Title: Assessing *in vivo* mutation frequencies and creating a high-resolution genome-wide map of fitness costs of Hepatitis C virus

(Short title: Genome-wide *in vivo* fitness map of Hepatitis C virus mutations)

Kaho H. Tisthammer[1*], Weiyan Dong[2], Jeffrey B. Joy[2,3,4], Pleuni S. Pennings[1]

[1] Department of Biology, San Francisco State University

[2] BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

[3] Division of Infectious Diseases, Department of Medicine, University of British Columbia, Vancouver, BC, Canada

[4] Bioinformatics Programme, University of British Columbia, Vancouver, BC, Canada

* Corresponding author

E-mail: ktist@sfsu.edu (KHT)

Abstract

Studying *in vivo* fitness costs of mutations in viruses provides important insights into their evolutionary dynamics, which can help decipher how they adapt to host immune systems and develop drug resistance. However, studying fitness costs in natural populations is difficult, and is often conducted *in vitro* where evolutionary dynamics differ from *in vivo*. We aimed to understand *in vivo* fitness costs of mutations in Hepatitis C virus using next generation sequencing data. Hepatitis C virus is a positive-sense single-stranded RNA virus, and like many RNA viruses, has extremely high mutation and replication rates, making it ideal for studying mutational fitness costs. Using the 'frequency-based approach', we estimated genome-wide *in vivo* mutation frequencies at mutation-selection equilibrium, and inferred fitness costs (selection coefficients) at every genomic position using data from 195 patients. We applied a beta regression model to estimate the effects and the magnitudes of different factors on fitness costs. We generated a high-resolution genome-wide map of fitness costs in Hepatitis C virus for the first time. Our results revealed that costs of nonsynonymous mutations are three times higher than those of synonymous mutations, and mutations at nucleotides A/T have higher costs than those at C/G. Genome location had a modest effect, which is a clear contrast from previously reported *in vitro* findings, and highlights host immune selection. We inferred the strongest negative selection on the Core and NS5B proteins. We also found widespread natural prevalence of known drug resistance-associated variants in treatment naive patients, despite high fitness costs of these resistance sites. Our results indicate that *in vivo* evolutionary patterns and associated mutational costs are dynamic and can be virus specific, reinforcing the utility of constructing *in vivo* fitness cost maps of viral genomes.

Author Summary

Understanding how viruses evolve within patients is important for combatting viral diseases, yet studying viruses within patients is difficult. Laboratory experiments are often used to understand the evolution of viruses, in place of assessing the evolution in natural populations (patients), but the dynamics will be different. In this study, we aimed to understand the within-patient evolution of Hepatitis C virus, which is an RNA virus that replicates and mutates extremely quickly, by taking advantage of high-throughput next generation sequencing. Here, we describe the evolutionary patterns of Hepatitis C virus from 195 patients: We analyzed mutation frequencies and estimated how costly each mutation was. We also assessed what factors made a mutation more costly, including the costs associated with drug resistance mutations. We were able to create a genome-wide fitness map of within-patient mutations in Hepatitis C virus which proves that, with technological advances, we can deepen our understanding of within-patient viral evolution, which can contribute to develop better treatments and vaccines.

54 **Introduction**

55 Understanding the fitness costs of mutations in pathogenic viruses is critically important
56 to improve strategies to prevent viral adaptation to host immune systems or to antiviral therapies
57 [1-3], and for effective vaccine design. Viruses also provide an opportunity to study entire
58 genomes to test hypotheses about the fitness costs of various types of mutations, allowing us to
59 learn more about the process of evolution more generally. However, most studies on the fitness
60 costs of mutations in viruses have been conducted *in vitro* [4-7]. Although studying viruses *in*
61 *vitro* is often more practical than studying them *in vivo*, especially for viruses that infect humans,
62 the results obtained may not reflect fitness costs in the real world. For example, viruses studied
63 *in vitro* do not have to contend with the strong, multifarious selection imposed by a host's
64 immune response, resulting in a fitness landscape that may differ from the *in vivo* landscape in
65 dramatic and important ways. Selective bias can also exist with culturing viruses itself [8]. In this
66 study, we overcame the usual barriers to *in vivo* studies by analyzing clinical data obtained from
67 195 patients infected with the hepatitis C virus (HCV). We used these *in vivo* data to create a
68 fitness cost map of the entire HCV genome.

69 To determine the *in vivo* fitness costs of mutations across the genome, we used
70 information about (1) the HCV mutation rate and (2) a given mutation's average frequency
71 across multiple populations (*i.e*. patient samples) to determine the selection pressure against a
72 given mutation at a given site. This 'frequency-based' approach [9] is based on a model of
73 mutation-selection balance. As an RNA virus, HCV is especially well suited for this approach,
74 and thus for testing hypotheses about how selection operates on the genome. Each patient
75 infected with a long-lasting RNA virus infection harbors a diverse viral population that is almost
76 perfectly independent of those in other patients [10]; thus, each patient sample in our study could

77 be treated as a different viral population, as required in the frequency-based approach [9,11]. In
78 addition, RNA viruses have mutation rates three to four orders of magnitude higher than those of
79 bacteria and eukaryotes, due to the lack of proofreading activity in viral RNA-dependent RNA
80 polymerases [12]. This allowed us to gather a great deal of information on fitness costs from a
81 relatively compact genome.

82 The frequency-based approach has successfully been applied to estimate the *in vivo*
83 fitness costs in RNA viruses: Theys et al. [9] showed fitness costs are higher for CpG creating
84 mutations in the *pol* gene of HIV using data from 160 patients. Zanini et al. [11] reported
85 genome wide *in vivo* fitness costs of mutations in HIV from nine patients, showing consistent
86 estimated costs across patients, further validating the approach. However, the sample sizes were
87 limited in both studies. In Theys et al. [9], the number of patients was relatively large (160) but
88 the number of viral sequences were limited to about 20 per patient, totaling approximately 3000
89 data points per site. Deep-sequencing was used in Zanini et al. [11] with coverage per site in the
90 order of thousands and higher, but the number of patients were limited to nine, totaling in the
91 order of tens of thousands of data points per site. Using NGS data from a large number (195) of
92 patients, our study overcame the sample size limitation by providing close to one million
93 datapoints per site.

94 Estimating site-level fitness costs using the frequency-based approach requires mutation
95 rates for each base substitution. The overall HCV mutation rate *in vivo* has been estimated to be
96 around 10^{-4} to 10^{-5} base substitutions per site per replication cycle [7,13,14], though this was
97 inferred from limited genome regions, and nucleotide-level substitution rates were not available.
98 To overcome this limitation, Geller et al [15] recently estimated the genome-wide HCV mutation
99 rate using ultradeep sequencing *in vitro*, and found the overall rate to be 3.5×10^{-5}. Although the

100 estimation was done on a modified replicon system with fragmented cistrons based on three cell
101 lines, the rate estimated from their study was relatively consistent with previous *in vivo*
102 estimates. Their ultradeep sequencing approach also allowed calculation of nucleotide-level
103 mutation rates. Since no other base substitution rates are available for HCV, we applied the *in*
104 *vitro* mutation rates from Geller et al. [15] to predict fitness costs.

105 By elucidating the *in vivo* fitness costs of point mutations in the coding regions of the
106 HCV genome, we aimed to test several important hypotheses about fitness costs in general, as
107 well as in HCV specifically. Using the HCV subtype 1a, first, we tested whether CpG-creating
108 mutations are costly in HCV, as previous studies have shown that CpG usage tends to be under-
109 represented in RNA viruses [16,17] and are highly costly in HIV [9]. Second, we determined
110 how costly mutations are that result in drastic amino acid changes. Third, we assessed whether
111 mutations that result in drug resistance are costly, and how prevalent the known resistance-
112 associated variants are in pre-treatment patients.

113

114 **Results**

115 **1. Mutation frequencies within the HCV genome**

116 First, we assessed *in vivo* mutation frequencies along the HCV1a genome using next-
117 generation sequencing of samples from 195 HCV1a infected patients. We started by calculating
118 genome-wide averages of mutation frequencies. The overall average genome-wide mutation
119 frequency at mutation-selection equilibrium (see Methods) was 5.8 (± 0.083) x 10^{-3} per
120 nucleotide, and transition mutations were around 10 times as common than transversion
121 mutations (4.8 x 10^{-3} *vs*. 9.7 x 10^{-4}) (Fig 1A and Table 1). For both transitions and transversions,
122 the average frequency of synonymous mutations (transition: 8.1 x 10^{-3}, transversion: 8.9 x 10^{-4})

123 was more than twice as high than that of nonsynonymous mutations (transition: 3.3×10^{-3},

124 transversion: 4.1×10^{-4}, Wilcoxon test, $P = 0 - 2.6 \times 10^{-185}$). For transitions, the average

125 frequency of nonsense mutations (2.0×10^{-3}) was significantly lower than that of

126 nonsynonymous mutations (Wilcoxon test, $P = 1.3 \times 10^{-30}$), but this was not the case for

127 transversions. This suggests that the frequencies around nonsynonymous (4.1×10^{-4}) and

128 nonsense (5.2×10^{-4}) transversions (*i.e.* approximately $4 \sim 5 \times 10^{-4}$) may have been at the lower

129 detection limit of our dataset, since nonsense mutations are expected to occur at a lower

130 frequency than nonsynonymous mutations. Since transitions are more frequent than

131 transversions, and thus offer greater statistical power, we focused on transition mutations in our

132 subsequent analysis.

133

134 **Table 1. Genome-wide average estimated frequencies of different types of mutation in the**

135 **HCV1a genome.**

Mutation type	Average	S.E.
Transition	4.8×10^{-3}	7.6×10^{-5}
Synonymous	8.1×10^{-3}	9.8×10^{-5}
Nonsynonymous	3.3×10^{-3}	6.4×10^{-5}
Nonsense	2.0×10^{-3}	4.8×10^{-5}
Transversion	9.7×10^{-4}	3.4×10^{-5}
Synonymous	8.9×10^{-4}	3.3×10^{-5}
Nonsynonymous	$4.1 \times 10^{-}$	2.2×10^{-5}
Nonsense	$5.2 \times 10^{-}$	2.5×10^{-5}

136 S.E., standard error.

137

138 We looked for trends in mutation frequencies across the genome based on transition

139 mutations. Synonymous mutations generally had higher frequencies than nonsynonymous

140 mutations, and mutations at nucleotides A/T (*i.e*. A→G / T→C) had higher frequencies than
141 those at C/G (*i.e*. C→T / G→A) (Fig 1B and C). Hypervariable Region 1 (HVR1), the most
142 variable section of the HCV genome [18,19], had the highest estimated average mutation
143 frequency of all HCV genes (Wilcoxon test, $P = 1.2 \times 10^{-7} \sim 0.004$), whereas the Core gene, the
144 genome's most conserved gene [18,19], had the lowest average mutation frequency (Wilcoxon
145 test, $P = 1.6 \times 10^{-12} \sim 0.04$) (Fig 2A). Additionally, HVR1 had high nonsynonymous mutation
146 frequencies, and was the only region that did not have significantly higher synonymous mutation
147 frequencies than nonsynonymous ones (Fig 2B).

148 **1.1. Model fitting shows large effects of nonsynonymous mutations, nucleotide types, and**
149 **genome location on mutation frequencies in HCV.** In order to simultaneously assess the
150 effects of multiple factors at the same time on HCV mutation frequencies, such as mutations that
151 create drastic amino acid change or CpG sites, we used beta regression models since parameters
152 of beta regression are interpretable and they allow us to quantitatively assess the effects of
153 specific factors. We specifically explored the effects of the following factors; ancestral
154 nucleotide, mutation type (synonymous/nonsynonymous mutations, CpG-creating mutations,
155 mutations that result in a drastic amino acid change), genomic location (*i.e*, genes), RNA
156 structure, and the interactions of these factors on transition mutation frequencies. The results of
157 the best-fit model, based on AIC, contained 18 variables and are shown in Fig 3 and S1 Table
158 S1. Of the 18 variables, 17 had significant effects on mutation frequencies. The ancestral
159 nucleotide had a large effect on mutation frequencies, with G→A and C→T mutations having
160 41~51 % lower frequencies than A→G and T→C mutations ($P < 0.0001$) (Fig 3B).
161 Nonsynonymous changes were also associated with a large reduction in mutation frequency of
162 over 53% ($P < 0.0001$) relative to synonymous changes. Genomic location had modest effects;

163 the best-fit model included eight genes out of 11, of which seven had significant effects. The
164 largest locational effect was with the HVR1 region, which had a frequency increase of 27 % (P
165 <0.0001), whereas being located in Core was associated with a frequency reduction of 20% (P
166 <0.0001). CpG-creating mutations and mutations that cause a drastic amino acid change were
167 associated with modest reductions in mutation frequency of 7% ($P < 0.0001$) and 12% ($P <$
168 0.0001), respectively. Mauger et al [20] identified 15 regions that likely form secondary
169 structures and are functionally conserved in HCV genotype 1a, 1b, and 2a. We tested the
170 aggregate effects of the 15 regions and found no significant effect on *in vivo* mutation
171 frequencies. RNA structure was associated with an increase of ~1.1% in mutation frequency ($P =$
172 0.39) and was not selected in the best fit-model.

173 **1.2. *In vivo vs. in vitro* comparison reveals similar trends at the nucleotide-level, but**
174 **different genome-wide patterns.** We compared observed *in vivo* mutation frequencies with a
175 study of *in vitro* mutation frequencies [15] (for transition mutations). Patterns of *in vivo* mutation
176 frequencies were congruent with the trends observed in *in vitro* mutation frequencies [15] at the
177 nucleotide level, but differed at the gene level. The *in vivo* estimated mutation frequencies were
178 an order of magnitude higher than those found *in vitro*. Mutation frequencies for A→G and
179 T→C mutations were higher both *in vivo* and *in vitro* than C→T and G→A mutations (S1A Fig).
180 At the gene level, genome location was found to have very little effect (0.5%) on *in vitro*
181 mutation frequencies [15], while our regression model results suggest that seven out of 11 genes
182 have significant effects on *in vivo* mutation frequencies (Fig 3B and S1 Table). The difference
183 was most prominent in HVR1; the mutation frequency peak observed in HVR1 *in vivo* was not
184 seen *in vitro* (S1B Fig). The *in vitro* study [15] also found clusters of six highly mutable and 13
185 less mutable sites along the HCV genome, but in our *in vivo* study, we found no clear trends in

186 mutation frequencies at the corresponding sites (S2 Fig). Similarly, the *in vitro* study [15]
187 reported the sites that had higher transversion than transition mutation frequencies (reversal sites)
188 at over 1000 sites (12.4% of the total sites), but far less such reversal sites were observed *in vivo*
189 (206 sites, 2.6%). The most salient reversal site *in vitro* (genome position C4080) again did not
190 show a higher frequency of transversion mutations than transition mutations *in vivo*.
191

192 **2. Genome wide fitness landscape of HCV**

193 **2.1 Estimated selection coefficients reveal high costs of A→G and T →C mutations,**
194 **nonsynonymous mutations, and mutations in NS5B.** Based on the observed mutation
195 frequencies, we next estimated selection coefficients (s) for each site in the HCV1a genome.
196 Selection coefficients can quantitatively reveal fitness consequences of new mutations. We
197 estimated s for each site based on the equation $f = \mu/s$ (see Methods), using the mutation
198 frequencies (f) and the HCV mutation rates (μ) calculated from the datasets of Geller et al. [15].
199 The genome-wide average selection coefficient for transition mutations in coding regions, using
200 the best available mutation rates [15], was 2.0 (± 0.5) x 10^{-3}. The average selection coefficient
201 for nonsynonymous mutations (2.5 ± 0.7 x 10^{-3}) was three times higher than that for synonymous
202 mutations (8.9 ± 6.0 x 10^{-4}) (Fig 4A).
203 Selection coefficients for each nucleotide also revealed interesting patterns. The selection
204 coefficients (s) and mutation frequency (f) generally have an inverse relationship (equation: $f =$
205 μ/s); that is, higher f implies lower s (*i.e*. a lower cost). If mutation rates (μ) at different
206 nucleotides vary widely, the inverse relationship may not hold when assessed at a nucleotide
207 level, which was observed in our results. The average mutation frequency at nucleotides A/T
208 (A→G/T →C mutations) was higher than at C/G (C →T/G →A mutations), leading one to

209 expect that selection coefficients for A/T would be lower than those for C/G sites. In reality,
210 however, for A/T sites, average selection coefficient values were higher despite the higher
211 average mutation frequencies because the mutation rates based on Geller's dataset [15] were so
212 much higher (1.1-1.3 x 10^{-5} for A→G /T →C *vs*. 3.2 -4.2 x 10^{-6} for C →T/G →A) (Fig 4B and
213 C). The differences in selection coefficients between nucleotides (excluding CpG-creating
214 mutations) were all statistically significant, with the values of selection coefficients in
215 descending order being A→G (1.5 x 10^{-3}) > T→C (1.3 x 10^{-3}) > G→A (8.4 x 10^{-4}) > C→T (8.1 x
216 10^{-4}) (Wilcoxon test, P = 0 - 7.1 x 10^{-7}). It is, however, important to note that the degree of
217 accuracy in estimating fitness costs (s) inevitably depends on the accuracy of estimated mutation
218 rates. Therefore, we confirmed the higher mutation rates for A→G /T →C mutations than for C
219 →T/G →A mutations by testing the differences (Wilcoxon test, all P-values = 0.0), as well as
220 calculating the 95% confidence intervals (S3 Fig).
221 Similarly, when comparing average selection coefficients for sites in different genes. The
222 average selection coefficient for sites in NS5B (RNA dependent RNA reverse transcriptase) was
223 significantly higher than the average selection coefficients of any other genes (S3A Fig,
224 Wilcoxon Test, P = 1.1 x 10^{-13} - 0.039), including sites in Core, despite the fact that Core's
225 average mutation frequency was the lowest among HCV genes (Table 2). This finding may
226 reflect the higher genomic AT content of NS5B relative to Core (S3 Fig), since higher AT
227 content can lead to a higher average selection coefficient due to A→G /T→C mutations having
228 higher selection coefficients. We, therefore, also quantified the effects of different factors on
229 selection coefficients using beta regression models, which followed highly similar trends as the
230 mutation frequency results, except for the effects of ancestral nucleotide C and G for the reason
231 described in the previous paragraph (S2 Table and S4 Fig). Model results also indicated that the

232 effects of Core alone on selection coefficients was slightly larger than those of NS5B (+6.1% vs.
233 +5.7%), likely reflecting the effects of higher AT content. As expected, sites in HVR1, which
234 had the highest *in vivo* mutation frequencies, had significantly lower selection coefficients than
235 those in the rest of the genes (Wilcoxon Test, $P = 1.5 \times 10^{-8} - 5.9 \times 10^{-3}$).
236
237 **Table 2. Average estimated selection coefficients per genic region of the HCV1a genome.**

Gene	*s*	S.E.
Core	2.11×10^{-3}	4.1×10^{-5}
E1	1.85×10^{-3}	3.1×10^{-5}
HVR1	1.39×10^{-3}	3.9×10^{-5}
E2	1.85×10^{-3}	3.9×10^{-5}
NS1	1.66×10^{-3}	3.7×10^{-5}
NS2	1.78×10^{-3}	4.2×10^{-5}
NS3	2.03×10^{-3}	5.6×10^{-5}
NS4A	2.00×10^{-3}	6.3×10^{-5}
NS4B	2.01×10^{-3}	6.9×10^{-5}
NS5A	1.97×10^{-3}	6.5×10^{-5}
NS5B	2.21×10^{-3}	6.5×10^{-5}

238 *s*, selection coefficient; S.E., standard error.
239
240 **2.2. Distribution of fitness effects and gamma parameter estimation.** Based on the estimated
241 fitness costs (*s*) for each nucleotide in our study, we assessed the distribution of fitness effects
242 (DFEs) of synonymous and nonsynonymous transition mutation for each nucleotide (Fig 5). All
243 synonymous mutations and 99.7% of nonsynonymous mutations had fitness costs <0.01.
244 Parameters for the gamma distribution were also estimated for the entire DFEs. The scale (k) and
245 shape (θ) parameters were estimated to be 8.32×10^{-4} (95% confidence interval [CI] = 8.11 -
246 8.55×10^{-4}) and 2.378 (95% CI = 2.318 -2.436), respectively, which can be applied to future
247 population simulation studies of HCV.

248 **2.3. High fitness costs of drug resistance sites.** A number of direct-acting antiviral agents have
249 been developed for HCV, all of which specifically target the NS3, NS5A, or NS5B proteins.
250 Drug resistance often develops in patients, and many HCV resistance-associated variants have
251 been reported [21-26]. Prevalence of baseline (pre-existing) resistance-associated variants in
252 treatment naive patients (*i.e*. natural prevalence), as well as the fitness costs of these variants
253 provide insights into the evolution of drug resistance traits, treatment outcome prediction, and
254 durability of resistance phenotypes [27-29]. They can also help guide treatment options, such as
255 combining antivirals for which resistance mutations have high fitness costs to lower the risk of
256 developing drug resistance.

257 We assessed the natural prevalence of resistance-associated variants using our 'unfiltered'
258 HCV datasets, which included sites with alternate majority nucleotides and transversion
259 mutations, as opposed to the filtered datasets used to estimate *in vivo* fitness costs, which
260 excluded the sites that were not in selection-mutation balance, including those with a nucleotide
261 different from the reference (H77, AF011753, see Methods). Of 87 known resistance-associated
262 variants across 26 sites we examined (S3 Table), natural prevalence was observed for all variants
263 in our datasets, including the 51 variants previously not observed or reported [23] (Fig 6). The
264 proportion of the samples with pre-existing resistance-associated variants varied from 3.6 to
265 99.5% (average 62%) depending on the resistance mutation. The result highlights that many
266 variants do occur naturally in treatment naive patients, making them more vulnerable to
267 developing drug resistance. Additionally, 20 of the 87 resistance-associated variants were
268 observed as a majority variant in one or more patients (Fig 5, '% fixed sample'). The percentages
269 of patients that had the resistant variant as the majority vary from 0.5 % to 54.6%, depending on
270 the variants, with Q80K in NS3 being most likely to be present as the majority. The average

271 fitness costs of all resistance-associated variants examined was 4.7 x 10^{-3}, with 2.5 x 10^{-3} for
272 transition mutations and 6.8 x 10^{-3} for transversion mutations. Estimated fitness costs of
273 resistance-associated variants (at transition sites only) were compared to those of non-resistance-
274 associated sites in the three genes, which revealed on average 21% higher fitness cost in
275 resistance-associated variants of HCV (Wilcoxon test, $P = 0.0021$).

276 **3. Validation of our frequency-based approach**

277 **3.1 Between-host variability is consistent with within-host variability in HCV.** The
278 frequency-based approach we used in this study assumes that site-level mutation frequencies are
279 relatively consistent across different viral populations (*i.e*. patients) at selection-mutation
280 equilibrium. To verify this assumption, we assessed the relationship between genetic variability
281 within hosts (*in vivo*) and between hosts by (1) testing the existence of correlation between the
282 within-host and between-host nucleotide diversity at each nucleotide site, (2) checking the
283 proportion of excluded sites through our filtering process to make sure that we did not eliminate
284 too many sites/samples from our analysis to bias our results, and (3) assessing the highly
285 conserved sites between-host were also conserved within-host to further confirm that patterns of
286 variability are consistent between-host and within-host.

287 First, we assessed nucleotide diversity levels between- and within-host to validate our
288 assumption, using (1) average mutation frequencies to represent within-host (*in vivo*) diversity
289 and (2) Shannon's entropy indices of the HCV genome sequences downloaded from NCBI
290 GenBank (n=423) along with consensus sequences of each viral population (*i.e*. each patient)
291 from our dataset to represent between-host diversity. The two measures of diversity were highly
292 correlated (Spearman's $\rho = 0.677$, $P < 2.2\text{x } 10^{-16}$) (S5 Fig), indicating that genome-wide
293 evolutionary patterns were consistent between within-host and between-host.

294 Next, we investigated potential effects of our filtering process, where we excluded the
295 sites that had a majority nucleotide different from the reference nucleotide (H77) from each viral
296 population (see Methods). If too many sites were excluded, and if these sites had high variability,
297 this could mask within-host variability. We first calculated the proportion of excluded sites due
298 to the nucleotide difference for each viral population, which was on average, only 6.6 (± 0.1) %.
299 We also compared the consensus (majority nucleotide) sequences of all viral populations to gain
300 further insights into the genetic variability between hosts. The overall sequence similarity (the
301 average pairwise identity of consensus sequences) was relatively high (91.8%), and was within
302 the range of the expected values (> 90% within a subtype) [17,29,30]. These results indicate that
303 the number of sites excluded from the analysis was quite small, and the filtering step probably
304 did not bias our results.
305 Lastly, we investigated if a group of highly conserved sites between-host also
306 corresponded to highly conserved sites within-host. Alignments of all consensus sequences
307 across the 195 genomes showed 49% of sites to be identical within the HCV coding regions.
308 These identical sites, or highly conserved sites, occurred relatively evenly throughout the
309 genome, except for in the HVR1 region, which had zero identical sites across the populations.
310 Relative to non-identical sites, identical sites had a significantly lower (-5%) average mutation
311 frequency (4.69×10^{-3} *vs*. 4.93×10^{-3}, Wilcoxon test, $P = 0.0003$) and a significantly higher
312 (+4%) average selection coefficient (2.68×10^{-3} *vs*. 2.58×10^{-3}, Wilcoxon test, $P = 0.032$),
313 confirming the correspondence between conserved sites between-host and within-host.
314 **3.2. In vitro mutation rates correspond to in vivo estimates of mutation rates.** The mutation
315 rates used to obtain fitness costs in our study came from *in vitro* data [15], since *in vivo*
316 nucleotide-level mutation rates are not currently available. If *in vivo* and *in vitro* mutation rates

317 differ drastically, it would affect our estimations of fitness costs. Therefore, we evaluated pseudo
318 *in vivo* mutation rates using nonsense mutation frequencies observed in our dataset, assuming
319 that the fitness cost (s) at a nonsense mutation is one [6]. There are nine base substitution
320 combinations (out of 12) that can create stop codons; specifically, since there are no Cs in stop
321 codons, A→C, T→C, or G→C cannot be nonsense mutations. All but G→T nonsense mutations
322 were observed in our datasets. The pseudo mutation rates obtained from our *in vivo* nonsense
323 mutation frequencies followed a similar pattern as *in vitro* rates, with a significant correlation
324 between the two groups (Spearman's $\rho = 0.79$, $P = 0.028$) (S6 Fig).
325

326 **Discussion**

327 The consistent genetic diversity patterns observed between and within hosts validate the
328 frequency-based approach used in this study, as well as the idea that samples from multiple hosts
329 can serve as biological replicates. The correlation between estimated *in vitro* mutation rates with
330 pseudo *in vivo* mutation rates also supports that applying *in vitro* rates to estimate *in vivo* fitness
331 costs was appropriate. Genome-wide maps of fitness costs at each locus are key to understanding
332 the evolutionary dynamics of pathogenic viruses, yet few such maps exist. Here, we present a
333 high-resolution map of inferred *in vivo* fitness costs for every mutation in the HCV1a genome.
334 Using deep sequencing data from 195 viral populations (*i.e*. sequencing data from different
335 infected hosts), we have successfully applied the frequency-based approach to estimate *in vivo*
336 fitness costs. Our results revealed the major effects on fitness costs by nonsynonymous mutation,
337 ancestral nucleotide, and genic location.

338 **Comparing to HIV**

339 Previous studies on fitness costs of mutations in HIV found an overall similar trend in
340 site-level fitness costs, such as a clear difference in synonymous and nonsynonymous mutations
341 and large effects of ancestral nucleotide [9,11]. However, CpG-creating mutations and mutations
342 that cause drastic amino acid changes, which have large effects on fitness costs in the HIV pol
343 gene (~50% reduction in mutation frequencies), showed markedly smaller effects in HCV,
344 suggesting that the magnitude of these effects likely vary between viruses. Many RNA and DNA
345 viruses display suppressed frequencies of CpG creating mutations, indicating that CpG creating
346 mutations are costly to a variety of viruses [17], and the magnitude of such effects appear to vary
347 greatly among viruses [16]. Here, we show an additional difference: Interestingly, in HCV, CpG
348 creating mutations come with a modest cost, whereas they are much costlier in HIV. This may be
349 due to different interactions of these viruses with our immune systems and merits further study.

350 **Comparing to *in vitro* study**

351 *In vitro* studies conducted in tightly controlled environments allow us to obtain detailed
352 information about viral evolutionary dynamics, including nucleotide level mutation rates.
353 However, the use of cell-lines in an environment without selective pressures from hosts creates
354 different dynamics from *in vivo* evolution. Understanding the difference and similarity between
355 *in vitro* and *in vivo* evolution offers crucial insights into how we can study viral adaptation and
356 evolution using these tools. Our study provided an ideal opportunity to compare the genome-
357 wide mutation patterns between *in vitro* and *in vivo* settings.

358 The *in vivo* mutation patterns from our study highlighted the selective pressures exerted
359 by the host immune system on different parts of the HCV genome, compared to those of *in vitro*.
360 Significant location effects were observed *in vivo,* but not *in vitro* [15]. First, the prominent peak
361 in mutation frequency in HVR1 was observed *in vivo*, while no such peak was observed *in vitro*.

362 HVR1 plays an important role in immune escape through positive selective pressure [18,30], and
363 we observed that HVR1 was the only genic region with similarly elevated nonsynonymous
364 mutation frequency relative to synonymous frequency (Fig 2B), indicating a strong selective
365 pressure. Based on our model, location effects were also observed for seven genes *in vivo*, which
366 likely reflect selective pressures from the host. Lastly, the sites with high mutability identified *in*
367 *vitro* did not show similar trends *in vivo*, suggesting that these sites are also under negative
368 selection *in vivo*.

369 However, intrinsic mutation patterns at the nucleotide-level appear to be consistent
370 between *in vivo* and *in vitro*; A and T have higher mutation frequencies than G and C in both
371 cases, and mutation rates for each nucleotide were also correlated between *in vivo* and *in vitro*
372 (S7 Fig) [15]. Additionally, there was no significant effects of RNA structure on either *in vivo*
373 and *in vitro* mutation frequencies [15]. We also compared *in vivo* and *in vitro* mutation rates –
374 where we used observed frequencies of nonsense mutations to estimate *in vivo* mutation rates.
375 We found that relative rates of different nucleotides were consistent between the *in vivo* and *in*
376 *vitro* mutation rates. This is important since it validates our application of *in vitro* mutation rates
377 to estimate *in vivo* fitness costs. However, the *in vivo* mutation rates were two orders of
378 magnitude higher than those of *in vitro*. We suspect that this is partly because of the costs of *in*
379 *vivo* nonsense mutations are not likely to be equal to one as we assume *in vivo*; defective viruses
380 with premature stop codons can be rescued by other defective or non-defective viruses [31-34]
381 due to complex quasispecies dynamics in viruses. This phenomenon may likely allow nonsense
382 mutations to be observed at higher than expected *in vivo* viral populations.

383 **Prevalence of drug resistance**

384 Natural prevalence of resistance associated variants was more frequently and widely

385 observed in our study than previously reported [23]. This highlights the importance of deep
386 sequencing, since many variants occur at a low frequency and would therefore not be captured
387 by traditional Sanger based methods. Although the fraction of patients fixed with the alternate
388 resistance-associated variants were low (average 5.7%), one variant Q80K was observed
389 especially at high prevalence (fixed in 54.6% of patients, observed in 94.3% of patients, Fig 4).
390 This is, however, consistent with previous observations, as Q80K is known to exist at a high
391 prevalence (up to 48%) in strains from North America [21,23,35]. Patients fixed with alternate
392 alleles indicates that they were infected with the resistant strains. The results provide important
393 insight into how common resistance associate variants are.

394 Despite HCV's extremely high level of genetic variation, our detailed analysis of genome
395 wide mutation patterns revealed *in vivo* mutations do not occur in a completely stochastic
396 manner, illuminating the mutant spectrum complexity of HCV. Swarms of mutants, often termed
397 viral quasispecies, are continuously generated in RNA virus populations. The mutation spectrum
398 compositions vary over time, producing transient equilibrium distributions that are affected by
399 internal and external factors [31]. However, even with these ever-changing mutant spectra, our
400 results showed that we could observe a situation close to mutation-selection equilibrium, likely
401 controlled by intrinsic constraints. RNA viruses are considered to replicate at a mutagenesis level
402 close to the lethal threshold, at which the error rates exceed the tolerance limit of genetic
403 information preservation [36]. A delicate balance between this threshold and host selective
404 pressures is likely responsible for shaping the observed *in vivo* mutation patterns; approximately
405 50% of sites were highly conserved while the remaining variable sites contribute to continuous
406 generation of variation. From a clinical perspective one of the most challenging aspects of the
407 treatment and management of pathogenic RNA viruses is their high genetic diversity. Targeting

408 the highly conserved, high-cost sites or regions may help improve treatment efficacy by
409 hampering the development of drug resistance.

410 **Strengths and limitations of this study**

411 Deleterious mutations that occur at a low frequency indicate high fitness costs. Our
412 results suggest that inferring nucleotide-level fitness costs from mutation frequencies, however,
413 should be done cautiously: Lower mutation frequencies may not necessary translate into higher
414 fitness costs, if substantial differences in mutation rates exist at the nucleotide level along with
415 biased base composition. In our dataset, sites with A/Ts had higher fitness costs than those with
416 C/Gs, even though their mutation frequencies were significantly higher. A similar trend was
417 observed in a previous study on HIV's *pol* gene; *in vivo* mutation frequencies were highest for
418 G→A mutations, due to the high mutation rate at these sites, despite the fact that fitness costs
419 were also highest for G→A mutations [9]. Thus, without nucleotide-level mutation rates,
420 differences in fitness frequencies at the nucleotide-level need to be interpreted carefully.

421 Our unique approach worked well, thanks to the fact that the average read depth at each
422 site was > 6,200 after filtering. This resulted in an average of > 877,000 data points at each
423 nucleotide site analyzed. The high correlation between *in vivo* and among-host genetic diversity
424 at each site supports our basic assumption that *in vivo* mutation patterns are largely dictated by
425 intrinsic factors at selection-mutation equilibrium. Adding > 400 genomes from NCBI had little
426 effect on the correlation between *in vivo* and among-host diversity, which suggests that we had
427 more than adequate sample size to capture the true genetic diversity of HCV1a. Additionally, we
428 have previously performed simulation studies to assess the optimal combination of sequencing
429 depth and sample size, which showed that > 150 samples with at least 1,000X sequencing depth
430 should yield reliable estimates (*i.e.*, a high correlation coefficient, > 0.95, between real and

431 estimated fitness costs) (S8 Fig). Here, we had 195 samples with 6200X sequencing depth, more
432 than enough to produce reliable data.
433
434 **Conclusion**
435 This study serves as an additional proof of concept for the frequency-based approach to
436 understanding *in vivo* evolutionary dynamics. This opens the door to opportunities to investigate
437 *in vivo* fitness costs of a wide variety of rapidly evolving pathogens, as deep sequencing of
438 pathogen populations becomes more common with the steep decline in of next generation
439 sequencing costs. We showed that there are aspects of viral evolution that we cannot infer from
440 *in vitro* studies, as well as from the between-host level of analysis. Practical applications of
441 better understanding *in vivo* evolution and the fitness landscape of viruses include more targeted
442 antiviral drug development, efficient treatment management, better treatment outcome
443 prediction, and informative for vaccine design. By leveraging the technological advances and
444 tremendous resources already available, we have the potential to deepen our understanding of *in*
445 *vivo* evolution of viruses.
446
447 **Methods**
448 **Samples**
449 We obtained stored baseline plasma samples from a selection (n = 406) of HCV treatment
450 naïve and viremic HCV/HIV co-infected participants from sites in the Canadian Co-infection
451 Cohort (CCC) [37]. Samples were collected between May 8, 2003 and November 20, 2014 from
452 acute or chronically infected patients attending 18 different clinics across Canada.

453 **Viral RNA extraction and sequencing**

454 For each baseline sample HCV RNA was extracted from 500 ul of plasma using a
455 NucliSens easyMag (BioMerieux Inc.). Extracted RNA was eluted in 60ul of elution buffer and
456 stored at -20C prior to amplification. Near full length amplicons were generated using an oligo
457 d(A)20 primer for cDNA synthesis. Two nested PCR reactions using pangenotypic primers were
458 utilized to generate an 8,991-bp amplicon spanning the HCV Core to NS5B regions. Subsequent
459 to second round PCR amplification MiSeq sequencing libraries were generated using Nextera
460 XT DNA Sample Preparation Kits (Illumina, Inc.). Amplicons for each sample were tagged
461 using Nextera XT Index Kits (Illumina, Inc.), multiplexed and sequenced on an Illumina MiSeq
462 next-generation sequencing platform (Illumina, Inc.).

463 **Data filtering and formatting steps**

464 Sequence reads of virus isolates were quality filtered and trimmed using BBTools
465 v38.38.00 [38]. First, the adapters were removed. To increase our confidence in identifying
466 minor variants, bases that had phred score <35 were all trimmed. Contaminants (PhiX) were then
467 filtered out and reads with a <35 average score were also eliminated. Potential duplicates were
468 identified and clumped, and filtered reads were merged. Both merged and unmerged reads were
469 mapped to a reference sequence using BWA 0.7.17 [39]. Mapped reads were hard clipped, and
470 the resulting SAM files were imported into R and filtered further to eliminate potential
471 contaminants and chimeric sequences. This was done using a custom script in which all reads
472 with a harmonic distance >10 from their own consensus sequences were eliminated, as they
473 likely represented contaminants or mis-mapped sequences. Filtered reads were remapped to their
474 own consensus sequences using BWA, and the resulting bam files were imported in R [40]. For
475 each patient sample (*i.e.*, viral population), frequency tables of nucleotides at each position were

476 created with Rsamtools and custom scripts in R.

477 Each frequency table was processed to calculate the frequencies of transition and
478 transversion mutations relative to the reference sequence (wildtype H77). Types of mutation
479 (synonymous/nonsynonymous, CpG-creating, those causing drastic amino acid changes, etc.)
480 were also identified for each site. To accurately estimate *in vivo* deleterious mutation frequencies
481 at the selection-mutation-drift equilibrium, we applied rigorous filtering, removing sites with a
482 mutation frequency >0.2 (*i.e.*, those likely under active selection), sites that had a different
483 majority nucleotide from the reference, and sites with <1000 read depth. Mutation frequencies of
484 all samples were then compiled into one table, and sites that were present in less than one-third
485 of the total sample (< 65) were removed. The filtering steps produced 7,958 usable sites in the
486 coding regions, and the averages of mutation frequencies were calculated for the remaining sites.
487 This process yielded a table containing the average transition mutation frequencies at each
488 position across the HCV genome. To assess the natural prevalence of RAVs, the frequency tables
489 were processed similar to the previous step, but further filtering steps were not performed except
490 for removing sites with <1000 read depth.

491 **Data validation**

492 Since the sequencing error rate of the Illumina Miseq is reported to be around 0.1% [41],
493 we took additional steps to validate our filtering methods and the resulting datasets. We were
494 concerned that if our previous filtering steps did not remove the errors efficiently, mutations
495 observed at low frequencies (<0.001) could actually represent sequencing errors. To evaluate
496 whether this posed a problem, we replaced all mutation frequencies <0.001 with either NA or
497 zero and assessed the resulting mutation frequency patterns. The results showed consistent
498 patterns. Therefore, we used the original datasets for further analysis. We also assessed whether

499 viral load biased observed mutation frequencies by comparing the average mutation frequencies
500 of patient samples with low and high viral loads. We found no difference in the average mutation
501 frequencies between the two groups. Therefore, all samples were included in the analysis.

502 **Calculation of selection coefficients and other statistics**

503 To gain a fine-grained understanding of the *in vivo* fitness effects of mutations in HCV
504 viruses, we used the *in vivo* frequency-based approach [9,11], in which fitness costs of mutations
505 are estimated from their frequencies across multiple natural populations (*i.e*. multiple patients).
506 This approach relies on translating *in vivo* mutation frequencies into fitness costs based on the
507 principle of mutation-selection balance. In infinitely large populations, the opposing forces of
508 mutation and selection cause costly mutations to be present at a constant frequency [42,43]. In
509 haploid systems, it is expressed as:

510 $$f = \mu/s \qquad (1)$$

511 where μ is the mutation rate from wildtype to mutant and s is the negative fitness effect of the
512 mutation (selection coefficient). In natural populations of finite size, the frequency of mutations
513 is not constant but fluctuates around the expected frequency of μ/s, because of the stochastic
514 nature of mutation and replication[44] [31]. It is thus impossible to accurately infer the selection
515 acting on individual mutations from a single observation of a single population. We solved this
516 problem by averaging frequencies of each individual mutation from multiple populations. Note
517 that the principle of mutation-selection balance holds even when bottlenecks or selective sweeps
518 occur, because such events do not change the expected frequencies of mutations. Also, this
519 method allows us to directly assess fitness costs at all sites in the genome and to calculate
520 selection coefficients rather than relative rates [9].

521 Selection coefficients were estimated for each mutation at each nucleotide based on the

522 equation (1) using estimated *in vivo* mutation frequencies and mutation rates calculated from raw
523 datasets of [15] in R. The nucleotide-level mutation rates are the keys for estimating selection
524 coefficients using the frequency-based approach. Therefore, we assessed the validity of *in vitro*
525 mutation rates from Geller et al. [15] using beta regression (see below), and found no major
526 effects of selective pressures seen in *in vivo* data (see S1 Appendix).

527 *In vitro* mutation frequencies used in comparison were also derived from the datasets of
528 [15]. Statistical analyses were all performed in R. Beta regression model fitting was conducted
529 using the R package *betareg* [45]: For predictor variables, we included ancestral nucleotide,
530 types of mutations (synonymous vs. nonsynonymous, create a CpG site or not, cause a drastic
531 amino acid change or not), location (11 genic regions), RNA structure (the aggregate effects of
532 15 regions that likely form secondary structure, identified by [20], and interactions. AIC values
533 were used as a criterion to select the best fit model. Parameters for DFEs were estimated using
534 the R package noplr [46]. To assess the diversity between in vivo and among hosts, additional
535 HCV1a genomes were downloaded from NCBI GenBank.

536

537 **Acknowledgments**

538 The authors thank Kristin Harper, PhD, of Harper Health & Science Communications,
539 LLC, for providing writing and editorial support in accordance with Good Publication Practice
540 (GPP3) guidelines.

541

542 **References**

543 1. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of
544 human influenza A. Science. 1999;286: 1921–1925. doi:10.1126/science.286.5446.1921

545 2. Łuksza M, Lässig M. A predictive fitness model for influenza. Nature. Nature Publishing
546 Group; 2014;507: 57–61. doi:10.1038/nature13087

547 3. Gaunt E, Wise HM, Zhang H, Lee LN, Atkinson NJ, Nicol MQ, et al. Elevation of CpG
548 frequencies in influenza A genome attenuates pathogenicity but enhances host response to
549 infection. Elife. 2016;5: e12735. doi:10.7554/eLife.12735

550 4. Crotty S, Cameron CE, Andino R. RNA virus error catastrophe: direct molecular test by
551 using ribavirin. Proc Natl Acad Sci USA. 2001;98: 6895–6900.
552 doi:10.1073/pnas.111085598

553 5. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral Mutation Rates. Journal
554 of Virology. 2010;84: 9733–9748. doi:10.1128/JVI.00694-10

555 6. Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus
556 revealed through population sequencing. Nature. Nature Publishing Group; 2014;505:
557 686–690. doi:10.1038/nature12861

558 7. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cellular and Molecular Life
559 Sciences (CMLS). Springer International Publishing; 2016;73: 4433–4448.
560 doi:10.1007/s00018-016-2299-6

561 8. Trouplin V, Salvatori F, Cappello F, Obry V, Brelot A, Heveker N, et al. Determination of
562 Coreceptor Usage of Human Immunodeficiency Virus Type 1 from Patient Plasma
563 Samples by Using a Recombinant Phenotypic Assay. Journal of Virology. 2001;75: 251–
564 259. doi:10.1128/JVI.75.1.251-259.2001

565 9. Theys K, Feder AF, Gelbart M, Hartl M, Stern A, Pennings PS. Within-patient mutation
566 frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in
567 HIV. Bloom JD, editor. PLoS Genet. 2018;14: e1007420–24.
568 doi:10.1371/journal.pgen.1007420

569 10. Pennings PS, Kryazhimskiy S, Wakeley J. Loss and Recovery of Genetic Diversity in
570 Adapting Populations of HIV. Fraser C, editor. PLoS Genet. 2014;10: e1004000–9.
571 doi:10.1371/journal.pgen.1004000

572 11. Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape
573 of fitness costs of HIV-1. Virus Evolution. 2017;3: 10312.–12. doi:10.1093/ve/vex003

574 12. Kautz T, Forrester N. RNA Virus Fidelity Mutants: A Useful Tool for Evolutionary
575 Biology or a Complex Challenge? Viruses. 2018;10: 600–17. doi:10.3390/v10110600

576 13. Cuevas JM, González-Candelas F, Moya A, Sanjuán R. Effect of Ribavirin on the
577 Mutation Rate and Spectrum of Hepatitis C Virus In Vivo. Journal of Virology. 2009;83:
578 5760–5764. doi:10.1128/JVI.00201-09

579 14. Ribeiro RM, Li H, Wang S, Stoddard MB, Learn GH, Korber BT, et al. Quantifying the
580 Diversification of Hepatitis C Virus (HCV) during Primary Infection: Estimates of the In
581 Vivo Mutation Rate. Wilke CO, editor. PLoS Pathog. 2012;8: e1002881–13.
582 doi:10.1371/journal.ppat.1002881

583 15. Geller R, Estada Ú, Peris JB, Andreu I, Bou J-V, Garijo R, et al. Highly heterogeneous
584 mutation rates in the hepatitis C virus genome. Nature Publishing Group; 2016;1: 16045.
585 doi:10.1038/nmicrobiol.2016.45

586 16. Cheng X, Virk N, Chen W, Ji S, Ji S, Sun Y, et al. CpG usage in RNA viruses: data and
587 hypotheses. Burk RD, editor. PLoS ONE. 2013;8: e74109.
588 doi:10.1371/journal.pone.0074109

589 17. Caudill VR, Qin S, Winstead R, Kaur J, Tisthammer K, Pineda EG, et al. CpG-creating
590 mutations are costly in many human viruses. Evol Ecol. Springer International Publishing;
591 2020;34: 339–359. doi:10.1007/s10682-020-10039-z

592 18. Kato N. Genome of Human Hepatitis C Virus (HCV): Gene organiazation, sequence
593 diveristy, and variation. Microbial Comparative Genomics. 2000;5: 129–151.

594 19. Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo
595 of hepatitis C virus evolution within and among hosts. BMC Evolutionary Biology.
596 2011;11. doi:10.1186/1471-2148-11-131

597 20. Mauger DM, Golden M, Yamane D, Williford S, Lemon SM, Martin DP, et al.
598 Functionally conserved architecture of hepatitis C virus RNA genomes. Proc Natl Acad
599 Sci USA. 2015;: 201416266–6. doi:10.1073/pnas.1416266112

600 21. Poveda E, Wyles DL, Mena Á, Pedreira JD, Castro-Iglesias Á, Cachay E. Update on
601 hepatitis C virus resistance to direct-acting antiviral agents. Antiviral Research. Elsevier
602 B.V; 2014;108: 181–191. doi:10.1016/j.antiviral.2014.05.015

603 22. Lontok E, Harrington P, Howe A, Kieffer T, Lennerstrand J, Lenz O, et al. Hepatitis C
604 virus drug resistance-associated substitutions: State of the art summary. Hepatology.
605 2015;62: 1623–1632. doi:10.1002/hep.27934

606 23. Sarrazin C. The importance of resistance to direct antiviral drugs in HCV infection in
607 clinical practice. European Association for the Study of the Liver; 2015;: 1–19.
608 doi:10.1016/j.jhep.2015.09.011

609 24. Kim S, Han K-H, Ahn SH. Hepatitis C Virus and Antiviral Drug Resistance. Gut and
610 Liver. 2016;10: 890–895. doi:10.5009/gnl15573

611 25. Wyles DL, Luetkemeyer AF. Understanding Hepatitis C Virus Drug Resistance: Clinical
612 Implications for Current and Future Regimens. Top Antivir Med. 2017;25: 103–109.

613 26. Sorbo MC, Cento V, Di Maio VC, Howe AYM, Garcia F, Perno CF, et al. Hepatitis C
614 virus drug resistance associated substitutions and their clinical relevance: Update 2018.
615 Drug Resistance Updates. 2018;37: 17–39. doi:10.1016/j.drup.2018.01.004

616 27. Johnson JA, Li JF, Wei X, Lipscomb J, Iribeck D, Craig C, et al. Minority HIV-1 drug
617 resistance mutations are present in antiretroviral treatment–naïve populations and
618 associate with reduced treatment efficacy. PLoS Medicine. 2008;5: e158.
619 doi:10.1371/journal. pmed.0050158

620 28. Li JZ, Paredes R, Ribaudo HJ, Svarovskaia ES, Metzner KJ, Kozal MJ, et al. Low-
621 frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral
622 treatment failure: a systematic review and pooled analysis. JAMA. 2011;305: 1327–1335.
623 doi:10.1001/jama.2011.375

624 29. Echeverría N, Betancour G, Gámbaro F, Hernández N, López P, Chiodi D, et al. Naturally
625 occurring NS3 resistance-associated variants in hepatitis C virus genotype 1: Their

626 relevance for developing countries. Virus Research. 2016;223: 140–146.
627 doi:10.1016/j.virusres.2016.07.008

628 30. Keck Z-Y, Girard-Blanc C, Wang W, Lau P, Zuiani A, Rey FA, et al. Antibody Response
629 to Hypervariable Region 1 Interferes with Broadly Neutralizing Antibodies to Hepatitis C
630 Virus. Ou JHJ, editor. Journal of Virology. 2016;90: 3112–3122. doi:10.1128/JVI.02458-
631 15

632 31. Perales C, Lorenzo-Redondo R, López-Galíndez C, Martínez MA, Domingo E. Mutant
633 spectra in virus behavior. Future Virology. 2010;5: 679–698. doi:10.2217/fvl.10.61

634 32. Quan Y, Xu H, Wainberg MA. Defective HIV-1 quasispecies in the form of multiply
635 drug-resistant proviral DNA within cells can be rescued by superinfection with different
636 subtype variants of HIV-1 and by HIV-2 and SIV. J Antimicrob Chemother. 2014;69: 21–
637 27. doi:10.1093/jac/dkt326

638 33. Díaz-Muñoz SL, Sanjuán R, West S. Sociovirology: Conflict, Cooperation, and
639 Communication among Viruses. Cell Host and Microbe. The Authors; 2017;22: 437–441.
640 doi:10.1016/j.chom.2017.09.012

641 34. Vignuzzi M, López CB. Defective viral genomes are key drivers of the virus–host
642 interaction. Nature Microbiology. Springer US; 2019;4: 1075–1087. doi:10.1038/s41564-
643 019-0465-y

644 35. McCloskey RM, Liang RH, Joy JB, Krajden M, Montaner JSG, Harrigan PR, et al. Global
645 Origin and Transmission of Hepatitis C Virus Nonstructural Protein 3 Q80K

646 Polymorphism. The Journal of Infectious Diseases. 2014;211: 1288–1295.
647 doi:10.1093/infdis/jiu613

648 36. Echeverría N. Hepatitis C virus genetic variability and evolution. WJH. 2015;7: 831–16.
649 doi:10.4254/wjh.v7.i6.831

650 37. Klein MB, Saeed S, Yang H, Cohen J, Conway B, Cooper C, et al. Cohort profile: the
651 Canadian HIV–hepatitis C co-infection cohort study. International Journal of
652 Epidemiology. Oxford University Press; 2010;39: 1162–1169.

653 38. Bushnell B. BBTools [Internet]. 2018. Available: https://sourceforge.net/projects/bbmap/

654 39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
655 Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

656 40. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
657 Vienna, Austria: R Foundation for Statistical Computing; 2014. Available: http://www.R-
658 project.org

659 41. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing
660 Platforms. Next Gener Seq Appl. 2014;1: 1000106. doi:10.4172/jngsa.1000106

661 42. Wright S. Evolution in Mendelian Populations. Genetics. 1931;16: 97–159.

662 43. Haldane J. The effect of variation on fitness. American naturalist. 1937;71: 337–349.

663 44. Hartl DL, Clark AG. Principles of population genetics. Third. Sunderland, MA: Sinauer
664 Associates, Inc; 1997.

665 45. Francisco Cribari-Neto AZ. Beta Regression in R. 2019; 1–24.

666 46. Ypma J, Borchers HW, Eddelbuettel D. Package "nloptr." 2018 Oct pp. 1–46.

667

668 **Figure Captions**

669 **Fig 1. Estimated *in vivo* mutation frequencies in HCV1a, as determined by analysis of 195**
670 **viral populations**. (A) Average estimated frequencies for all, synonymous (Syn), non-
671 synonymous (Nonsyn), and nonsense mutations, stratified by transitions vs transversions.
672 Transition mutation frequencies are much higher for every class of mutation. (B) Genome-wide
673 transition mutation frequencies, ordered by mutation frequency and colored by mutation type,
674 show that synonymous mutations are more common than non-synonymous mutations in the
675 HCV genome. (C) Transition mutation frequencies along the HCV genome, colored by mutation
676 type, show that average mutation frequency is roughly consistent across the genome, with
677 synonymous mutations more common than nonsynonymous mutations. Each dot represents the
678 average mutation frequency at a nucleotide position, across the 195 viral populations. The line
679 represents the sliding window average of 100 bases.

680

681 **Fig 2. Estimated transition mutation frequencies of HCV by gene**. Aggregated observed
682 frequencies by gene and type of mutations; dots and lines indicate the averages and the standard
683 errors across 195 samples. * denotes statistically significant difference ($P<0.05$) between
684 synonymous and nonsynonymous mutations by Wilcoxon test.

685

686 **Fig 3. Various factors that affect the frequency of mutations in HCV, based on analysis of**
687 **195 viral populations, each derived from a different patient.** (A) Predicted effects of ancestral

688 nucleotide (T, C, or G, *vs*. A); CpG-creating status (*vs*. non-CpG-creating status);
689 nonsynonymous (Nonsyn; *vs*. synonymous); amino acid-changing (AAChange; *vs*. non-amino
690 acid-changing); presence in the Core, HVR1, E2, NS1, NS2, NS4A, and NS5B genes (*vs*. the
691 NS3/NS4B/NS5A regions) on mutation frequencies in the HCV genome. Beta regression models
692 were used to determine the effects of the different factors on mutation frequencies across the
693 genome, and this figure reflects the results of the best-fit model based on AIC. (B) Effects of
694 ancestral nucleotides upon synonymous (syn) and nonsynonymous (nonsyn) mutation
695 frequencies in 195 patients infected with HCV.
696
697 **Fig 4. Estimated genome-wide selection coefficients (fitness costs) in the HCV genome**. (A)
698 Selection coefficients along the HCV genome, colored by mutation type (syn, synonymous;
699 nonsynon, nonsynonymous); each dot represents the average at each position across 195 patient
700 samples, and lines represent the sliding window average for 50 bases. (B) Selection coefficients
701 stratified by nucleotide and syn/nonsyn status, colored by mutation type. (C) Estimated mutation
702 frequencies stratified by starting nucleotide and syn/nonsyn status, colored by mutation type.
703 Comparison of (B) and (C) shows higher estimated selection coefficients in A /T than C/G,
704
705 **Fig 5. Distribution of fitness effects for each nucleotide (A, T, C, G) in the coding regions of**
706 **the HCV genome.** Stratified by synonymous and nonsynonymous mutation status, using the
707 estimated fitness costs (*s*) calculated in this study.
708
709 **Fig 6. Natural prevalence and fitness costs of resistance-associated variants in the HCV**
710 **genome.** Estimated fitness costs (top) and natural prevalence (bottom) of resistance-associated

711 variants, where each dot represents a mutation frequency observed in a patient. The first numbers
712 below the cost plot (top) show the percentage of samples (patients) with resistance-associated
713 variants observed, and the second numbers show the percentage of patients that had given
714 resistance-associated variants as majority nucleotides ('fixed sample'). Variant names in black are
715 created by transition and those in purple are created by transversion mutations.
716
717 **Supporting Information**
718 **S1 Fig. Comparison of *in vivo* and *in vitro* transition mutation frequencies in the HCV1a**
719 **genome.** *In vivo* frequencies are from this study, and *in vitro* frequencies are from Geller et al.
720 (2016). (A) Observed mutation frequencies by nucleotide, and (B) by gene.
721
722 **S2 Fig. Transition mutation frequencies of the sites corresponding to the highly mutable**
723 **(HM) and less mutable (LM) sites identified in Geller et al. (2016)**. The dashed line
724 represents the genome-wide mean mutation frequency in our 195 patient samples.
725
726 **S3 Fig. Mutation rates estimated from the dataset obtained from Geller et al. (2016)**. Blue
727 dots represent the estimated transition mutation rates for each nucleotide, and bars represent the
728 95% confidential intervals.
729
730 **S4 Fig. Average selection coefficients (A) and AT contents (B) of each genic region in the**
731 **HCV genome.**
732

733 **S5 Fig. Predicted effects of various factors on mutation frequency and selection coefficient,**
734 **based on the best fit model from beta regression.** For details, see Fig 3 description.
735
736 **S6 Fig. Within-host vs. between-host nucleotide diversity comparison.**
737
738 **S7 Fig. HCV mutation rates estimated from Gellers et al. (2016)'s *in vitro* dataset and our**
739 ***in vivo* dataset, using nonsense mutation frequencies.**
740
741 **S8 Fig. Simulations results showing the relationship between number of samples, coverage,**
742 **and accuracy of inference, as determined by the correlation coefficient between simulated**
743 **fitness costs and their estimates based on a mean-field approach.**
744
745 **S9 Fig. Phylogenetic tree of HCV 1a used in this study.** The numbers at the node represent
746 the Bayesian posterior probability values.
747
748 **S1 Table. The results of beta regression with the estimated effects of different factors on**
749 **transition mutation frequencies.** The intercept is estimated for synonymous, non CpG creating
750 A→G mutations. The results are from the best fit model selected using the AIC scores. Effect
751 size represents the estimated increase of decrease in mutation frequencies compared to the
752 intercept.
753
754 **S2 Table. The results of beta regression with the estimated effects of different factors on**
755 **inferred selection coefficients.** The same model as the best fit model for estimating mutation

756 frequencies was used for comparison. See S1. Table legend for details. Effect size represents the
757 estimated increase of decrease in selection coefficients compared to the intercept.
758
759 **S3 Table. Resistance-associated variants in the NS3, NS5A, and NS5B genes examined in**
760 **our study.** AA: amino acid, NT: nucleotide, wt: wildtype (reference), res: resistant, Ts: transition
761 mutation, Tv: transversion mutation
762
763 **S1 File. Supplementary methods.**
764
765 **Data availability**
766 All relevant data are within the manuscript and its Supporting Information files,

Figure

Fig. 1. Estimated *in vivo* mutation frequencies in HCV1a, as determined by analysis of 195 viral populations. (A) Average estimated frequencies for all, synonymous (Syn), non-synonymous (Nonsyn), and nonsense mutations, stratified by transitions vs transversions. Transition mutation frequencies are much higher for every class of mutation. (B) Genome-wide transition mutation frequencies, ordered by mutation frequency and colored by mutation type, show that synonymous mutations are more common than non-synonymous mutations in the HCV genome. (C) Transition mutation frequencies along the HCV genome, colored by mutation type, show that average mutation frequency is roughly consistent across the genome, with synonymous mutations more common than nonsynonymous mutations. Each dot represents the average mutation frequency at a nucleotide position, across the 195 viral populations. The line represents the sliding window average of 100 bases.

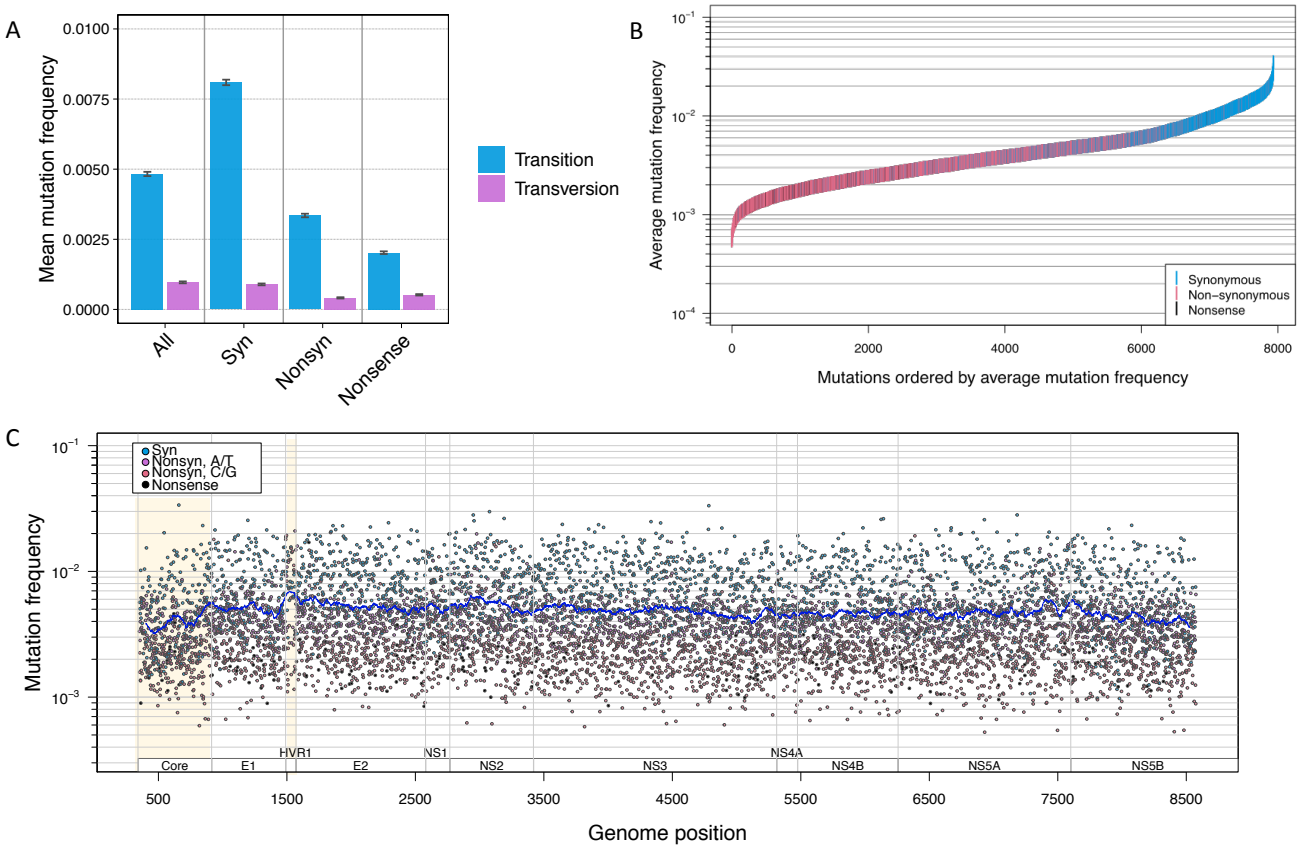


Fig. 2. Estimated transition mutation frequencies of HCV by gene. Aggregated observed frequencies by gene and type of mutations; dots and lines indicate the averages and the standard errors across 195 samples. * denotes statistically significant difference (P<0.05) between synonymous and nonsynonymous mutations by Wilcoxon test.

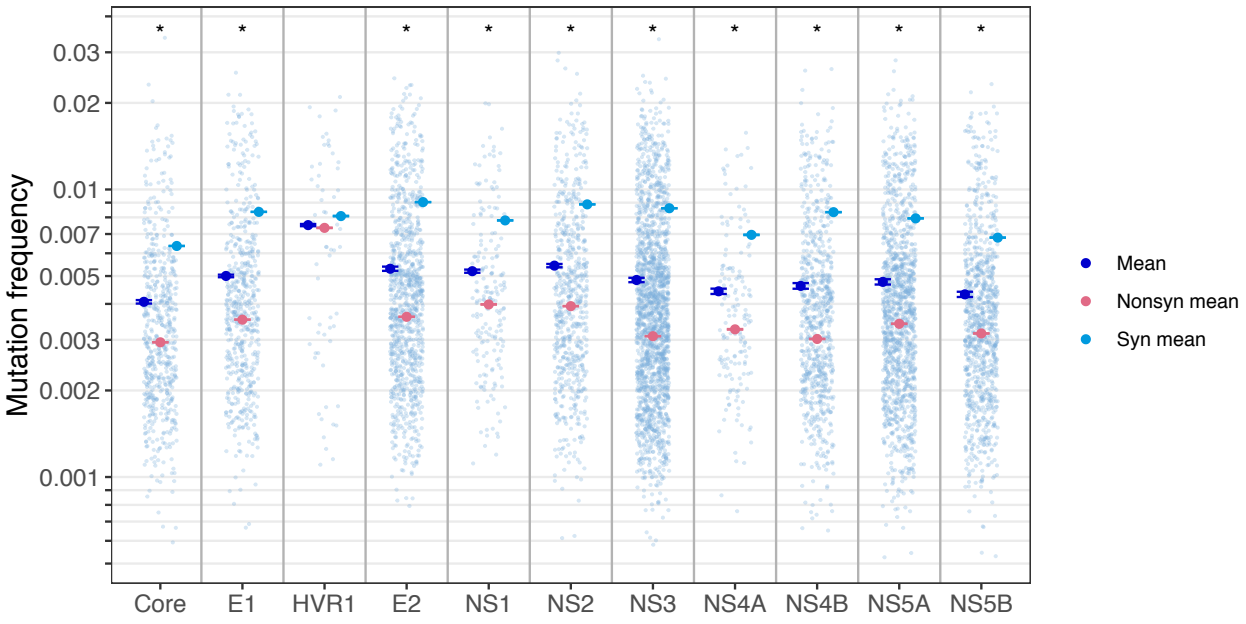


Fig. 3. Various factors that affect the frequency of mutations in HCV, based on analysis of 195 viral populations, each derived from a different patient. (A) Predicted effects of ancestral nucleotide (T, C, or G, *vs.* A); CpG-creating status (*vs.* non-CpG-creating status); nonsynonymous (Nonsyn; *vs.* synonymous); amino acid-changing (AAChange; *vs.* non-amino acid-changing); presence in the Core, HVR1, E2, NS1, NS2, NS4A, and NS5B genes (*vs.* the NS3/NS4B/NS5A regions) on mutation frequencies in the HCV genome. Beta regression models were used to determine the effects of the different factors on mutation frequencies across the genome, and this figure reflects the results of the best-fit model based on AIC. (B) Effects of ancestral nucleotides upon synonymous (syn) and nonsynonymous (nonsyn) mutation frequencies in 195 patients infected with HCV.

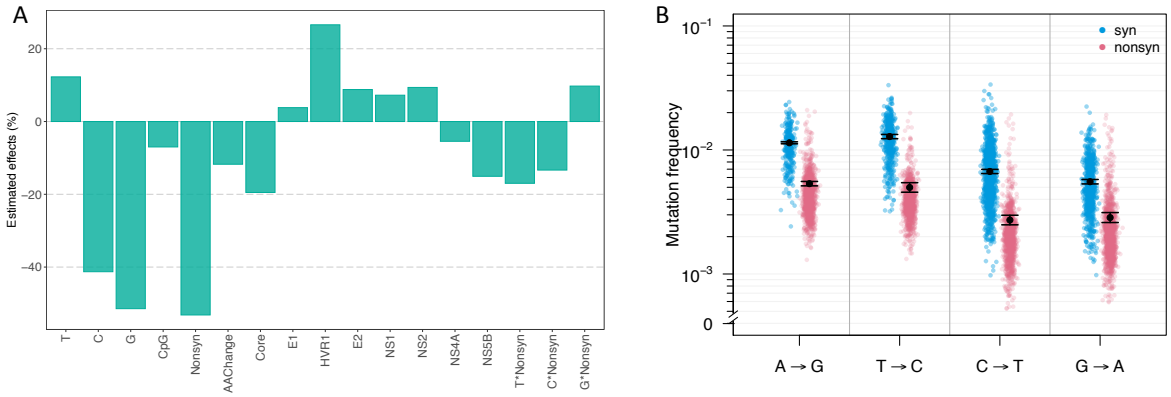


Fig. 4. Estimated genome-wide selection coefficients (fitness costs) in the HCV genome. (A) Selection coefficients along the HCV genome, colored by mutation type (syn, synonymous; nonsynon, nonsynonymous); each dot represents the average at each position across 195 patient samples, and lines represent the sliding window average for 50 bases. (B) Selection coefficients stratified by nucleotide and syn/nonsyn status, colored by mutation type. (C) Estimated mutation frequencies stratified by starting nucleotide and syn/nonsyn status, colored by mutation type. Comparison of (B) and (C) shows higher estimated selection coefficients in A /T than C/G, though their mutation frequencies were higher than C/G (hence considered costly).

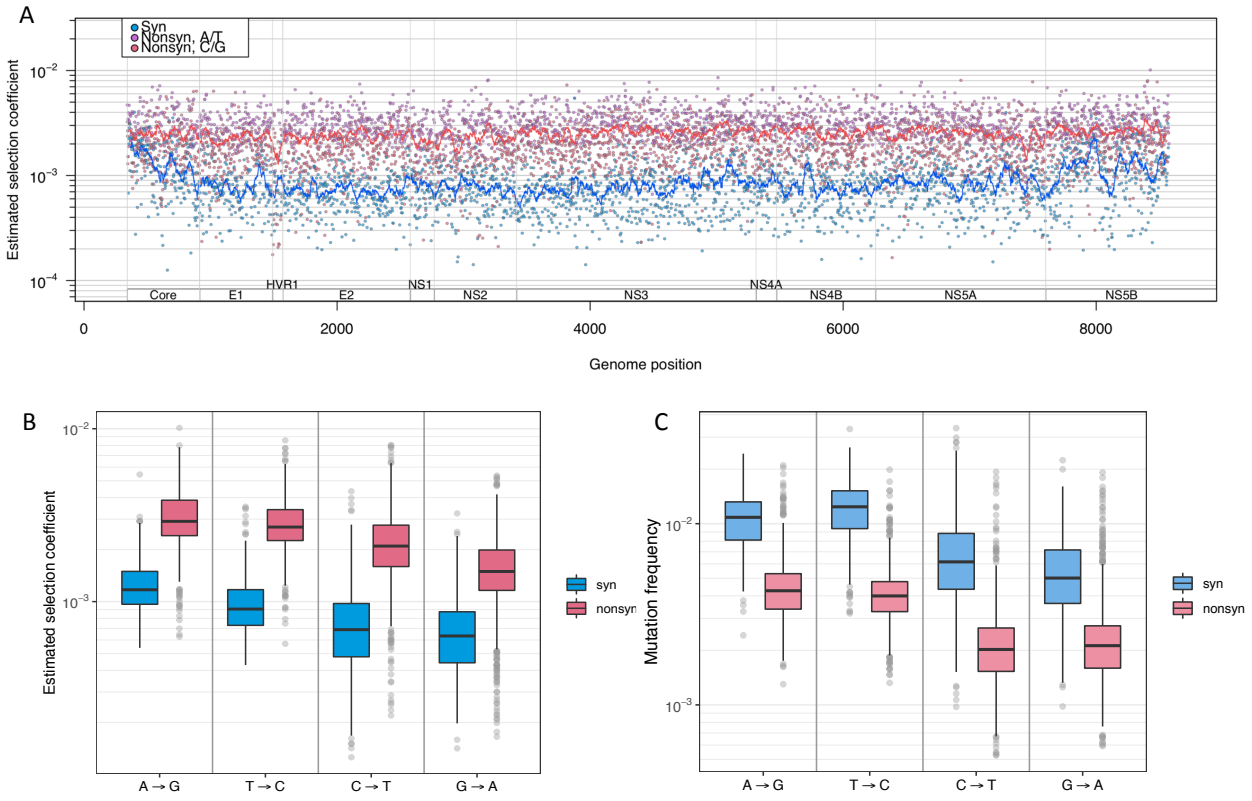


Fig. 5. Distribution of fitness effects for each nucleotide (A, T, C, G) in the coding regions of the HCV genome, stratified by synonymous and nonsynonymous mutation status, using the estimated fitness costs (s) calculated in this study.

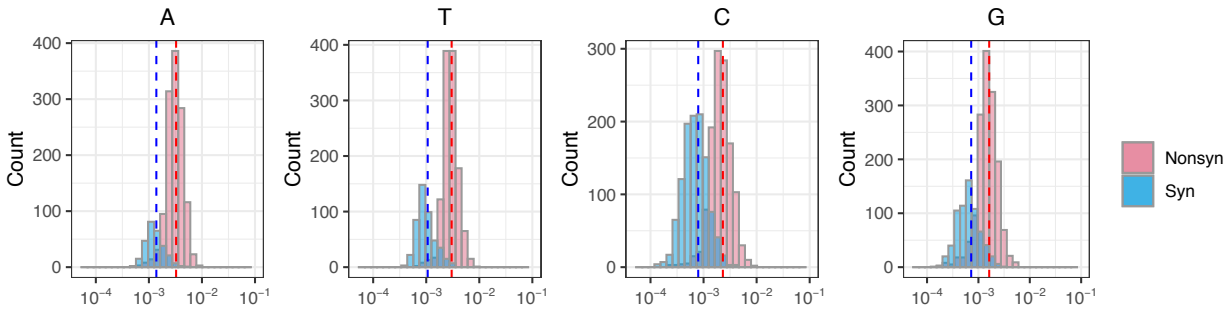
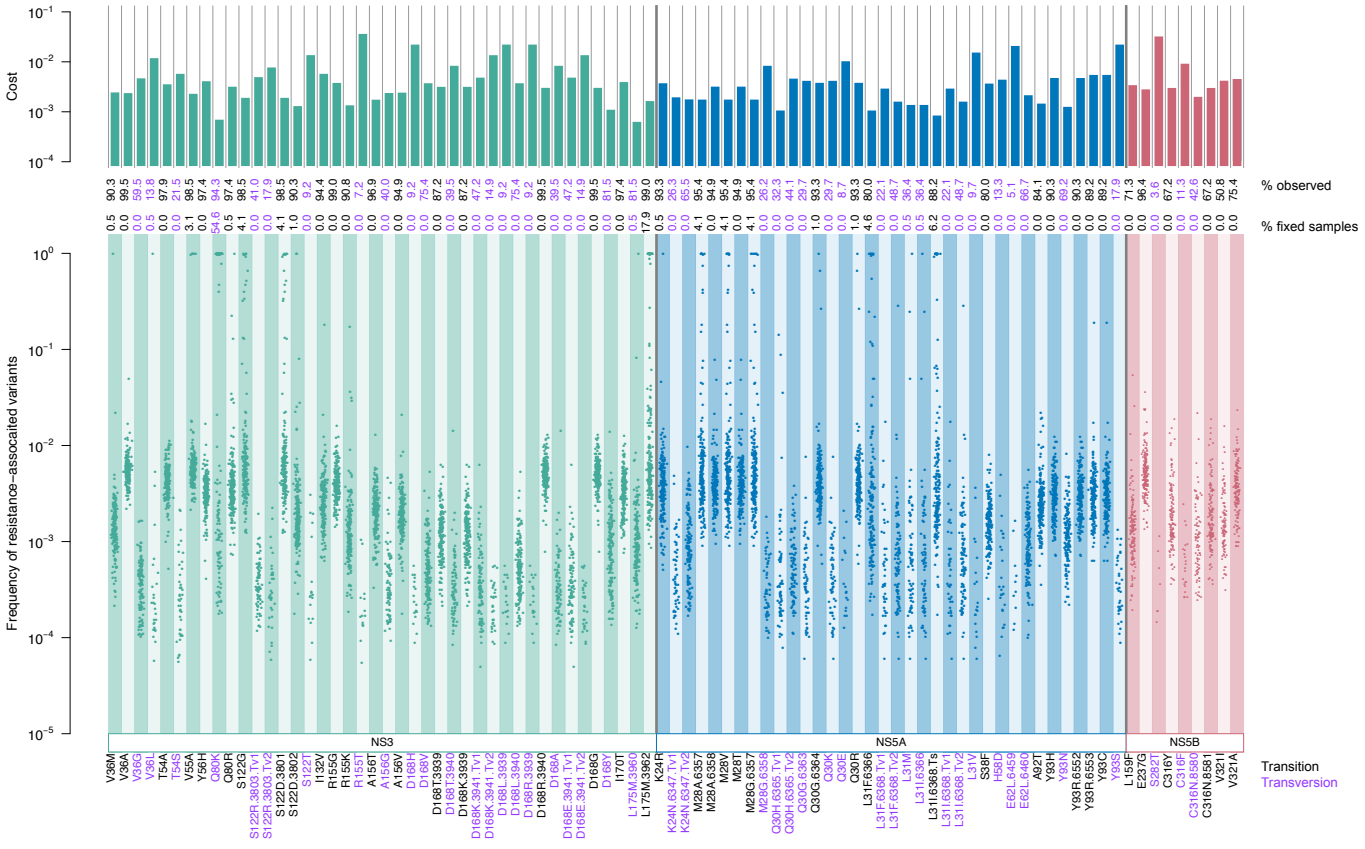


Fig. 6. Natural prevalence and fitness costs of resistance-associated variants in the HCV genome. Estimated fitness costs (top) and natural prevalence (bottom) of resistance-associated variants, where each dot represents a mutation frequency observed in a patient. The first numbers below the cost plot (top) show the percentage of samples (patients) with resistance-associated variants observed, and the second numbers show the percentage of patients that had given resistance-associated variants as majority nucleotides ('fixed sample') . Variant names in black are created by transition and those in purple are created by transversion mutations.



S1 Table. The results of beta regression with the estimated effects of different factors on transition mutation frequencies. The intercept is estimated for synonymous, non CpG creating A→G mutations. The results are from the best fit model selected using the AIC scores. Effect size represents the estimated increase of decrease in mutation frequencies compared to the intercept.

	Estimate	Z-value	*P*-value	Effects
Intercept	-4.3767	-243.55	0	
T	0.1157	5.55	<0.0001	**12.3%**
C	-0.5323	-25.96	<0.0001	**-41.3%**
G	-0.7222	-31.47	<0.0001	**-51.4%**
CpG	-0.0725	-5.39	<0.0001	**-7.0%**
Nonsyn	-0.7582	-33.94	<0.0001	**-53.2%**
bigAA Change	-0.1249	-8.89	<0.0001	**-11.7%**
Core	-0.2169	-10.96	<0.0001	**-19.5%**
E1	0.0375	2.06	<0.0001	**3.8%**
HVR1	0.2357	5.09	<0.0001	**26.6%**
E2	0.0843	5.94	<0.0001	**8.8%**
NS1	0.0699	2.32	<0.0001	**7.2%**
NS2	0.0895	5.27	<0.0001	**9.4%**
NS4A	-0.0559	-1.61	0.107	-5.4%
NS5B	-0.1632	-10.48	<0.0001	**-15.1%**
T:Nonsyn	-0.1861	-6.94	<0.0001	**-17.0%**
C:Nonsyn	-0.1432	-5.16	<0.0001	**-13.3%**
G:Nonsyn	0.0930	3.24	0.0012	**9.7%**

S2 Table. The results of beta regression with the estimated effects of different factors on inferred selection coefficients. The same model as the best fit model for estimating mutation frequencies was used for comparison. See S1. Table legend for details. Effect size represents the estimated increase of decrease in selection coefficients compared to the intercept.

	Estimate	Z-value	*P*-value	Effects
Intercept	-6.5497	-242.851	0.0000	
T	-0.2358	-6.928	<0.0001	-21.0%
C	-0.5067	-16.425	<0.0001	-39.7%
G	-0.5772	-17.072	<0.0001	-43.9%
CpG	0.0677	5.773	<0.0001	7.0%
Nonsyn	0.7292	25.988	<0.0001	107.3%
bigAA Change	0.1298	12.367	<0.0001	13.9%
Core	0.0593	3.503	0.0005	6.1%
E1	-0.0878	-4.933	<0.0001	-8.4%
HVR1	-0.4570	-7.670	<0.0001	-36.7%
E2	-0.1040	-7.284	<0.0001	-9.9%
NS1	-0.1672	-5.424	<0.0001	-15.4%
NS2	-0.1346	-7.725	<0.0001	-12.6%
NS4A	-0.0535	-4.206	<0.0001	-5.2%
NS5B	0.0558	4.094	0.0001	5.7%
T:Nonsyn	0.1685	4.666	<0.0001	18.4%
C:Nonsyn	0.1716	5.185	<0.0001	18.7%
G:Nonsyn	-0.0891	-2.474	0.0134	-8.5%

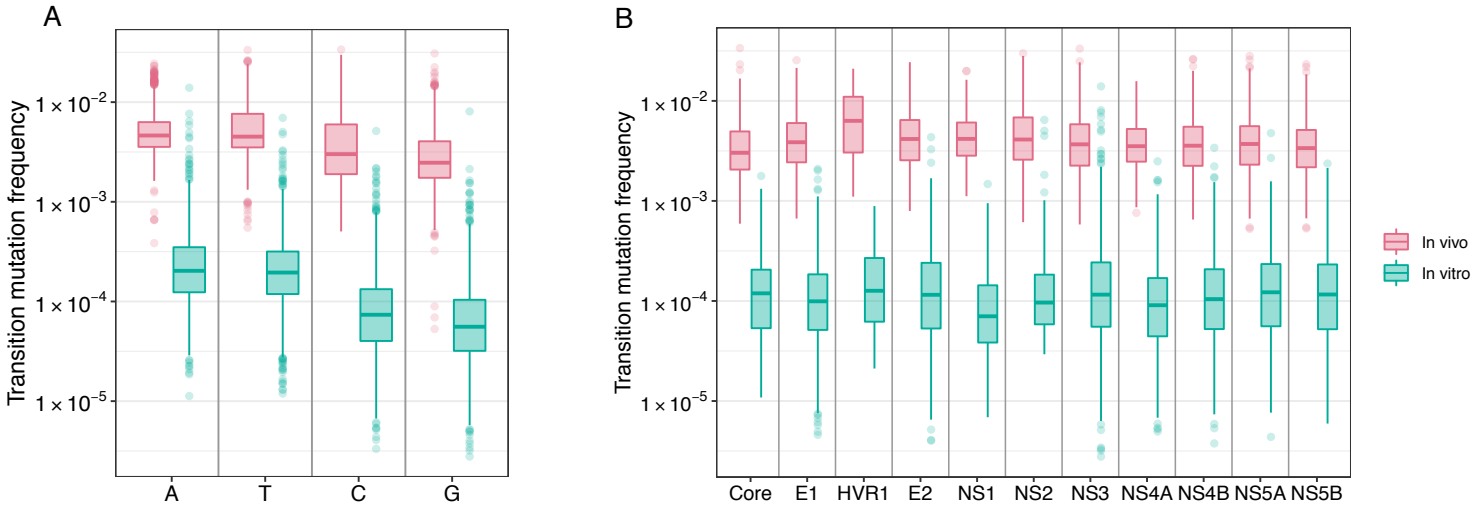
S3 Table. Resistance-associated variants in the NS3, NS5A, and NS5B genes examined in our study. AA: amino acid, NT: nucleotide, wt: wildtype (reference), res: resistant, Ts: transition mutation, Tv: transversion mutation

Variant	*AA Position*	*wtAA*	*resAA*	*Gene*	*NT Position*	*Mutation Type*	*wtNT*	*resNT*
V36M	36	V	M	NS3	3543	Ts	g	a
V36A	36	V	A	NS3	3544	Ts	t	c
V36G	36	V	G	NS3	3544	Tv	t	g
V36L	36	V	L	NS3	3543	Tv	g	c
T54A	54	T	A	NS3	3597	Ts	a	g
T54S	54	T	S	NS3	3598	Tv	c	g
V55A	55	V	A	NS3	3601	Ts	t	c
Y56H	56	Y	H	NS3	3603	Ts	t	c
Q80K	80	Q	K	NS3	3675	Tv	c	a
Q80R	80	Q	R	NS3	3676	Ts	a	g
S122G	122	S	G	NS3	3801	Ts	a	g
S122R	122	S	R	NS3	3803	Tv	c	a
S122R	122	S	R	NS3	3803	Tv	c	g
S122D	122	S	D	NS3	3801	Ts	a	g
S122D	122	S	D	NS3	3802	Ts	g	a
S122T	122	S	T	NS3	3802	Tv	g	c
I132V	132	I	V	NS3	3831	Ts	a	g
R155G	155	R	G	NS3	3900	Ts	a	g
R155K	155	R	K	NS3	3901	Ts	g	a
R155T	155	R	T	NS3	3901	Tv	g	c
A156T	156	A	T	NS3	3903	Ts	g	a
A156G	156	A	G	NS3	3904	Tv	c	g
A156V	156	A	V	NS3	3904	Ts	c	t
D168H	168	D	H	NS3	3939	Tv	g	c
D168V	168	D	V	NS3	3940	Tv	a	t
D168T	168	D	T	NS3	3939	Ts	g	a
D168T	168	D	T	NS3	3940	Tv	a	c
D168K	168	D	K	NS3	3939	Ts	g	a
D168K	168	D	K	NS3	3941	Tv	c	a
D168K	168	D	K	NS3	3941	Tv	c	g

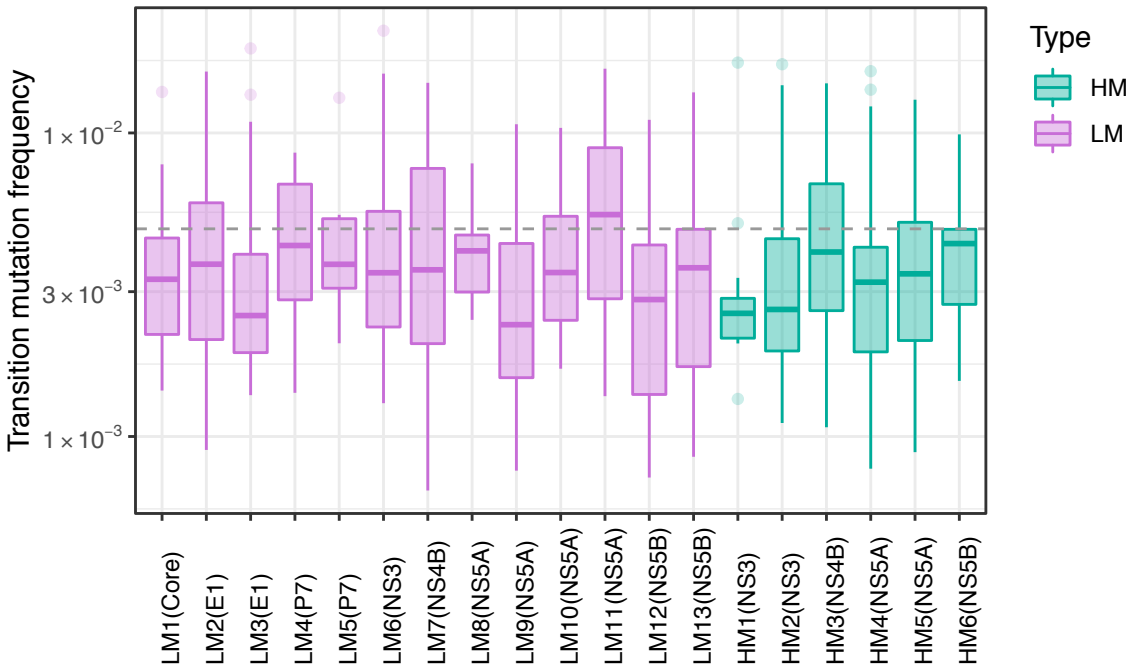
D168L	168	D	L	NS3	3939	Tv	g	c
D168L	168	D	L	NS3	3940	Tv	a	t
D168R	168	D	R	NS3	3939	Tv	g	c
D168R	168	D	R	NS3	3940	Ts	a	g
D168A	168	D	A	NS3	3940	Tv	a	c
D168E	168	D	E	NS3	3941	Tv	c	a
D168E	168	D	E	NS3	3941	Tv	c	g
D168G	168	D	G	NS3	3940	Ts	a	g
D168Y	168	D	Y	NS3	3939	Tv	g	t
I170T	170	I	T	NS3	3946	Ts	t	c
L175M	175	L	M	NS3	3960	Tv	c	a
L175M	175	L	M	NS3	3962	Ts	a	g
K24R	24	K	R	NS5A	6346	Ts	a	g
K24N	24	K	N	NS5A	6347	Tv	g	c
K24N	24	K	N	NS5A	6347	Tv	g	t
M28A	24	M	A	NS5A	6357	Ts	a	g
M28A	24	M	A	NS5A	6358	Ts	t	c
M28V	24	M	V	NS5A	6357	Ts	a	g
M28T	24	M	T	NS5A	6358	Ts	t	c
M28G	28	M	G	NS5A	6357	Ts	a	g
M28G	28	M	G	NS5A	6358	Tv	t	g
Q30H	30	Q	H	NS5A	6365	Tv	a	c
Q30H	30	Q	H	NS5A	6365	Tv	a	t
Q30G	30	Q	G	NS5A	6363	Tv	c	g
Q30G	30	Q	G	NS5A	6364	Ts	a	g
Q30K	30	Q	K	NS5A	6363	Tv	c	a
Q30E	30	Q	E	NS5A	6363	Tv	c	g
Q30R	30	Q	R	NS5A	6364	Ts	a	g
L31F	31	L	F	NS5A	6366	Ts	c	t
L31F	31	L	F	NS5A	6368	Tv	g	c
L31F	31	L	F	NS5A	6368	Tv	g	t
L31M	31	L	M	NS5A	6366	Tv	c	a
L31I	31	L	I	NS5A	6366	Tv	c	a
L31I	31	L	I	NS5A	6368	Ts	g	a
L31I	31	L	I	NS5A	6368	Tv	g	c
L31I	31	L	I	NS5A	6368	Tv	g	t

L31V	31	L	V	NS5A	6366	Tv	c	g
S38F	38	S	F	NS5A	6388	Ts	c	t
H58D	58	H	D	NS5A	6447	Tv	c	g
E62L	62	E	D	NS5A	6459	Tv	g	c
E62L	62	E	D	NS5A	6460	Tv	a	t
A92T	92	A	T	NS5A	6549	Ts	g	a
Y93H	93	Y	H	NS5A	6552	Ts	t	c
Y93N	93	Y	N	NS5A	6552	Tv	t	a
Y93R	93	Y	R	NS5A	6552	Ts	t	c
Y93R	93	Y	R	NS5A	6553	Ts	a	g
Y93C	93	Y	C	NS5A	6553	Ts	a	g
Y93S	93	Y	S	NS5A	6553	Tv	a	c
L159F	159	L	F	NS5B	8109	Ts	c	t
E237G	237	E	G	NS5B	8344	Ts	a	g
S282T	282	S	T	NS5B	8479	Tv	g	c
C316Y	316	C	Y	NS5B	8581	Ts	g	a
C316F	315	C	F	NS5B	8581	Tv	g	t
C316N	316	C	N	NS5B	8580	Tv	t	a
C316N	316	C	N	NS5B	8581	Ts	g	a
V321I	321	V	I	NS5B	8595	Ts	g	a
V321A	321	V	A	NS5B	8596	Ts	t	c

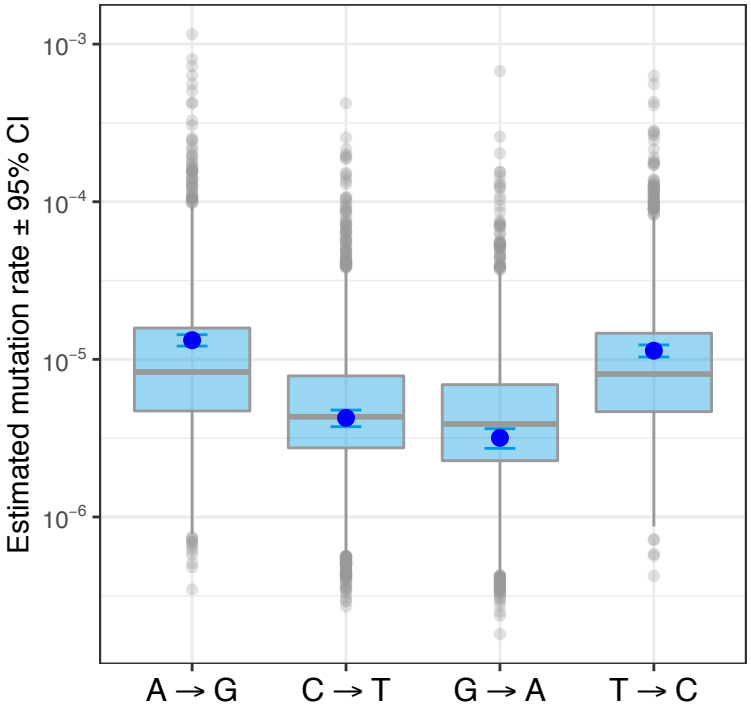
S1 Fig. Comparison of *in vivo* and *in vitro* transition mutation frequencies in the HCV1a genome. *In vivo* frequencies are from this study, and *in vitro* frequencies are from Geller et al. (2016). (A) Observed mutation frequencies by nucleotide, and (B) by gene.



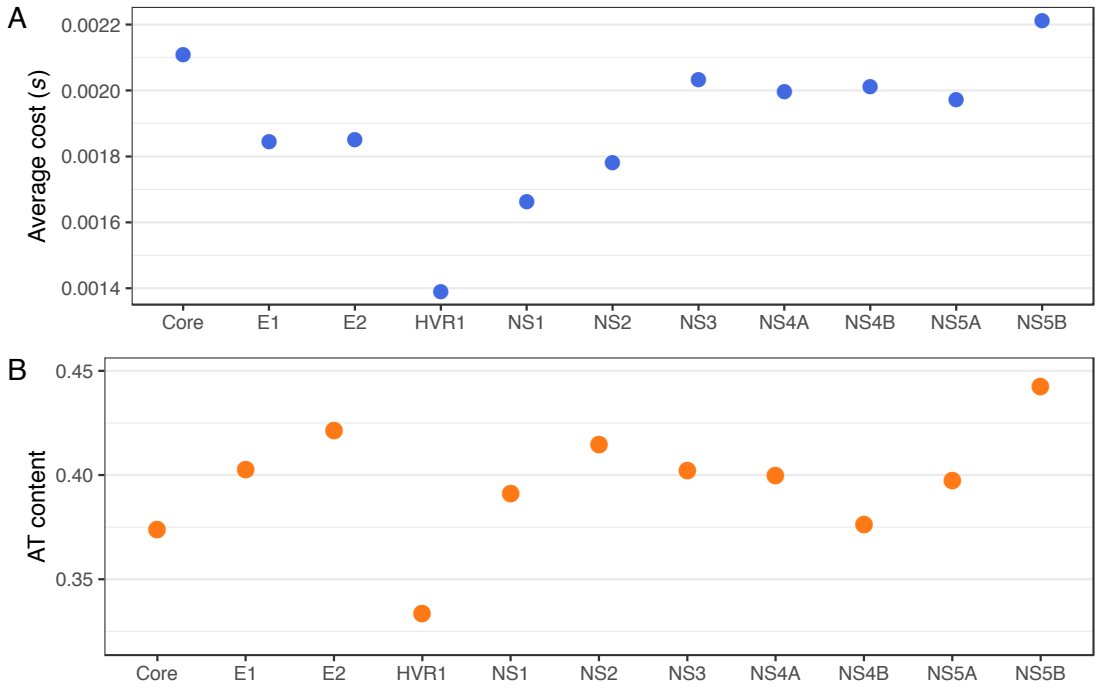
S2 Fig. Transition mutation frequencies of the sites corresponding to the highly mutable (HM) and less mutable (LM) sites identified in Geller et al. (2016). The dashed line represents the genome-wide mean mutation frequency in our 195 patient samples.



S3 Fig. Mutation rates estimated from the dataset obtained from Geller et al. (2016). Blue dots represent the estimated transition mutation rates for each nucleotide, and bars represent the 95% confidence intervals.

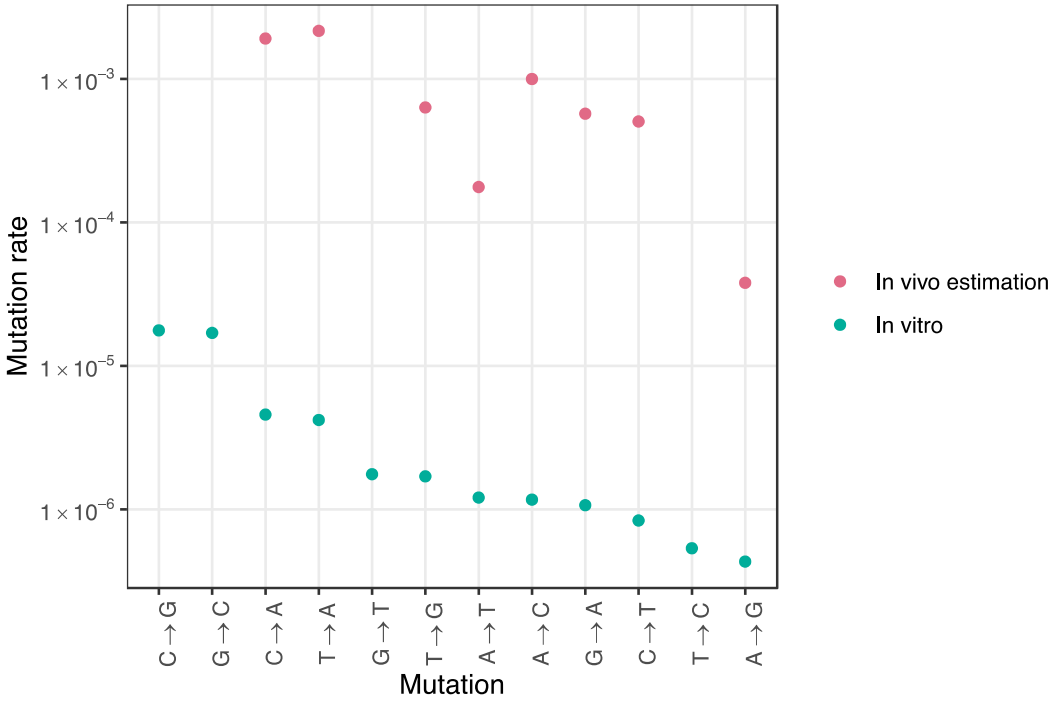


S4 Fig. Average selection coefficients (A) and AT contents (B) of each genic region in the HCV genome.

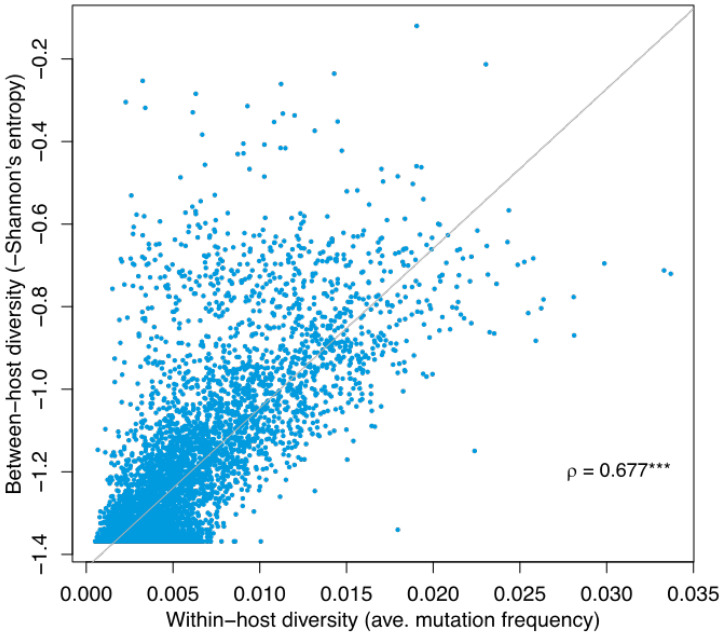


S5 Fig. Predicted effects of various factors on mutation frequency and selection coefficient, based on the best fit model from beta regression. For details, see Fig. 3 description.

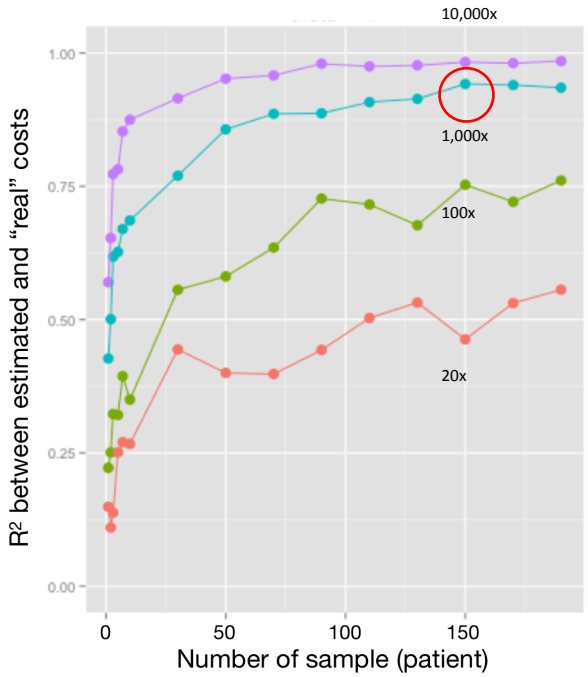
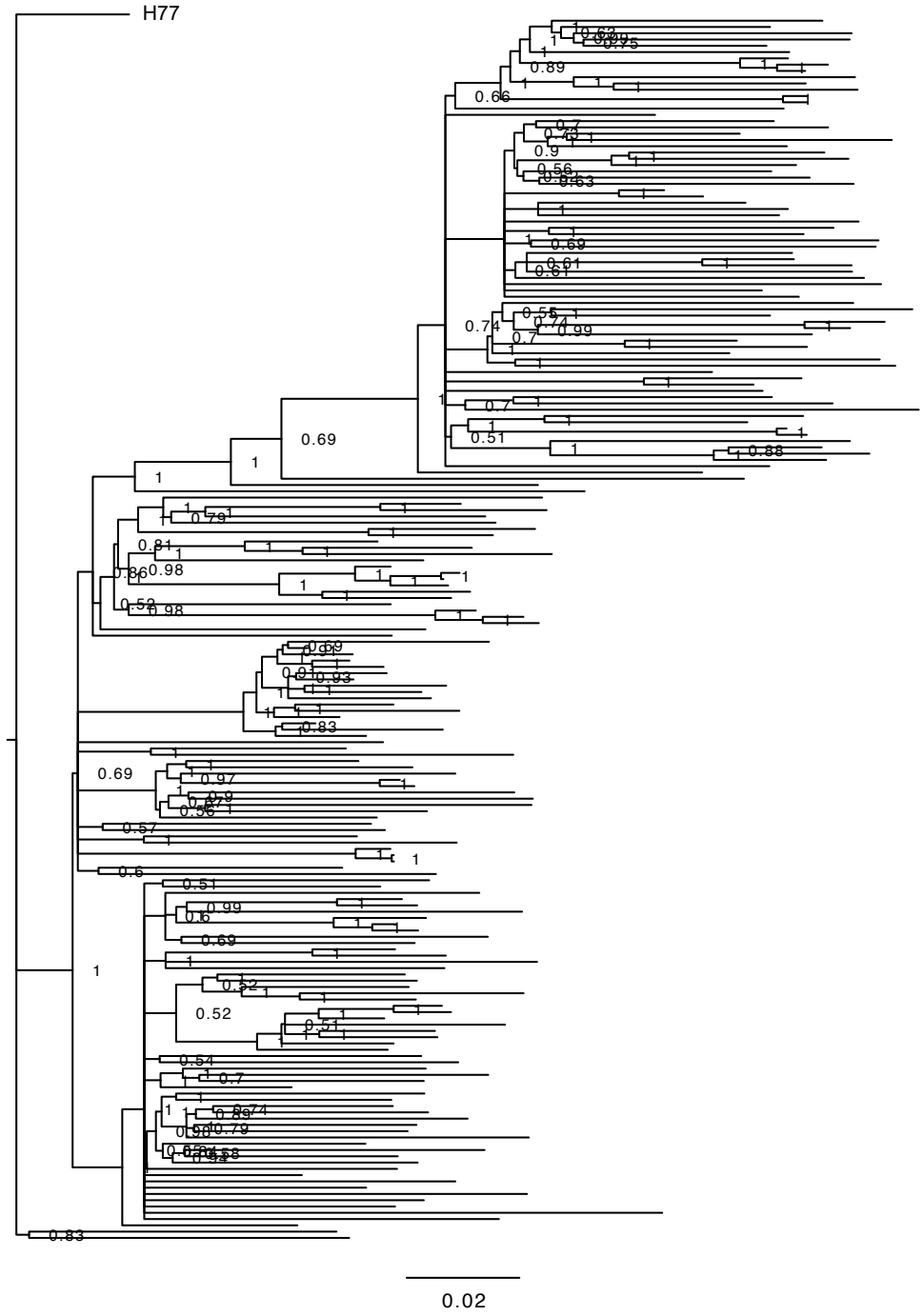
S6 Fig. Within-host (*in vivo*) *vs*. between-host nucleotide diversity comparison



S7 Fig. HCV mutation rates estimated from Gellers et al. (2016)'s *in vitro* dataset and our *in vivo* dataset, using nonsense mutation frequencies.



S8 Fig. Simulations results showing the relationship between number of samples, coverage, and accuracy of inference, as determined by the correlation coefficient between simulated fitness costs and their estimates based on a mean-field approach.



S9 Fig. Phylogenetic tree of HCV 1a used in this study. The numbers at the node represent the Bayesian posterior probability values.