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 Dong², Jeffrey B. Joy^{2,3,4}, Pleuni S. Pennings¹

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.

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

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

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

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

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

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

Results

1. Mutation frequencies within the HCV genome

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

was more than twice as high than that of nonsynonymous mutations (transition: 3.3×10^{-3} , transversion: 4.1×10^{-4} , Wilcoxon test, $P = 0 - 2.6 \times 10^{-185}$). For transitions, the average frequency of nonsense mutations (2.0×10^{-3}) was significantly lower than that of nonsynonymous mutations (Wilcoxon test, $P = 1.3 \times 10^{-30}$), but this was not the case for transversions. This suggests that the frequencies around nonsynonymous (4.1×10^{-4}) and nonsense (5.2×10^{-4}) transversions (*i.e.* approximately $4\sim 5 \times 10^{-4}$) may have been at the lower detection limit of our dataset, since nonsense mutations are expected to occur at a lower frequency than nonsynonymous mutations. Since transitions are more frequent than transversions, and thus offer greater statistical power, we focused on transition mutations in our subsequent analysis.

Table 1. Genome-wide average estimated frequencies of different types of mutation in the 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	4.1×10^{-4}	2.2×10^{-5}
Nonsense	5.2×10^{-4}	2.5×10^{-5}

S.E., standard error.

We looked for trends in mutation frequencies across the genome based on transition mutations. Synonymous mutations generally had higher frequencies than nonsynonymous

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

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

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

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

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

2. Genome wide fitness landscape of HCV

2.1 Estimated selection coefficients reveal high costs of A→G and T→C mutations,

nonsynonymous mutations, and mutations in NS5B. Based on the observed mutation frequencies, we next estimated selection coefficients (s) for each site in the HCV1a genome. Selection coefficients can quantitatively reveal fitness consequences of new mutations. We estimated s for each site based on the equation $f = \mu/s$ (see Methods), using the mutation frequencies (f) and the HCV mutation rates (μ) calculated from the datasets of Geller et al. [15]. The genome-wide average selection coefficient for transition mutations in coding regions, using the best available mutation rates [15], was $2.0 (\pm 0.5) \times 10^{-3}$. The average selection coefficient for nonsynonymous mutations ($2.5 \pm 0.7 \times 10^{-3}$) was three times higher than that for synonymous mutations ($8.9 \pm 6.0 \times 10^{-4}$) (Fig 4A).

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). If mutation rates (μ) at different nucleotides vary widely, the inverse relationship may not hold when assessed at a nucleotide level, which was observed in our results. The average mutation frequency at nucleotides A/T (A→G/T→C mutations) was higher than at C/G (C→T/G→A mutations), leading one to

expect that selection coefficients for A/T would be lower than those for C/G sites. In reality, however, for A/T sites, average selection coefficient values were higher despite the higher average mutation frequencies because the mutation rates based on Geller's dataset [15] were so much higher ($1.1-1.3 \times 10^{-5}$ for A→G / T→C vs. $3.2-4.2 \times 10^{-6}$ for C→T/G→A) (Fig 4B and C). The differences in selection coefficients between nucleotides (excluding CpG-creating mutations) were all statistically significant, with the values of selection coefficients in descending order being A→G (1.5×10^{-3}) > T→C (1.3×10^{-3}) > G→A (8.4×10^{-4}) > C→T (8.1×10^{-4}) (Wilcoxon test, $P = 0 - 7.1 \times 10^{-7}$). It is, however, important to note that the degree of accuracy in estimating fitness costs (s) inevitably depends on the accuracy of estimated mutation rates. Therefore, we confirmed the higher mutation rates for A→G / T→C mutations than for C→T/G→A mutations by testing the differences (Wilcoxon test, all P -values = 0.0), as well as calculating the 95% confidence intervals (S3 Fig).

Similarly, when comparing average selection coefficients for sites in different genes. The average selection coefficient for sites in NS5B (RNA dependent RNA reverse transcriptase) was significantly higher than the average selection coefficients of any other genes (S3A Fig, Wilcoxon Test, $P = 1.1 \times 10^{-13} - 0.039$), including sites in Core, despite the fact that Core's average mutation frequency was the lowest among HCV genes (Table 2). This finding may reflect the higher genomic AT content of NS5B relative to Core (S3 Fig), since higher AT content can lead to a higher average selection coefficient due to A→G / T→C mutations having higher selection coefficients. We, therefore, also quantified the effects of different factors on selection coefficients using beta regression models, which followed highly similar trends as the mutation frequency results, except for the effects of ancestral nucleotide C and G for the reason described in the previous paragraph (S2 Table and S4 Fig). Model results also indicated that the

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

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}

s , selection coefficient; S.E., standard error.

2.2. Distribution of fitness effects and gamma parameter estimation. Based on the estimated fitness costs (s) for each nucleotide in our study, we assessed the distribution of fitness effects (DFEs) of synonymous and nonsynonymous transition mutation for each nucleotide (Fig 5). All synonymous mutations and 99.7% of nonsynonymous mutations had fitness costs <0.01 . Parameters for the gamma distribution were also estimated for the entire DFEs. The scale (k) and shape (θ) parameters were estimated to be 8.32×10^{-4} (95% confidence interval [CI] = $8.11 - 8.55 \times 10^{-4}$) and 2.378 (95% CI = 2.318 - 2.436), respectively, which can be applied to future population simulation studies of HCV.

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

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

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

3. Validation of our frequency-based approach

3.1 Between-host variability is consistent with within-host variability in HCV. The

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

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

Next, we investigated potential effects of our filtering process, where we excluded the sites that had a majority nucleotide different from the reference nucleotide (H77) from each viral population (see Methods). If too many sites were excluded, and if these sites had high variability, this could mask within-host variability. We first calculated the proportion of excluded sites due to the nucleotide difference for each viral population, which was on average, only 6.6 (± 0.1) %. We also compared the consensus (majority nucleotide) sequences of all viral populations to gain further insights into the genetic variability between hosts. The overall sequence similarity (the average pairwise identity of consensus sequences) was relatively high (91.8%), and was within the range of the expected values ($> 90\%$ within a subtype) [17,29,30]. These results indicate that the number of sites excluded from the analysis was quite small, and the filtering step probably did not bias our results.

Lastly, we investigated if a group of highly conserved sites between-host also corresponded to highly conserved sites within-host. Alignments of all consensus sequences across the 195 genomes showed 49% of sites to be identical within the HCV coding regions. These identical sites, or highly conserved sites, occurred relatively evenly throughout the genome, except for in the HVR1 region, which had zero identical sites across the populations. Relative to non-identical sites, identical sites had a significantly lower (-5%) average mutation frequency (4.69×10^{-3} vs. 4.93×10^{-3} , Wilcoxon test, $P = 0.0003$) and a significantly higher (+4%) average selection coefficient (2.68×10^{-3} vs. 2.58×10^{-3} , Wilcoxon test, $P = 0.032$), confirming the correspondence between conserved sites between-host and within-host.

3.2. In vitro mutation rates correspond to in vivo estimates of mutation rates. The mutation rates used to obtain fitness costs in our study came from *in vitro* data [15], since *in vivo* nucleotide-level mutation rates are not currently available. If *in vivo* and *in vitro* mutation rates

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

Discussion

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

Comparing to HIV

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

Comparing to *in vitro* study

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

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

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

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

Prevalence of drug resistance

Natural prevalence of resistance associated variants was more frequently and widely

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

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

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

Strengths and limitations of this study

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

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

estimated fitness costs) (S8 Fig). Here, we had 195 samples with 6200X sequencing depth, more than enough to produce reliable data.

Conclusion

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

Methods

Samples

We obtained stored baseline plasma samples from a selection (n = 406) of HCV treatment naïve and viremic HCV/HIV co-infected participants from sites in the Canadian Co-infection Cohort (CCC) [37]. Samples were collected between May 8, 2003 and November 20, 2014 from acute or chronically infected patients attending 18 different clinics across Canada.

Viral RNA extraction and sequencing

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

Data filtering and formatting steps

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

created with Rsamtools and custom scripts in R.

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

Data validation

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

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

Calculation of selection coefficients and other statistics

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

$$f = \mu/s \quad (1)$$

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

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

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

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

Acknowledgments

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

References

1. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of 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

9. Theys K, Feder AF, Gelbart M, Hartl M, Stern A, Pennings PS. Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. Bloom JD, editor. PLoS Genet. 2018;14: e1007420–24. doi:10.1371/journal.pgen.1007420
10. Pennings PS, Kryazhimskiy S, Wakeley J. Loss and Recovery of Genetic Diversity in Adapting Populations of HIV. Fraser C, editor. PLoS Genet. 2014;10: e1004000–9. doi:10.1371/journal.pgen.1004000
11. Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evolution. 2017;3: 10312.–12. doi:10.1093/ve/vex003
12. Kautz T, Forrester N. RNA Virus Fidelity Mutants: A Useful Tool for Evolutionary Biology or a Complex Challenge? Viruses. 2018;10: 600–17. doi:10.3390/v10110600
13. Cuevas JM, González-Candelas F, Moya A, Sanjuán R. Effect of Ribavirin on the Mutation Rate and Spectrum of Hepatitis C Virus In Vivo. Journal of Virology. 2009;83: 5760–5764. doi:10.1128/JVI.00201-09
14. Ribeiro RM, Li H, Wang S, Stoddard MB, Learn GH, Korber BT, et al. Quantifying the Diversification of Hepatitis C Virus (HCV) during Primary Infection: Estimates of the In Vivo Mutation Rate. Wilke CO, editor. PLoS Pathog. 2012;8: e1002881–13. doi:10.1371/journal.ppat.1002881
15. Geller R, Estada Ú, Peris JB, Andreu I, Bou J-V, Garijo R, et al. Highly heterogeneous mutation rates in the hepatitis C virus genome. Nature Publishing Group; 2016;1: 16045. 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 organization, sequence
593 diversity, 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

23. Sarrazin C. The importance of resistance to direct antiviral drugs in HCV infection in clinical practice. *European Association for the Study of the Liver*; 2015;: 1–19. doi:10.1016/j.jhep.2015.09.011
24. Kim S, Han K-H, Ahn SH. Hepatitis C Virus and Antiviral Drug Resistance. *Gut and Liver*. 2016;10: 890–895. doi:10.5009/gnl15573
25. Wyles DL, Luetkemeyer AF. Understanding Hepatitis C Virus Drug Resistance: Clinical Implications for Current and Future Regimens. *Top Antivir Med*. 2017;25: 103–109.
26. Sorbo MC, Cento V, Di Maio VC, Howe AYM, Garcia F, Perno CF, et al. Hepatitis C virus drug resistance associated substitutions and their clinical relevance: Update 2018. *Drug Resistance Updates*. 2018;37: 17–39. doi:10.1016/j.drug.2018.01.004
27. Johnson JA, Li JF, Wei X, Lipscomb J, Iribeck D, Craig C, et al. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. *PLoS Medicine*. 2008;5: e158. doi:10.1371/journal.pmed.0050158
28. Li JZ, Paredes R, Ribaldo HJ, Svarovskaia ES, Metzner KJ, Kozal MJ, et al. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA*. 2011;305: 1327–1335. doi:10.1001/jama.2011.375
29. Echeverría N, Betancour G, Gámbaro F, Hernández N, López P, Chiodi D, et al. Naturally occurring NS3 resistance-associated variants in hepatitis C virus genotype 1: Their

relevance for developing countries. *Virus Research*. 2016;223: 140–146.

doi:10.1016/j.virusres.2016.07.008

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

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

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

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

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

35. McCloskey RM, Liang RH, Joy JB, Krajden M, Montaner JSG, Harrigan PR, et al. Global 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-](http://www.R-project.org)
658 [project.org](http://www.R-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.

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

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

Figure Captions

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; nonsyn, 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,

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.

Supporting Information

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% confidential 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 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.

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.

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.

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

S1 File. Supplementary methods.

Data availability

All relevant data are within the manuscript and its Supporting Information files,

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; nonsyn, 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

<i>Variant</i>	<i>AA Position</i>	<i>wtAA</i>	<i>resAA</i>	<i>Gene</i>	<i>NT Position</i>	<i>Mutation Type</i>	<i>wtNT</i>	<i>resNT</i>
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.