

 Open access • Posted Content • DOI:10.1101/2021.04.22.441053

A two-level, dynamic fitness landscape of hepatitis C virus revealed by self-organized haplotype maps — [Source link](#)

Soledad Delgado, Celia Perales, Celia Perales, Celia Perales ...+12 more authors

Institutions: Technical University of Madrid, Spanish National Research Council, Carlos III Health Institute, Autonomous University of Madrid ...+1 more institutions

Published on: 24 Apr 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Fitness landscape

Related papers:

- [Viral Fitness as a Measure of Adaptation](#)
- [Measuring ruggedness in fitness landscapes](#)
- [Evolutionary advantage of small populations on complex fitness landscapes.](#)
- [The rugged adaptive landscape of an emerging plant RNA virus](#)
- [Percolation on fitness landscapes: effects of correlation, phenotype, and incompatibilities](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/a-two-level-dynamic-fitness-landscape-of-hepatitis-c-virus-mzjatv81pm>

1

2 **A two-level, dynamic fitness landscape of hepatitis C**
3 **virus revealed by self-organized haplotype maps**

4

5 **Soledad Delgado^{1#*}, Celia Perales^{2,3,4#}, Carlos García-Crespo³, María Eugenia**
6 **Soria^{2,3}, Isabel Gallego^{3,4}, Ana Isabel de Ávila³, Brenda Martínez-González², Lucía**
7 **Vázquez-Sirvent², Cecilio López-Galíndez⁵, Federico Morán⁶ and Esteban**
8 **Domingo^{3,4*}**

9 *¹Departamento de Sistemas Informáticos, Escuela Técnica Superior de Ingeniería de Sistemas*
10 *Informáticos (ETSISI). Universidad Politécnica de Madrid, 28031, Madrid, Spain, ²Department*
11 *of Clinical Microbiology, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz*
12 *University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM) Av. Reyes Católicos 2,*
13 *28040 Madrid, Spain, ³Centro de Biología Molecular “Severo Ochoa” (CSIC-UAM), Consejo*
14 *Superior de Investigaciones Científicas (CSIC), Campus de Cantoblanco, 28049, Madrid,*
15 *Spain, ⁴Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas*
16 *(CIBERehd), Instituto de Salud Carlos III, 28029, Madrid, Spain, ⁵Unidad de Virología*
17 *Molecular. Laboratorio de Referencia e Investigación en retrovirus. Centro Nacional de*
18 *Microbiología. Instituto de salud Carlos III. Majadahonda, Madrid. Spain, ⁶Departamento de*
19 *Bioquímica y Biología Molecular, Universidad Complutense de Madrid, 28040, Madrid, Spain*

20

21 [#]Both authors have contributed equally to this work; the names are written in
22 alphabetical order.

23 *Addresses for correspondence:

24 mariasoledad.delgado@upm.es; edomingo@cbm.csic.es

25

26 Keywords: viral quasispecies; mutant spectrum; genome diversification; NS5A-NS5B
27 amplicons; haplotype frequency; SOM network; prototype vector; fitness platform

28

29 **ABSTRACT**

30 Fitness landscapes reflect the adaptive potential of viruses. There is no
31 information on how fitness peaks evolve when a virus replicates extensively in a
32 controlled cell culture environment. Here we report the construction of Self-Organized
33 Maps (SOMs), based on deep sequencing reads of three amplicons of the NS5A-NS5B-
34 coding region of hepatitis C virus (HCV). A two-dimensional neural network was
35 constructed and organized according to sequence relatedness. The third dimension of
36 the fitness profile was given by the haplotype frequencies at each neuron. Fitness maps
37 were derived for 44 HCV populations that share a common ancestor that was passaged
38 up to 210 times in human hepatoma Huh-7.5 cells. As the virus increased its adaptation
39 to the cells, the number of fitness peaks expanded, and their distribution shifted in
40 sequence space. The landscape consisted of an extended basal platform, and a lower
41 number of protruding higher fitness peaks. The function that relates fitness level and
42 peak abundance corresponds a power law, a relationship observed with other complex
43 natural phenomena. The dense basal platform may serve as spring-board to attain high
44 fitness peaks. The study documents a highly dynamic, double-layer fitness landscape of
45 HCV when evolving in a monotonous cell culture environment. This information may
46 help interpreting HCV fitness landscapes in complex in vivo environments.

47 **IMPORTANCE**

48 The study provides for the first time the fitness landscape of a virus in the course
49 of its adaptation to a cell culture environment, in absence of external selective
50 constraints. The deep sequencing-based self-organized maps document a two-layer
51 fitness distribution with an ample basal platform, and a lower number of protruding,
52 high fitness peaks. This landscape structure offers potential benefits for virus resilience
53 to mutational inputs.

54 **INTRODUCTION**

55 High viral mutation rates lead to generation of complex and dynamic mutant
56 spectra termed viral quasispecies, which are important for adaptability to changing
57 environments (1). In the case of hepatitis C virus (HCV), quasispecies complexity in
58 infected patients (quantified as the number of different genomes estimated to be present
59 in the replicating mutant ensembles, as sampled from serum and liver samples) can
60 exert an influence on disease progression and response to antiviral treatment [(2-4);

61 reviewed in (5, 6)]. An understanding of the mechanisms that modify mutant
62 distributions *in vivo* can be facilitated by minimizing the number of selective constraints
63 during viral replication. This can be approached with cell culture systems that sustain
64 long-term virus replication, as is the case of HCV replicating in Huh-7.5 cells (7-10).

65 The objective of the present work has been to determine fitness landscapes of
66 sequential HCV populations replicating in a non-coevolving cellular environment,
67 devoid of externally applied selective constraints. Only perturbations inherent to the cell
68 culture and the changing mutant spectrum of the replicating virus were present (11). The
69 study analyzes the haplotype relatedness and frequencies in a clonal HCV population,
70 and its derivatives resulting from up to 210 serial passages in human hepatoma Huh-7.5
71 cells (equivalent to about 730 days of continuous replication) (11-14). The starting,
72 clonal HCV population was generated by transcription of plasmid Jc1FLAG2(p7-
73 nsGluc2A) (15), followed by RNA electroporation into Huh-Lunet cells, and minimum
74 amplification of the progeny virus in Huh-7.5 cells (12). In this experimental design,
75 fresh cells were infected with the virus shed into the cell culture medium of the previous
76 infection, so that cellular evolution was prevented. Each passage involved infection of
77 4×10^5 Huh-7.5 reporter cells with 4×10^4 to 4×10^6 HCV TCID₅₀ units (depending on
78 the passage number) (11, 13). The multiplicity of infection (MOI) was 0.1 to 10
79 TCID₅₀/cell. Under these conditions, possible distorting effects of stochasticity on
80 quasispecies structure should be limited (16), and accumulation of defective genomes
81 was largely avoided. This was suggested by the constancy of specific infectivity along
82 200 passages, and the fact that biological and molecular clones retrieved from the
83 passaged virus displayed similar sequence diversification (11).

84 Our previous comparative analyses of the initial HCV population (termed HCV
85 p0) and the populations at passages 100 and 200 (termed HCV p100 and HCV p200,
86 respectively) revealed several phenotypic modifications, concomitantly with the process
87 of adaptation to cell culture. The modifications included enhanced resistance to antiviral
88 agents in the absence of specific inhibitor-escape amino acid substitutions (12, 17-21),
89 as well as increases in virus particle density, in capacity to kill host cells, and in the
90 extent of shutoff of host cell protein synthesis (13). Both, HCV p100 and HCV p200
91 exhibited a 2.3-fold increase in replicative fitness, relative to the initial population HCV
92 p0 arbitrarily assigned a fitness of 1.0, as measured by growth-competition experiments
93 in Huh-7.5 cells (13, 17). The reason why fitness did not increase from passage 100 to
94 200 may lie in viral population size limitations, as previously documented with

95 vesicular stomatitis virus (22, 23). However, not all replicative parameters —calculated
96 for each population individually— plateaued at passage 100. In a five serial passage test
97 in Huh-7.5 cells, the intracellular exponential growth rate was 17- and 45-fold larger for
98 HCV p100 and HCV p200, respectively, than for HCV p0 (13). In contrast, the
99 maximum extracellular infectious progeny attained was 1.17-fold higher for both HCV
100 p100 and HCV p200 than for HCV p0, in agreement with the leveling-off of fitness
101 values denoted by the competition experiments (13).

102 Deep-sequencing of the genome of populations that were sampled to monitor the
103 evolution from HCV p0 to HCV p200 revealed a large number of mutations that varied
104 in frequency even between successive passages; we referred to the effect of these types
105 of mutations as mutational waves (13). Strikingly, the waves did not subside when the
106 population had increased its adaptation to the cellular environment since they were even
107 more pronounced at late than at early viral passages (11, 24, 25). This seemingly
108 paradoxical observation begged for an examination of the possible modifications
109 underwent by the fitness landscape of individual HCV populations in their transition
110 from HCV p0 to HCV p200.

111 Previous studies have afforded evidence that fitness landscapes for RNA viruses
112 replicating in their natural environments are rugged and variable (16, 26-33). Fitness
113 effects of mutations or amino acid substitutions have often been inferred from predicted
114 or experimentally verified activity or stability of virus-coded proteins, or from the
115 replicative performance of reconstructed viruses (34-39). An alternative approach has
116 been to derive fitness landscapes from mutation frequencies calculated either from
117 standard (consensus) sequences or from deep-sequencing data (27, 31, 38, 40).

118 There is no information on fitness landscapes of viral populations that have been
119 extensively passaged in a cell culture environment, in absence of external selective
120 constraints, as is the case of the evolution from HCV p0 to HCV p200. The abundance
121 of genome types in a mutant spectrum is ranked according to relative fitness [reviewed
122 in (16)]. This has been the rationale to investigate the fitness landscape of individual
123 HCV populations, based on haplotype abundance derived from ultra-deep sequencing
124 (UDS) reads. To this aim, we have applied an Artificial Neural Network (ANN)
125 procedure as a learning method to derive Self-Organized Maps (SOM) (41, 42). The
126 Kohonen's SOM algorithm classifies a set of input data-vectors (in our case viral
127 genomic sequences) in a bi-dimensional map. By an unsupervised process, it groups
128 data vectors by similarity, projecting those vectors that have similar content in

129 neighboring regions of the map (two-dimensional grid). In the case of viral genome
130 sequences, the SOM algorithm generates an ordered grid in which each node (neuron) is
131 associated with a reference RNA sequence (30, 43). Each neuron of the network maps
132 all input sequences that fall within a distance from its reference vector which is smaller
133 than the distance to the rest of reference vectors. Since vectors represent viral genomic
134 sequences, to calculate numerical distances between vectors, a codification algorithm
135 has been used, as previously described (43) (details are given in Materials and Methods
136 and Fig. S1 in <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). The SOM analysis of
137 44 HCV populations derived from HCV p0, HCV p100 and HCV p200 has disentangled
138 the fitness landscape during long-term adaptation of HCV to Huh-7.5 cells. The SOM
139 display has defined a remarkable HCV fitness topology consisting of a discrete number
140 of high fitness peaks emerging from a lower fitness layer that approximates a fitness
141 platform. The landscape is highly dynamic as evidenced by an almost complete shift in
142 the region of sequence space occupied by the analyzed amplicons during the last 100
143 serial passages. Implications of the two-level fitness topology are discussed.

144

145 **RESULTS**

146 **Self-organized maps and fitness landscape of mutant spectra of HCV populations,**
147 **obtained from haplotype abundances.** A clonal HCV p0 population derived from
148 plasmid Jc1FLAG2(p7-nsGluc2A) (12, 15) was subjected to 200 serial passages in
149 Huh-7.5 reporter cells, and samples from the initial population and from HCV p100 and
150 HCV p200 were further passaged up to ten times in two separate experiments and
151 several replicas, thus providing a total of 44 HCV populations for deep sequencing
152 analysis (Fig. 1A). Three amplicons (termed A1, A2 and A3), extending from HCV
153 genome residues 7649 to 8653 (residue numbering according to isolate JFH-1;
154 accession number #AB047639) were analyzed (Fig. 1B). Amplicon A1 spans residues
155 7649 to 7960 (that correspond to amino acids 461 of NS5A to amino acid 98 of NS5B).
156 A2 spans residues 7940 to 8257 (amino acids 92 to 197 of NS5B), and A3 covers
157 residues 8231 to 8653 (amino acids 189 to 329 of NS5B). The number of processed,
158 clean reads and the deduced number of haplotypes (number of identical reads
159 represented by a nucleotide sequence) for the three amplicons are given in Table 1. For
160 each of the 44 viral populations and amplicon, a FASTA file with the haplotype
161 sequences, including the HCV genomic sequence contained in plasmid Jc1FLAG2(p7-

162 nsGluc2A) —which is also used as reference for mutation counting— was prepared (Fig.
163 S2 in <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). The sequence of each
164 haplotype was labeled with a name, the number of identical sequences that define it, and
165 its frequency in each population (Fig. S3 in
166 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). We employed the 3D irregular
167 codification (43) to transform nucleotide sequences into numerical vectors. In this
168 procedure, each nucleotide is located at a vertex of an irregular tetrahedron with
169 distance 1 between A-G and C-U vertices, and distance 2 between the rest of pairs; that
170 is, a distinction is made among mutation types. The codified sequences in each
171 amplicon were used to train a SOM that comprised a set of neurons, each with a
172 prototype vector, organized in a 15x15 two-dimensional (2D) neuron grid; a different
173 SOM was trained for each amplicon. In this way, training sequences were mapped
174 around the neuron with the prototype vector that best matched in terms of Euclidean
175 distance [the “best matching” unit or bmu (41, 42); see Materials and Methods for
176 additional information on SOM derivation)]. In our sequence analysis and processing,
177 no insertion-deletions (*indels*) were recorded. [Their exclusion is justified by our
178 evidence that they may arise artifactually in homopolymeric tracts upon RNA
179 amplification (13)]. On the 2D neuron grid, the third dimension is given by the
180 frequency of each group of sequences mapped around each neuron, thereby unfolding
181 into a three-dimensional (3D) fitness map for the 44 HCV populations depicted in Fig.
182 1A. The 15 x 15 2D grids with sequence identification of each neuron, 3D maps, and
183 tabulated numerical values for each population, experiment and amplicon are compiled
184 in Figs. S4 to S6 in <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>, with additional
185 numerical information in links quoted therein). Since no major differences were
186 observed between experiment 1 and 2 and among parallel passage replicas, composite
187 fitness maps that included all the populations derived either from HCV p0, HCV p100
188 or HCV p200, were obtained for each amplicon. The resulting maps (Fig. 2) reveal an
189 expansion of the total number of haplotypes and fitness peaks in the evolution from
190 HCV p0 to either HCV p100 or HCV p200 (fold-increase range of 2.1 to 5.3 for
191 haplotypes, and of 2.0 to 4.0 for fitness peaks). Concerning the number of different
192 haplotypes (given inside each panel of Fig. 2), the increase was significant in the
193 evolution from HCV p0 to HCV p100 ($p = 0.0439$; t-test), and from HCV p0 to HCV
194 p200 ($p = 0.0015$; t-test). The difference between HCV p100 and HCV p200 did not
195 reach statistical significance ($p = 0.2687$; t-test).

196 The difference in the number of fitness peaks between HCV p0 and HCV p100
197 and between HCV p0 and HCV p200 was statistically significant ($p = 0.0295$ and $p =$
198 0.0004 , respectively; t-test). The difference between HCV p100 and HCV p200 did not
199 reach statistical significance ($p = 0.1857$; t-test). The distribution of the number of peaks
200 as a function of peak height was indistinguishable for the three viral populations and
201 amplicons ($p = 1$; chi-square test). In all cases, there is an accumulation of the number
202 of fitness peaks within the peak height range 0-1 (graphics in Fig. 2). Considering the
203 three amplicons together, the number of fitness peaks in range 0-1, 1-2, and all other
204 range values was 18, 4, 5, respectively, for HCV p0; the corresponding values were 43,
205 10, 25 for HCV p100, and 66, 11, 17 for HCV p200 (data in Fig. 2). The bias is also
206 evidenced by the ratio of fitness peaks with the minimal range (0-1) sequence frequency
207 (the third dimension of the fitness maps color coded in Fig. 2) relative to the number of
208 peaks that fall into any other range. The average ratio for all amplicons and populations
209 was 0.66 (range 0.46-0.78). The dominance is also recapitulated in the function that
210 relates the number of peaks with their sequence abundance at each neuron (third
211 dimension in the fitness plot) (equations given in the legend for Fig. 2). Interestingly,
212 the functions identify a power law for each virus and amplicon, unveiling for fitness
213 maps a type of relationship found with other complex phenomena in physics and
214 biology (see Discussion).

215 A shift in the occupation of sequence space (position of peaks in the 2D grid)
216 was observed in all cases (Fig. 2). The position of the fitness peaks moved in their
217 location in the three populations, except for the most prominent peak of amplicon 2 in
218 HCV p0 that was also present in HCV p100. This shared peak was represented by a
219 haplotype of identical sequence in each of the HCV p0 and HCV p100 populations that
220 were integrated into the maps depicted in Fig. 2; the sequence was coincident with that
221 present in the parental plasmid Jc1FLAG2(p7-nsGluc2A) (Table S1 in
222 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>); haplotype alignments are available in
223 (<https://saco.csic.es/index.php/s/586L2f9jJQtbRXq>). The consistency of peak display
224 among replicas of the same population (data given in Figs. S4 to S6 in
225 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>) validates the differences observed
226 among different populations. Therefore, the replicative fitness increase in the evolution
227 from HCV p0 to either HCV p100 or HCV p200 was reflected mainly in the number of
228 low fitness peaks that occupied an increased, albeit shifting, portion of sequence space.

229 **Shared and unique fitness peaks among HCV populations.** To express quantitatively
230 the spread of mutant spectra in sequence space upon evolution from HCV p0 to HCV
231 p100 and HCV p200, the number of shared and unique fitness peaks was recorded (Fig.
232 3). The ratio of number of unique peaks in HCV p0 relative to the number of peaks
233 shared by the three populations was 2.5, 6 and 2 for amplicons A1, A2 and A3,
234 respectively; the corresponding ratios were 14, 14, 11 for HCV p100, and 14, 30, 12.5
235 for HCV p200. The values were similar when the ratio was calculated relative to the
236 number of peaks shared with any of the other populations; in this case, the ratios for
237 HCV p0 were 2.5, 3, 4 for amplicons A1, A2, A3, respectively, and increased to 14, 7,
238 22 for HCV p100, and to 14, 30, 25 for HCV p200. For HCV p0, the difference between
239 the number of unique versus shared peaks was not statistically significant: $p = 0.50$, $p =$
240 0.17 , and $p = 0.50$ for amplicons A1, A2, and A3, respectively (proportion test). In
241 contrast, for HCV p100 and HCV p200, the bias in favor of unique versus shared peaks
242 was highly significant. For HCV p100 the p values obtained were $p = 1.76 \times 10^{-7}$, $p =$
243 0.00135 , and $p = 1.209 \times 10^{-6}$ for amplicons A1, A2, and A3, respectively (proportion
244 test). For HCV p200 the p values obtained were $p = 1.76 \times 10^{-7}$, $p = 7.392 \times 10^{-12}$, and p
245 $= 9.972 \times 10^{-9}$ for amplicons A1, A2, and A3, respectively (proportion test). Therefore,
246 the diversification and progressive occupation of sequence space by clonal HCV upon
247 replication in Huh-7.5 cells is confirmed by the number of unique fitness maxima in
248 HCV p100 and HCV p200. The largest increase was scored by amplicon 2, in
249 agreement with the quantification of haplotypes and fitness peaks (compare Figs. 2 and
250 3).

251 **Fused amplicons.** To produce a global image of the fitness landscape of the genomic
252 region analyzed by incorporating the information of the three amplicons in a single
253 graphic, it was necessary to equalize their length in nucleotides. Since the amplicons
254 have overlapping sequences (Fig. 1B), we completed for each amplicon a length of
255 1005 nucleotides using the missing information provided by the other amplicons from
256 the same population (procedure detailed in Materials and Methods). Then a 25×25
257 Kohonen's ANN was trained using all the fused haplotypes. Fitness maps were built for
258 each of the 44 populations, based on haplotype frequencies mapped around each neural
259 unit (Figs. S7 to S9 in <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). Since no
260 major differences were noted among the individual fitness maps, a composite landscape
261 was recapitulated for HCV p0, HCV p100 and HCV p200, each together with its

262 derived populations. The results (Fig. 4) illustrate the peak dispersion upon evolution
263 from HCV p0 to HCV p100 and HCV p200, and renders evident a striking location
264 displacement of fitness peak abundance within the 2D grid. In particular, peaks in HCV
265 p100 and HCV p200 clumped at opposite grid localities. Interestingly, in the process of
266 amplicon fusion the power law that related number of fitness peaks with peak height for
267 individual amplicons was no longer found for HCV p100 and HCV p200 (equations
268 given in the legend for Fig. 4) (see Discussion). In conclusion, despite prolonged
269 replication in a non-evolving cellular environment the HCV fitness landscape appears as
270 remarkably broad, rugged, dynamic, and that approximates a two-layer peak height
271 distribution.

272 **Mutation types and amino acid substitution tolerance in haplotypes from low and**
273 **high fitness peaks.** To analyze a possible difference in mutation types and amino acid
274 substitution tolerance between sequences found in low and high fitness peaks, the peaks
275 were divided in two groups: one with the sequences that populate fitness peaks of height
276 range 0-1, and another group with sequences in peaks of height range 2-3 and higher
277 (peak height distributions given in Fig. 2). Using the HCV sequence in plasmid
278 Jc1FLAG2(p7-nsGluc2A) as reference (15), the ratio of transition versus transversion
279 mutations increased in a similar proportion for the haplotypes present in low and high
280 fitness values of HCV p100 and HCV p200. A similar increase was found for the ratio
281 of synonymous versus non-synonymous mutations (Fig. S10A, B in
282 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). Amino acid acceptability was
283 determined with the PAM 250 matrix (44). The low acceptability substitution group
284 (PAM 250 < 0) was less abundant in the haplotypes of the high fitness peaks of
285 population HCV p200, but the difference with low fitness peaks was not statistically
286 significant (Fig S10C in <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). The
287 comparisons of mutation types and amino acid substitution tolerance mark only
288 tendencies in the diversification process. The fitness of the genomes whose mutations
289 conform the haplotypes sampled in low or high fitness peaks may be dictated by
290 mutations located anywhere in the genome. The decrease of amino acid substitutions
291 with PAM250 < 0 in haplotypes of the high fitness peaks of HCV p200 may be due to
292 negative selection acting on the genomes harboring them. Such a decrease is the only
293 distinctive feature that we have identified in the mutation repertoire of the populations

294 examined (compiled in Table S2 to S4 in
295 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>).

296

297 **DISCUSSION**

298 Genetic variability of RNA (and many DNA) viruses is a major feature of their
299 biology, and an obstacle for disease control. Numerous analyses of clinical and
300 laboratory isolates have indicated that HCV is one of the most genetically variable RNA
301 viral pathogens. Its plasticity results in considerable phenotypic heterogeneity that can
302 influence disease progression and the effectiveness of antiviral interventions [reviews in
303 (5, 6)]. Yet, the information on HCV fitness landscapes is very limited, and it has been
304 largely restricted to genomic sequences from infected patients and centered on the effect
305 of antiviral interventions on viral population composition. In this line, a HCV sequence
306 database was translated into an empirical fitness landscape to design vaccines that might
307 simultaneously decrease viral fitness and avoid selection of escape mutants (45).

308 Our previous studies evidenced wide and dynamic diversification of mutant
309 spectra in the evolution from the initial HCV p0 to HCV p100 and HCV p200 (11, 13,
310 18); haplotype alignments are available in
311 <https://saco.csic.es/index.php/s/586L2f9jJQtbRXq>). As an interpretation of the results,
312 we proposed broadly diversifying selection as an attribute of viral quasispecies
313 dynamics, manifested when viruses replicate in environments that do not experience
314 external perturbations (24). A likely driver of broadly diversifying selection is the
315 modification of mutant spectrum composition due to mutational input (11, 24, 46). The
316 diversity indices quantified in previous studies did not inform of the relationships
317 among the sequences present in mutant spectra. Such relationships have been
318 approached in the present study with the ANN method SOM developed by Kohonen
319 and colleagues (41, 42, 47). In this manner, the sequence information has yielded a
320 fitness landscape of each HCV population. The SOM procedure has been previously
321 used to determine the fitness landscape of HIV-1 clones and populations (30). Other
322 applications have included the interpretation of patterns of cellular gene expression (48,
323 49), or the analysis of taxonomic clustering of cellular and viral RNA sequences (43). In
324 connection with HCV, SOM clustering was used to investigate hepatocellular
325 carcinoma (HCC) development as the basis for tumor differentiation and invasiveness,
326 from expression levels of 12,600 genes in 50 HCC samples from patients with positive

327 HCV serology (50). Also, Kohonen's ANN were trained to predict undiagnosed HCV
328 infections and infection risk (51).

329 The SOM analysis of HCV populations has revealed a two-layer fitness
330 landscape. The first layer consists of multiple low fitness peaks that tend to form a
331 broad platform, covering multiple points in sequence space. Although of limited
332 extension, this platform was discernible in population HCV p0, implying that it was
333 already initiated with the rounds of genome multiplication that followed the initial RNA
334 transfection to produce HCVcc, and then the limited number of infection cycles to
335 obtain population HCV p0 (Fig. 1A) (12). In the evolution towards HCV p100 and
336 HCV p200, the populations maintained the same pattern, with low fitness peaks
337 expanded towards larger areas of sequence space. The basal fitness platform is adorned
338 with a limited number of protruding fitness peaks that resemble the standard
339 representation of a rugged fitness landscape in the Wrightian sense (52). Interestingly,
340 the function that relates the number of fitness peaks with peak height corresponds to a
341 power law since it has the form $y = ax^{-b}$ (the equations for different amplicons and viral
342 populations are given in the legend for Fig. 2, and the confirmation of a straight line in a
343 log-log plot of the same data is shown in Fig. S11 in
344 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>). A power law describes scale-free
345 (non-random) processes in physics and biology that have some underlying dynamic
346 event in their construction (53, 54). In the power law discovered with the HCV fitness
347 landscape, the underlying force may be mutation, with the power law reflecting far
348 more frequent pathways to reach the first fitness platform than high fitness peaks, with
349 the latter requiring organized, non-random, clusters of mutations. Interestingly, the
350 power law relationship was lost for HCV p100 and HCV p200 when the fused
351 amplicons were used for the graphics (equations in the legend for Fig. 4). This may be
352 due to a number of ambiguous genome positions that were generated in the amplicon
353 fusion process (described in Materials and Methods). This point, as well as further
354 penetration in the significance of this particular power law, require further research.

355 Despite the similar landscape morphology, the fitness maps of HCV p0, HCV
356 p100 and HCV p200 show differences in peak distribution, either considering individual
357 amplicons or a fused single 2D SOM network that recapitulates the information from
358 the three amplicons (Figs. 2 and 4). In particular, the comparison evidences dynamics of
359 peak movements, with striking differences between HCV p100 and HCV p200 despite
360 the two populations having reached the same fitness value as measured by the standard

361 growth-competition assays (13, 17). The only biochemical parameter that we identified
362 —and that may fuel the dynamics of change from HCV p100 to HCV p200— is a 2.6-
363 fold larger intracellular exponential growth displayed by HCV p200 relative to HCV
364 p100 (13). However, the three NS5B amplicons did not follow the same trajectory of
365 fitness modification. While for amplicons 1 and 3 the ratio of peaks or haplotypes
366 unique to the population to those shared by other populations was the same for HCV
367 p100 and HCV p200, for amplicon 2 it was two times higher for HCV p200 than HCV
368 p100 (derived from the graphics of Figs. 2 and 4, and included in Table S1 in
369 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>).

370 The comparison of fitness landscapes has not revealed traceable evolutionary
371 trajectories. No minority haplotypes in the ancestral HCV populations are the ancestors
372 of most haplotypes that stand as dominant in subsequent populations. Rather, the picture
373 obtained is that of a network of interconnected, transient sequences that do not define
374 linear evolutionary events. Despite the absence of sub-lineages with temporal
375 continuity, the number of identical fitness peaks that arose in independent passage
376 replicas of the same starting population is remarkable [50.4% (range 37.5% - 72.7%) of
377 the total for replicas (a), (b), and (c) of populations HCV p0, HCV p100, and HCV
378 p200, subjected to four serial passages in Huh-7.5 cells (Table S5 in
379 <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>)]. Similarity of behavior in separate
380 evolutionary viral lineages suggests a component of determinism (predictability) in a
381 system whose evolution should be strongly directed by stochastically arising mutations.
382 This paradoxical behavior has been previously observed in different studies with other
383 RNA viruses, and a number of possible underlying mechanisms have been proposed
384 (55-59).

385 A more realistic perception of the complexity of the HCV fitness landscape can
386 be obtained by considering that the SOM maps have been constructed with haplotypes
387 from amplicons that cover only 10% of the entire HCV genome. This is a limitation of
388 our study, although achieving a similar depth of mutation detection for whole genome
389 amplicons than short amplicons is still technically challenging. A more populated basal
390 platform than displayed in the SOM graphics of Figs. 2 and 4 is predicted if the analysis
391 of haplotype frequencies were extended to additional genomic sites. The reason is that
392 the sites of heterogeneity —defined as those with more than one nucleotide, revealed by
393 Sanger sequencing— were found along the entire genome of the same HCV populations
394 (11).

395 The fitness landscapes resulting from HCV replication in a monotonous
396 environment can serve as a basis for comparison with the landscapes acquired when a
397 selective constraint is applied to the evolving population. In particular, the analysis
398 should reveal if alternative mutational pathways are available to the virus to respond to
399 a specific constraint. Also, how the HCV fitness is shaped in patients versus the cell
400 culture environment may be informative of adaptive mechanisms, and such a work is
401 now in progress.

402 A two-layer fitness distribution may have biological consequences. The first
403 layer or platform may prevent mutations from driving genomes into low fitness pits. It
404 may also act as a spring-board for viral populations to reach higher fitness peaks. This
405 should reduce the transition time between fitness peaks which is a limitation of
406 adaptability recognized in general evolutionary genetics (60-62).

407

408 MATERIALS AND METHODS

409 **Origin of the HCV populations, and serial passages in Huh-7.5 reporter cells.** The
410 initial HCVcc population was obtained by *in vitro* transcription of plasmid
411 Jc1FLAG2(p7-nsGluc2A) (15), followed by RNA electroporation into Huh-Lunet cells,
412 and further amplification in Huh-7.5 reporter cell monolayers to yield the parental
413 population HCV p0 (12). HCV p0 was further passaged in Huh-7.5 reporter cells to
414 obtain HCV p100 and HCV p200 (HCV p0 that has been propagated 100 and 200 times
415 in Huh-7.5 reporter cells, respectively), as has been previously described (13). The
416 sequences to derive the SOM-based fitness landscape were obtained from the three
417 parental HCV p0, HCV p100 and HCV p200 populations subjected to further passages
418 in two different experiments (experiment 1, and experiment 2) and several replicas. This
419 yielded the 44 populations for which a fitness landscape was determined (populations
420 depicted as empty circles in Fig. 1A). The sequences on which the present study is
421 based have been previously reported (11, 18), and are available in
422 (<https://saco.csic.es/index.php/s/586L2f9jJQtbRXq>). To control for the absence of
423 cross-contamination with virus from another population or replica, mock-infected cells
424 were maintained in parallel with each infected culture, and each supernatant was
425 titrated; no infectivity in the mock-infected cultures was detected in any of the
426 experiments.

427 Experiments of short-term evolution (up to 10 serial passages) starting from
428 HCV p0, HCV p100 and HCV p200 (Fig. 1A) were carried out also in Huh-7.5 reporter
429 cells. To initiate serial passages, 4×10^5 Huh-7.5 reporter cells were infected with HCV
430 p0, HCV p100 and HCV p200 at a multiplicity of infection (MOI) of 0.03 TCID₅₀/cell;
431 for subsequent passages 4×10^5 fresh, Huh-7.5 reporter cells were infected with the
432 virus contained in 0.5 ml of the cell culture medium from the previous infection of the
433 same lineage; the multiplicity of infection (MOI) ranged from 0.1 to 0.5 TCID₅₀ per
434 cell. In all passages, infections were allowed to proceed for 72h to 96h. Additional
435 procedures, including titration of infectivity to determine TCID₅₀ values, and viral RNA
436 quantification have been previously described (11-13).

437

438 **RNA extraction, viral RNA amplification and ultra-deep sequencing of cell culture**
439 **populations.** Intracellular viral RNA was extracted from the initial HCV p0, HCV p100
440 and HCV p200 populations, and their passaged derivatives using the Qiagen RNeasy kit
441 (Qiagen, Valencia, Ca, USA). HCV RNA was amplified by RT-PCR using Accuscript
442 (Agilent), and specific HCV oligonucleotide primers that have been previously
443 described [Table S10 of (11)]. Agarose gel electrophoresis was used to analyze the
444 amplification products, using Gene Ruler 1 Kb Plus DNA ladder (Thermo Scientific) as
445 molar mass standard. To ascertain absence of contaminating templates, all experiments
446 included negative controls without template RNA. To avoid sequence representation
447 biases due to redundant amplifications of the same initial RNA templates due to
448 template molecule limitations, amplifications were carried out with template
449 preparations diluted 1:10, 1:100 and 1:1000; only when at least the 1:100 diluted
450 template produced a visible DNA band, was molecular cloning performed using the
451 DNA amplified from the undiluted template sample. PCR products were purified
452 (QIAquick Gel Extraction Kit, QIAGEN), quantified (Pico Green assay), and tested for
453 quality (Bioanalyzer DNA 1000, Agilent Technologies) prior to Illumina deep
454 sequencing analysis (MiSeq platform, with the 2x300-bp mode with v3 chemistry).

455 Several control experiments were performed in preparation of the ultra-deep
456 sequencing procedure to ensure the reliability of the mutations derived from clean reads,
457 for proper mutant spectrum characterization. They have been previously described (63-
458 66), and they were as follows: first, we determined the basal error of the amplification
459 and sequencing process, using an infectious HCV cDNA clone to perform RT-PCR, the
460 nested PCR, and ultra-deep sequencing using Illumina MiSeq. Second, we quantified

461 the PCR recombination frequency during the amplification steps using mixtures of *wt*
462 and a mutant clone to perform RT-PCR, and the nested PCR and ultra-deep sequenced
463 using Illumina MiSeq. Third, we ascertained the similarity of read composition in
464 different RT-PCR amplifications and sequencing runs, using different samples of the
465 same RNA preparation. We concluded that mutations identified with a frequency above
466 the 0.5% cut-off value, and that were consistently found in the two DNA strands were
467 considered for the analyses. For additional details of the read cleaning procedures,
468 criteria for mutation acceptance, and experimental controls with reconstructed HCV
469 RNA mutant mixtures, see (63-66).

470 **SOM derivation.** Detailed description of the SOM algorithm has been published
471 elsewhere (43). The ANN model (41) exhibits an architecture consisting of a set of
472 neurons arranged in a rectangular grid that define a neighborhood relationship. The map
473 size has been chosen to ensure sufficiently dispersion of the sequences mapped in the
474 grid, while preserving the grouping of those that are similar; the resulting size is a
475 function of the size of the data set. In the case of mapping each amplicon, a 15 x 15 size
476 grid was selected as suitable, while for the map generated with all the fused amplicons,
477 the selected size was 25 x 25, due to the greater number of sequences.

478 Every neuron has an associated prototype vector with the same nature and
479 dimension that the input data set (in this work, the amplicon sequences). SOM generates
480 a projection or mapping of the input space, usually high-dimensional, in the two-
481 dimensional topological structure of the network. The SOM training algorithm
482 determines the way in which this mapping is created. This process iteratively modifies
483 the SOM prototype vectors to fit them to the distribution of the input data space, using a
484 methodology similar to a regression. During the SOM training, each input vector is
485 associated with the neuron that best matches with the pattern in any metric (the so-
486 called ‘best matching unit’, bmu). As a result of this process, the prototype vectors
487 associated with the bmu, and all the neurons located in a neighborhood area around it,
488 are modified in order to move them closer to the input vector. In this work the bmu has
489 been calculated in terms of Euclidean distance. In the case of classification of vectors
490 with sequence data, the algorithm requires a previous transformation into equivalent
491 numeric vectors. This has been done using the previously described codification (30,
492 43). Each nucleotide is transformed into the corresponding 3D numerical coordinates in
493 an irregular tetrahedron (Fig. S1 in <https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>).

494 In this way, each RNA sequence is transformed into a numerical vector of a dimension
495 which is three times the length of the sequence, and this is the vector that is used by the
496 SOM algorithm during the training process. After the training, SOM can determine
497 similarities over the input vectors (amplicon sequences), in the sense that similar
498 sequences will be mapped by the same neuron or by a neighboring neuron.

499 With the dataset of each amplicon, 25 SOM networks of size 15 x 15 were
500 trained, and the one with the lowest Kaski-Lagus error (ϵ_{k-l}) was selected. The same
501 training factors were used: number of input neurons (N); dimension of the dataset
502 vectors (length of the sequence times 3); size of the output map 15 rows times 15
503 columns; hexagon neighborhood connection (each neuron has six neighbors around it:
504 two at the top, two at the bottom, one on the left, and one on the right); initial
505 neighborhood of 14 rows and 14 columns, with neighborhood decrement at the end of
506 each epoch (equivalent to the number of sequences in the dataset); learning factor $\alpha(t)$
507 = $\alpha_1 (1 - t / \alpha_2)$, with $\alpha_1 = 0.1$ and α_2 equal to the total iterations of the training
508 algorithm; the total number of iterations is equal to the total number of epochs times the
509 number of sequences. The total of epochs is determined by the initial neighborhood + 5,
510 that is, the algorithm carries out the necessary epochs so that the neighborhood area
511 decreases until it affects only the bmu, plus 5 additional fine-tuning epochs (total
512 iterations: 19 times number of dataset sequences).

513 Finally, a labeling process was applied to each map. Using as a basis the
514 network selected for each dataset, the 3D fitness maps labeling was generated with the
515 accumulated frequencies for each haplotype, so that each neuron was assigned the sum
516 of frequencies of the haplotypes for which it is the bmu. This value represents the
517 cumulative frequency of sequences that fall in the Voronoi region of the neuron.
518 Although the SOM map is generated or trained with all the sequences of each dataset,
519 the 3D maps can be obtained with the subset of sequences to be represented.

520 **Amplicon fusion method.** Based on the fact that the amplicons have overlapping
521 sequences (Fig. 1B), we completed for each haplotype of an amplicon a length of 1005
522 nucleotides using the missing information provided by the haplotypes of the other two
523 amplicons in the same population, passage number and experiment. To achieve this for
524 amplicon A1 (original length 312 bases), the amplicon A2 haplotypes with initial
525 overlapping sequences matching the last 21 bases of haplotype A1 were located. The
526 same operation was conducted with the amplicon A3 haplotypes whose initial

527 overlapping sequences matched the last 27 bases of any of the A2 haplotypes found in
528 the previous step. Fusion sequences were obtained for the A2 and A3 haplotype lists.
529 To generate the final fusion sequence, the bases of each position were compared,
530 keeping the base for any position with identical nucleotide in all sequences, or the
531 IUPAC nucleotide ambiguity code associated with the combination of the bases when a
532 position had more than one nucleotide. The 312 bases of the amplicon A1 haplotype
533 were completed by adding the last 297 bases of the A2 fusion sequence followed by the
534 last 396 bases of the A3 fusion sequence. A similar procedure was used to derive the
535 amplicon A2 (original length of 318 bases) haplotype fusion sequence. The amplicon
536 A1 haplotypes with final overlapping sequences matching the first 21 bases of
537 haplotype A2, and the amplicon A3 haplotypes with initial overlapping sequences
538 matching the last 27 bases of haplotype A2 were located. The fusion sequence was
539 obtained for the A1 haplotype list and for the A3 haplotype list. The 318 bases of the
540 amplicon A2 haplotype were completed including the first 291 bases of the A1 fusion
541 sequence at the beginning and the last 396 bases of the A3 fusion sequence at the end,
542 as described to complete amplicon 1. Likewise, for amplicon A3 (original length of 423
543 bases), the amplicon A2 haplotypes with final overlapping sequences matching the first
544 27 bases of haplotype A3, and the amplicon A1 haplotypes with final overlapping
545 sequences matching the initial 21 bases of any of the A2 haplotypes found in the
546 previous step were located. The fusion sequence was obtained for the A1 haplotype list
547 and for the A2 haplotype list. The 423 bases of the amplicon A3 haplotype were
548 completed including at the beginning the first 291 bases of the A1 fusion sequence
549 followed by the first 291 bases of the A2 fusion sequence. When no haplotypes with
550 matching overlapping sequences were found in any of the other two amplicons, the full
551 list of haplotypes of the mismatched amplicon was used to equalize the length.

552 **Statistics.** The statistical significance of differences among the number of fitness peaks
553 and among the number of haplotypes of HCV p0, HCV p100 and HCV p200 was
554 calculated with the t-test since the data follow a normal distribution ($p > 0.05$; Shapiro-
555 Wilk test). The differences between the distribution of fitness peaks of HCV p0, HCV
556 p100 and HCV p200 for each amplicon was calculated with the Pearson's chi-square
557 test. The comparison between the number of unique peaks and the number of shared
558 peaks of HCV p0, HCV p100 and HCV p200 for each amplicon, was calculated with a
559 proportion test. All calculations were carried out using software R version 4.0.2.

560 **Data availability.** The Illumina data have been deposited in the NCBI BioSample
561 database under accession numbers SAMN18645452, SAMN18645453,
562 SAMN18645456, SAMN18645457, SAMN18645460, SAMN18645463,
563 SAMN18645464 and SAMN18645467 (BioProject accession number PRJNA720288)
564 for experiment 1 and SAMN13531332 to SAMN13531367 (BioProject accession
565 number PRJNA593382) for experiment 2. The haplotypes alignments are available in
566 <https://saco.csic.es/index.php/s/586L2f9jJQtbRXq>. The fasta files are termed according
567 to the experiment (Exp1, Exp2a, Exp2b or Exp2c), population (HCV p0, HCV p100 or
568 HCV p200), passage (initial, p1, p2, p3, p4 or p10) and amplicon (A1, A2 or A3).

569

570 **ACKNOWLEDGMENTS**

571 The work at UPM was supported by grants TIN2017-085727-C4-3-P (DeepBio)
572 from Ministerio de Ciencia, Innovación y Universidades (MCIU) and PID2019-
573 104903RB-100 (Funded by the EU under the FEDER program). Work at CBMSO was
574 supported by grants SAF2014-52400-R from Ministerio de Economía y Competitividad
575 (MINECO), SAF2017-87846-R and BFU2017-91384-EXP from MCIU, PI18/00210
576 from Instituto de Salud Carlos III (ISCIII), S2013/ABI-2906 (PLATESA from
577 Comunidad de Madrid/FEDER), and S2018/BAA-4370 (PLATESA2 from Comunidad
578 de Madrid/FEDER). C.P. is supported by the Miguel Servet program of the Instituto de
579 Salud Carlos III (CP14/00121 and CPII19/00001), cofinanced by the European
580 Regional Development Fund (ERDF). CIBERehd (Centro de Investigación en Red de
581 Enfermedades Hepáticas y Digestivas) is funded by Instituto de Salud Carlos III.
582 Institutional grants from the Fundación Ramón Areces and Banco Santander to the
583 CBMSO are also acknowledged. The team at CBMSO belongs to the Global Virus
584 Network (GVN). Work at Centro Nacional de Microbiología (ISCIII) was supported by
585 grants SAF2016-77894-R from MINECO and PI13/02269 from ISCIII. C.G.-C. is
586 supported by predoctoral contract PRE2018-083422 from MCIU. B.M.-G. is supported
587 by predoctoral contract PFIS FI19/00119 from ISCIII, cofinanced by Fondo Social
588 Europeo (FSE). The work at Universidad Complutense Madrid has been supported by
589 research grant CTQ2017-87864-C2-2-P, from the Ministerio de Economía y
590 Competitividad (MINECO), Spain.

591

592 FIGURE LEGENDS

593 **FIG 1** Experimental design and HCV amplicon analysis. (A) Schematic representation
594 of the passages underwent by HCV p0 [derived from HCVcc (12); Materials and
595 Methods] in Huh-7.5 reporter cells. Populations are depicted as empty circles and
596 passage number is indicated by p (HCV p100, p3 means population HCV p100
597 subjected to three passages in Huh-7.5 cells). Experiment 1 (upper part) and experiment
598 2 (lower part) were performed starting with samples of the same HCV p0, HCV p100
599 and HCV p200 populations. In experiment 2, (a), (b) and (c) indicate triplicate passage
600 series carried out in parallel. A total of 44 HCV populations (corresponding to the
601 empty circles) were analyzed by deep sequencing. The mutations (and deduced amino
602 acid substitutions) identified in the populations from experiment 1 were reported in
603 (18), and those in the populations from experiment 2 in (11). (B) HCV genomic
604 residues 8261 (NS5A-coding region) to 9265 (NS5B-coding region) (genome
605 numbering according to reference isolate JFH-1), and length in base pairs (bp) of
606 amplicons A1, A2 and A3 analyzed by Illumina MiSeq sequencing. Note that the 21
607 most 3'terminal nucleotides of A1 are redundant with the 21 most 5'terminal
608 nucleotides of A2, and that the 27 most 3'terminal nucleotides of A2 are redundant with
609 the most 5'terminal nucleotides of A3. Further details on virus origin, GenBank
610 accession numbers, and sequencing procedures are given in Materials and Methods.

611 **FIG 2** SOM-derived fitness maps, and number of fitness peaks distributed according to
612 haplotype abundance. The amplicon number is indicated at the top of each panel and
613 graphics group. The three 15 x 15 neuron grids for all populations derived either from
614 HCV p0, HCV p100, or HCV p200 (displayed in Fig. 1A), with the total number of
615 haplotypes that entered the analysis are indicated in each panel. Peak height is
616 determined by sequence abundance, which is color coded with a scale included at the
617 right of each fitness graph. The distribution of number of fitness peaks (ordinate) versus
618 peak height (sequence abundance in unit range displayed in abscissa) is described by the
619 following functions: Amplicon 1: HCV p0: $y=3.7934x^{-0.403}$ ($R^2=0.7672$); HCV p100:
620 $y=8.2657x^{-0.77}$ ($R^2=0.6463$); HCV p200: $y=6.6996x^{-0.709}$ ($R^2=0.5974$). Amplicon 2:
621 HCV p0: $y=3.1334x^{-0.281}$ ($R^2=0.3804$); HCV p100: $y=3.4527x^{-0.395}$ ($R^2=0.3649$); HCV
622 p200: $y=7.0358x^{-0.638}$ ($R^2=0.5818$). Amplicon 3: HCV p0: $y=2.5728x^{-0.233}$ ($R^2=0.3807$);
623 HCV p100: $y=7.453x^{-0.723}$ ($R^2=0.7755$); HCV p200: $y=6.97x^{-0.629}$ ($R^2=0.5886$). Note

624 that scales are not the same in different panels. The origin of the sequences, derived
625 haplotypes, and procedures are described in Materials and Methods.

626 **FIG 3** Distribution of fitness peaks among HCV populations. Venn diagrams indicating
627 for each amplicon the number of peaks unique to one HCV population and those shared
628 by two or more HCV populations. Populations are color coded. Peak identity was
629 determined according to data summarized in Table 1, Fig. 2, and Supplemental Material
630 (<https://saco.csic.es/index.php/s/7TgiQcCr9ifpnt5>).

631 **FIG 4** Fitness maps constructed with the fused NS5B amplicons. The HCV population
632 and number of haplotypes used for the 25x25 neuron graphic are indicated on the left of
633 each fitness map. Peak height is determined by sequence abundance, which is color
634 coded with a scale included at the right of each map. The distribution of number of
635 fitness peaks (ordinate) versus peak height (sequence abundance in unit range displayed
636 in abscissa) is described by the following functions: HCV p0: $y=15.659x^{-1.001}$
637 ($R^2=0.6519$). HCV p100: $y=93.588 x^{2.441}$ ($R^2=0.9931$). HCV p200: $-94.031\ln(x) +$
638 108.16 ($R^2=0.9931$). Note that scales are not the same in different panels. Procedures
639 are described in Materials and Methods.

640 REFERENCES

- 641 1. Domingo E, Schuster P. 2016. Quasispecies: from theory to experimental
642 systems. *Current Topics in Microbiology and Immunology*. Vol. 392. Springer.
- 643 2. Farci P, Shimoda A, Coiana A, Diaz G, Peddis G, Melpolder JC, Strazzer A,
644 Chien DY, Munoz SJ, Balestrieri A, Purcell RH, Alter HJ. 2000. The outcome
645 of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science*
646 288:339-44.
- 647 3. Farci P, Strazzer A, Alter HJ, Farci S, Degioannis D, Coiana A, Peddis G, Usai
648 F, Serra G, Chessa L, Diaz G, Balestrieri A, Purcell RH. 2002. Early changes in
649 hepatitis C viral quasispecies during interferon therapy predict the therapeutic
650 outcome. *Proc Natl Acad Sci U S A* 99:3081-6.
- 651 4. Kumar D, Malik A, Asim M, Chakravarti A, Das RH, Kar P. 2008. Influence of
652 quasispecies on virological responses and disease severity in patients with
653 chronic hepatitis C. *World J Gastroenterol* 14:701-8.
- 654 5. Farci P. 2011. New insights into the HCV quasispecies and
655 compartmentalization. *Semin Liver Dis* 31:356-74.
- 656 6. Domingo E, Sheldon J, Perales C. 2012. Viral quasispecies evolution. *Microbiol*
657 *Mol Biol Rev* 76:159-216.
- 658 7. Lindenbach BD, Evans MJ, Syder AJ, Wolk B, Tellinghuisen TL, Liu CC,
659 Maruyama T, Hynes RO, Burton DR, McKeating JA, Rice CM. 2005. Complete
660 replication of hepatitis C virus in cell culture. *Science* 309:623-6.

- 661 8. Zhong J, Gastaminza P, Cheng G, Kapadia S, Kato T, Burton DR, Wieland SF,
662 Uprichard SL, Wakita T, Chisari FV. 2005. Robust hepatitis C virus infection in
663 vitro. *Proc Natl Acad Sci U S A* 102:9294-9.
- 664 9. Wakita T, Pietschmann T, Kato T, Date T, Miyamoto M, Zhao Z, Murthy K,
665 Habermann A, Krausslich HG, Mizokami M, Bartenschlager R, Liang TJ. 2005.
666 Production of infectious hepatitis C virus in tissue culture from a cloned viral
667 genome. *Nat Med* 11:791-6.
- 668 10. Gottwein JM, Pham LV, Mikkelsen LS, Ghanem L, Ramirez S, Scheel TKH,
669 Carlsen THR, Bukh J. 2018. Efficacy of NS5A Inhibitors Against Hepatitis C
670 Virus Genotypes 1-7 and Escape Variants. *Gastroenterology* 154:1435-1448.
- 671 11. Gallego I, Soria ME, Garcia-Crespo C, Chen Q, Martinez-Barragan P, Khalfaoui
672 S, Martinez-Gonzalez B, Sanchez-Martin I, Palacios-Blanco I, de Avila AI,
673 Garcia-Cehic D, Esteban JI, Gomez J, Briones C, Gregori J, Quer J, Perales C,
674 Domingo E. 2020. Broad and Dynamic Diversification of Infectious Hepatitis C
675 Virus in a Cell Culture Environment. *J Virol* 94:e01856-19.
- 676 12. Perales C, Beach NM, Gallego I, Soria ME, Quer J, Esteban JI, Rice C,
677 Domingo E, Sheldon J. 2013. Response of hepatitis C virus to long-term passage
678 in the presence of alpha interferon: multiple mutations and a common
679 phenotype. *J Virol* 87:7593-607.
- 680 13. Moreno E, Gallego I, Gregori J, Lucia-Sanz A, Soria ME, Castro V, Beach NM,
681 Manrubia S, Quer J, Esteban JI, Rice CM, Gomez J, Gastaminza P, Domingo E,
682 Perales C. 2017. Internal Disequilibria and Phenotypic Diversification during
683 Replication of Hepatitis C Virus in a Noncoevolving Cellular Environment. *J*
684 *Virol* 91:e02505-16.
- 685 14. Garcia-Crespo C, Soria ME, Gallego I, Avila AI, Martinez-Gonzalez B,
686 Vazquez-Sirvent L, Gomez J, Briones C, Gregori J, Quer J, Perales C, Domingo
687 E. 2020. Dissimilar Conservation Pattern in Hepatitis C Virus Mutant Spectra,
688 Consensus Sequences, and Data Banks. *J Clin Med* 9.
- 689 15. Marukian S, Jones CT, Andrus L, Evans MJ, Ritola KD, Charles ED, Rice CM,
690 Dustin LB. 2008. Cell culture-produced hepatitis C virus does not infect
691 peripheral blood mononuclear cells. *Hepatology* 48:1843-50.
- 692 16. Schuster P. 2016. Quasispecies on fitness landscapes. In: E. Domingo and P.
693 Schuster, eds. *Quasispecies: From Theory to Experimental Systems.*, *Curr Top*
694 *Microbiol Immunol* 392: 61-120.
- 695 17. Sheldon J, Beach NM, Moreno E, Gallego I, Pineiro D, Martinez-Salas E,
696 Gregori J, Quer J, Esteban JI, Rice CM, Domingo E, Perales C. 2014. Increased
697 replicative fitness can lead to decreased drug sensitivity of hepatitis C virus. *J*
698 *Virol* 88:12098-111.
- 699 18. Gallego I, Gregori J, Soria ME, Garcia-Crespo C, Garcia-Alvarez M, Gomez-
700 Gonzalez A, Valiergue R, Gomez J, Esteban JI, Quer J, Domingo E, Perales C.
701 2018. Resistance of high fitness hepatitis C virus to lethal mutagenesis. *Virology*
702 523:100-109.
- 703 19. Gallego I, Sheldon J, Moreno E, Gregori J, Quer J, Esteban JI, Rice CM,
704 Domingo E, Perales C. 2016. Barrier-Independent, Fitness-Associated
705 Differences in Sofosbuvir Efficacy against Hepatitis C Virus. *Antimicrob*
706 *Agents Chemother* 60:3786-93.
- 707 20. Perales C. 2018. Quasispecies dynamics and clinical significance of hepatitis C
708 virus (HCV) antiviral resistance. *Int J Antimicrob Agents*
709 doi:10.1016/j.ijantimicag.2018.10.005;doi: 10.1016/j.ijantimicag.2018.10.005.

- 710 21. Domingo E, de Avila AI, Gallego I, Sheldon J, Perales C. 2019. Viral fitness:
711 history and relevance for viral pathogenesis and antiviral interventions. *Pathog*
712 *Dis* 77:ftz021.
- 713 22. Novella IS, Duarte EA, Elena SF, Moya A, Domingo E, Holland JJ. 1995.
714 Exponential increases of RNA virus fitness during large population
715 transmissions. *Proc Natl Acad Sci USA* 92:5841-5844.
- 716 23. Novella IS, Quer J, Domingo E, Holland JJ. 1999. Exponential fitness gains of
717 RNA virus populations are limited by bottleneck effects. *J Virol* 73:1668-71.
- 718 24. Domingo E, Soria ME, Gallego I, de Avila AI, Garcia-Crespo C, Martinez-
719 Gonzalez B, Gomez J, Briones C, Gregori J, Quer J, Perales C. 2020. A new
720 implication of quasispecies dynamics: Broad virus diversification in absence of
721 external perturbations. *Infect Genet Evol* 82:104278.
- 722 25. García-Crespo C, Gallego I, Soria ME, De Ávila AI, Martínez-González B,
723 Vázquez-Sirvent L, Lobo-Vega R, Moreno E, Gómez J, Briones C, Gregori J,
724 Quer J, Domingo E, Perales C. 2021. Population disequilibrium as promoter of
725 adaptive explorations in hepatitis C virus. *Viruses* 13:616.
- 726 26. Kouyos RD, Leventhal GE, Hinkley T, Haddad M, Whitcomb JM, Petropoulos
727 CJ, Bonhoeffer S. 2012. Exploring the complexity of the HIV-1 fitness
728 landscape. *PLoS Genet* 8:e1002551.
- 729 27. Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of
730 an RNA virus revealed through population sequencing. *Nature* 505:686-90.
- 731 28. Lorenzo-Redondo R, Borderia AV, Lopez-Galindez C. 2011. Dynamics of in
732 vitro fitness recovery of HIV-1. *J Virol* 85:1861-1870.
- 733 29. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM,
734 Petropoulos CJ, Bonhoeffer S. 2011. A systems analysis of mutational effects in
735 HIV-1 protease and reverse transcriptase. *Nat Genet* 43:487-489.
- 736 30. Lorenzo-Redondo R, Delgado S, Moran F, Lopez-Galindez C. 2014. Realistic
737 three dimensional fitness landscapes generated by self organizing maps for the
738 analysis of experimental HIV-1 evolution. *PLoS One* 9:e88579.
- 739 31. Seifert D, Beerenwinkel N. 2016. Estimating Fitness of Viral Quasispecies from
740 Next-Generation Sequencing Data. *Curr Top Microbiol Immunol* 392:181-200.
- 741 32. de Visser JA, Krug J. 2014. Empirical fitness landscapes and the predictability
742 of evolution. *Nat Rev Genet* 15:480-90.
- 743 33. Munoz-Moreno R, Martinez-Romero C, Blanco-Melo D, Forst CV,
744 Nachbagauer R, Benitez AA, Mena I, Aslam S, Balasubramaniam V, Lee I,
745 Panis M, Ayllon J, Sachs D, Park MS, Krammer F, tenOever BR, Garcia-Sastre
746 A. 2019. Viral Fitness Landscapes in Diverse Host Species Reveal Multiple
747 Evolutionary Lines for the NS1 Gene of Influenza A Viruses. *Cell Rep* 29:3997-
748 4009 e5.
- 749 34. Sanjuan R, Moya A, Elena SF. 2004. The distribution of fitness effects caused
750 by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A*
751 101:8396-401.
- 752 35. Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new
753 mutations. *Nat Rev Genet* 8:610-8.
- 754 36. Fernandez G, Clotet B, Martinez MA. 2007. Fitness landscape of human
755 immunodeficiency virus type 1 protease quasispecies. *J Virol* 81:2485-96.
- 756 37. Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts
757 for most mutational fitness effects in viruses. *Proc Natl Acad Sci U S A*
758 108:9916-21.

- 759 38. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, Truong S, Remenyi R,
760 Chen Z, Du Y, Su SY, Al-Mawsawi LQ, Wu TT, Chen SH, Lin CY, Zhong W,
761 Lloyd-Smith JO, Sun R. 2014. A quantitative high-resolution genetic profile
762 rapidly identifies sequence determinants of hepatitis C viral fitness and drug
763 sensitivity. *PLoS Pathog* 10:e1004064.
- 764 39. Peris JB, Davis P, Cuevas JM, Nebot MR, Sanjuan R. 2010. Distribution of
765 fitness effects caused by single-nucleotide substitutions in bacteriophage ϕ 1.
766 *Genetics* 185:603-9.
- 767 40. Quadeer AA, Barton JP, Chakraborty AK, McKay MR. 2020. Deconvolving
768 mutational patterns of poliovirus outbreaks reveals its intrinsic fitness landscape.
769 *Nat Commun* 11:377.
- 770 41. Kohonen T. 2001. *Self-Organizing Maps*, vol 501 p. Springer-Verlag.
- 771 42. Kohonen T, Kaski S, Lagus K, Salojarvi J, Honkela J, Paatero V, Saarela A.
772 2000. Self organization of a massive document collection. *IEEE Trans Neural*
773 *Netw* 11:574-85.
- 774 43. Delgado S, Moran F, Mora A, Merelo JJ, Briones C. 2015. A novel
775 representation of genomic sequences for taxonomic clustering and visualization
776 by means of self-organizing maps. *Bioinformatics* 31:736-44.
- 777 44. Feng DF, Doolittle RF. 1996. Progressive alignment of amino acid sequences
778 and construction of phylogenetic trees from them. *Methods in Enzymol*
779 266:368-82.
- 780 45. Hart GR, Ferguson AL. 2018. Computational design of hepatitis C virus
781 immunogens from host-pathogen dynamics over empirical viral fitness
782 landscapes. *Phys Biol* 16:016004.
- 783 46. Kupperts BO. 2016. The Nucleation of Semantic Information in Prebiotic Matter.
784 *Curr Top Microbiol Immunol* 392:23-42.
- 785 47. Kohonen T. 1982. Self-organized formation of topologically correct feature
786 maps. *Biological Cybernetics* 43:59-69.
- 787 48. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander
788 ES, Golub TR. 1999. Interpreting patterns of gene expression with self-
789 organizing maps: methods and application to hematopoietic differentiation. *Proc*
790 *Natl Acad Sci U S A* 96:2907-12.
- 791 49. Toronen P, Kolehmainen M, Wong G, Castren E. 1999. Analysis of gene
792 expression data using self-organizing maps. *FEBS Lett* 451:142-6.
- 793 50. Iizuka N, Oka M, Yamada-Okabe H, Mori N, Tamesa T, Okada T, Takemoto N,
794 Sakamoto K, Hamada K, Ishitsuka H, Miyamoto T, Uchimura S, Hamamoto Y.
795 2005. Self-organizing-map-based molecular signature representing the
796 development of hepatocellular carcinoma. *FEBS Lett* 579:1089-100.
- 797 51. Reiser M, Wiebner B, Hirsch J, German Liver F. 2019. Neural-network analysis
798 of socio-medical data to identify predictors of undiagnosed hepatitis C virus
799 infections in Germany (DETECT). *J Transl Med* 17:94.
- 800 52. Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97-159.
- 801 53. Domingo E. 2020. *Virus as Populations*. Academic Press, Elsevier, Amsterdam.
802 Second Edition.
- 803 54. Newman MEJ. 2005. Power laws, Pareto distributions and Zipf's law.
804 *Contemporary Physics* 46:325-351.
- 805 55. Quer J, Huerta R, Novella IS, Tsimring L, Domingo E, Holland JJ. 1996.
806 Reproducible nonlinear population dynamics and critical points during
807 replicative competitions of RNA virus quasispecies. *J Mol Biol* 264:465-471.

- 808 56. Quer J, Hershey CL, Domingo E, Holland JJ, Novella IS. 2001. Contingent
809 neutrality in competing viral populations. *J Virol* 75:7315-20.
- 810 57. Herrera M, Grande-Perez A, Perales C, Domingo E. 2008. Persistence of foot-
811 and-mouth disease virus in cell culture revisited: implications for contingency in
812 evolution. *J Gen Virol* 89:232-244.
- 813 58. Ruiz-Jarabo CM, Miller E, Gómez-Mariano G, Domingo E. 2003. Synchronous
814 loss of quasispecies memory in parallel viral lineages: a deterministic feature of
815 viral quasispecies. *J Mol Biol* 333:553-563.
- 816 59. Tsimring LS, Levine H, Kessler DA. 1996. RNA virus evolution via a fitness-
817 space model. *Phys Rev Lett* 76:4440-4443.
- 818 60. Lande R. 1985. Expected time for random genetic drift of a population between
819 stable phenotypic states. *Proc Natl Acad Sci U S A* 82:7641-5.
- 820 61. Kirkpatrick M. 1996. Genes and adaptation: a pocket guide to the theory, p. 125-
821 146. In MR Rose, G V Lauder (eds.), *Adaptation*. Academic Press, San Diego.
- 822 62. Bell G. 2008. *Selection: The mechanism of evolution*. Oxford University Press
823 Inc., New York.
- 824 63. Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R,
825 Alvarez-Tejado M, Rodriguez-Frias F, Guardia J, Domingo E, Quer J. 2013.
826 Ultra-Deep Pyrosequencing (UDPS) Data Treatment to Study Amplicon HCV
827 Minor Variants. *PLoS ONE* 8:e83361.
- 828 64. Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. 2016.
829 Viral quasispecies complexity measures. *Virology* 493:227-237.
- 830 65. Gregori J, Salicru M, Domingo E, Sanchez A, Esteban JI, Rodriguez-Frias F,
831 Quer J. 2014. Inference with viral quasispecies diversity indices: clonal and
832 NGS approaches. *Bioinformatics* 30:1104-1111
833 doi:10.1093/bioinformatics/btt768.
- 834 66. Soria ME, Gregori J, Chen Q, Garcia-Cehic D, Llorens M, de Avila AI, Beach
835 NM, Domingo E, Rodriguez-Frias F, Buti M, Esteban R, Esteban JI, Quer J,
836 Perales C. 2018. Pipeline for specific subtype amplification and drug resistance
837 detection in hepatitis C virus. *BMC Infect Dis* 18:446.

838

839

840 **Table 1.** Number of reads and haplotypes derived from MiSeq Illumina sequencing of HCV
841 amplicons A1, A2 and A3.

Experiment	Virus	Passage	Number of reads (Number of haplotypes ^a)		
			Amplicon 1 (7649-7960) ^b	Amplicon 2 (7940-8257) ^b	Amplicon 3 (8231-8653) ^b
Experiment 1	HCV p0	Initial	240,376 (8)	189,190 (4)	122,836 (3)
		p3	243,755 (6)	273,783 (2)	119,705 (4)
	HCV p100	Initial	225,977 (18)	251,949 (10)	108,548 (14)
		p3	201,355 (14)	282,139 (10)	87,596 (15)
		p10	188,215 (8)	197,078 (7)	79,412 (9)
	HCV p200	Initial	50,502 (15)	166,060 (11)	50,111 (11)
		p3	53,462 (13)	160,179 (10)	62,872 (12)
		p10	57,730 (16)	149,758 (12)	51,910 (10)
	Experiment 2a	HCV p0	p1	18,817 (4)	45,378 (4)
p2			25,866 (4)	61,599 (3)	9,693 (3)
p3			33,964 (5)	112,247 (4)	6,690 (4)
p4			46,180 (3)	159,699 (3)	6,695 (5)
HCV p100		p1	32,729 (15)	119,433 (5)	6,698 (15)
		p2	34,670 (14)	138,215 (5)	8,040 (15)
		p3	29,787 (16)	135,219 (5)	7,621 (17)
		p4	35,007 (13)	115,195 (5)	7,101 (16)
HCV p200		p1	24,630 (13)	52,247 (15)	8,862 (18)
		p2	31,553 (19)	136,972 (13)	8,917 (14)
		p3	32,166 (20)	102,391 (13)	7,137 (15)
		p4	54,605 (19)	179,432 (15)	13,687 (14)
Experiment 2b	HCV p0	p1	148,212 (9)	149,731 (2)	64,912 (2)
		p2	195,318 (5)	137,365 (3)	74,195 (3)
		p3	138,832 (6)	120,688 (4)	60,089 (4)
		p4	122,525 (4)	177,166 (5)	83,941 (4)
	HCV p100	p1	128,168 (15)	149,817 (8)	52,351 (15)
		p2	116,120 (16)	84,408 (6)	75,494 (17)
		p3	120,016 (17)	117,116 (5)	53,153 (17)
		p4	117,114 (16)	120,914 (6)	85,711 (17)
	HCV p200	p1	155,347 (17)	124,204 (15)	53,455 (16)
		p2	152,250 (18)	116,427 (14)	62,190 (11)
		p3	122,481 (12)	95,088 (14)	44,088 (13)
		p4	85,906 (19)	135,373 (14)	73,119 (13)
Experiment 2c	HCV p0	p1	104,149 (8)	90,216 (3)	134,903 (2)
		p2	209,396 (8)	130,509 (3)	54,018 (2)
		p3	86,558 (5)	69,588 (3)	124,363 (3)
		p4	155,428 (5)	144,968 (6)	44,764 (3)
	HCV p100	p1	73,080 (18)	72,919 (9)	123,388 (16)
		p2	186,831 (17)	79,384 (8)	55,436 (18)
		p3	98,930 (18)	75,155 (6)	125,049 (18)
		p4	137,944 (17)	115,719 (7)	61,177 (18)
	HCV p200	p1	85,384 (20)	106,672 (16)	117,945 (14)

p2	141,307 (17)	107,340 (16)	75,365 (14)
p3	106,261 (21)	138,379 (12)	94,545 (15)
p4	141,033 (21)	159,476 (17)	58,184 (15)

842

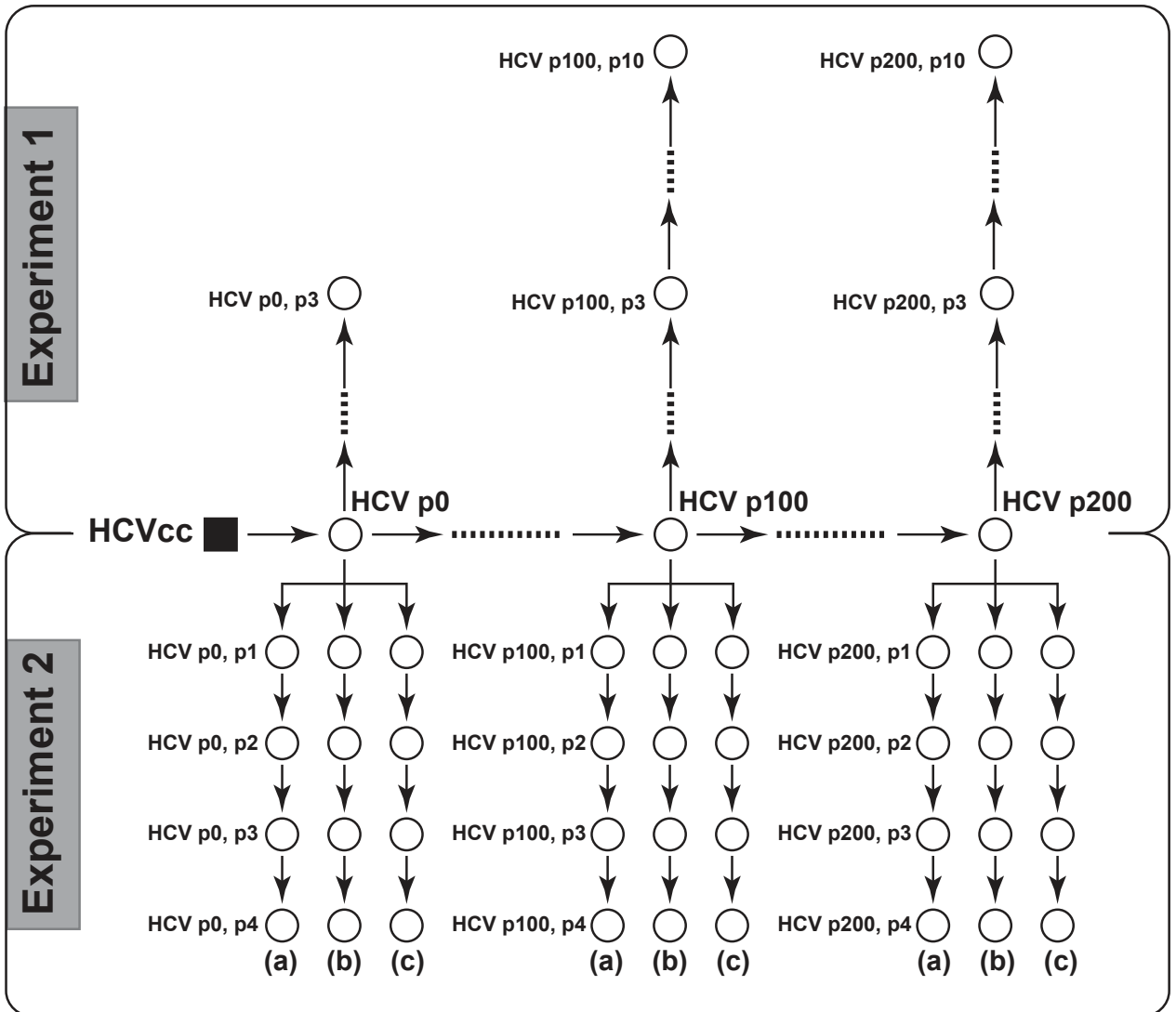
843 ^aThe experiments, HCV populations and amplicon numbers are those described in Fig.1.
844 Mutations were counted relative to the HCV sequence encoded in plasmid Jc1FLAG2(p7-
845 nsGluc2A), as previously described (11, 18). The total number of reads and haplotypes (in
846 parenthesis) were derived as detailed in Materials and Methods.

847 ^bThe HCV genomic residues spanned by each amplicon are: 7649-7960 (A1), 7940-8257 (A2),
848 and 8231-8653 (A3) (numbering according to isolate JFH-1; accession number #AB047639).

849

Figure 1

A



B

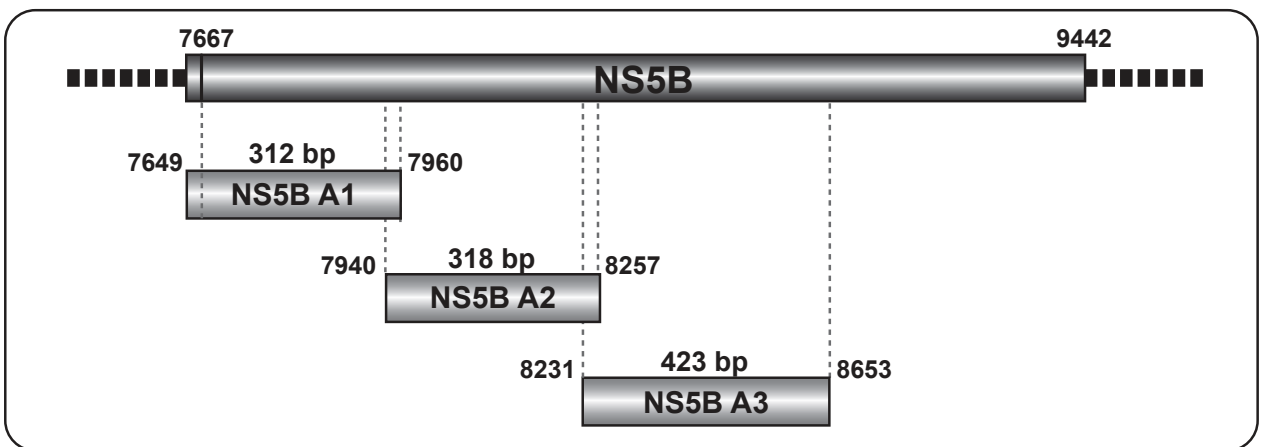


Figure 2

Amplicon 1

HCV p0

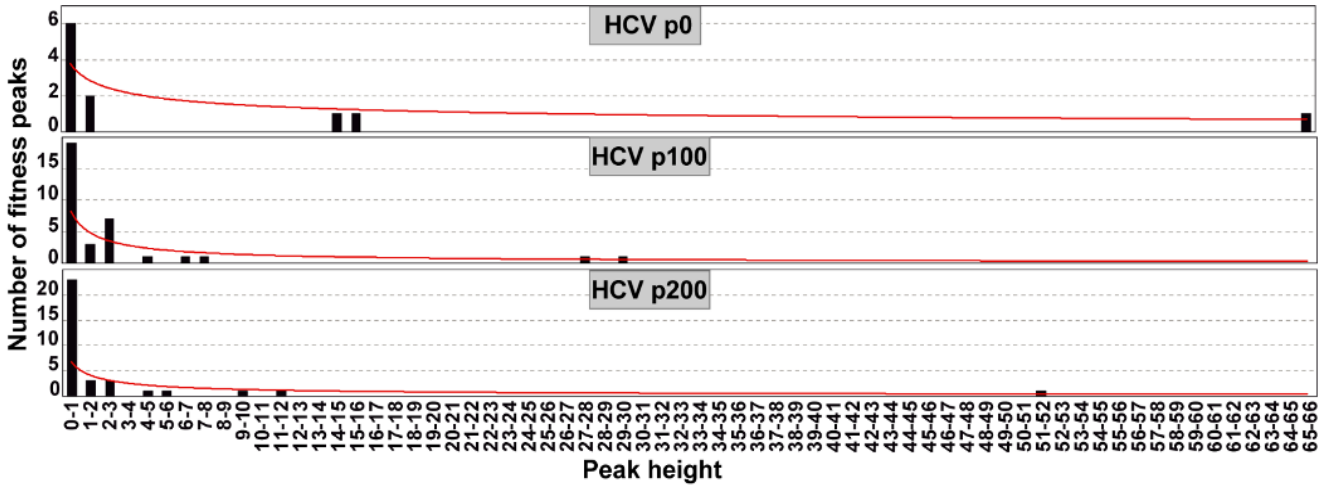
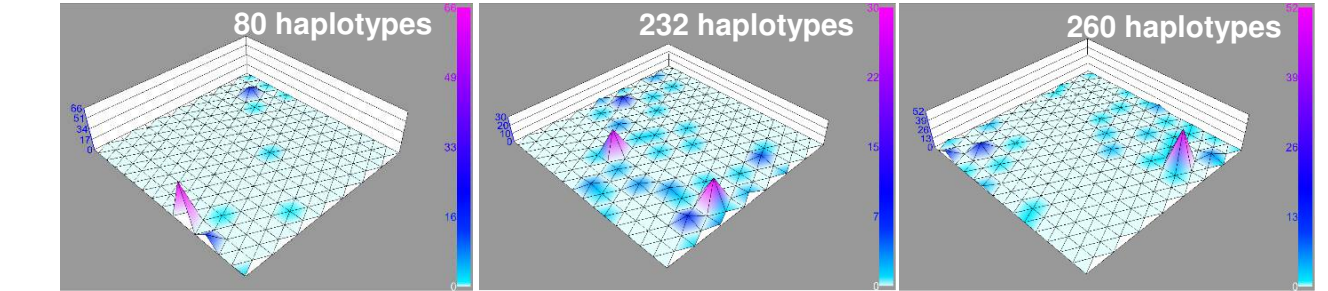
HCV p100

HCV p200

80 haplotypes

232 haplotypes

260 haplotypes



Amplicon 2

HCV p0

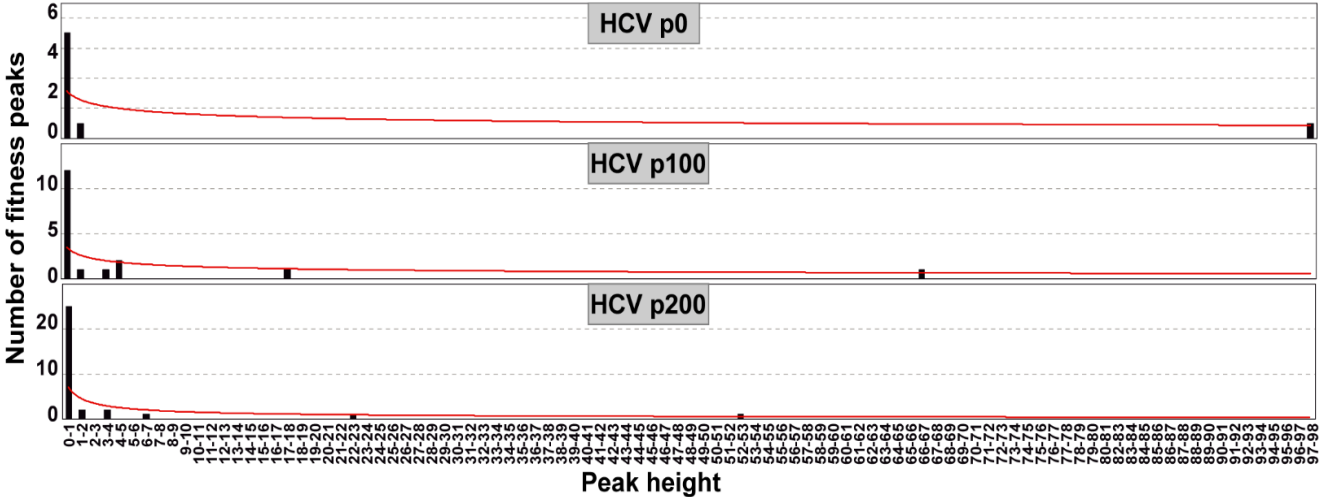
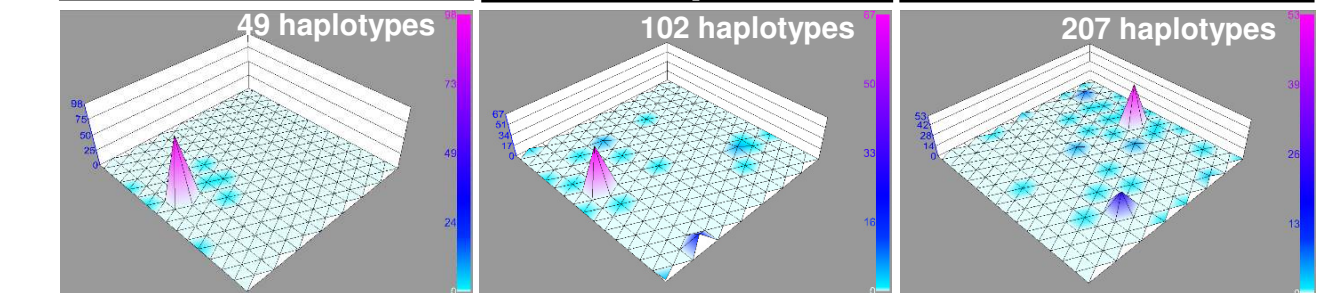
HCV p100

HCV p200

49 haplotypes

102 haplotypes

207 haplotypes



Amplicon 3

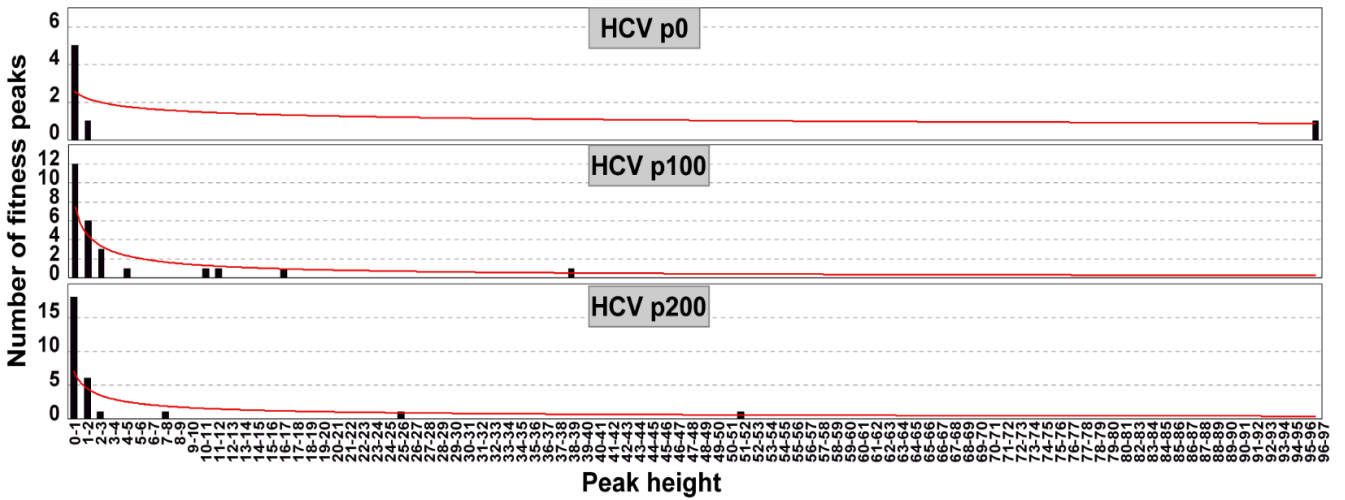
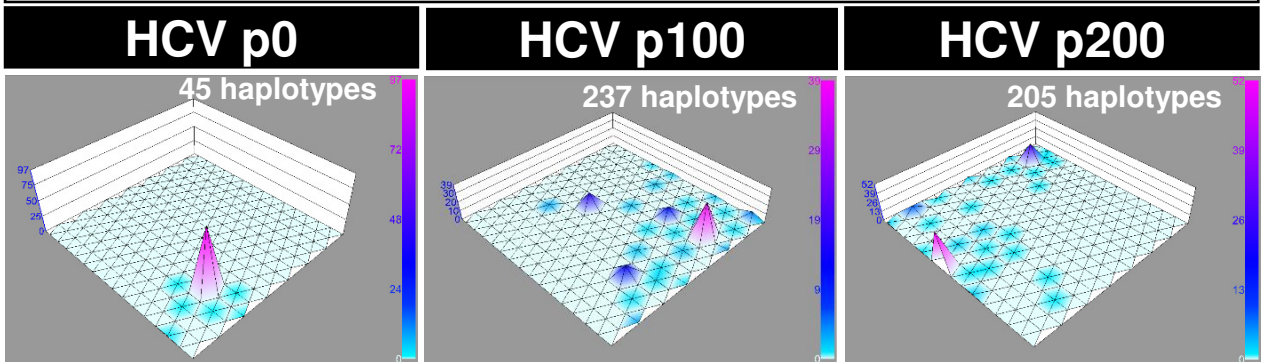


Figure 3

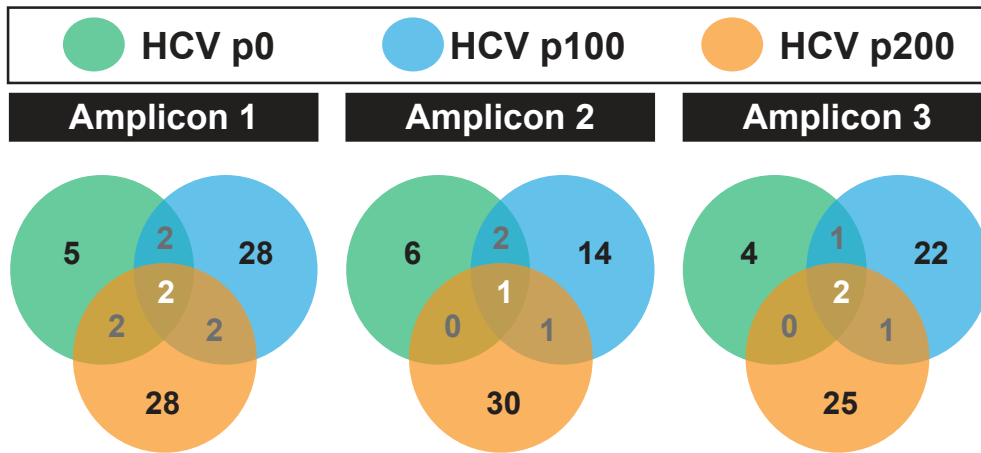


Figure 4

