1

2 **Title: Assessing *in vivo* mutation frequencies and creating a high-resolution genome-wide**

3 **map of fitness costs of Hepatitis C virus**

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

5

6

7 Kaho H. Tisthammer[1*], Caroline Solis[1], Faye Oracles[1], Madu Nzerem [1,2], Ryan Winstead[1],

8 Weiyan Dong[3], Jeffrey B. Joy[3,4,5], Pleuni S. Pennings[1]

9

10 [1] Department of Biology, San Francisco State University, San Francisco, California, United

11 States of America

12 [2] Current Address: New York University Langone Medical Center, New York, New York,

13 United States of America

14 [3] BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

15 [4] Division of Infectious Diseases, Department of Medicine, University of British Columbia,

16 Vancouver, BC, Canada

17 [5] Bioinformatics Programme, University of British Columbia, Vancouver, BC, Canada

18

19 * Corresponding author

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

21

Abstract

Like many viruses, Hepatitis C Virus (HCV) has a high nutation rate, which helps the virus adapt quickly, but mutations come with fitness costs. Fitness costs can be studied by different approaches, such as experimental or frequency-based approaches. The frequency-based approach is particularly useful to estimate *in vivo* fitness costs, but this approach works best with deep sequencing data from many hosts, are. In this study, we applied the frequency-based approach to a large dataset of 195 patients and estimated the fitness costs of mutations at 7957 sites along the HCV genome. We used beta regression and random forest models to better understand how different factors influenced fitness costs. Our results revealed that costs of nonsynonymous mutations were three times higher than those of synonymous mutations, and mutations at nucleotides A or T had higher costs than those at C or G. Genome location had a modest effect, with lower costs for mutations in HVR1 and higher costs for mutations in Core and NS5B. Resistance mutations were, on average, costlier than other mutations. Our results show that *in vivo* fitness costs of mutations 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.

Introduction

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

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

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

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

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

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

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

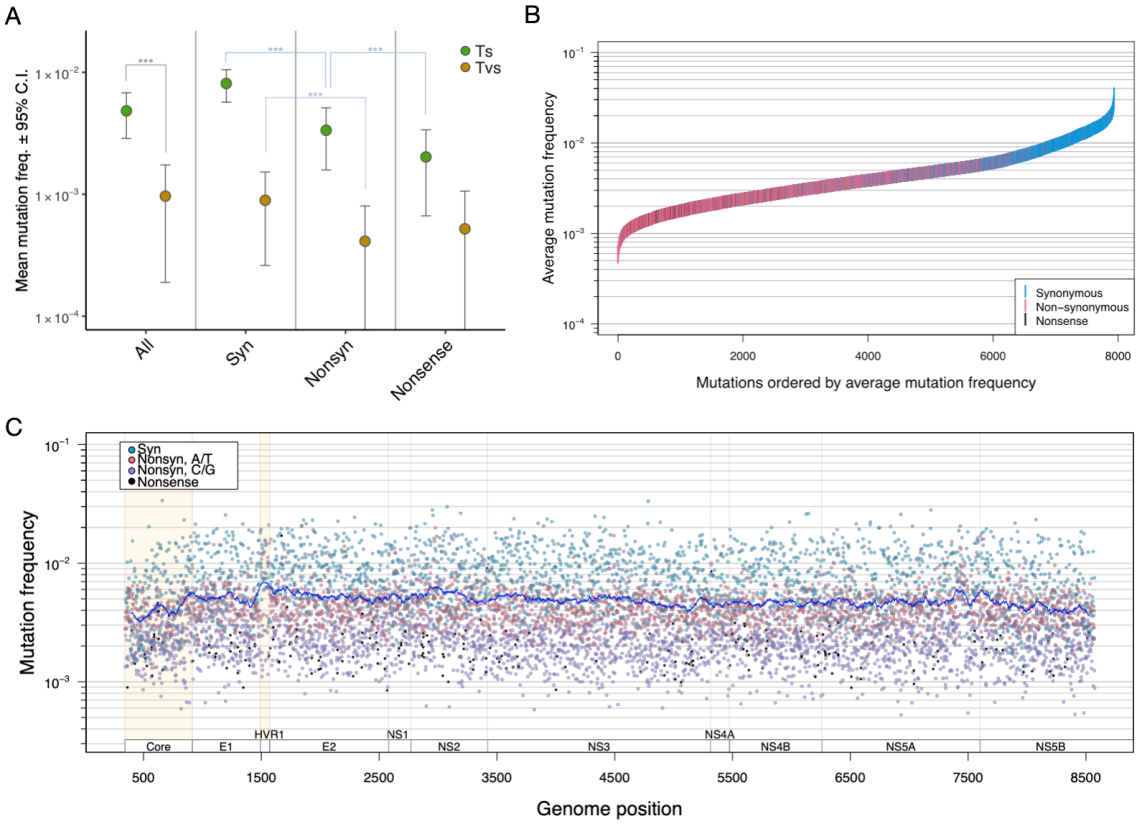
113

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 was 5.8 (± 1.3) x 10^{-3} per nucleotide, and transition mutations were around five times

120 more common than transversion mutations (4.8×10^{-3} *vs*. 9.7×10^{-4}) (Fig 1A and Table 1). For
121 both transitions and transversions, the average frequency of synonymous mutations (transition:
122 8.1×10^{-3}, transversion: 8.9×10^{-4}) was more than twice as high than that of nonsynonymous
123 mutations (transition: 3.3×10^{-3}, Mann-Whitney test, corrected $P = 0$, transversion: 4.1×10^{-4},
124 Mann-Whitney test, corrected $P = 4.7 \times 10^{-175}$). For transitions, the average frequency of
125 nonsense mutations (2.0×10^{-3}) was significantly lower than that of nonsynonymous mutations
126 (Mann-Whitney test, corrected $P = 3.9 \times 10^{-41}$), but this was not the case for transversions. This
127 suggests that the frequencies of nonsynonymous (4.1×10^{-4}) and nonsense (5.2×10^{-4})
128 transversions (*i.e*., approximately $4\sim5 \times 10^{-4}$) may reflect the lowest detection limit of our dataset
129 (see Methods), since nonsense mutations are expected to occur at a lower frequency than
130 nonsynonymous mutations. Since transitions mutations are observed at higher frequencies than
131 transversions (likely because of their higher mutation rate), they offer greater statistical power
132 and we decided to focus on transition mutations in all our subsequent analysis.
133



134

135 **Fig. 1. Estimated *in vivo* mutation frequencies in HCV1a, as determined by analysis of 195**

136 **viral populations**. (A) Average estimated frequencies for all, synonymous (Syn), non-

137 synonymous (Nonsyn), and nonsense mutations, stratified by transitions (Ts) *vs.* transversions

138 (Tvs). Transition mutation frequencies are much higher for every class of mutation. *** denotes

139 statistical significance with adjusted P-values <0.001. (B) Genome-wide transition mutation

140 frequencies, ordered by mutation frequency and colored by mutation type, show that

141 synonymous mutations are more common than non-synonymous mutations in the HCV genome.

142 (C) Transition mutation frequencies along the HCV genome, colored by mutation type, show that

143 average mutation frequency is roughly consistent across the genome, with synonymous

144 mutations more common than nonsynonymous mutations. Each dot represents the average

145 mutation frequency at a nucleotide position, across the 195 viral populations. The line represents
146 the sliding window average of 100 bases.
147
148 **Table 1. Genome-wide average estimated frequencies of different types of mutation in the**
149 **HCV1a genome.**

Mutation type	Average	95% C.I.
Transition	4.8×10^{-3}	$\pm 1.9 \times 10^{-3}$
Synonymous	8.1×10^{-3}	$\pm 2.4 \times 10^{-3}$
Nonsynonymous	3.3×10^{-3}	$\pm 1.7 \times 10^{-3}$
Nonsense	2.0×10^{-3}	$\pm 1.4 \times 10^{-3}$
Transversion	9.7×10^{-4}	$\pm 4.4 \times 10^{-4}$
Synonymous	8.9×10^{-4}	$\pm 6.3 \times 10^{-4}$
Nonsynonymous	4.1×10^{-4}	$\pm 3.9 \times 10^{-4}$
Nonsense	5.2×10^{-4}	$\pm 5.4 \times 10^{-4}$

150 C.I. = Estimated confidence interval.
151
152 Next, we looked for trends in transition mutation frequencies across the genome. As
153 expected, synonymous mutations generally had higher frequencies than nonsynonymous
154 mutations. Nonsense mutations on the other hand were not clearly separated from the
155 nonsynonymous mutations (Fig 1B). Throughout the genome, mutations at nucleotides A/T (*i.e.*
156 A→G / T→C) had higher frequencies on average than those at C/G (*i.e.* C→T / G→A) (Fig 1C).
157 The Hypervariable Region 1 (HVR1), the most variable section of the HCV genome [18,19], had
158 the highest estimated average mutation frequency of all HCV genes (7.52×10^{-3}, Mann-Whitney
159 test (against each gene), corrected $P = 1.1 \times 10^{-5}$ ~0.014, indicated in light yellow in Fig 1C,
160 Table), whereas the Core gene, the genome's most conserved gene [18,19], had the lowest
161 average mutation frequency (4.07×10^{-3}, statistically significant in 7 out of 10 genes, Mann-

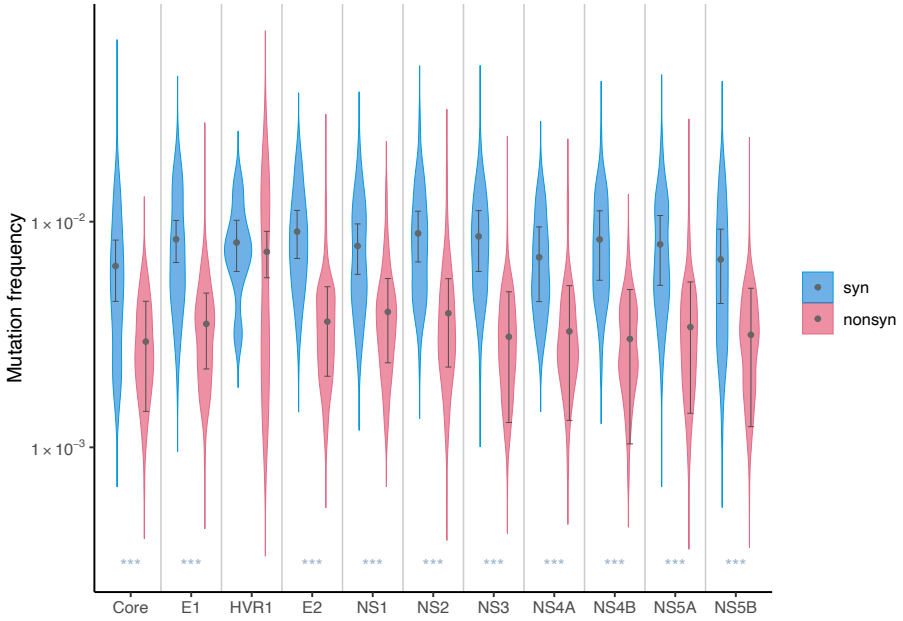
162 Whitney test, corrected $P = 1.1 \times 10^{-10} \sim 9.3 \times 10^{-4}$, indicated in light yellow in Fig 1C, also see

163 Fig 2). The high average mutation frequency of the HVR1 region is due to higher

164 nonsynonymous mutation frequencies than in other genes. Also, HVR1 was the only genic

165 region with nonsynonymous mutation frequencies as high as synonymous frequencies (Fig 2).

166



167

168 **Fig. 2. Estimated transition mutation frequencies of HCV by gene**. Aggregated observed frequencies

169 by gene and type of mutations; dots indicate the averages and the error bars represent the estimated 95%

170 C.I. from 195 samples. *** denotes statistically significant difference (adjusted P<0.0001) between

171 synonymous and nonsynonymous mutations by Mann-Whitney test.

172

173 **1.1. Beta regression and random forest models show large effects of nonsynonymous**

174 **mutations, nucleotide types, and genome location on mutation frequencies in HCV.** In order

175 to simultaneously assess the effects of multiple factors at the same time on HCV mutation

176 frequencies, such as mutations that create drastic amino acid change or CpG sites, we used 1) a

177 beta regression models since parameters of a beta regression are interpretable and they allow us
178 to quantitatively assess the effects of specific factors, and 2) a random forest regression since the
179 predictive machine learning (ML) approach allows us to investigate much larger numbers of
180 factors (i.e. 'features' in ML) simultaneously and a random forest can give the relative
181 importance of assessed features (but not the direction of the effects). A total of 83 features were
182 include in our random forest regression analysis (S2 Table).

183 For the beta regression, we specifically explored the effects of the following factors;
184 ancestral nucleotide, mutation type (synonymous/nonsynonymous mutations, CpG-creating
185 mutations, mutations that result in a drastic amino acid change), genomic location (*i.e.*, genes),
186 RNA structure, and the interactions between the ancestral nucleotides and the synonymous or
187 nonsynonymous mutation type on transition mutation frequencies. The results of the best-fit
188 model, based on AIC, contained 18 variables and are shown in Fig 3A and S3 Table. Of the 18
189 variables, 17 had significant effects on mutation frequencies. The ancestral nucleotide had a
190 large effect on mutation frequencies, with G→A and C→T mutations having 41~51 % lower
191 frequencies than A→G and T→C mutations ($P < 0.0001$) (Fig 3B). Nonsynonymous changes
192 were also associated with a large reduction in mutation frequency of over 53% ($P < 0.0001$)
193 relative to synonymous changes. Our model also suggested relatively modest effects of genomic
194 location when assessed by gene; the best-fit model included eight genes out of 11, of which
195 seven had significant effects. The largest locational effect was with the HVR1 region, which had
196 a frequency increase of 27 % ($P < 0.0001$), whereas being located in Core was associated with a
197 frequency reduction of 20% ($P < 0.0001$). CpG-creating mutations and mutations that cause a
198 drastic amino acid change were associated with modest reductions in mutation frequency of 7%
199 ($P < 0.0001$) and 12% ($P < 0.0001$), respectively. Mauger et al. [20] identified 15 regions that

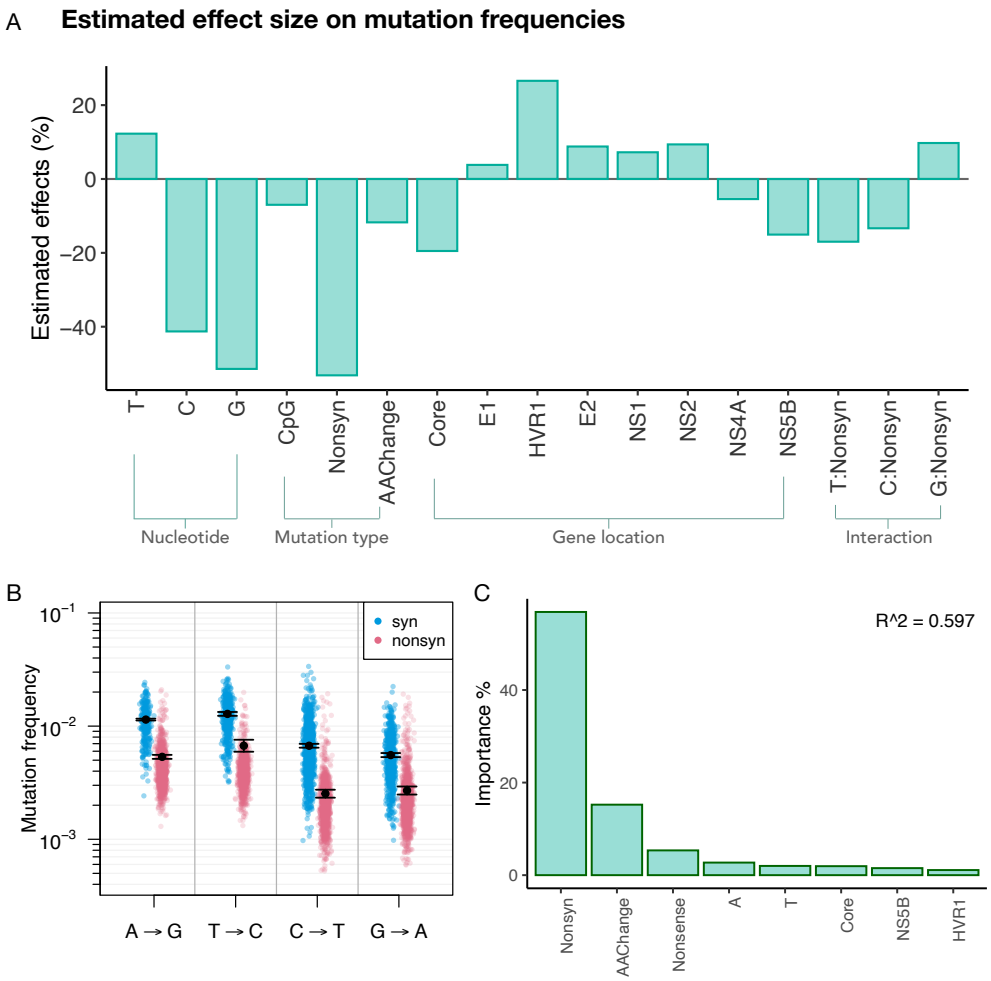
200 likely form secondary structures and are functionally conserved in HCV genotype 1a, 1b, and 2a.

201 We tested the aggregate effects of the 15 regions and found no significant effect on *in vivo*

202 mutation frequencies. RNA structure was associated with an increase of ~1.1% in mutation

203 frequency ($P = 0.39$) and was not selected in the best fit-model. The interactions assessed

204 showed modest effects sizes (-17 – 9.7%).



205

206 **Fig. 3. Various factors that affect the frequency of mutations in HCV, based on analysis of 195 viral**

207 **populations, each derived from a different patient.** (A) Predicted effects of ancestral nucleotide (T, C,

208 or G, *vs.* A); CpG-creating status (*vs.* non-CpG-creating status); nonsynonymous (Nonsyn; *vs.*

209 synonymous); amino acid-changing (AAChange; *vs.* non-amino acid-changing); presence in the Core,

210 HVR1, E2, NS1, NS2, NS4A, and NS5B genes (*vs.* the NS3/NS4B/NS5A regions) on mutation
211 frequencies in the HCV genome. Beta regression models were used to determine the effects of the
212 different factors on mutation frequencies across the genome, and this figure reflects the results of the best-
213 fit model based on AIC. (B) Estimated average transition mutation frequencies from the beta regression
214 model (black dots with standard errors) and the actual observed frequencies from 195 patients infected
215 with HCV (in colors). (C) Top 8 important features identified from the predictive random forest
216 regression model on mutation frequencies (for all features tested, see S1 Table).
217
218 The results of the random forest regression model were highly congruent with the beta
219 regression results, even though the predictability of our model reached a modest R^2 score of
220 0.597. The top features identified from the model were, in the order of importance,
221 nonsynonymous changes, drastic amino acid changes, nonsense mutations (stop codons),
222 ancestral nucleotide A and T, gene Core, NS5B and HVR1 in the order (Fig 3C, S1 Fig).
223 **1.2. *In vivo vs. in vitro* comparison reveals similar trends at the nucleotide-level, but**
224 **different genome-wide patterns.** We compared observed *in vivo* mutation frequencies with a
225 study of *in vitro* mutation frequencies [15] (for transition mutations). Patterns of *in vivo* mutation
226 frequencies were congruent with the trends observed in *in vitro* mutation frequencies [15] at the
227 nucleotide level, but differed at the gene level. The *in vivo* estimated mutation frequencies were
228 an order of magnitude higher than those found *in vitro.* Mutation frequencies for A→G and
229 T→C mutations were higher both *in vivo* and *in vitro* than C→T and G→A mutations (S2A Fig).
230 At the gene level, genome location was found to have very little effect (0.5%) on *in vitro*
231 mutation frequencies [15], while our regression model results suggest that seven out of 11 genes
232 have significant effects on *in vivo* mutation frequencies (Fig 3B and S1 Table). The difference
233 was most prominent in HVR1; the mutation frequency peak observed in HVR1 *in vivo* was not

234 seen *in vitro* (S2B Fig). The *in vitro* study [15] also found clusters of six highly mutable and 13
235 less mutable sites along the HCV genome, but in our *in vivo* study, we found no clear trends in
236 mutation frequencies at the corresponding sites (S3 Fig). Similarly, the *in vitro* study [15]
237 reported sites that had higher transversion than transition mutation frequencies (reversal sites) at
238 over 1000 sites (12.4% of the total sites), but far less such reversal sites were observed *in vivo*
239 (206 sites, 2.6%). The most salient reversal site *in vitro* (genome position C4080) again did not
240 show a higher frequency of transversion mutations than transition mutations *in vivo*.
241
242 **2. Genome wide fitness landscape of HCV**
243 **2.1 Estimated selection coefficients reveal high costs of A→G and T →C mutations,**
244 **nonsynonymous mutations, and mutations in NS5B.** Based on the observed mutation
245 frequencies, we next estimated selection coefficients (s) for each site in the HCV1a genome.
246 Selection coefficients can quantitatively reveal fitness consequences of new mutations. We
247 estimated s for each site based on the equation $f = \mu/s$ (see Methods), using the mutation
248 frequencies (f) and the HCV mutation rates (μ) calculated from the datasets of Geller et al. [15].
249 The genome-wide average selection coefficient for transition mutations in coding regions, using
250 the best available mutation rates [15], was 2.0 (± 0.7) x 10^{-3}. The average selection coefficient
251 for nonsynonymous mutations (2.5 ± 0.8 x 10^{-3}) was three times higher than that for synonymous
252 mutations (8.9 ± 4.8 x 10^{-4}) (Fig 4A).

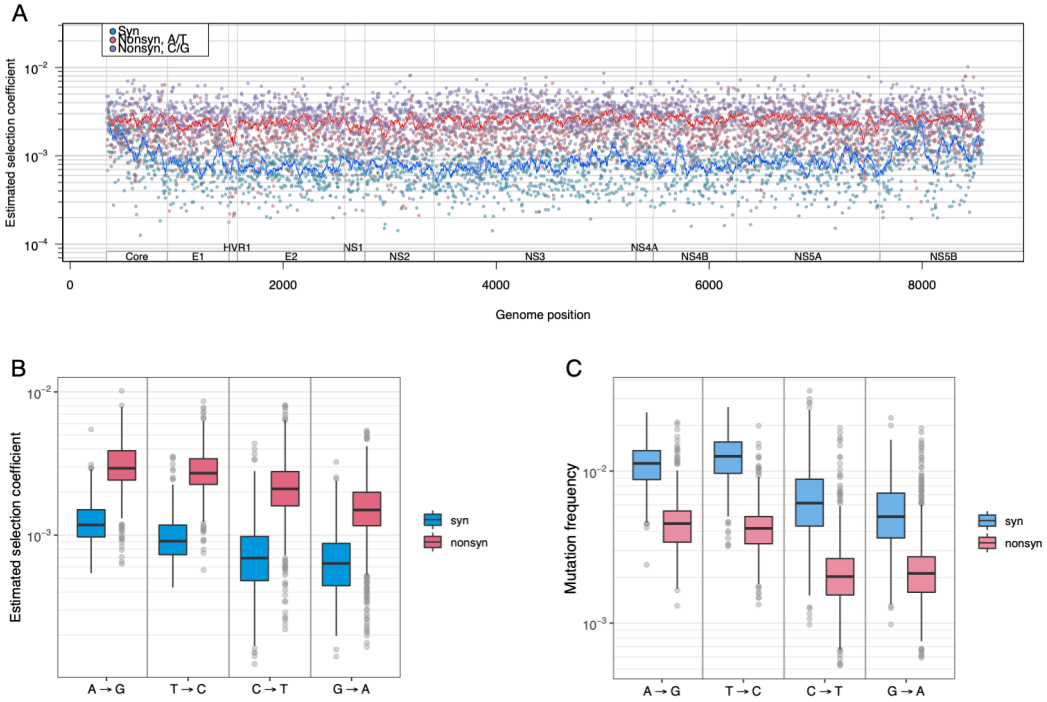


Fig. 4. Estimated genome-wide selection coefficients (fitness costs) in the HCV genome. (A) Selection coefficients (1/replication cycle) along the HCV genome, colored by mutation type (syn = synonymous; nonsynon = nonsynonymous); each dot represents the average at each position across 195 patient samples, and lines represent the sliding window average for 50 bases. (B) Selection coefficients (1/replication cycle) 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 at A and T sites than at C and G sites, even though mutation frequencies were higher at A and T sites compared to C and G sites.

Selection coefficients for each nucleotide also revealed interesting patterns. The selection coefficients (s) and mutation frequency (f) generally have an inverse relationship (equation: $f = \mu/s$); that is, higher f implies lower s (*i.e.*, a lower cost). However, if mutation rates (μ) at different nucleotides vary widely, the inverse relationship may not hold when assessed at a

268 nucleotide level, which was observed in our results. The average mutation frequency at
269 nucleotides A and T sites (A→G and T →C mutations) was higher than at C and G sites (C →T
270 and G →A mutations), leading one to expect that selection coefficients for A and T sites would
271 be lower than those for C and G sites. In reality, however, for A and T sites, average selection
272 coefficient values were higher despite the higher average mutation frequencies because the
273 mutation rates based on Geller's dataset [15] were so much higher (1.1-1.3 x 10^{-5} for A→G and
274 T →C *vs.* 3.2 -4.2 x 10^{-6} for C →T and G →A) (Fig 4B and C). The differences in selection
275 coefficients between nucleotides (excluding CpG-creating mutations) were all statistically
276 significant, with the values of selection coefficients in descending order being A→G (1.5 x 10^{-3})
277 > T→C (1.3 x 10^{-3}) > G→A (8.4 x 10^{-4}) > C→T (8.1 x 10^{-4}) (Mann-Whitney test, corrected *P*-
278 values = 0 – 3.5 x 10^{-6}). It is, however, important to note that the degree of accuracy in
279 estimating fitness costs (*s*) inevitably depends on the accuracy of estimated mutation rates.
280 Therefore, we confirmed the that A→G and T →C mutations have a higher mutation rate than C
281 →T and G →A mutations in the Geller data through a statistical test (Mann-Whitney test, all
282 corrected *P*-values = 0.0), as well as calculating the 95% confidence intervals (S4 Fig).
283 Similarly, when comparing average selection coefficients for sites in different genes, the
284 average selection coefficient for sites in NS5B (RNA dependent RNA polymerase) was the
285 highest among all genes (statistically significant in 7 out of 9 genes, S5A Fig, Mann-Whitney
286 test, corrected *P*-values = 1.2 x 10^{-11} - 0.077), despite the fact that Core's average mutation
287 frequency was the lowest among all genes (Table 2). This finding may reflect the higher genomic
288 AT content of NS5B relative to Core (S5 Fig), since higher AT content can lead to a higher
289 average selection coefficient due to A→G and T→C mutations having higher selection
290 coefficients. Therefore, we also quantified the effects of different factors on selection

291 coefficients using the two statistical models used for mutation frequencies: a beta regression and

292 a random forest model.

293 **Table 2. Average estimated selection coefficients per genic region of the HCV1a genome.**

Gene	*s*	95% C.I.
Core	2.11×10^{-3}	$\pm 1.1 \times 10^{-3}$
E1	1.85×10^{-3}	$\pm 8.2 \times 10^{-4}$
HVR1	1.39×10^{-3}	$\pm 1.0 \times 10^{-3}$
E2	1.85×10^{-3}	$\pm 9.0 \times 10^{-4}$
NS1	1.66×10^{-3}	$\pm 9.8 \times 10^{-4}$
NS2	1.78×10^{-3}	$\pm 1.1 \times 10^{-3}$
NS3	2.03×10^{-3}	$\pm 1.5 \times 10^{-3}$
NS4A	2.00×10^{-3}	$\pm 1.6 \times 10^{-3}$
NS4B	2.01×10^{-3}	$\pm 1.8 \times 10^{-3}$
NS5A	1.97×10^{-3}	$\pm 1.7 \times 10^{-3}$
NS5B	2.21×10^{-3}	$\pm 1.7 \times 10^{-3}$

294 *S* = Selection coefficient; C.I. = Estimated confidence interval.

295

296 **2.2. Beta regression and random forest models reveal nonsynonymous mutations, drastic**

297 **amino acid changes, ancestral nucleotides, and key gene regions as determinant factors of**

298 **fitness costs.** We first applied a beta regression in a similar manner as to mutation frequencies,

299 which followed highly similar trends as the mutation frequency results, except for the effects of

300 ancestral nucleotide C and G for the reason described in the previous paragraph (Fig 5A, S4

301 Table). Model results also indicated that the effects of Core alone on selection coefficients was

302 slightly larger than those of NS5B (+6.1% *vs*. +5.7%), likely reflecting the effects of higher AT

303 content. As expected, sites in HVR1, which had the highest *in vivo* mutation frequencies, had the

304 largest effect (-36.7%) out of all genes .

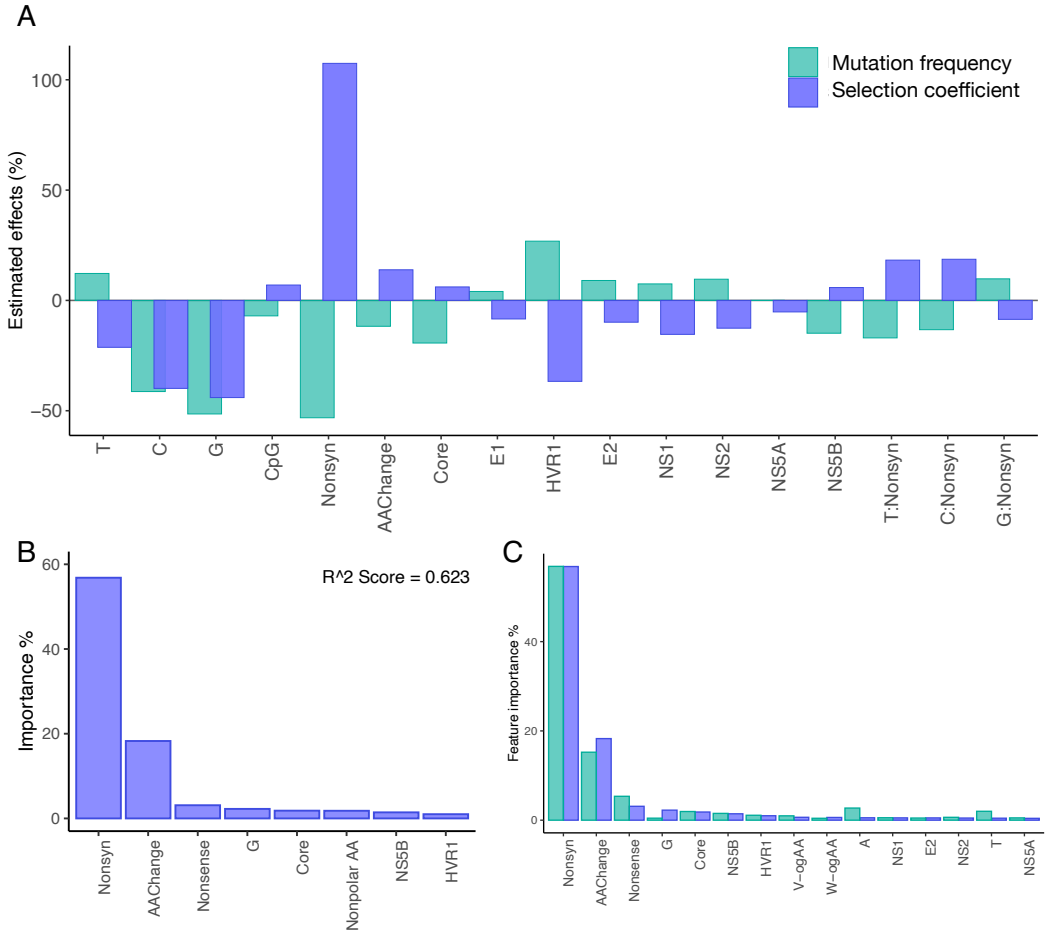


Fig. 5. Factors that affect the fitness cost (selection coefficient) in HCV, based on analysis of 195 viral populations, each derived from a different patient. (A) Predicted effects from beta regression models on selection coefficients (SC) in the HCV genome, shown together with predicted effects on mutation frequencies (See Fig. 3A for description of different factors in the figure.) (B) Top 8 important features identified from the predictive random forest regression model (for all features tested, see S1 Table). (C) Overlapping features ranked in the top 20 in the random forest regression models for mutation frequencies (MF) and selection coefficients (SC).

314 We also applied the RF regression models to further elucidate contributing factors for
315 selection coefficients in the same manner as for mutation frequencies. Features identified as
316 important for predicting selection coefficient values were in agreement with our beta regression
317 outcomes. The results identified 'nonsynonymous mutation' as the most determinant factor (Fig.
318 5B), followed by drastic amino acid changes, nonsense mutations, ancestral nucleotide G. Gene
319 locations of Core, NS5B and HVR1 were also in the top eight features of importance. We then
320 compared the features that ranked top 20 in both RF models for mutation frequency and selection
321 coefficients, which showed highly consistent results, revealing the three most important
322 determinants as nonsynonymous mutations, drastic amino acid changes and nonsense mutations
323 in both models (Fig 5C).

324 **2.3. Distribution of fitness effects and gamma parameter estimation.** Based on the estimated
325 fitness costs (*s*) for each nucleotide in our study, we assessed the distribution of fitness effects
326 (DFEs) of synonymous and nonsynonymous transition mutation for each nucleotide (S6 Fig). All
327 synonymous mutations and 99.7% of nonsynonymous mutations had fitness costs <0.01.
328 Parameters for the gamma distribution were also estimated for the entire DFEs. The scale (k) and
329 shape (θ) parameters were estimated to be 8.32×10^{-4} (95% confidence interval [CI] = 8.11 –
330 8.55×10^{-4}) and 2.378 (95% CI = 2.318 -2.436), respectively, which can be applied to future
331 population simulation studies of HCV.

332 **2.4. High fitness costs of drug resistance sites.** A number of direct-acting antiviral agents have
333 been developed for HCV, all of which specifically target the NS3, NS5A, or NS5B proteins.
334 Drug resistance often develops in patients, and many HCV resistance-associated variants have
335 been reported [21–26]. Prevalence of baseline (pre-existing) resistance-associated variants in
336 treatment naive patients (*i.e.* natural occurrence), as well as the fitness costs of these variants

337 provide insights into the evolution of drug resistance traits, treatment outcome prediction, and
338 durability of resistance phenotypes [27–29]. They can also help guide treatment options, such as
339 combining antivirals for which resistance mutations have high fitness costs to lower the risk of
340 developing drug resistance.

341 We assessed the natural occurrence of resistance-associated variants using‘our 'unfil’ered'
342 HCV datasets, which included sites with alternate majority nucleotides and transversion
343 mutations, as opposed to the filtered datasets used to estimate *in vivo* fitness costs, which
344 excluded the sites that were not in selection-mutation balance, including those with a nucleotide
345 different from the reference (H77, AF011753, see Methods). Of 87 known resistance-associated
346 variants across 26 sites we examined (S5 Table), natural occurrence was observed for all variants
347 in our datasets, including the 51 variants previously not observed or reported [23] (Fig 6). The
348 proportion of the samples with pre-existing resistance-associated variants varied from 3.6 to
349 99.5% (average 62%) depending on the resistance mutation (S7 Fig.). The result highlights that
350 many variants do occur naturally in treatment naive patients, making them more vulnerable to
351 developing drug resistance. Additionally, 20 of the 87 resistance-associated variants were
352 observed as a majority variant in one or more patients (S7 Fig , '% fixed sample'). The
353 percentages of patients that had the resistant variant as the majority vary from 0.5 % to 54.6%,
354 depending on the variants, with Q80K in NS3 being most likely to be present as the majority.
355 The average fitness costs of all resistance-associated variants examined was 4.7×10^{-3}, with $2.5 \times$
356 10^{-3} for transition mutations and 6.8×10^{-3} for transversion mutations. Estimated fitness costs of
357 resistance-associated variants (at transition sites only) were compared to those of non-resistance-
358 associated sites in the three genes, which revealed on average 21% higher fitness cost in
359 resistance-associated variants of HCV (Mann-Whitney test, $P = 0.0021$).
360

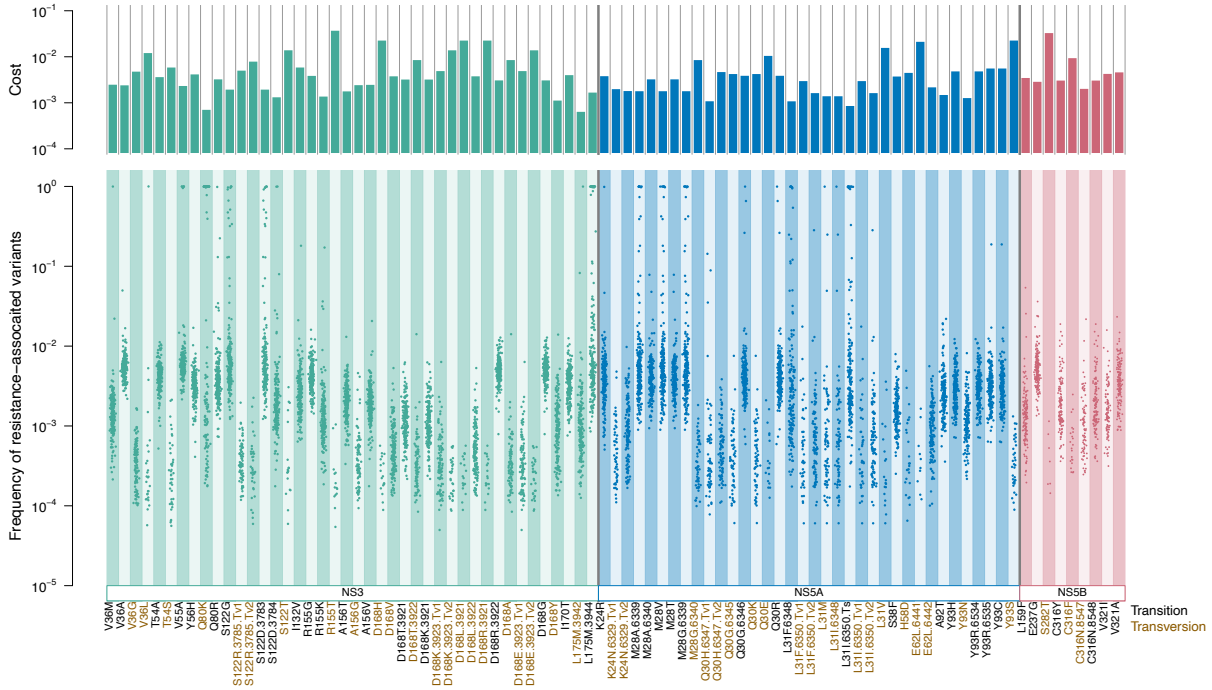


Fig. 6. Occurrence and fitness costs of resistance-associated variants in the HCV genome. Estimated fitness costs (top) and natural prevalence (bottom) of resistance-associated variants, where each dot represents a mutation frequency observed in a patient. Variant names in black are created by transition and those in brown are created by transversion mutations. RAVs that can be created by different mutations are specified in names, in a format of variant name, nucleotide position, and type of mutation (Tv1 stands for transition mutations that result in C or A and Tv2 stands for transition mutations that results in G or T).

3. Validation of our frequency-based approach

The frequency-based approach we used in this study assumes that site-level mutation frequencies are governed by consistent mutation rates and selection coefficients (mutation-selection equilibrium) across different viral populations (*i.e.*, patients). To confirm this assumption, we ran simulations to generate site-based mutation frequencies and compared these to observed frequencies. We also compared the patterns of *in vitro* mutation rates from the study of [1] with

376 the patterns of *in vivo* mutation rates estimated in our study. Additional analyses that we
377 performed to further to validate the frequent based approach are summarized in S2 Text (SI File).
378 **3.1. Simulated mutation frequencies based on the estimated selection coefficients are**
379 **consistent with observed mutation frequency patterns**. To confirm that the data are
380 consistent with HCV populations being in mutation-selection balance, we created simulated
381 datasets under the assumption of mutation-selection balance and found that simulated and
382 observed frequency distributions were very similar. Specifically, we ran simulations to generate
383 mutation frequencies using the nucleotide specific mutation rate estimates from [1], selection
384 coefficient estimates from our dataset for each site, and a within-host viral population size of
385 100,000 using code written previously [1]. Using these input parameters, for each site separately,
386 we simulated 195 viral populations (one for each patient). In these simulated populations, we
387 follow a population of 100,000 virus-infected cells. Each simulated generation, 100,000 new
388 cells are infected using a random sample of virus from the previous generation. Mutations occur
389 during replication, and virus with mutations has a slightly lower probability to infect new cells in
390 the next generation (quantified by the site-specific fitness cost). We then "harvest" samples from
391 these simulations of mutation-selection balance and determine how many cells carry mutated
392 virus, to calculate the mutation frequency for that site for that patient. For each of the 7957 sites,
393 we compare the simulated mutation frequencies with the observed mutation frequencies using a
394 Mann-Whitney test. We found that for 92.6 % of the sites the distribution of frequencies was not
395 significantly different than the simulated distribution of frequencies (*i.e.*, adjusted P-values >
396 0.05 by Mann-Whitney Test with Holm correction). The remaining 7.4 % of sites showed
397 significantly difference between observed and real data. This could be due to these sites not
398 being in mutation-selection equilibrium, due noise in the NGS data which is not reflected in the

399 simulated data, or due to epistasis or patient-specific fitness effects. This result demonstrates that
400 most of the data we used in this paper are consistent with a model of mutation-selection-
401 equilibrium (S8 Fig., S1 Datafile).

402 **3.2. Patterns of *in vitro* mutation rates correspond to those of *in vivo* estimates of mutation**
403 **rates.** The mutation rates used to obtain fitness costs in our study came from *in vitro* data [15],
404 since *in vivo* nucleotide-level mutation rates are not currently available. If *in vivo* and *in vitro*
405 mutation rates differ drastically, it would affect our estimations of fitness costs. Therefore, we
406 evaluated pseudo *in vivo* mutation rates using nonsense mutation frequencies observed in our
407 dataset, assuming that the fitness cost (s) of a nonsense mutation is one [6]. There are nine base
408 substitution combinations (out of 12) that can create stop codons; specifically, since there are no
409 Cs in stop codons, A→C, T→C, or G→C cannot be nonsense mutations. All but G→T nonsense
410 mutations were observed in our datasets. The pseudo mutation rates obtained from our *in vivo*
411 nonsense mutation frequencies followed a similar pattern as *in vitro* rates, with a significant
412 correlation between the two groups (Spearman's $\rho = 0.79$, $P = 0.028$) (S9 Fig).

413 **Discussion**

414 Genome-wide maps of fitness costs at each locus are key to understanding the
415 evolutionary dynamics of pathogenic viruses, yet few such maps exist. Here we used deep
416 sequencing data from 195 viral populations (*i.e.,* sequencing data from 195 infected hosts) and
417 applied the frequency-based approach to estimate *in vivo* fitness costs. We present a high-
418 resolution map of inferred *in vivo* fitness costs for transition mutations at 7957 sites in the
419 HCV1a genome. Our results revealed that whether a mutation is nonsynonymous, the ancestral
420 nucleotide, and genic location are the most important factors that determine fitness costs.

421 **Comparing to HIV**

422 Previous studies on fitness costs of mutations in HIV found similar trends in site-level
423 fitness costs, such as a clear difference in synonymous and nonsynonymous mutations and large
424 effects of the ancestral nucleotide [9,11]. However, there are also clear differences between HIV
425 and HCV. CpG-creating mutations and mutations that cause drastic amino acid changes have
426 large effects on fitness costs in the HIV pol gene, have markedly smaller effects in HCV,
427 suggesting that the magnitude of these effects likely vary between viruses. Specifically, CpG-
428 creating mutations are 1.5 times more costly than non-CpG creating mutations in HIV, but only
429 1.07 times more costly than non-CpG creating mutations in HCV. Many RNA and DNA viruses
430 display suppressed frequencies of CpG creating mutations, indicating that CpG creating
431 mutations are costly to a variety of viruses [17], and the magnitude of such effects appear to vary
432 greatly among viruses [16]. This may be due to different interactions of these viruses with our
433 immune systems or differences in their replication cycles and merits further study.

434 **Comparing *in vivo* mutation frequencies to *in vitro* mutation frequencies**

435 *In vitro* studies conducted in tightly controlled environments allow us to obtain detailed
436 information about viral evolutionary dynamics, including nucleotide level mutation rates.
437 However, the use of cell-lines in an environment without selective pressures from hosts creates
438 different dynamics from *in vivo* evolution. Understanding the difference and similarity between
439 *in vitro* and *in vivo* evolution can offer crucial insights into which forces are important for viral
440 evolution *in vivo*. Here we compare the *in vivo* mutation frequencies from our study with *in vitro*
441 mutation frequencies from the study of [1].

442 The *in vivo* mutation patterns from our study highlighted the selective pressures exerted
443 by the host immune system on different parts of the HCV genome, compared to those of *in vitro*.
444 Significant location effects were observed *in vivo,* but not *in vitro* [15]. First, the prominent peak

445 in mutation frequency in HVR1 was observed *in vivo*, while no such peak was observed *in vitro.*
446 HVR1 plays an important role in immune escape through positive selective pressure [18,30], and
447 we observed that HVR1 was the only genic region with similarly elevated nonsynonymous
448 mutation frequency relative to synonymous frequency (Fig 2B). In addition to the effect of
449 HVR1, we also observed location effects for seven other genic regions, which likely reflects
450 selective pressures from the host, which are not present in the *in vitro* study. Lastly, the sites with
451 high mutability identified by [1] *in vitro* did not show similar trends *in vivo*, suggesting that these
452 sites are also under strong negative selection *in vivo*.

453 We found several important similarities between the *in vivo* and *in vitro* settings. There
454 was no significant effects of RNA structure on either *in vivo* and *in vitro* mutation frequencies
455 [15]. We found that mutation frequency patterns at the nucleotide-level appear to be consistent
456 between *in vivo* and *in vitro*; A and T have higher mutation frequencies than G and C in both
457 cases [15]. Mutation rates for each nucleotide were also correlated between *in vivo* and *in vitro*
458 (*i.e.*, relative rates of different nucleotides were consistent), where we used observed frequencies
459 of nonsense mutations to estimate *in vivo* mutation rates (Fig S9). This is important since it
460 validates our application of *in vitro* mutation rates to estimate *in vivo* fitness costs. However, the
461 *in vivo* mutation rates were two orders of magnitude higher than those of *in vitro.* We suspect
462 that this is partly because of the costs of *in vivo* nonsense mutations are not likely to be equal to
463 one as we assume *in vivo*; defective viruses with premature stop codons can be rescued by other
464 defective or non-defective viruses [14,32–35] due to complex quasispecies dynamics in viruses
465 [36,37]. This phenomenon may allow nonsense mutations to be observed at higher-than-expected
466 frequencies *in vivo* viral populations. Also, how we calculated nonsense mutation frequencies
467 was not based on counting actual stop codons; we calculated the site-level nonsense mutation

468 frequencies, which would create stop codons if other two nucleotides in a codon were the same
469 as the reference. This would slightly overestimate the actual nonsense mutation frequencies.
470 Ribeiro et al. [14] estimated the *in vivo* mutation rate using a model and validated with nonsense
471 by counting all stop codons. Their estimation (2.5×10^{-5} mutations per nucleotide per replication
472 cycle) aligned with the *in vitro* mutation rate estimation [15], which may be due to their PCR-
473 based sequencing, which cannot capture low frequency mutations as NGS can and their stricter
474 method of calculating stop codons (Text S1 in [14]).

475 **Prevalence of drug resistance**

476 Natural occurrence of resistance associated variants was more frequently and widely
477 observed in our study than previously reported [23]. This highlights the importance of deep
478 sequencing, since many variants occur at a low frequency and would not have been captured by
479 traditional Sanger based methods. Although the fraction of patients in which the resistance-
480 associated variant was fixed was fairly low (average 5.7% of patients), one variant Q80K was
481 observed at an especially high prevalence (fixed in 54.6% of patients, the variant was observed
482 in 94.3% of patients, S7 Fig). This is consistent with previous observations, as Q80K is known to
483 exist at a high prevalence (up to 48%) in strains from North America [21,23,38]. The results
484 provide important insight into how common resistance associate variants are.

485 **Strengths and limitations of this study**

486 Deleterious mutations that occur at a low frequency indicate high fitness costs. However,
487 our results suggest that inferring nucleotide-level fitness costs from mutation frequencies should
488 be done cautiously: Lower mutation frequencies may not necessarily translate into higher fitness
489 costs, if substantial differences in mutation rates exist at the nucleotide level along with biased
490 base composition. In our dataset, mutations at A or T sites had higher fitness costs than

491 mutations at C or G sites, even though their mutation frequencies were significantly higher. A
492 similar trend was observed in a previous study on HIV's *pol* gene; *in vivo* mutation frequencies
493 were highest for G→A mutations even though fitness costs were also highest for G→A
494 mutations, due to the high mutation rate at these sites [9]. Thus, without nucleotide-level
495 mutation rates, differences in fitness frequencies at the nucleotide-level need to be interpreted
496 carefully.

497 Our unique approach worked well, thanks to the fact that the average read depth at each
498 site was > 6,200 after filtering, which resulted in an average of > 877,000 data points at each
499 nucleotide site analyzed. The correlation between *in vivo* and among-host genetic diversity at
500 each site supports our basic assumption that *in vivo* mutation patterns are largely dictated by
501 intrinsic factors at selection-mutation equilibrium. Adding > 400 genomes from NCBI had little
502 effect on the correlation between *in vivo* and among-host diversity, which suggests that we had
503 more than adequate sample size to capture the true genetic diversity of HCV1a. Additionally, we
504 have previously performed simulation studies to assess the optimal combination of sequencing
505 depth and sample size, which showed that > 150 samples with at least 1,000X sequencing depth
506 should yield reliable estimates (*i.e.*, a high correlation coefficient, > 0.95, between real and
507 estimated fitness costs, S10 Fig). Here, we had 195 samples with 6200X sequencing depth, more
508 than enough to produce reliable data.

509

510 **Conclusion**

511 Despite HCV's extremely high level of genetic variation, our detailed analysis of genome
512 wide mutation patterns revealed *in vivo* mutations do not occur in a completely stochastic
513 manner, illuminating the mutant spectrum complexity of HCV. Swarms of mutants, often termed

514 viral quasispecies [36,37], are continuously generated in RNA virus populations. The mutation
515 spectrum composition varies over time, producing transient equilibrium distributions that are
516 affected by internal and external factors [32]. However, even with these ever-changing mutant
517 spectra, our results showed that we could observe a situation close to mutation-selection
518 equilibrium. RNA viruses are considered to replicate at a mutagenesis level close to the lethal
519 threshold, at which the error rates exceed the tolerance limit of genetic information preservation
520 [39]. A delicate balance between this threshold and host selective pressures is likely responsible
521 for shaping the observed *in vivo* mutation patterns; approximately 50% of sites were highly
522 conserved while the remaining variable sites contribute to continuous generation of variation.
523 From a clinical perspective, one of the most challenging aspects of the treatment and
524 management of pathogenic RNA viruses is their high genetic diversity. Targeting the highly
525 conserved, high-cost sites or regions may help improve treatment efficacy by hampering the
526 development of drug resistance.

527 This study serves as an additional proof of concept for the frequency-based approach to
528 understanding *in vivo* fitness costs. This opens the door to opportunities to investigate *in vivo*
529 fitness costs of a wide variety of rapidly evolving pathogens, as deep sequencing of pathogen
530 populations becomes more common with the steep decline in next generation sequencing costs.
531 We showed that there are aspects of viral evolution that we cannot infer from *in vitro* studies, as
532 well as from the between-host level of analysis. Practical applications of better understanding *in*
533 *vivo* evolution and the fitness landscape of viruses include more targeted antiviral drug
534 development, efficient treatment management, better treatment outcome prediction, and
535 informative for vaccine design. By leveraging the technological advances and tremendous

536 resources already available, we have the potential to deepen our understanding of *in vivo*
537 evolution of viruses.
538
539 **Methods**
540 **Samples**
541 We obtained stored baseline plasma samples from a selection of HCV treatment naive
542 and viremic HCV/HIV co-infected participants from sites in the Canadian Co-infection Cohort
543 (CCC) [40]. Samples were collected between May 8, 2003 and November 20, 2014 from acute or
544 chronically infected patients attending 18 different clinics across Canada. We used 195 samples
545 identified as genotype 1a out of the dataset for our study.
546 **Viral RNA extraction and sequencing**
547 For each baseline sample HCV RNA was extracted from 500 ul of plasma using a
548 NucliSens easyMag (BioMerieux Inc.). Extracted RNA was eluted in 60ul of elution buffer and
549 stored at -20C prior to amplification. Near full length amplicons were generated using an oligo
550 d(A)20 primer for cDNA synthesis. Two nested PCR reactions using pangenotypic primers were
551 utilized to generate an 8,991-bp amplicon spanning the HCV Core to NS5B regions. Subsequent
552 to second round PCR amplification MiSeq sequencing libraries were generated using Nextera
553 XT DNA Sample Preparation Kits (Illumina, Inc.). Amplicons for each sample were tagged
554 using Nextera XT Index Kits (Illumina, Inc.), multiplexed and sequenced on an Illumina MiSeq
555 next-generation sequencing platform (Illumina, Inc.).
556 **Data filtering and formatting steps**
557 Sequence reads of virus isolates were quality filtered and trimmed using BBTools
558 v38.38.00, following its suggested steps [41]. First, the adapters were removed. To increase our

559 confidence in identifying minor variants, bases that had phred score <35 were all trimmed from
560 both ends. Contaminants (PhiX) were then filtered out and reads with a <35 average score were
561 also eliminated. Potential duplicates were identified and clumped, and filtered reads were
562 merged. To retain maximum read depths in order to confidently estimate minor variant
563 frequencies, both merged and unmerged reads were mapped to a reference sequence using BWA
564 0.7.17 [42]. Mapped reads were hard clipped, and the resulting SAM files were imported into R
565 and filtered further to eliminate potential contaminants and chimeric sequences. This was done
566 using a custom script in which all reads with a harmonic distance >10 from their own consensus
567 sequences were eliminated, as they likely represented contaminants or mis-mapped sequences.
568 This cut off value was determined empirically; histograms of read counts per harmonic distance
569 in each sample showed that the first peak mostly flattened out by 10, and the second peak, which
570 represents mis-mapping, did not appear until past 10. Filtered reads were remapped to their own
571 consensus sequences using BWA, and the resulting bam files were imported in R [43]. For each
572 patient sample (*i.e.*, viral population), frequency tables of nucleotides at each position were
573 created with Rsamtools and custom scripts in R.

574 Each frequency table was processed to calculate the frequencies of transition and
575 transversion mutations relative to the reference sequence (wildtype H77). Types of mutation
576 (synonymous/nonsynonymous, CpG-creating, those causing drastic amino acid changes, etc.)
577 were also identified for each site. To accurately estimate *in vivo* deleterious mutation frequencies
578 at the selection-mutation-drift equilibrium, we applied rigorous filtering, removing sites with a
579 mutation frequency >0.2 (*i.e.*, those likely under active selection), sites that had a different
580 majority nucleotide from the reference, and sites with <1000 read depth. Mutation frequencies of
581 all samples were then compiled into one table, and sites that were present in less than one-third

582 of the total sample (< 65) were removed. The filtering steps produced 7,958 usable sites in the
583 coding regions, and the averages of mutation frequencies were calculated for these sites. This
584 process yielded a table containing the average transition mutation frequencies at each position
585 across the HCV genome. To assess the natural occurrence of RAVs, the frequency tables were
586 processed similar to the previous step, but further filtering steps were not performed except for
587 removing sites with <1000 read depth. The 95% confidence intervals for mutation
588 frequencies/selection coefficients were estimated by calculating standard errors (SE, $\sqrt{\frac{p(1-p)}{n}}$) at
589 each site for each patient based on the number of reads (n), and then means of SEs for each site
590 were calculated across all patients to obtain the global estimates of standard errors at each site.
591 Using these estimated standard errors, upper and lower 95% confidence intervals were estimated
592 by $z_{\alpha=0.05}$(=1.96) x SE for each site. We used this estimation approach as aggregating at a patient
593 level was not optimal because of resulting in losing sample size and thus leading to
594 overestimation of standard errors [44].

595 **Data validation**

596 Since the sequencing error rate of the Illumina Miseq is reported to be around 0.1% [45],
597 we took additional steps to validate our filtering methods and the resulting datasets. We were
598 concerned that if our previous filtering steps did not remove the errors efficiently, mutations
599 observed at low frequencies (<0.001) could actually represent sequencing errors. To evaluate
600 whether this posed a problem, we replaced all mutation frequencies <0.001 with either NA or
601 zero and assessed the resulting mutation frequency patterns. The results showed consistent
602 patterns. Therefore, we used the original datasets for further analysis. Additionally, we have used
603 the homogeneous sample of RNA virus (SIV) from a different study to estimate the error rates
604 from our pipeline [46]. The results were consistent with published previous MiSeq error rates:

605 The average mutation frequency of the 'homogeneous viral sample' was 0.00156, and its median
606 was 0.00148. The error rate estimations confirmed that our approach can estimate mutation
607 frequency lower than these values since our transversion mutation frequencies were 0.00089,
608 clearly lower than 0.001, and biologically makes sense. We also assessed whether viral load
609 biased observed mutation frequencies by comparing the average mutation frequencies of patient
610 samples with low and high viral loads. We found no difference in the average mutation
611 frequencies between the two groups. Therefore, all samples were included in the analysis.

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

613 To gain a fine-grained understanding of the *in vivo* fitness effects of mutations in HCV
614 viruses, we used the *in vivo* frequency-based approach [9,11], in which fitness costs of mutations
615 are estimated from their frequencies across multiple natural populations (*i.e.* multiple patients).
616 This approach relies on translating *in vivo* mutation frequencies into fitness costs based on the
617 principle of mutation-selection balance. In infinitely large populations, the opposing forces of
618 mutation and selection cause costly mutations to be present at a constant frequency [47,48]. In
619 haploid systems, it is expressed as:

620 $$f = \mu/s \qquad (1)$$

621 where μ is the mutation rate from wildtype to mutant and s is the negative fitness effect of the
622 mutation (selection coefficient). In natural populations of finite size, the frequency of mutations
623 is not constant but fluctuates around the expected frequency of μ/s, because of the stochastic
624 nature of mutation and replication [49,32]. It is thus impossible to accurately infer the selection
625 acting on individual mutations from a single observation of a single population. We solved this
626 problem by averaging frequencies of each individual mutation from multiple populations
627 (sampling vertically, rather than horizontally, S11 Fig.). Note that the principle of mutation-

628 selection balance holds even when bottlenecks or selective sweeps occur, because such events do
629 not change the expected frequencies of mutations. Also, this method allows us to directly assess
630 fitness costs at all sites in the genome and to calculate selection coefficients rather than relative
631 rates [9].

632 Selection coefficients were estimated for each mutation at each nucleotide based on the
633 equation (1) using estimated *in vivo* mutation frequencies and mutation rates calculated from raw
634 datasets of [15] in R, using the same method Geller et al. [15] used to calculate the global
635 mutation rate. Briefly, mutation rates for each nucleotide-level mutation were obtained by
636 dividing the total mutations observed during the experimental period by the generation cycle
637 with adjustment for errors. The nucleotide-level mutation rates are the keys for estimating
638 selection coefficients using the frequency-based approach. Therefore, we assessed the validity of
639 *in vitro* mutation rates from Geller et al. [15] using beta regression (see below), and found no
640 major effects of selective pressures seen in *in vivo* data (see S2 Text SI File). *In vitro* mutation
641 frequencies used in comparison were also derived from the datasets of [15].

642 Statistical analyses were performed in R and Python. Beta regression model fitting was
643 conducted using the R package *betareg* [50]: For predictor variables, we included ancestral
644 nucleotide, types of mutations (synonymous vs. nonsynonymous, create a CpG site or not, cause
645 a drastic amino acid change or not), location (11 genic regions), RNA structure (the aggregate
646 effects of 15 regions that likely form secondary structure, identified by [20]), and interactions.
647 AIC values were used as a criterion to select the best fit model. Parameters for DFEs were
648 estimated using the R package noplr [51]. To assess the diversity between *in vivo* and among
649 hosts, additional HCV1a genomes were downloaded from NCBI GenBank. Random forest
650 regression analyses were conducted in Python 3.9 (Python Software Foundation,

651 https://www.python.org/) using Scikit-learn [52] module. After constructing a matrix of all
652 features to be used in RM (83 features including mutation types, nucleotides, amino acids, types
653 of amino acids, location, and RNA structure, see S3 Table), highly correlated features were
654 removed, and hyperparameter tuning was performed using the GridSearchCV and
655 RandomizedSearchCV functions in Scikit-learn. Models fitted with the best hyperparameters
656 reached R^2 of 0.597 for mutation frequencies and 0.623 for selection coefficients.

657 Simulation of site-specific mutation frequencies were performed using the nucleotide
658 specific mutation rates estimated from [1], site-specific selection coefficients estimated in our
659 current study, and a population size of Ne = 100,000 to generate mutation frequencies for the
660 total number of samples available for the particular site (ranged from 65 to 189). The simulation
661 program is compiled in C++ and available at the GitHub repository, along with all coding scripts
662 (https://github.com/kahot/HCV_FitnessCost). We then sampled from these simulated
663 populations using on the average read depths of each patient. The resulting simulated sample
664 frequencies and the observed frequencies were statistically compared using the Mann-Whitney
665 test with the Holm correction.

666

667 **Acknowledgments**

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

671

672 **References**

673 1. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the Evolution of Human
674 Influenza A. Science. 1999;286: 1921–1925.

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

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

680 4. Crotty S, Cameron CE, Andino R. RNA virus error catastrophe: Direct molecular test by
681 using ribavirin. Proceedings of the National Academy of Sciences. 2001;98: 6895–6900.
682 doi:10.1073/pnas.111085598

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

685 6. Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus
686 revealed through population sequencing. 2014; 686–90.
687 doi:http://dx.doi.org/10.1038/nature12861

688 7. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci. 2016;73:
689 4433–4448. doi:10.1007/s00018-016-2299-6

690 8. Trouplin V, Salvatori F, Cappello F, Obry V, Brelot A, Heveker N, et al. Determination of
691 coreceptor usage of human immunodeficiency virus type 1 from patient plasma samples by
692 using a recombinant phenotypic assay. J Virol. 2001;75: 251–259.
693 doi:10.1128/JVI.75.1.251-259.2001

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

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

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

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

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

708 14. Ribeiro RM, Li H, Wang S, Stoddard MB, Learn GH, Korber BT, et al. Quantifying the
709 Diversification of Hepatitis C Virus (HCV) during Primary Infection: Estimates of the In

710 Vivo Mutation Rate. Wilke CO, editor. PLoS Pathog. 2012;8: e1002881-13.
711 doi:10.1371/journal.ppat.1002881

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

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

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

721 18. Kato N. Genome of Human Hepatitis C Virus (HCV): Gene organization, sequence
722 diversity, and variation. Microbial Comparative Genomics. 2000;5: 129–151.

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

726 20. Mauger DM, Golden M, Yamane D, Williford S, Lemon SM, Martin DP, et al. Functionally
727 conserved architecture of hepatitis C virus RNA genomes. Proc Natl Acad Sci USA.
728 2015;112: 3962–3697. doi:10.1073/pnas.1416266112

729 21. Poveda E, Wyles DL, Mena Á, Pedreira JD, Castro-Iglesias Á, Cachay E. Update on
730 hepatitis C virus resistance to direct-acting antiviral agents. Antiviral Research. 2014;108:
731 181–191. doi:10.1016/j.antiviral.2014.05.015

732 22. Lontok E, Harrington P, Howe A, Kieffer T, Lennerstrand J, Lenz O, et al. Hepatitis C virus
733 drug resistance-associated substitutions: State of the art summary. Hepatology. 2015;62:
734 1623–1632. doi:10.1002/hep.27934

735 23. Sarrazin C. The importance of resistance to direct antiviral drugs in HCV infection in
736 clinical practice. 2015; 1–19. doi:10.1016/j.jhep.2015.09.011

737 24. Kim S, Han K-H, Ahn SH. Hepatitis C Virus and Antiviral Drug Resistance. Gut and Liver.
738 2016;10: 890–895. doi:10.5009/gnl15573

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

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

744 27. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP
745 Discovery and Genetic Mapping Using Sequenced RAD Markers. Fay JC, editor. PLoS
746 ONE. 2008;3: e3376. doi:10.1371/journal.pone.0003376.g003

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

751 29. Echeverría N, Betancour G, Gámbaro F, Hernández N, López P, Chiodi D, et al. Naturally
752 occurring NS3 resistance-associated variants in hepatitis C virus genotype 1: Their
753 relevance for developing countries. Virus Research. 2016;223: 140–146.
754 doi:10.1016/j.virusres.2016.07.008

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

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

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

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

767 34. Vignuzzi M, López CB. Defective viral genomes are key drivers of the virus-host
768 interaction. Nat Microbiol. 2019;4: 1075–1087. doi:10.1038/s41564-019-0465-y

769 35. Andino R, Domingo E. Viral quasispecies. Virology. 2015;479–480: 46–51.
770 doi:10.1016/j.virol.2015.03.022

771 36. Domingo E, Perales C. Viral quasispecies. PLoS Genet. 2019;15: e1008271.
772 doi:10.1371/journal.pgen.1008271

773 37. McCloskey RM, Liang RH, Joy JB, Krajden M, Montaner JSG, Harrigan PR, et al. Global
774 Origin and Transmission of Hepatitis C Virus Nonstructural Protein 3 Q80K
775 Polymorphism. The Journal of Infectious Diseases. 2014;211: 1288–1295.
776 doi:10.1093/infdis/jiu613

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

39. Klein MB, Saeed S, Yang H, Cohen J, Conway B, Cooper C, et al. Cohort profile: the Canadian HIV-hepatitis C co-infection cohort study. Int J Epidemiol. 2010;39: 1162–1169. doi:10.1093/ije/dyp297

40. Bushnell B. BBTools. 2018.

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

42. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021.

43. McCarthy CJ, Whittaker TA, Boyle LH, Eyal M. Quantitative Approaches to Group Research: Suggestions for Best Practices. J SPEC GROUP WORK. 2017;42: 3–16. doi:10.1080/01933922.2016.1264520

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

45. Tisthammer KH, Kline C, Rutledge T, Diedrich CR, Ita S, Lin PL, et al. SIV evolutionary dynamics in cynomolgus macaques during SIV-Mycobacterium tuberculosis co-infection. Viruses. 2021;In press.

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

47. Haldane J. The effect of variation on fitness. The American Naturalist. 1937;71: 337–349.

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

49. Cribari-Neto F, Zeileis A. Beta Regression in R. Journal of Statistical Software. 2010;34: 1–24. doi:10.18637/jss.v034.i02

50. Ypma J, Borchers HW, Eddelbuettel D. Package 'nloptr.' 2018 Oct pp. 1–46. Available: https://cran.r-project.org/web/packages/nloptr/nloptr.pdf

51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12: 2825–2830.

Supporting Information

S1 Fig. Top 20 features identified as important from the random forest regression model for predicting *in vivo* HCV mutation frequencies.

810 **S2 Fig. Comparison of *in vivo* and *in vitro* transition mutation frequencies in the HCV1a**
811 **genome.**
812
813 **S3 Fig. Transition mutation frequencies of the sites corresponding to the highly mutable**
814 **(HM) and less mutable (LM) sites identified in Geller et al. (2016)**.
815
816 **S4 Fig. Mutation rates estimated from the dataset obtained from Geller et al. (2016)**.
817
818 **S5 Fig. Average selection coefficients (A) and AT contents (B) of each genic region in the**
819 **HCV genome.**
820
821 **S6 Fig. Distribution of fitness effects for each nucleotide (A, T, C, G) in the coding regions**
822 **of the HCV genome**, **stratified by synonymous and nonsynonymous mutation status, using**
823 **the estimated fitness costs (s) calculated in this study.**
824
825 **S7 Fig. The percentage of 195 HCV patient samples with resistance-associated variants**
826 **(RAVs) observed (Top), and the percentage of patients that had given RAVs as majority**
827 **nucleotide ('fixed sample') (Bottom).**
828
829 **S8 Fig. Comparison of simulated and observed *in vivo* mutation frequencies of HCV from**
830 **195 patients.**
831
832 **S9 Fig. Comparisons between HCV mutation rates estimated from Gellers et al. (2016)'s *in***

833 ***vitro* dataset and our *in vivo* dataset using nonsense mutation frequencies (A).** Correlation

834 between estimated in vivo and in vitro mutation rates are shown in B.

835

836 **S10 Fig. Simulations results showing the relationship between number of samples,**

837 **coverage, and accuracy of inference, as determined by the correlation coefficient between**

838 **simulated fitness costs and their estimates based on a mean-field approach.**

839

840 **S11 Fig. Diagram showing our approach to observe *in vivo* evolution of viruses.**
841

842 **S12 Fig. Phylogenetic tree of HCV 1a used in this study.**

843

844 **S1 Table. Results of statistical tests comparing synonymous and non-synonymous mutation**

845 **frequencies within each gene.**

846

847 **S2 Table. List of features included in random forest classification and regression models**

848 **run on mutation frequencies and selection coefficients.**

849

850 **S3 Table. The results of beta regression with the estimated effects of different factors on**

851 **transition mutation frequencies.**

852

853 **S4 Table. The results of beta regression with the estimated effects of different factors on**

854 **inferred selection coefficients.**

855

856 **S5 Table. Resistance-associated variants in the NS3, NS5A, and NS5B genes examined in**

857 **our study.**

858

859 **S1 Text (SI File). Validation of our frequency based methods: Between-host variability is**

860 **consistent with within-host variability in HCV.**

861

862 **S2 Text (SI File). Supplementary Methods: Validating the estimated mutation rates.**
863

864

865 **Data availability**

866 Processed bam files of HCV from each patient are available at

867 https://figshare.com/articles/dataset/HCV_Bam_Files/13239491?file=25498772. All other

868 relevant data are within the manuscript and its Supporting Information files.