We then limited the results to subnetworks composed solely of covariances with the 10% highest covariance scores to ensure that we were using the most robust regions of the networks. Figure 6 shows 3 examples for each response class. These subnetworks had 4–8 nodes, were 100% accurate, had no false associations, and covered 95.5%–100% of the samples. Therefore, we conclude that patterns of amino acid covariance in a handful of positions are closely associated with the outcome of antiviral therapy, and these covariances are potential biomarkers for prediction of the outcome of therapy.

Discussion

We previously found that HCV sequences from patients who responded well to pegylated IFN-α and ribavirin were more variable than were poor responders in genes implicated in counteracting the type 1 IFN response (17). We interpreted this to mean that viral isolates with a relatively tight genetic distribution around an optimum sequence were more able to withstand the pressures induced by therapy, and those that were more distant from this optimum were less able to survive, presumably as a result of the presence of multiple variations that each reduced the overall efficacy of the viral proteins. Here, we found that genome-wide networks of covarying amino acids existed, that the connections within the networks (connectivity) were different in the responders and nonresponders, and that the nonresponder networks had many more hydrophobic amino acid pairs than did the responder networks.

The covariance networks covered all 10 HCV proteins and all had hub-and-spoke architectures, which indicates that a few residues covaried with many other residues but that most covaried with only a few other positions. The network connectivity was very different between the Marked and Poor and between the SVR and Non-SVR response classes. Therefore, the genetic and functional interactions represented by the covariances in the response-specific networks may represent HCV genetic differences that affect the ability of the viruses to withstand the pressures of therapy. There was a large overlap in the covariances in the networks from the responder Marked and SVR classes, and a similar overlap was found in the nonresponder Poor and Non-SVR classes, for both genotypes (Figure 4). Therefore, the viral variables reflected in these networks that affect the day-28 response to therapy were similar to those affecting the outcome of therapy.

Genome-wide covariance analysis has very recently been used by Campo and colleagues to assess coordinated evolution of residues throughout the HCV genome (37). This work was performed independently of our analysis and used a different method to identify the covariances, but the results from the 2 studies were very similar. The algorithm used by Campo et al. assessed the physiochemical properties of residues at the 10% of most variable positions in an alignment of 114 genotype 1b HCV amino acid sequences. Similar to our results, the covariances they identified linked into a hub-and-spoke network that encompassed all 10 of the proteins encoded in the HCV polyprotein; this network was analogous to our 1b All network. Furthermore, many of the most highly connected hubs in the Campo network were also found in our 1b networks that were generated without regard to the physiochemical properties of the amino acids. Campo and colleagues concluded that the network was a tightly coordinated unit that was functionally and/or structurally connected (37), in full agreement with our present conclusions. Although the Campo sequences were not stratified by outcome of antiviral therapy, and thus their network cannot be used to evaluate differential sensitivity to IFN-α–based therapy, their results are important to our work because they provide an independent validation of the existence of an All network. By extension, they also support the validity of the response-specific networks, because every covariance and node in our All network was also found in one or more of the response-specific networks.

Genetic covariance indicates a functional interaction between the covarying residues, but it does not identify the nature of the interaction. The functional linkages could involve direct binding between the covarying residues, compensatory allosteric changes within a protein, and/or compensatory changes on the surface of the HCV proteins where they interact with host or other viral proteins. Examples of all of these mechanisms are likely to be present among the large number of covariances we identified, but a major mechanism by which the differences between the responder and nonresponder networks may contribute to differential response to IFN-α–based therapy was revealed by the chemical nature of the covarying residues. The covariance networks from the nonresponder Poor and Non-SVR classes had greater than 3-fold more hydrophobic residue pairs than did sequences from the responder Marked and SVR classes (Table 5). In contrast, the responders had many more hydrogen bond donors or acidic-basic residue pairs. Hydrophobic interactions contribute much more to protein stability in an aqueous environment than do hydrophilic interactions. Therefore, the potential for greater stability provided by the higher hydrophobic nature of the interactions may allow some of the viruses in the population to better survive the pressures introduced by therapy. However, because the covariant residues were rarely close enough to bind to each other directly, we predict that in most cases the increased hydrophobicity provided by the covariant pairs would stabilize multiprotein complexes rather than the structure of a given protein.

IFN-α activates a multitude of host barriers that limit the spread of infection (10), and ribavirin has at least 3 proposed effects against HCV (38). Therefore, it is highly unlikely that the generalized increase in the hydrophobic nature of the covarying residue pairs in viruses from nonresponders acts through a few discrete intermolecular interactions. Rather, the simplest explanation is

Genotype and response class	Subnetwork	Coverage
1a SVR	1583_{NS3}–401_{E2}–384_{E2}–885_{NS2}	21/22 (95.5%)
	885_{NS2}–384_{E2}–2153_{NS5A}–386_{E2}	21/22 (95.5%)
	1974_{NS5A} / 2375_{NS5A}–2282_{NS5A}–395_{E2}–384_{E2}–401_{E2} / 885_{NS2}	21/22 (95.5%)
1a Non-SVR	2518_{NS5B}–384_{E2}–401_{E2}–2079_{NS5A} / 2189_{NS5A}–395_{E2}	24/25 (96%)
	395_{E2}–386_{E2}–500_{E2}–2375_{NS5A}–626_{E2} / 528_{E2} / 2153_{NS5A}–242_{E1}	24/25 (96%)
	941_{NS2} / 522_{E2}–610_{E2}–400_{E2} / 404_{E2}–216_{E1} / 386_{E2}	24/25 (96%)
1b SVR	2009_{NS5A}–397_{E2}–478_{E2}–444_{E2}–384_{E2}	26/26 (100%)
	476_{E2}–75_{core}–1805_{NS4B}–444_{E2}–384_{E2}	26/26 (100%)
	2637_{NS5B}–837_{NS2}–827_{NS2}–397_{E2}–461_{E2}	26/26 (100%)
1b Non-SVR	2283_{NS5A}–464_{E2}–479_{E2}–391_{E2}–478_{E2}	21/21 (100%)
	837_{NS2}–384_{E2}–478_{E2} / 1011_{NS2}	21/21 (100%)
	479_{E2}–478_{E2}–1805_{NS4B}–2375_{NS5A}–464_{E2} / 397_{E2}	21/21 (100%)

Figure 6
Example subnetworks associated with outcome of anti-HCV therapy, with 100% accuracy and 0% false coverage rates, that are potential biomarkers for prediction of therapy outcome. The subnetworks contain 1 or more covariances that together are found in the indicated fraction (coverage) of the appropriate response class and are never found in the opposing response class. Polyprotein residue numbers are shown for each subnetwork, and the identity of the mature HCV protein is indicated in subscript.

that the sum of these interactions strengthened complexes involving viral proteins. In the structural proteins that form the virion (core, E1, and E2), the greater number of hydrophobic interactions would be predicted to stabilize the virus particle and to somehow increase its infectivity and/or resistance to degradation by IFN-induced mechanisms. The predicted increase in the stability of complexes including the nonstructural proteins (P7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B) would presumably both stabilize the replicase complex to reduce its sensitivity to effectors of the IFN-α response and improve the ability of the viral proteins to interdict the cellular type 1 IFN response (e.g., the ability of NS3/NS4A to block sensing of double-stranded RNA by TLR3 and RIG-I; ref. 10). The mechanisms by which the structural and nonstructural proteins function during viral replication and the generalized increase of hydrophobic residue pairs in the responder networks together imply that clearance of the virus during IFN-α–based therapy may be aided by both lower cell-to-cell infectivity of the virions and higher sensitivity of the viral components within cells to the drugs.

Furthermore, the majority of the covariances probably reflect compensatory variations among multiprotein complexes composed primarily of viral proteins. The justification for this prediction is that host proteins do not vary with the high frequency observed among HCV sequences, with the exception of the antigen-binding regions of the immunoglobulins and the T cell receptors. However, escape from cell-mediated immunity is unlikely to be the dominant force driving development of the covariance network because sets of T cell epitopes would need to coevolve, but T cell epitopes are short linear peptides, and very few of the covariances were adjacent to one another in the linear amino acid sequence. Effective antibody-mediated selective pressures would also be unable to generate the genome-wide covariance networks because these pressures would be largely limited to the E1 and E2 surface glycoproteins that form the exterior of the virion, but covariances were found among all 10 HCV proteins. Therefore, the high degree of variation — and covariation — among the HCV sequences is not needed to accommodate the limited sequence diversity present in the vast majority of human genes. However, some of the covariances between different HCV proteins could represent compensatory adaptations between HCV proteins to maintain a common interaction with a third partner that may be of host origin.

The presence of amino acid covariance networks in the HCV genome specific to the outcome of antiviral therapy has 3 practical implications for personalized medicine. First, the nonoverlapping regions of the covariance networks from the Marked and Poor response classes (Table 3 and Figure 4) may provide a basis for a sequence-based test that could predict the susceptibility of individual HCV isolates to IFN-α–based therapies. Our initial assessment of such biomarker positions (Figure 6) is very promising, because hundreds of subnetworks providing 100% accuracy and greater than 90% coverage for prediction of both SVR and Non-SVR were found.

However, the precision by which the networks may be able to predict the outcome of therapy must be viewed with some reserve, because although the All networks have been validated in external data sets by us and others (37), we cannot yet externally validate the response-based networks. This is because no non–Virahep-C sequence set exists for which the outcome of IFN-based therapy is available. We anticipate that the ability of the networks to predict treatment outcome will be less robust outside of this training set

because the relatively small number of sequences available could have led to overestimation of the degree of separation between the treatment outcome networks. However, the large genetic diversity differences between the response classes (17), the extensive overlap between the congruent response-based networks (Marked with SVR and Poor with Non-SVR), and the largely nonoverlapping nature of the contrasting networks (Marked versus Poor and SVR versus Non-SVR) all imply that HCV sequence variation has a major role in determining the outcome of therapy. Therefore, the large number of potential covariance biomarkers available and the ability to simultaneously consider multiple subnetworks strongly imply that a clinically useful predictive test could be designed.

The differences between the 1a and 1b networks indicates that predictive tests based on the covariance networks will need to be customized for each subtype. However, even with customization, such tests would be highly cost effective because chip-based assays could be designed for about $100 per sample, whereas the drugs used in a failed course of therapy can cost up to $30,000 (39). We anticipate that the ability to predict nonresponse would be the most practical form of the assay, because treatment of a susceptible HCV isolate could still result in Non-SVR through drug intolerance or noncompliance. In this context, physicians could counsel against IFN-α–based therapy, avoiding tens of thousands of dollars in expenses and painful side effects for the patient. For example, more than 250 HCV patients are treated at Saint Louis University Hospital per year, and if futile treatment of just half of the nonresponders (approximately 62 patients) was eliminated at a cost of about $25,000 for the screening assay, the savings could be up to $1.8 million in drug costs alone.

The second medical implication for these networks is that the highly connected hub residues have a large number of functional interactions with other residues; hence, disrupting a hub would be predicted to weaken this web of interactions. Therefore, the hubs may be valuable antiviral drug targets. This is an attractive concept because knockout of hubs in interaction networks has previously been shown to be lethal in several different organisms (40–42). Targeting variable sites for drug design is counterintuitive, but it should be feasible for anti-HCV therapy, because new anti-HCV drugs are likely to be used in conjunction with IFN-α. Therefore, an anti-hub drug would be designed to inhibit the IFN-resistant hub configuration, leaving variant viruses with the IFN-sensitive configuration to be eliminated by IFN-α. Targeting the hubs would be especially attractive because evolution of resistant mutants should be slow, as a result of the high genetic cost of mutating a highly interconnected residue without simultaneously mutating many of its covarying partners.

Finally, the high error rate of RNA synthesis that is a fundamental feature of RNA virus replication leads to high genetic diversity among these viruses. Therefore, covariance network analysis should be applicable to essentially all RNA viruses. If similar networks correlating with virulence or drug sensitivity exist in other viruses, covariance network analysis should open a wide range of diagnostic and therapeutic options in medical, veterinary, and agricultural settings.

Methods

Patients and response classes. We used 94 viral pretreatment sequences derived from participants in the Virahep-C clinical study (16). Each patient received full doses of pegylated IFN-α and ribavirin for the first 28 days. This research was conducted in accordance with the Helsinki principles: all patients gave informed, written consent prior to their participation in Virahep-C and its associated basic science components, and all components of Virahep-C were approved by their local Institutional Review Boards. The Virahep-C viral genetics cohort (17) was selected to be evenly stratified by genotype, 1a versus 1b, and by day-28 responses to therapy. Marked responders had a decline in HCV titres greater than 3.5 $\log_{10}$ or to undetectable between baseline and day 28 of therapy, Intermediate responders had declines of 3.5–1.4 $\log_{10}$, and Poor responders had declines less than 1.4 $\log_{10}$. In the analyses reported here, the samples were stratified by either genotype (1a or 1b), genotype plus treatment outcome (SVR or Non-SVR), or genotype plus the extremes of early response to therapy (Marked or Poor). The 118 non–Virahep-C HCV genotype 1a sequences used for external validation were downloaded from the Hepatitis C Database of the Broad Institute (http://www.broad.mit.edu/annotation/viral/HCV/Home.html).

Sequencing. Consensus sequences for the HCV ORF were previously obtained by directly sequencing overlapping RT-PCR amplicons as described previously (GenBank accession nos. EF407411–EF407504; refs. 17, 43). Mixed-base positions caused by the HCV quasispecies were resolved by identifying the predominant base at each position. The 3′ end of the HCV ORF could not be amplified in a few samples, hence the C-terminal 56 amino acids of NS5B were excluded from the analyses so that all sequences were represented equally. The numbering system for 1a sequences was identical to the H77 isolate (GenBank accession no. AF009606), and numbering for the 1b sequences was the same as the J4 isolate (GenBank accession no. AF054247).

Calculation of covariance pairs. The sequences were aligned using Clustal W (44) as previously described (17). For each genotype, 5 different alignments were created and independently subjected to covariance analysis (Table 2).

Algorithms that calculate covariance require an intermediate level of conservation. Positions that are invariant do not contain sufficient information to assess correlated variations, and positions that display random differences can generate spurious correlations. Therefore, we evaluated 3 algorithms to identify covarying positions (24, 29, 45). In accord with the results of Fodor and Aldrich (46), we found that the OMES method (29) performed well on this data set. The other methods gave spurious correlations as a result of the relatively high degree of conservation in the HCV sequences or because their clustering methods did not perform well in this context. We used complete clustering with the OMES method because no assumptions (e.g., number or size of the clusters) are needed.

To identify the covarying pairs, we calculated for every possible pair of columns i and j, a score S using observed and expected pairs:

$$S = \frac{\sum_{1}^{L}(N_{obs} - N_{exp})^2}{N_{valid}} \quad \text{(Equation 1)}$$

where L is the list of all observed pairs and N_{obs} is the number of occurrences for a pair of residues. The expected number for the pair (N_{exp}) is given by:

$$N_{exp} = \frac{C_{xi}C_{yj}}{N_{valid}} \quad \text{(Equation 2)}$$

in which N_{valid} is the number of sequences in the alignment that are nongap residues, C_{xi} is the observed number of residues x at position i, and C_{yj} is the observed number of residues y at position j. The expected number of column pairs calculated in this manner provides a null model for comparisons of the observed pairs.

An OMES score of 0.5 was used as the cutoff for all analyses. This cutoff was chosen for 4 reasons. First, at this value, the All network formed a complete graph (all nodes connected by at least 1 edge), but at a cut-

off value of 0.9, 20% of the nodes had no edges. Therefore, 0.5 was not so stringent as to exclude possibly informative data. Second, the number of edges increased dramatically at cutoff values lower than 0.5 (Supplemental Figure 1). A cutoff value of 0.5 corresponded to 3 of 16 possible differences in the pair of columns in the day-28 alignments. [This value is calculated as $S = \Sigma(N_{obs} - N_{exp})^2/N_{valid}$; with S of 0.5 and N_{valid} of 16, then $0.5 \times 16 = 8 = \Sigma(N_{obs} - N_{exp})^2$; the square root of 8 is 2.8, the maximum difference between N_{obs} and N_{exp}.] Therefore, covariances with scores less than 0.5 are weak, and using values below 0.5 would greatly increase the number of spurious associations. Third, we observed relatively small effects on the response-specific networks at less than 25% shuffling in the permutation analysis (Figure 4). A cutoff value of 0.5 corresponded to at least 3 of 16 (18.75%) possible differences in the day-28 response–specific alignments, hence this value was well within the stable range of the network. Finally, the number of edges in the Marked and Poor responder networks began to show a difference at a value of 0.5 (Supplemental Figure 1). This was an asset because one of our goals was to find differences between the response classes.

Network generation. Graphs were generated and rendered for the covarying positions in each response class using Cytoscape (version 2.6.1; ref. 47). Such complete clustering is usually not performed on large data sets because it is computationally intensive. However, complete clustering was easily achievable in this case because the covarying positions were obviously linked together. Therefore, we did not need to use the covariance score to generate the clusters. Instead, we simply chose a cutoff value for the score; this is the equivalent of single-linkage clustering.

Statistics. Race and sex distributions were compared across response groups using Pearson's χ^2 test for association. Normally distributed characteristics were compared across response groups using ANOVA, whereas the Kruskal-Wallis nonparametric test was used when distributions were not normally distributed. The fraction of true positives relative to the unshuffled network was compared using a 2-tailed Student's t test. Fisher's exact test was used to compare the proportions of hydrophobic pairs between response groups. The program to calculate covariance scores and associated parameters was custom written. Statistical tests were performed using R (version 2.6.2; ref. 48). In all cases, a P value of 0.05 or less was considered significant.

Calculation of solvent exposure. Solvent exposure was calculated using the method of Lee and Richards (49) as described previously (50). The raw exposed surface area was normalized using the tripeptide as a standard state (51). The normalized exposure was averaged over all side-chain atoms, and the residue was considered to be solvent exposed if it had greater than 40% exposed side-chain atoms.

Topology and structural analyses. The transmembrane regions were predicted using TMHMM (52, 53). These regions were then mapped onto the positions of signal sequences and combined with the reported location of proteins (e.g., cytoplasm, endoplasmic reticulum lumen, or transmembrane; refs. 54, 55). Inter-residue distances were calculated using the PDB coordinates (2HD0, 1CU1, 1ZH1, 2GIR, and 2A4Q) using a custom-written program and RASMOL version 2.7 (56).

Acknowledgments

We thank Abdus Wahed for helpful discussions. The members of the Virahep-C Study Group are listed in ref. 16. This study was supported by NIH grant DK60345.

Received for publication August 8, 2008, and accepted in revised form October 22, 2008.

Address correspondence to: Rajeev Aurora or John E. Tavis, Department of Molecular Microbiology and Immunology, Saint Louis University School of Medicine, 1100 South Grand Blvd., St. Louis, Missouri 63104, USA. Phone: (314) 977-8891; Fax: (314) 977-8717; E-mail: aurorar@slu.edu (R. Aurora). Phone: (314) 977-8893; Fax: (314) 977-8717; E-mail: tavisje@slu.edu (J.E. Tavis).

1. Armstrong, G.L., et al. 2006. The prevalence of hepatitis C virus infection in the United States, 1999 through 2002. *Ann. Intern. Med.* **144**:705–714.
2. Hadziyannis, S.J., et al. 2004. Peginterferon-alpha2a and ribavirin combination therapy in chronic hepatitis C: a randomized study of treatment duration and ribavirin dose. *Ann. Intern. Med.* **140**:346–355.
3. Manns, M.P., et al. 2001. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *Lancet.* **358**:958–965.
4. McHutchison, J.G., et al. 1998. Interferon alfa-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C. *N. Engl. J. Med.* **339**:1485–1492.
5. Poynard, T., et al. 1998. Randomised trial of interferon alpha2b plus ribavirin for 48 weeks or for 24 weeks versus interferon alpha2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. International Hepatitis Interventional Therapy Group (IHIT). *Lancet.* **352**:1426–1432.
6. Davis, G.L., et al. 1998. Interferon alfa-2b alone or in combination with ribavirin for treatment of relapse of chronic hepatitis C. *N. Engl. J. Med.* **339**:1493–1499.
7. Lemon, S.M., Walker, C., Alter, M.J., and Yi, M. 2007. Hepatitis C Virus. In *Fields virology*. D.M. Knipe, et al., editors. Lippincott, Williams & Wilkins. Philadelphia, Pennsylvania, USA. 1253–1304.
8. Lindenbach, B.D., and Rice, C.M. 2005. Unravelling hepatitis C virus replication from genome to function. *Nature.* **436**:933–938.
9. Moradpour, D., Penin, F., and Rice, C.M. 2007. Replication of hepatitis C virus. *Nat. Rev. Microbiol.* **5**:453–463.
10. Gale, M., Jr., and Foy, E.M. 2005. Evasion of intracellular host defence by hepatitis C virus. *Nature.* **436**:939–945.
11. Simmonds, P., et al. 1993. Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *J. Gen. Virol.* **74**:2391–2399.
12. Bukh, J., Miller, R., and Purcell, R. 1995. Genetic heterogeneity of hepatitis c virus: quasispecies and genotypes. *Semin. Liver Dis.* **15**:41–63.
13. Robertson, B., et al. 1998. Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. *Arch. Virol.* **143**:2493–2503.
14. Simmonds, P. 2004. Genetic diversity and evolution of hepatitis C virus--15 years on. *J. Gen. Virol.* **85**:3173–3188.
15. Fried, M.W., et al. 2002. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N. Engl. J. Med.* **347**:975–982.
16. Conjeevaram, H.S., et al. 2006. Peginterferon and ribavirin treatment in african american and caucasian american patients with hepatitis C genotype 1. *Gastroenterology.* **131**:470–477.
17. Donlin, M.J., et al. 2007. Pretreatment sequence diversity differences in the full-length Hepatitis C Virus open reading frame correlate with early response to therapy. *J. Virol.* **81**:8211–8224.
18. Cannon, N.A., Donlin, M.J., Fan, X., Aurora, R., and Tavis, J.E. 2008. Hepatitis C virus diversity and evolution in the full open-reading frame during antiviral therapy. *PLoS ONE.* **3**:e2123.
19. Eyal, E., Frenkel-Morgenstern, M., Sobolev, V., and Pietrokovski, S. 2007. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins.* **67**:142–153.
20. Frenkel-Morgenstern, M., Magid, R., Eyal, E., and Pietrokovski, S. 2007. Refining intra-protein contact prediction by graph analysis. *BMC Bioinformatics.* **8**(Suppl. 5):S6.
21. Gobel, U., Sander, C., Schneider, R., and Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins.* **18**:309–317.
22. Altschuh, D., Lesk, A.M., Bloomer, A.C., and Klug, A. 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**:693–707.
23. Larson, S.M., Di Nardo, A.A., and Davidson, A.R. 2000. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.* **303**:433–446.
24. Olmea, O., Rost, B., and Valencia, A. 1999. Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**:1221–1239.
25. Wang, Y.E., and DeLisi, C. 2006. Inferring protein-protein interactions in viral proteins by co-evolution of conserved side chains. *Genome Inform.* **17**:23–35.
26. Lockless, S.W., and Ranganathan, R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* **286**:295–299.
27. Suel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**:59–69.
28. Lee, B.C., Park, K., and Kim, D. 2008. Analysis of the residue-residue coevolution network and the functionally important residues in proteins. *Pro-*

teins. **72**:863–872.
29. Kass, I., and Horovitz, A. 2002. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins.* **48**:611–617.
30. Blight, K., and Norgard, E.A. 2006. HCV replicon systems. In *Hepatitis C viruses: Genomes and molecular biology*. S.L. Tan, editor. Horizon Bioscience. Wymondham, United Kingdom. 311–351.
31. Blight, K.J. 2007. Allelic variation in the hepatitis C virus NS4B protein dramatically influences RNA replication. *J. Virol.* **81**:5724–5736.
32. Dustin, L.B., and Rice, C.M. 2007. Flying under the radar: the immunobiology of hepatitis C. *Annu. Rev. Immunol.* **25**:71–99.
33. Barabasi, A.L. 2002. *Linked: The new science of networks*. Perseus Books Group. Cambridge, Massachusetts, USA. 256 pp.
34. Albert, R., and Barabasi, A.L. 2000. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* **85**:5234–5237.
35. Zhou, D., et al. 2007. Separation of near full-length hepatitis C virus quasispecies variants from a complex population. *J. Virol. Methods.* **141**:220–224.
36. Xu, Z., Fan, X., Xu, Y., and Di Bisceglie, A.M. 2008. Comparative analysis of nearly full-length hepatitis C virus quasispecies from patients experiencing viral breakthrough during antiviral therapy: clustered mutations in three functional genes, E2, NS2, and NS5a. *J. Virol.* **82**:9417–9424.
37. Campo, D.S., Dimitrova, Z., Mitchell, R.J., Lara, J., and Khudyakov, Y. 2008. Coordinated evolution of the hepatitis C virus. *Proc. Natl. Acad. Sci. U. S. A.* **105**:9685–9690.
38. Lau, J.Y., Tam, R.C., Liang, T.J., and Hong, Z. 2002. Mechanism of action of ribavirin in the combination treatment of chronic HCV infection. *Hepatology.* **35**:1002–1009.
39. McHutchison, J.G., Bacon, B.R., and Owens, G.S. 2007. Making it happen: managed care considerations in vanquishing hepatitis C. *Am. J. Manag. Care.* **13**(Suppl. 12):S327–S336.
40. Hahn, M.W., and Kern, A.D. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**:803–806.
41. Ingolia, N.T. 2004. Topology and robustness in the Drosophila segment polarity network. *PLoS Biol.* **2**:e123.
42. Wagner, A. 2000. Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24**:355–361.
43. Yao, E., and Tavis, J.E. 2005. A general method for nested RT-PCR amplification and sequencing the complete HCV genotype 1 open reading frame. *Virol. J.* **2**:88.
44. Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., and Gibson, T.J. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**:403–405.
45. Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., and Dress, A.W. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* **17**:164–178.
46. Fodor, A.A., and Aldrich, R.W. 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins.* **56**:211–221.
47. Shannon, P., et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**:2498–2504.
48. R. Development Core Team. 2004. *A language and environment for statistical computing*. The R Foundation for Statistical Computing. Vienna, Austria. http://www.r-project.org/.
49. Lee, B., and Richards, F.M. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**:379–400.
50. Aurora, R., Srinivasan, R., and Rose, G.D. 1994. Rules for alpha-helix termination by glycine. *Science.* **264**:1126–1130.
51. Lesser, G.J., and Rose, G.D. 1990. Hydrophobicity of amino acid subgroups in proteins. *Proteins.* **8**:6–13.
52. Kahsay, R.Y., Gao, G., and Liao, L. 2005. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics.* **21**:1853–1858.
53. Sonnhammer, E.L., von Heijne, G., and Krogh, A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**:175–182.
54. Penin, F. 2003. Structural biology of hepatitis C virus. *Clin. Liver Dis.* **7**:1–21, vii.
55. Reed, K.E., and Rice, C.M. 2000. Overview of hepatitis C virus genome structure, polyprotein processing, and protein properties. *Curr. Top. Microbiol. Immunol.* **242**:55–84.
56. Sayle, R.A., and Milner-White, E.J. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**:374.