Supporting Information

Campo *et al.* 10.1073/pnas.0801774105

SI Text

Sequences. Two hundred and eight complete genome HCV sequences were obtained from the Los Alamos HCV Sequence Database (1). Of these 208 sequences, those in the following categories were excluded: recombinants, chimeras, patents, nonhuman hosts, a genotype other than 1b, and epidemiologically related sequences. This process left 114 different HCV 1b complete genome sequences with the following GenBank accession numbers: X61596, D90208, D11355, M96362, L02836, M84754, M58335, U01214, D10934, D30613, D63857, D50484, D50481, D50480, D14484, U45476, D45172, D89872, D85516, AJ000009, D89815, AJ132997, AJ238799, AF176573, AJ238800, AB016785, AF165046, AF165048, AF165050, AF165052, AF165054, AF165056, AF165058, AF165060, AF165062, AF165064, AF208024, AF207752, AF207753, AF207754, AF207755, AF207756, AF207757, AF207758, AF207759, AF207760, AF207761, AF207762, AF207763, AF207764, AF207765, AF207766, AF207767, AF207768, AF207769, AF207770, AF207771, AF207772, AF207773, AF207774, AB049087, AB049088, AB049089, AB049090, AB049091, AB049092, AB049093, AB049094, AB049095, AB049096, AB049097, AB049098, AB049099, AB049100, AB049101, AF356827, AY045702, AF313916, AF483269, AF139594, AB080299, AY460204, AY587844, AY587016, AB154177, AB154178, AB154179, AB154180, AB154181, AB154182, AB154183, AB154184, AB154185, AB154186, AB154187, AB154188, AB154189, AB154190, AB154191, AB154192, AB154193, AB154194, AB154195, AB154196, AB154197, AB154198, AB154199, AB154200, AB154201, AB154202, AB154203, AB154204, AB154205 and AB154206.

Alignment. The 114 HCV sequences were aligned by using ClustalW (2). The HCV strain H77 polyprotein (GenBank accession no. NC_004102) was used as a reference sequence throughout this study.

Detection of Selection by Single-Likelihood Ancestor Counting (SLAC). SLAC involves counting the number of nonsynonymous (dN) and synonymous changes (dS) and testing whether dN is significantly different from dS. SLAC is a substantially modified and improved derivative of the Suzuki–Gojobori method (3). SLAC is implemented in the program HyPhy (4), which is available in a parallel computing fashion at the Datamonkey web interface (5). The algorithm processes an alignment by using likelihood-based branch lengths, nucleotide and codon substitution parameters, and ancestral sequence reconstructions. SLAC has sufficient power to detect nonneutral evolution in large (>50 sequences) alignments. This algorithm consists of the following steps:

1. An initial estimate of the phylogeny is obtained by neighbor-joining (6) by using the Tamura–Nei distance (7).

2. Fitting of a Codon-Based Substitution Model. To obtain a reasonable ancestral reconstruction, we use a codon-based substitution model based on the MG model (8) but augmented by the best-fitting nucleotide model. The general rate matrix element for this model defines the instantaneous rate of substituting a non-stop codon *x* with a non-stop codon *y*. The MG model is slightly different from the GY model (9) in that GY models use the frequency of the target codon, whereas MG models use the frequency of the target nucleotide in the appropriate codon position. Although very similar, the models, in general, are not equivalent. MG models tend to yield better likelihood scores for most datasets, especially for some codon alignments with significant compositional biases (4, 10). This was confirmed in the preset work. A likelihood ratio test was performed to find a codon model that better fits the data (MG vs. GY). For every protein it was found that MG is better than GY ($P = 0.00001$).

3. Estimation of Substitution Rates. Given an estimate of the phylogeny (topology and branch lengths) and the codon-based substitution model, the number of changes that have occurred within the phylogeny was inferred. The simplest method involves reconstructing ancestral sequences by using the joint likelihood reconstruction method in the codon-state space. This reconstruction strategy avoids most of the problems in the original Suzuki–Gojobori method (3) by disallowing stop codons, recovering a unique state at each internal tree node, and fully incorporating synonymous and nonsynonymous substitution structure into ancestral state reconstruction. Treating the reconstructed sequences as known, dS and dN are computed (excluding stop codons). Ambiguous codons in the data are defined by averaging the counts over all possible codon states, weighting by the relative frequency of each state. Given the rate bias parameters inferred earlier, the rate matrix of the jump chain component of the substitution process is built (10).

4. Testing for Positive or Negative Selection. Given an estimate of dS and dN at a codon, it was tested whether dN is significantly different from dS. It was assumed that the observed dS follows a continuous extension of the binomial distribution. The site is classified as positively selected if $T = dN - dS > 0$, and the probability of observing T by using the extended binomial distribution is <0.05. Sites where no substitutions were inferred cannot be classified with this scheme. Of the 3,010 HCV 1b codon sites, only 212 (7.04%) did not show any substitutions.

Comparison with Previous Methods. Suzuki and Gojobori (11) also analyzed selection in HCV 1b sequences. They found evidence of negative selection in 1,560 sites and positive selection in only 13 of 3,010 aa sites, eight of which were also identified by the algorithm used in the present study (SLAC). The number of POSI sites that were found by using SLAC is larger than that found by Suzuki and Gojobori (11), which is explained by the following differences between the two methods:

1. Suzuki and Gojobori (11) did not take into account the transition/transversion rate bias and the base/codon frequency bias. Those biases affect the accuracy of the estimates, overestimating dS and underestimating dN, which results in a lower power for detecting positive selection (12).

2. SLAC uses maximum likelihood ancestral reconstructions, unlike Suzuki and Gojobori (11), who used parsimony. Maximum likelihood ancestral reconstruction avoids most of the problems in the original Suzuki–Gojobori method (3, 11) by disallowing stop codons, recovering a unique state at each internal tree node, and fully incorporating synonymous and nonsynonymous substitution structure into ancestral state reconstruction (13).

3. To obtain a reasonable phylogenetic reconstruction, SLAC uses a codon substitution model based on the MG model but augmented by the best-fitting nucleotide model. The method of Suzuki and Gojobori (3, 11) does not fit any substitution model to the data.

4. Suzuki and Gojobori (11) used a lower sample size (average of 83 sequences: 129 for Core, 135 for E1, 59 for E2, 74 for NS2, 68 for NS3, 70 for NS4, 69 for NS5A, and 61 for NS5B). Simulations have shown that methods for detecting selection are strongly influenced by sample size (10, 13).

Physicochemical Factors. In the present study, each amino acid in the 114 HCV sequences was transformed into a string of different values of five physicochemical factors. These five factors are multidimensional patterns of highly intercorrelated physicochemical variables created by factor analysis of 494 amino acid properties (14, 15) that can be used toward understanding the evolutionary, structural, and functional aspects of protein variation. The factors are the following: POLARF1, which reflects polarity, portion of exposed residues, free energy, the number of hydrogen bond donors, and hydrophobicity; HELIXF2, a secondary structure factor with an inverse relationship of relative propensity for the amino acid in various secondary structural configurations such as a coil, turn, or bend versus the frequency in an α-helix; SIZEF3, which relates to molecular size or volume with high factor coefficients for bulkiness, residue volume, average volume of a buried residue, side-chain volume, and molecular weight; CODONF4, which reflects relative aa composition in proteins; and CHARGEF5, which refers to electrostatic charge with high coefficients on isoelectric point and net charge (14).

Sources of Covariation. There are three different sources of covariation in related biological sequences (such as the dataset of HCV sequences): (*i*) chance, (*ii*) common ancestry, and (*iii*) structural or functional constraints. Effectively discriminating among these underlying causes is a difficult task with many statistical and computational difficulties (16). There are many methods to test whether a correlation value reflects a significant association (possibly because of structural and functional constraints), or results from evolutionary history and stochastic events (background covariation) (17), but no single method has demonstrated general utility or achieved widespread acceptance (18). We used the following four criteria to define the pairs of significantly correlated pairs of sites.

1. Each one of the two sites has an entropy >0.2370, which is 10% of the highest entropy found in the HCV polyprotein. Only 448 amino acid sites of the 3,010 amino acid sites of the HCV polyprotein are above this entropy cutoff. This cutoff was chosen because prior modeling of protein coevolution showed that it is difficult to identify sites that are coevolving if they are highly conserved (18, 19). The 448 polymorphic sites include 60 (13.39%) positively selected, 105 (23.44%) negatively selected, and 283 (63.17%) neutral sites. Although the use of highly polymorphic sites potentially reduces effects of a recent common ancestry in the correlation analysis, it also disproportionately reduces the number of negatively selected sites because of their limited heterogeneity. Nevertheless, because our goal was to identify the phenomenon of substitution coordination rather than to conduct a detailed analysis of all potential contributions in coordinated amino acid changes along the polyprotein, such a conservative approach to site selection seems justified.

2. A permutation procedure was performed, whereby the amino acids at each site in the sequence alignment were vertically shuffled. Ten thousand random alignments were created this way, simulating the distribution of correlation values under the null hypothesis that substitutions of amino acids at two sites are statistically independent. For each physicochemical factor, a pair of sites was considered significantly correlated if its correlation value in the observed dataset was higher than the correlation value for those two sites in any of the random datasets ($P = 0.0001$). We addressed the multiple comparisons problem with the False Discovery Rate approach, which controls the expected proportion of false-positive results (20). The False Discovery Rate in our study has a *q*-value of 0.0116, which means a maximum of 50 false links are expected.

3. Related sequences (such as the dataset of HCV sequences) are part of a hierarchically structured phylogeny and, therefore, for statistical purposes, cannot be regarded as being drawn independently from the same distribution. We used a data weighting approach based on the assumption that the lower the time of divergence of two sequences from their common ancestor, the higher is the covariation between these two sequences (21, 22). The one-dimensional weights were calculated by using a distance matrix among sequences built by using the synonymous sites of the full HCV genome.

4. We also used a modified version of the method of Martin *et al.* (18) and Gloor *et al.* (19). This method makes the assumption that each position in a multiple sequence alignment is affected equally by background correlation, and that the majority of positions in the alignment covary only because of common ancestry. On the basis of these assumptions, each alignment is used as its own null model for the identification of covarying positions. A critical correlation threshold was calculated by using the value of the Student's distribution at a given significance level ($P = 0.01$) with a sample size of 114 sequences, according to Afonnikov *et al.* (23). Interestingly, 4,007 of the 4,317 significant links (92.81%) are also significant with this approach.

Invariance of the Physicochemical Properties. An important feature of coordinated substitutions is their contribution to the invariance of the physicochemical characteristics of a protein (23). The program CRASP (24) was used to estimate the contribution of the coordinated substitutions to the evolutionary invariance in the protein physicochemical characteristics of HVR1.

Other Networks. Seven networks were built in total. The one shown in this article was built by using significant links for one or more physicochemical correlations (4,317 links and 404 sites in the giant component) with an average correlation of 0.4389 at 2.4869 factors; five more networks were built for each physicochemical factor: POLARF1 (1,963 links and 284 sites), HELIXF2 (2,032 links and 288 sites), SIZEF3 (2,214 links and 276 sites), CODONF4 (2,198 links and 292 sites), and CHARGEF5 (2,275 links and 291 sites). A final network was built by using mutual information instead of the correlation of physicochemical properties (3,260 links and 324 sites). All networks have very similar topology and hierarchical structure. Of all pairwise comparisons, 1,453 are significant at only 1 factor, 1,037 at 2, 728 at 3, 464 at 4, 635 at 5, and 614 at all factors and mutual information. The number of links of each site is highly correlated with its entropy ($r = 0.4126$, $P = 0.0001$), which is to be expected because two variables must exhibit variation before they can exhibit shared or common variation, as demonstrated algebraically by Atchley *et al.* (25).

Power-Law Distribution. The degree distribution was analyzed to tell whether the power law is a reasonable fit to the data by using maximum likelihood methods and the Kolmogorov–Smirnov statistic (26). The hypothesis that this degree distribution follows a power law was not rejected ($P = 0.9541$).

Efficiency of Small-World Networks. We calculated the global and local efficiency of the network (27), assuming that the efficiency between two vertices is proportional to the reciprocal of the length of the shortest path between them. Global efficiency is calculated for the whole graph and local efficiency is calculated for the subgraph around each vertex. The efficiency measure gives a definition of small world with a clear physical meaning and allows a precise quantitative analysis of the information flow. Small-world networks are systems that are both globally and locally efficient (27), which is the case for the HCV network, with a global efficiency of 0.3966 and a local efficiency of 0.3966.

1. Kuiken C, Yusim K, Boykin L, Richardson R (2005) The Los Alamos hepatitis C sequence database. *Bioinformatics* 21:379–384.
2. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
3. Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16:1315–1328.
4. Kosakovsky S, Frost S, Muse S (2005) HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
5. Kosakovsky S, Frost S (2005) Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21:2531–2533.
6. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
7. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial- DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526.
8. Muse S, Gaut B (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715–724.
9. Goldman N, Yang Z (1994) Codon-based model of nucleotide substitution for protein-coding DNA-sequences. *Mol Biol Evol* 11:725–736.
10. Kosakovsky S, Frost S (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 22:1208–1222.
11. Suzuki Y, Gojobori T (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. *Gene* 276:83–87.
12. Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 46:409–418.
13. Kosakovsky S, Frost S (2005) A simple hierarchical approach to modeling distributions of substitution rates. *Mol Biol Evol* 22:223–234.
14. Atchley W, Zhao J, Fernandes A, Druke T (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci USA* 102:6395–6400.
15. Kawashima S, Kanehisa M (2000) AAindex: Amino acid index database. *Nucleic Acids Res* 28:374.
16. Wollenberg K, Atchley W (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci USA* 97:3288–3291.
17. Atchley W, Wollenberg K, Fitch W, Terhalle W, Dress A (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol* 17:164–178.
18. Martin L, Gloor G, Dunn S, Wahl L (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21:4116–4124.
19. Gloor G, Martin L, Wahl L, Dunn S (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44:156–165.
20. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300.
21. Vingron M, Sibbald P (1993) Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci USA* 90:8777–8781.
22. Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1–15.
23. Afonnikov D, Oshchepkov D, Kolchanov N (2001) Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. *Bioinformatics* 17:1035–1046.
24. Afonnikov D, Kolchanov N (2004) CRASP: A program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res* 32:W64–W68.
25. Atchley W, Terhalle W, Dress A (1999) Positional dependance, cliques and predictive motifs in the bHLH protein domain. *J Mol Evol* 48:501–506.
26. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data, arXiv:0706.1062v1 [physics.data-an].
27. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. *Phys Rev Lett* 87:1–14.

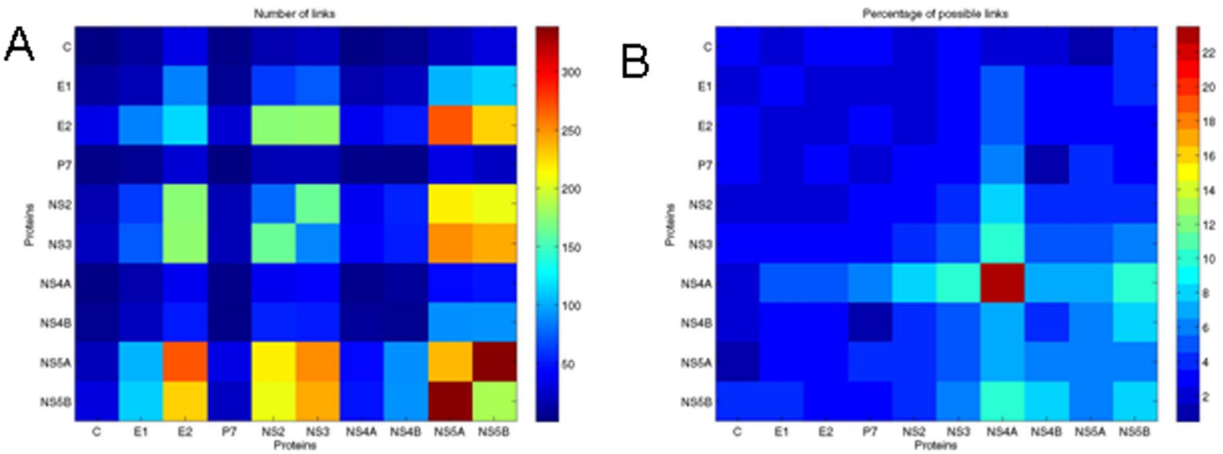

Fig. S1. Distribution of links amoung HCV proteins. (*A*) Number of links between proteins (upper and lower triangular halves of the matrix) and links inside the same proteins (diagonal). C is involved in 1.89% of all links, E1 in 6.44%, E2 in 15.20%, P7 in 1.79%, NS2 in 13.01%, NS3 in 14.25%, NS4A in 3.13%, NS4B in 4.87%, NS5A in 20.33%, and NS5B in 19.08%. The number of links of each protein is highly correlated with the number of polymorphic sites included in the analysis ($r = 0.9257$, $P = 0.00001$). B) Percentage of possible links between proteins given the number of polymorphic amino acid sites (ratio of actual to possible links).

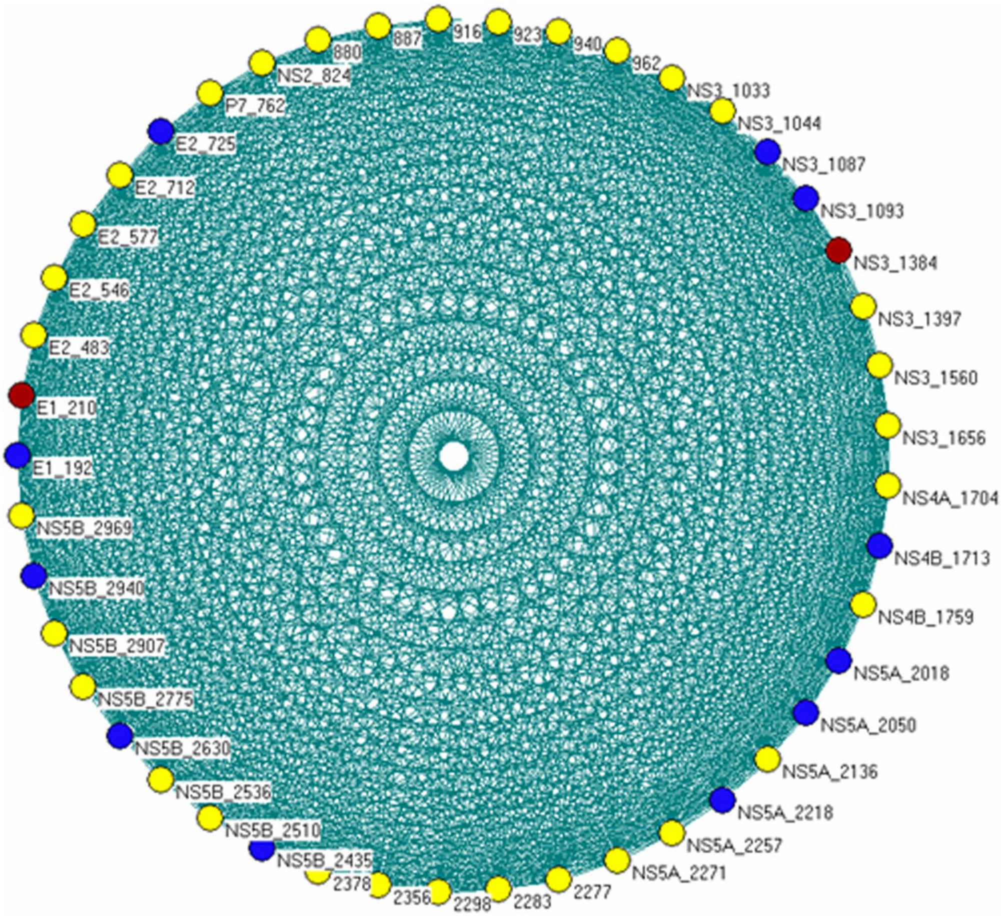

Fig. S2. Sites in the nucleus of the HCV network of coordinated substitutions. Sites correlated by one or more physicochemical properties are linked by green lines. Positively selected sites are shown in red, negatively selected sites in blue, and neutral sites in yellow.

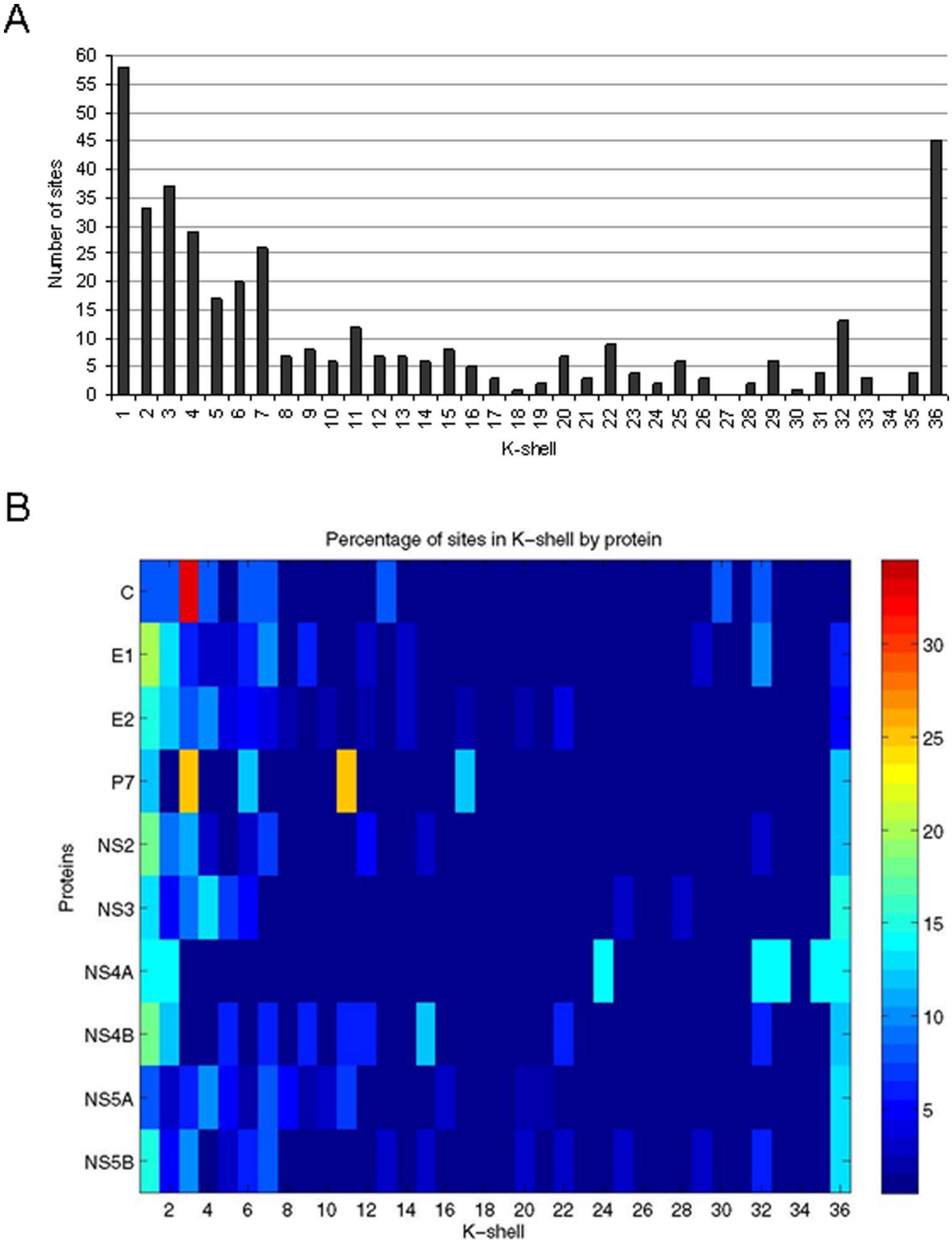

Fig. S3. *k*-shell distribution. (*A*) Number of sites in each *k*-shell. (*B*) Percentage of the sites of a protein in each *k*-shell.

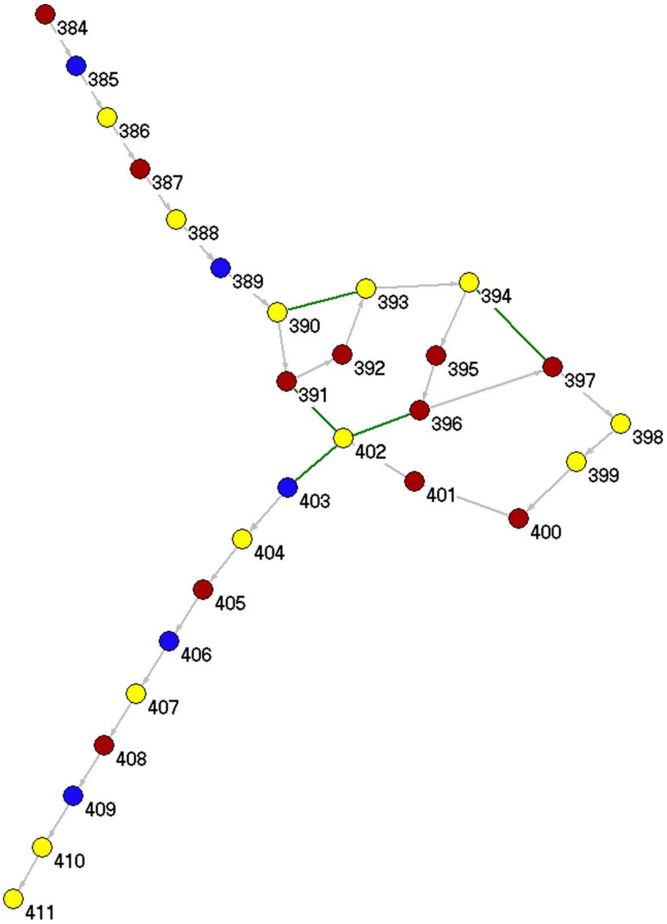

Fig. S4. The HVR1 (site 384–410) of the structural E2 protein. Contiguous sites in the sequence are linked by gray arrows and sites correlated by one or more physicochemical properties are linked by green lines. Positively selected sites are shown in red, negatively selected sites in blue, and neutral sites in yellow.

Table S1. Positively selected sites

Site	PROTE	CONAA	NAAVAR	dS	dN	*H*	*P*
75	C	T	4	0.0000	9.5000	1.1569	0.000451
161	C	G	2	0.0000	4.9838	0.4578	0.017914
210	E1	A	3	1.0000	6.4253	1.0435	0.030194
256	E1	T	4	2.0403	7.4297	0.7535	0.049441
299	E1	E	10	1.1461	11.3377	1.1803	0.002113
344	E1	V	4	1.0048	7.4823	0.7955	0.014141
345	E1	V	5	0.5658	11.5862	1.1088	0.001006
384	E2	G	14	29.5029	51.1128	3.3881	0.007910
387	E2	T	7	5.3423	23.4155	1.6168	0.000160
391	E2	T	10	11.6147	32.5702	2.6925	0.000430
392	E2	Q	12	18.4486	40.1153	2.4581	0.001812
395	E2	A	12	14.1786	27.0321	2.3507	0.019224
396	E2	A	5	8.3737	19.2868	1.4449	0.015870
397	E2	Q	14	23.9660	46.9899	3.1238	0.002118
400	E2	T	4	14.6642	25.5451	1.5480	0.037842
401	E2	S	7	11.6985	26.0400	1.3224	0.007101
405	E2	P	10	17.6685	41.2061	2.8140	0.000427
408	E2	S	9	14.5865	27.9279	1.6866	0.015730
434	E2	R	6	8.1469	29.0872	1.6997	0.000471
446	E2	R	7	4.1747	19.2944	1.8528	0.001038
449	E2	A	6	11.5960	21.2848	1.4368	0.042321
461	E2	P	6	3.0045	8.9933	1.0507	0.046611
464	E2	K	9	4.4742	16.9728	1.8067	0.005824
476	E2	E	11	6.4082	18.4766	2.2273	0.010578
478	E2	P	13	5.3996	29.9302	3.1419	0.000011
480	E2	L	6	7.8954	18.9501	1.8105	0.010650
501	E2	Q	6	2.6488	16.8311	1.2420	0.001387
528	E2	S	6	7.7502	20.2781	1.9013	0.011616
557	E2	S	5	4.3884	14.3095	1.3086	0.017058
625	E2	S	3	0.0000	6.9949	0.7401	0.003461
720	E2	L	4	5.0098	12.9942	1.0773	0.018119
781	P7	R	3	2.1147	8.4204	0.8439	0.032037
827	NS2	T	5	6.3462	16.5486	1.4798	0.014667
957	NS2	R	3	3.7921	12.8756	0.9472	0.010893
998	NS2	L	5	6.7012	19.0292	1.1586	0.002951
1008	NS2	R	3	0.5175	10.2809	0.8870	0.001517
1384	NS3	T	7	4.0125	18.4711	1.7966	0.000498
1409	NS3	G	6	7.0433	16.4529	1.7093	0.022446
1644	NS3	F	3	2.2119	10.3117	1.1492	0.013059
1647	NS3	A	2	1.0000	7.5000	0.6700	0.013702
2016	NS5A	R	2	1.1236	6.9851	0.6700	0.030939
2143	NS5A	D	3	1.1566	10.0059	0.9576	0.005090
2146	NS5A	T	5	5.0490	12.9634	0.9289	0.029782
2155	NS5A	P	6	6.3926	14.7076	1.5070	0.019639
2187	NS5A	K	3	1.1201	6.9741	0.7202	0.030762
2372	NS5A	P	6	6.0000	13.5000	1.9217	0.043438
2375	NS5A	D	8	1.1619	7.4946	1.1959	0.024792
2530	NS5B	N	2	0.0000	7.0471	0.7254	0.005813
2534	NS5B	K	3	0.0000	8.9081	0.8120	0.001293
2540	NS5B	H	6	0.0000	11.8204	1.1462	0.000085
2544	NS5B	K	2	0.0000	5.4290	0.6892	0.018404
2550	NS5B	T	5	1.0046	5.9863	0.6305	0.039461
2558	NS5B	I	2	0.0000	6.3629	0.5120	0.006779
2633	NS5B	N	5	3.4324	14.3471	1.6887	0.004994
2720	NS5B	S	3	1.0000	6.5000	0.7967	0.027404
2729	NS5B	Q	2	0.0000	9.2509	0.5374	0.001760
2755	NS5B	S	5	1.1444	8.4682	0.8554	0.012689
2758	NS5B	V	2	0.0000	5.0000	0.4855	0.017342
2964	NS5B	Q	4	1.0656	11.1896	0.9931	0.002288
2984	NS5B	V	3	3.3653	9.0635	0.7309	0.038281

PROTE, protein; CONAA, Consensus aa; NAAVAR, number of aa variants; dS, observed synonymous changes/expected synonymous changes; dN, observed nonsynonymous changes/expected nonsynonymous changes; *H*, entropy; *P*, probability of observing as many or fewer synonymous changes, interpreted as the *P* value for positively selected sites.