Adaptive potential of a drug-targeted viral protein as a function of positive selection

Lei Dai[1, 2]*, Yushen Du[1]*, Hangfei Qi[1], Nicholas C. Wu[1,3], Ergang Wang[1], James O. Lloyd-Smith[2], Ren Sun[1]

[1]Department of Molecular and Medical Pharmacology

[2]Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, United States

[3]Current address: Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, United States

*These authors contributed equally to this work.

Corresponding author:

Lei Dai, Ph.D.

Email: leidai@ucla.edu

Ren Sun, Ph.D.

Email: rsun@mednet.ucla.edu

Abstract

RNA viruses are notorious for their ability to evolve rapidly under positive selection in novel environments. It is known that the high mutation rate of RNA viruses can generate huge genetic diversity to facilitate viral adaptation. However, less attention has been paid to the underlying fitness landscape that represents the selection forces on viral genomes. Here we systematically quantified the distribution of fitness effects (DFE) of about 1,600 single amino acid substitutions in the drug-targeted region of NS5A protein of Hepatitis C Virus (HCV). We found that the majority of non-synonymous substitutions incur large fitness costs, suggesting that NS5A protein is highly optimized in natural conditions. We characterized the adaptive potential of HCV by subjecting the mutant viruses to positive selection by the NS5A inhibitor Daclatasvir. Both the selection coefficient and the number of beneficial mutations are found to increase with the strength of positive selection, which is modulated by the concentration of Daclatasvir. The shift in the spectrum of beneficial mutations in NS5A protein can be explained by a pharmacodynamics model describing viral fitness as a function of drug concentration. Finally, our large-scale fitness data of mutant viruses also provide insights into the biophysical basis of evolutionary constraints in protein evolution.

Introduction

In our evolutionary battles with microbial pathogens, RNA viruses are among the most formidable foes. HIV-1 and Hepatitis C Virus acquire drug resistance in patients under antiviral therapy. Influenza and Ebola virus cross the species barrier to infect human hosts. Understanding the evolution of RNA viruses is therefore of paramount importance for developing antivirals and vaccines and assessing the risk of future emergence events (Goldberg *et al.* 2012; Domingo *et al.* 2012; Metcalf *et al.* 2015). Comprehensive characterization of viral fitness landscapes, and the principles underpinning them, will provide us with a map of evolutionary pathways accessible to viruses and guide our design of effective strategies to limit antiviral resistance, immune escape and cross-species transmission (Turner and Elena 2000; Ke *et al.* 2015; Barton *et al.* 2016).

Although the concept of fitness landscapes has been around for a long time (Wright 1932), we still know little about their properties in real biological systems. Previous empirical studies of fitness landscapes have been constrained by very limited sampling of sequence space. In a typical study, mutants are generated by site-directed mutagenesis and assayed for growth rate individually. We and others have recently developed a high-throughput technique, often referred to as “deep mutational scanning” or “quantitative high-resolution genetics”, to profile the fitness effect of mutations by integrating deep sequencing with selection experiments in vitro or in vivo (Wu *et al.* 2013; Thyagarajan and Bloom 2014; Qi *et al.* 2014; Fowler and Fields 2014). This novel application of next generation sequencing has raised an exciting prospect of large-scale fitness measurements (Olson *et al.* 2014; Puchta *et al.* 2015; Li *et al.* 2016; Wu *et al.* 2016) and a revolution in our understanding of molecular evolution (He and Liu 2016).

The distribution of fitness effects (DFE) of mutations is a fundamental entity in genetics and reveals the local structure of a fitness landscape (Burch and Chao 2000; Eyre-Walker and Keightley 2007; Hietpas *et al.* 2011; Desai 2013; Jacquier *et al.* 2013; Bataillon and Bailey 2014; Chevereau *et al.* 2015; Bank *et al.* 2015). Deleterious mutations are usually abundant and impose severe constraints on the accessibility of fitness landscapes. In contrast, beneficial mutations are rare and provide the raw materials of adaptation. Quantifying the DFE of viruses is crucial for understanding how these pathogens evolve to acquire drug resistance and surmount other evolutionary challenges.

A central challenge is to characterize the DFE, and its determinants, in the fluctuating environments where evolution typically occurs (e.g. varying levels of selection pressure as drug concentration fluctuates between doses) (Hietpas *et al.* 2013). One recent empirical study has demonstrated that the strength of purifying selection modulates the shape of the DFE and determines the evolvability under new environments

(Stiffler *et al.* 2015). The effect of positive selection on the DFE, however, has not been investigated systematically. In this study, we profile the DFE of a drug-targeted viral protein under varying levels of positive selection by tuning the concentration of an antiviral drug. In addition, we show that viral evolution under drug selection is constrained by the need to maintain protein stability.

Results

Profiling the fitness landscape of the drug-interacting domain of HCV NS5A protein

The system used in our study is Hepatitis C Virus (HCV), a positive sense single-stranded RNA virus with a genome of ~9.6 kb. HCV has been studied extensively in the past two decades in patients and in laboratory and provides an excellent model system to study viral evolution. We applied high-throughput assays to map the fitness effects of all single amino acid substitutions in domain IA (amino acid 18-103) of HCV NS5A protein (Methods). This domain is the target of several directly-acting antiviral drugs, including the potent HCV NS5A inhibitor Daclatasvir (DCV) (Gao *et al.* 2010).

To study the DFE of mutations of HCV NS5A protein, we conducted new selection experiments using a previously constructed saturation mutagenesis library of mutant viruses (Qi *et al.* 2014). Briefly, each codon in the mutated region was randomized to cover all possible single amino acid substitutions. We observed 2520 non-synonymous mutations in the plasmid library, as well as 105 synonymous mutations. After transfection to reconstitute mutant viruses, we performed selection in an HCV cell culture system (Lindenbach *et al.* 2005; Wakita *et al.* 2005). The relative fitness of a mutant virus to the wild-type virus was calculated based on the changes in frequency of the mutant virus and the wild-type virus after one round of selection in cell culture (Supplementary Figure 1). In our selection experiment, we grew 5 small sub-libraries (~500 mutants each) separately to reduce the noise in fitness measurements (Methods). The fitness data reported in this study is highly correlated to an independent experiment using the same plasmid library (Supplementary Figure 2) (Qi *et al.* 2014).

Our experiment provides a comprehensive profiling of the fitness effect of single amino acid substitutions (1565 out of 1634 possible substitutions, after filtering out low frequency mutants in the plasmid library). We grouped together non-synonymous mutations leading to the same amino acid substitution. As expected, the fitness effects of synonymous mutations were nearly neutral, while most non-synonymous mutations were deleterious (Figure 1). We found that the majority of single amino acid mutations had fitness costs and more than half of them were found to be significantly deleterious, or “lethal” (Methods). The fraction of lethal mutations (not shown explicitly in Figure 1) is 57.0% (932/1634) for single amino acid substitutions, 1.0%

(1/105) for synonymous mutations and 90.6% (77/85) for nonsense mutations. The low tolerance of non-synonymous mutations in HCV NS5A, which is an essential protein for viral replication, is consistent with previous small-scale mutagenesis studies of RNA viruses (Sanjuan *et al.* 2004). Our data support the view that RNA viruses are very sensitive to the effect of deleterious mutations, possibly due to the compactness of their genomes (Elena *et al.* 2006; Rihn *et al.* 2013).

Using the distribution of fitness effects of synonymous mutations as a benchmark for neutrality, we identified that only 2.4% (39/1634) of single amino acid mutations are beneficial (Methods). The estimated fraction of beneficial mutations is consistent with previous small-scale mutagenesis studies in viruses including bacteriophages, vesicular stomatitis virus, etc. (Sanjuan *et al.* 2004; Burch *et al.* 2007; Silander *et al.* 2007; Eyre-Walker and Keightley 2007). Our results indicate that HCV NS5A protein is under strong purifying selection, suggesting that viral proteins are highly optimized in their natural conditions.

Adaptive potential as a function of positive selection

Beneficial mutations are the raw materials of protein adaptation (Eyre-Walker and Keightley 2007). In this study, we aimed to study the role of positive selection in modulating the adaptive potential of drug-targeted viral proteins. In an independent study (Qi *et al.* 2014), the mutant library of HCV NS5A protein was selected under a single drug concentration ([DCV]=20 pM) to profile the effects of mutations on drug resistance. In this study, we selected the mutant library at 10, 40 and 100 pM of DCV. The drug concentrations were chosen based on in vitro IC_{50} of wild type HCV virus (~20 pM) to represent different levels of positive selection (mild, intermediate and strong).

By tuning the concentration of DCV, we observed a shift in the DFE of beneficial mutations (Figure 2A). At higher drug concentrations, we observed an increase in the median selection coefficient (Figure 2B) as well as the total number of beneficial mutations (Figure 2C). We further tested whether the shape of this distribution changed under drug selection. Previous empirical studies supported the hypothesis that the DFE of beneficial mutations is exponential (Imhof and Schlötterer 2001; Sanjuan *et al.* 2004; Rokyta *et al.* 2005; Cowperthwaite *et al.* 2005; Kassen and Bataillon 2006; Burch *et al.* 2007; Carrasco *et al.* 2007; MacLean and Buckling 2009; Peris *et al.* 2010; Bataillon *et al.* 2011). Following a maximum likelihood approach, we fit the DFE of beneficial mutations to the Generalized Pareto Distribution (Supplementary Figure 3, Methods). The fitted distribution is described by two parameters: a scale parameter (τ), and a shape parameter (κ) that determines the behavior of the distribution's tail. Using a likelihood-ratio test (Beisel *et al.* 2007), we found that our data are consistent with the null hypothesis that the DFE of beneficial mutations is exponential(κ = 0) (Supplementary Table 1).

The effects of mutations on drug resistance and replication fitness

Our results show that the adaptive potential of proteins is modulated by the strength of positive selection, in parallel to earlier findings for purifying selection (Stiffler *et al.* 2015). The changing spectra of beneficial mutations upon drug treatment can be explained by a pharmacodynamics model describing viral fitness as a function of drug concentration (i.e. phenotype-fitness mapping) (Figure 3A). Mutations that reduce a protein's binding affinity to drug molecules (i.e. with a higher inhibitory concentration than wild-type) may come with a fitness cost (Wu *et al.* 2013). Thus, a drug-resistant mutant that is deleterious in the absence of drug may become beneficial under drug selection, leading to an increase in the number of beneficial mutations. Moreover, the relative fitness of the drug-resistant mutant is expected to increase with stronger selection pressure (Figure 3A, dashed line). The dose response curves were previously measured for a set of mutants constructed by site-directed mutagenesis (Supplementary Figure 4) (Qi *et al.* 2014). Indeed, we found that the relative fitness of validated drug-resistant mutants increased at higher drug concentration (Figure 3B); in contrast, drug-sensitive mutants became less fit under drug selection.

Furthermore, we showed that the effects of mutations on drug resistance can be estimated from the fitness data and the results were generally consistent with estimates based on the dose response curves (Supplementary Figure 5, Methods). Among all the non-lethal single amino acid substitutions profiled in our HCV NS5A protein library, we found that roughly half of the mutations increased resistance to DCV (i.e. improved new function) at the expense of replication fitness without drug (Figure 3C, Spearman's ρ= -0.13, $p=8.3\times10^{-4}$). This group of resistance mutations (lower right section in Figure 3C) can become beneficial when the positive selection imposed by the antiviral drug is strong, leading to an increase in the supply of beneficial mutations at higher drug concentrations.

Deleterious mutations as evolutionary constraints

While beneficial mutations open up adaptive pathways to genotypes with higher fitness, mutations that severely reduce replication fitness impose constraints on the evolution of viruses and are less likely to contribute to adaptation through gain of function. We analyzed sequence diversity of HCV sequences identified in patients from the HCV sequence database of Los Alamos National Lab (Methods). As expected, we found that amino acid sites with high fitness costs are often highly conserved (Figure 4A). The sequence diversity at each site was highly correlated to the replication fitness measured in our study (Spearman's $\rho=0.82$, $p=1.8\times10^{-21}$).

To understand the biophysical basis of mutational effects (Liberles *et al.* 2012), we took advantage of the

available structural information (Supplementary Figure 6). The crystal structure of NS5A domain I is available excluding the amphipathic helix at N-terminus (Tellinghuisen *et al.* 2005; Love *et al.* 2009). We found that the fitness effects of deleterious mutations at buried sites (i.e. with lower solvent accessibility) were more pronounced than those at surface exposed sites (Figure 4B, Spearman's ρ=0.51, $p=5.1\times10^{-6}$) (Ramsey *et al.* 2011). Moreover, we performed simulations of protein stability for individual mutants using PyRosetta (Methods) (Das and Baker 2008; Chaudhury *et al.* 2010). A mutation with $\Delta\Delta G>0$, i.e. shifting the free energy difference to favor the unfolded state, is expected to destabilize the protein. We found that mutations that decreased protein stability led to reduced viral fitness (Figure 4C, Spearman's ρ= -0.57, $p=1.5\times10^{-7}$). For example, mutations at a stretch of highly conserved residues (F88-N91) that run through the core of NS5A protein tended to destabilize the protein and significantly reduced the viral fitness. Mutations that increase $\Delta\Delta G$ beyond a threshold (~5 Rosetta Energy Unit) were mostly lethal. This is consistent with the threshold robustness model, which predicts that proteins become unfolded after using up the stability margin (Bloom *et al.* 2005; Wylie and Shakhnovich 2011; Olson *et al.* 2014).

Discussion

Mutation accumulation experiments (Levy *et al.* 2015) and site-directed mutagenesis (Visher *et al.* 2016) are traditional approaches to examine the DFE. Both methods provide pivotal insights into the shape of the DFE, yet with limitations. The site-directed mutagenesis approach requires fitness assays for each individual mutant and can only provide a sparse sampling of mutations. The sampling of sequence space in a mutation accumulation experiment is biased towards large-effect beneficial mutations, as they are more likely to fix in the population. In contrast, the deep mutational scanning approach (Wu *et al.* 2013; Fowler and Fields 2014), which utilizes high-throughput sequencing to simultaneously assay the fitness or phenotype of a library of mutants, allows for unbiased and large-scale sampling of fitness landscapes and thus is ideal for studying the characteristics of empirical DFE. The downside of this high-throughput approach is that the fitness measurements can be noisy, especially for large mutant libraries (Matuszewski *et al.* 2016). In our experiment, we divided the mutant library into smaller sub-libraries (~500 mutants) in selection experiments. We compared the data to an independent experiment and found that the fitness estimates were largely reproducible (Supplementary Figure 2). We also showed that the observed shift in the DFE under different conditions was consistent with validation experiments (Figure 3). Since this study is focused on the properties of the entire distribution of mutations rather than the effects of specific mutations, our findings on the general patterns of DFE are robust to the errors in fitness estimates.

The shape of the DFE determines mutational robustness (Visser *et al.* 2003; Draghi *et al.* 2010; Visher *et al.* 2016). Our study quantified the fitness effects of single amino acid substitutions in the drug-targeted region of an essential viral protein. In general, the empirical DFE of HCV NS5A was consistent with previous findings that viral proteins were highly optimized in the natural condition and very sensitive to the effects of deleterious mutations. One crucial but often overlooked point is that DFE will vary as a function of selection pressure (Martin and Lenormand 2006; Lalić *et al.* 2011; Stiffler *et al.* 2015). For example, mutations that impair function would become more deleterious with increasing pressure of purifying selection, thus leading to reduced protein evolvability (Stiffler *et al.* 2015). In this study, we have focused on gain-of-function mutations in a novel environment. The pleiotropic effect of mutations causes the spectrum of beneficial mutations to shift among the natural condition and the conditions with drug selection. Moreover, mutations enabling the new function (e.g. drug resistance) become more beneficial with increasing pressure of positive selection.

Although different systems have distinct protein-drug interactions that lead to different resistance profiles (Robinson *et al.* 2011), the results in our study provide a general framework to study DFE of drug-targeted proteins. Future studies along this line will further our understanding of how proteins evolve new functions under the constraint of maintaining their original function (Soskine and Tawfik 2010), as exemplified in the evolution of resistance to directly-acting antiviral drugs (Rosenbloom *et al.* 2012). We have also demonstrated that the fitness data could be utilized to infer drug resistance of mutants and inform predictive modeling of within-patient viral dynamics (Ke *et al.* 2015). Quantifying the characteristics of DFE of drug-targeted proteins under different environments (e.g. varying levels of selection pressure, or conflicting selection pressures), would allow us to assess repeatability in the outcomes of viral evolution (de Visser and Krug 2014) and guide the design of therapies to minimize drug resistance (Ogbunugafor *et al.* 2016).

Conclusions

Many viruses adapt rapidly to novel selection pressures, such as antiviral drugs. Understanding how pathogens evolve under drug selection is critical for the success of antiviral therapy against human pathogens. By combining deep sequencing with selection experiments in cell culture, we have quantified the distribution of fitness effects of mutations in the drug-targeted domain of Hepatitis C Virus NS5A protein. Our results indicate that the majority of single amino acid substitutions in NS5A protein incur large fitness costs. By subjecting the mutant viruses to positive selection under an antiviral drug, we find that the evolutionary potential of viral proteins in a novel environment is modulated by the strength of selection pressure. Combined

with stability predictions based on protein structure, our fitness data further reveal the biophysical constraints underlying the evolution of viral proteins.

Materials and Methods

Mutagenesis

The mutant library of HCV NS5A protein domain IA (86 amino acids) was constructed using saturation mutagenesis as previously described (Qi *et al.* 2014). In brief, the entire region was divided into five sub-libraries each containing 17-18 amino acids (~500 mutants in each sub-library). NNK (N: A/T/C/G, K: T/G) was used to replace each amino acid. The oligos, each of which contains one random codon, were synthesized by IDT. The mutated region was ligated to the flanking constant regions, subcloned into the pFNX-HCV plasmid and then transformed into bacteria. The pFNX-HCV plasmid carrying the viral genome was synthesized in Dr. Ren Sun's lab based on the chimeric sequence of genotype 2a HCV strains J6/JFH1.

Cell culture

The human hepatoma cell line (Huh-7.5.1) was provided by Dr. Francis Chisari from the Scripps Research Institute, La Jolla. The cells were cultured in T-75 tissue culture flasks (Genesee Scientific) at 37 ºC with 5% CO_2. The complete growth medium contained Dulbecco's Modified Eagle's Medium (Corning Cellgro), 10% heat-inactivated Fetal Bovine Serum (Omega Scientific), 10 mM HEPES (Life Technologies), 1x MEM Non-Essential Amino Acids Solution (Life Technologies) and 1x Penicillin-Streptomycin-Glutamine (Life Technologies).

Selection of mutant viruses

Plasmid mutant library was transcribed *in vitro* using T7 RiboMAX Express Large Scale RNA Production System (Promega) and purified by PureLink RNA Mini Kit (Life Technologies). 10 µg of *in vitro* transcribed RNA was used to transfect 4 million Huh-7.5.1 cells via electroporation by Bio-Rad Gene Pulser (246 V, 950 µF). The supernatant was collected 6 days post transfection and virus titer was determined by immunofluorescence assay. The viruses collected after transfection were used to infect ~2 million Huh-7.5.1 cells with an MOI at around 0.1-0.2. The five sub-libraries were passaged for selection separately. For the three different levels of selection pressure, the growth media was supplemented with 10 pM, 40 pM and 100 pM HCV NS5A inhibitor Daclatasvir (BMS-790052), respectively. The supernatant was collected at 6 days post infection.

Preparation of Illumina sequencing samples

For each sample, viral RNA was extracted from 700 µl supernatant collected after transfection and after

selection using QIAamp Viral RNA Mini Kit (Qiagen). Extracted RNA was reverse transcribed into cDNA by SuperScript III Reverse Transcriptase Kit (Life Technologies). The targeted region in NS5A (51-54 nt) was PCR amplified using KOD Hot Start DNA polymerase (Novagen). The Eppendorf thermocycler was set as following: 2 min at 95 °C; 25 to 35 three-step cycles of 20 s at 95 °C,15 s at 52-56 °C (sub-library #1, 52 °C; #2, 52 °C; #3, 52 °C; #4, 56 °C; #5, 54 °C) and 25s at 68 °C; 1 min at 68 °C. The number of PCR cycles are chosen based on the copy number of cDNA templates as determined by qPCR (Bio-Rad). The PCR primers are listed in Supplementary Table 2. The PCR products were purified using PureLink PCR Purification Kit (Life Technologies) and prepared for Illumina HiSeq 2000 sequencing (paired-end 100 bp) following 5'-phosphorylation using T4 Polynucleotide Kinase (New England BioLabs), 3' dA-tailing using dA-tailing module (New England BioLabs), and TA ligation of the adapter using T4 DNA ligase (Life Technologies). Each sample was tagged with a unique 3-bp customized barcodes, which were part of the adapter sequence and were sequenced as the first three nucleotides in both the forward and reverse reads (Wu *et al.* 2015) (Supplementary Table 3).

Analysis of Illumina sequencing data

The sequencing data were parsed by SeqIO function of BioPython. The reads from different samples were de-multiplexed by the barcodes and mapped to the entire mutated region in NS5A by allowing at maximum 5 mismatches with the reference genome (Supplementary Table 3) (Qi *et al.* 2014). Since both forward and reverse reads cover the whole amplicon, we used paired reads to correct for sequencing errors. A mutation was called only if it was observed in both reads and the quality score at the corresponding position was at least 30. Sequencing reads containing mutations not supposed to appear in the mutant library were excluded from downstream analysis. The sequencing depth for each sub-library is at least ~10^5 and two orders of magnitude higher than the library complexity.

Calculation of relative fitness

For each condition of selection experiments (i.e. different concentration of Daclatasvir [DCV]), the relative fitness (RF) of a mutant virus to the wild-type virus is calculated by the relative changes in frequency after selection,

$$RF_{mut}([DCV]) = \left(\frac{f_{mut}^{T=2}}{f_{mut}^{T=1}} \right) \Big/ \left(\frac{f_{WT}^{T=2}}{f_{WT}^{T=1}} \right)$$

where $f_{mut}^{T=round}$ and $f_{WT}^{T=round}$ is the frequency of the mutant virus and the wild-type virus at round 1 (after transfection) or round 2 (after infection). The fitness of wild-type virus is normalized to 1. The fitness values estimated from one round (round 1 to round 2) have been shown to be highly consistent to estimated based

round 0 to round 1 (Supplementary Figure 2), and estimates from multiple rounds of selection (Qi *et al.* 2014). A mutant was labeled as "missing" if the mutant's frequency in the plasmid library was less than 0.0005 (RF=NaN, see Supplementary Data 1 and 2). A mutant was labeled as "lethal" if the mutant's frequency after transfection was less than 0.0005, or its frequency after infection was 0 (RF=0) (Qi *et al.* 2014).

The selection coefficient is defined in the context of discrete generations (Chevin 2010)

$$s_{mut} = \log(RF_{mut})$$

The thresholds for beneficial mutations were defined as $2\sigma_{silent}$, where σ_{silent} is the standard deviation of the selection coefficients of synonymous mutations (Figure 1). The fitness effects of non-synonymous mutations leading to the same amino acid substitution were averaged to estimate the fitness effect of the given single amino acid substitution.

Fitting the distribution of fitness effects of beneficial mutations

The distribution of selection coefficients of beneficial mutations were fitted to a Generalized Pareto Distribution following a maximum likelihood approach (Beisel *et al.* 2007),

$$F(x|\kappa,\tau) = \begin{cases} 1-(1+\frac{\kappa}{\tau}x)^{-1/\kappa}, x \geq 0, \ \ if\, \kappa > 0 & \text{(Frechet)} \\ 1-(1+\frac{\kappa}{\tau}x)^{-1/\kappa}, 0 \leq x < -\frac{\tau}{\kappa}, \ \ if\, \kappa < 0 & \text{(Weibull)} \\ 1-e^{-x/\tau}, x \geq 0, \ \text{if } \kappa = 0 & \text{(Gumbel)} \end{cases}$$

Only mutations with selection coefficients higher than the beneficial threshold $2\sigma_{silent}$ were included in the distribution of beneficial mutations. The selection coefficients were normalized to the beneficial threshold. The shape parameter κ determines the tail behavior of the distribution, which can be divided into three domains of attraction: Gumbel domain (exponential tail, κ = 0), Weibull domain (truncated tail, κ < 0) and Fréchet domain (heavy tail, κ > 0). For each selection condition, a likelihood ratio test is performed to evaluate whether the null hypothesis κ = 0 (exponential distribution) can be rejected.

Inferring drug resistance from fitness data

We can quantify the drug resistance of each mutant in the library by computing its fold change in relative fitness,

$$W([DCV]) = \frac{RF_{mut}([DCV])}{RF_{mut}}$$

Here RF_{mut} is the relative fitness of a mutant under the natural condition (i.e. no drug). W is the fold change in relative fitness and represents the level of drug resistance relative to the wild type. W > 1 indicates drug resistance, and W < 1 indicates drug sensitivity.

This empirical measure of drug resistance can be directly linked to a simple pharmacodynamics model (Rosenbloom *et al.* 2012), where the viral replicative fitness is modeled as a function of drug dose,

$$W_{predict}([DCV]) = \left(\frac{IC_{mut}}{[DCV] + IC_{mut}} \right) \Big/ \left(\frac{IC_{wt}}{[DCV] + IC_{wt}} \right)$$

Here IC denotes the half-inhibitory concentration. The Hill coefficient describing the sigmoidal shape of the dose response curve is fixed to 1, as used in fitting the dose response curves of wild-type virus and validated mutant viruses (Supplementary Figure 4). The drug resistance score W inferred from fitness data is consistent with the drug resistance score $W_{predict}$ predicted from dose response curves of validated mutants (Supplementary Figure 5).

Calculation of relative solvent accessibility

DSSP (http://www.cmbi.ru.nl/dssp.html) was used to compute the Solvent Accessible Surface Area (SASA) (Kabsch and Sander 1983) from the HCV NS5A protein structure (PDB: 3FQM) (Love *et al.* 2009). SASA was then normalized to Relative Solvent Accessibility (RSA) using the empirical scale reported in (Tien *et al.* 2013).

Predictions of protein stability

ΔΔG (in Rosetta Energy Unit) of HCV NS5A mutants was predicted by PyRosetta (version: "monolith.ubuntu.release-104") as the difference in scores between the monomer structure of mutants (single amino acid mutations from site 32 to 103) and the reference (PDB: 3FQM). The score is designed to capture the change in thermodynamic stability caused by the mutation (ΔΔG) (Das and Baker 2008). The reference sequence of NS5A in the PDB file (PDB: 3FQM) is different from the WT sequence in our experiment by 20 amino acid substitutions. Thus instead of directly comparing ΔΔG to fitness effects of individual mutations, we used the median ΔΔG caused by amino acid substitutions at each site.

The PDB file of NS5A dimer was cleaned and trimmed to a monomer (chain A). Next, all side chains were repacked (sampling from the 2010 Dunbrack rotamer library (Shapovalov and Dunbrack 2011)) and minimized for the reference structure using the talaris2014 scoring function. After an amino acid mutation was introduced, the mutated residue was repacked, followed by quasi-Newton minimization of the backbone

and all side chains (algorithm: "lbfgs_armijo_nonmonotone"). This procedure was performed 50 times, and the predicted ΔG of a mutant structure is the average of the three lowest scoring structures.

We note that predictions based on NS5A monomer structure were only meant to provide a crude profile of how mutations at each site may impact protein stability. Potential structural constraints at the dimer interface have been ignored, which is further complicated by the observations of two different NS5A dimer structures (Tellinghuisen *et al.* 2005; Love *et al.* 2009).

Diversity of HCV sequences identified in patients

Aligned nucleotide sequences of HCV NS5A protein were downloaded from Los Alamos National Lab database (Kuiken *et al.* 2005) (all HCV genotypes, ~2600 sequences total) and clipped to the region of interest (amino acid 18-103 of NS5A). Sequences that caused gaps in the alignment of H77 reference genome were manually removed. After translation to amino acid sequences, sequences with ambiguous amino acids were removed (~2300 amino acid sequences after filtering). The sequence diversity at each amino acid site was quantified by Shannon entropy.

Data and reagent availability

All research materials are available upon request. Raw sequencing data have been submitted to the NIH Short Read Archive (SRA) under accession number: BioProject PRJNA395730. All scripts have been deposited to https://github.com/leidai-evolution/DFE-HCV.

Ethics Statement

The use of human cell lines and infectious agents in this paper is approved by Institutional Biosafety Committee at University of California, Los Angeles (IBC #40.10.2-f).

Acknowledgements

L.D. was supported by HHMI Postdoctoral Fellowship from Jane Coffin Childs Memorial Fund for Medical Research. N.C.W. was supported by Croucher Foundation Fellowship. R.S. was supported by NSFC 81172314, NIH DE023591 and NIH CA177322.

Author contributions

L.D., Y.D., H.Q. and R.S. designed the experiments. L.D., H.Q. and Y.D. performed the experiments. L.D. and Y.D. analyzed the experimental data. L.D., E.W. and Y.D. performed the bioinformatics analyses. L.D. wrote the first draft of the manuscript, with revisions from Y.D., J.O.L-S., and R.S.. All authors discussed the results and commented on the manuscript.

References

Bank C., Hietpas R. T., Jensen J. D., Bolon D. N. A., 2015 A systematic survey of an intragenic epistatic landscape. Mol. Biol. Evol. 32: 229–38.

Barton J. P., Goonetilleke N., Butler T. C., Walker B. D., McMichael A. J., *et al.*, 2016 Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat. Commun. 7: 11660.

Bataillon T., Zhang T., Kassen R., 2011 Cost of adaptation and fitness effects of beneficial mutations in Pseudomonas fluorescens. Genetics 189: 939–949.

Bataillon T., Bailey S., 2014 Effects of new mutations on fitness: insights from models and data. Ann. N. Y. Acad. Sci. 1320: 76–92.

Beisel C. J., Rokyta D. R., Wichman H. A., Joyce P., 2007 Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics 176: 2441–9.

Bloom J. D., Silberg J. J., Wilke C. O., Drummond D. A., Adami C., *et al.*, 2005 Thermodynamic prediction of protein neutrality. Proc. Natl. Acad. Sci. U. S. A. 102: 606–11.

Burch C. L., Chao L., 2000 Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406: 625–8.

Burch C., Guyader S., Samarov D., Shen H., 2007 Experimental estimate of the abundance and effects of nearly neutral mutations in the RNA virus ϕ6. Genetics 476: 467–476.

Carrasco P., Iglesia F. de la, Elena S., 2007 Distribution of fitness and virulence effects caused by single-nucleotide substitutions in Tobacco etch virus. J. Virol. 81: 12979–12984.

Chaudhury S., Lyskov S., Gray J. J., 2010 PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 26: 689–91.

Chevereau G., Dravecká M., Batur T., Guvenek A., Ayhan D. H., *et al.*, 2015 Quantifying the Determinants of Evolutionary Dynamics Leading to Drug Resistance. PLoS Biol. 13: e1002299.

Chevin L.-M., 2010 On measuring selection in experimental evolution. Biol. Lett.

Cowperthwaite M. C., Bull J. J., Meyers L. A., 2005 Distributions of beneficial fitness effects in RNA. Genetics 170: 1449–57.

Das R., Baker D., 2008 Macromolecular Modeling with Rosetta. Annu. Rev. Biochem. 77: 363–382.

Desai M. M., 2013 Statistical questions in experimental evolution. J. Stat. Mech. Theory Exp. 2013: P01003.

Domingo E., Sheldon J., Perales C., 2012 Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. 76: 159–216.

Draghi J. A., Parsons T. L., Wagner G. P., Plotkin J. B., 2010 Mutational robustness can facilitate adaptation.

Nature 463: 353–355.

Elena S. F., Carrasco P., Daròs J.-A., Sanjuán R., 2006 Mechanisms of genetic robustness in RNA viruses. EMBO Rep. 7: 168–73.

Eyre-Walker A., Keightley P. D., 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 610–8.

Fowler D. M., Fields S., 2014 Deep mutational scanning: a new style of protein science. Nat. Methods 11: 801–807.

Gao M., Nettles R. E., Belema M., Snyder L. B., Nguyen V. N., *et al.*, 2010 Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. Nature 465: 96–100.

Goldberg D. E., Siliciano R. F., Jacobs W. R., 2012 Outwitting evolution: fighting drug-resistant TB, malaria, and HIV. Cell 148: 1271–83.

He X., Liu L., 2016 Toward a prospective molecular evolution. Science 352: 769–70.

Hietpas R. T., Jensen J. D., Bolon D. N. A., 2011 Experimental illumination of a fitness landscape. Proc. Natl. Acad. Sci. U. S. A. 108: 7896–901.

Hietpas R. T., Bank C., Jensen J. D., Bolon D. N. A., 2013 SHIFTING FITNESS LANDSCAPES IN RESPONSE TO ALTERED ENVIRONMENTS. Evolution (N. Y). 67: 3512–3522.

Imhof M., Schlötterer C., 2001 Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc. Natl. Acad. Sci. U. S. A. 98: 1113–1117.

Jacquier H., Birgy A., Nagard H. Le, Mechulam Y., Schmitt E., *et al.*, 2013 Capturing the mutational landscape of the beta-lactamase TEM-1. Proc. Natl. Acad. Sci. U. S. A. 110: 13067–72.

Kabsch W., Sander C., 1983 Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.

Kassen R., Bataillon T., 2006 Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat. Genet. 38: 484–8.

Ke R., Loverdo C., Qi H., Sun R., Lloyd-Smith J. O., 2015 Rational Design and Adaptive Management of Combination Therapies for Hepatitis C Virus Infection. PLoS Comput. Biol. 11: e1004040.

Kuiken C., Yusim K., Boykin L., Richardson R., 2005 The Los Alamos hepatitis C sequence database. Bioinformatics 21: 379–384.

Lalić J., Cuevas J. M., Elena S. F., 2011 Effect of host species on the distribution of mutational fitness effects for an RNA virus. PLoS Genet. 7: e1002378.

Levy S. F., Blundell J. R., Venkataram S., Petrov D. A., Fisher D. S., *et al.*, 2015 Quantitative evolutionary

421 dynamics using high-resolution lineage tracking. Nature 519: 181–6.

422 Li C., Qian W., Maclean C. J., Zhang J., 2016 The fitness landscape of a tRNA gene. Science 352: 837–840.

423 Liberles D. A., Teichmann S. A., Bahar I., Bastolla U., Bloom J., *et al.*, 2012 The interface of protein structure,
424 protein biophysics, and molecular evolution. Protein Sci. 21: 769–785.

425 Lindenbach B. D., Evans M. J., Syder A. J., Wölk B., Tellinghuisen T. L., *et al.*, 2005 Complete Replication of
426 Hepatitis C Virus in Cell Culture. Science 309: 623–626.

427 Love R. A., Brodsky O., Hickey M. J., Wells P. A., Cronin C. N., 2009 Crystal structure of a novel dimeric form
428 of NS5A domain I protein from hepatitis C virus. J. Virol. 83: 4395–403.

429 MacLean R. C., Buckling A., 2009 The distribution of fitness effects of beneficial mutations in Pseudomonas
430 aeruginosa. PLoS Genet. 5: e1000406.

431 Martin G., Lenormand T., 2006 The fitness effect of mutations across environments: a survey in light of fitness
432 landscape models. Evolution 60: 2413–2427.

433 Matuszewski S., Hildebrandt M. E., Ghenu A.-H., Jensen J. D., Bank C., 2016 A Statistical Guide to the
434 Design of Deep Mutational Scanning Experiments. Genetics.

435 Metcalf C. J. E., Birger R. B., Funk S., Kouyos R. D., Lloyd-Smith J. O., *et al.*, 2015 Five challenges in
436 evolution and infectious diseases. Epidemics 10: 40–44.

437 Ogbunugafor C. B., Wylie C. S., Diakite I., Weinreich D. M., Hartl D. L., 2016 Adaptive Landscape by
438 Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. PLoS Comput.
439 Biol. 12: e1004710.

440 Olson C. A., Wu N. C., Sun R., 2014 A comprehensive biophysical description of pairwise epistasis throughout
441 an entire protein domain. Curr. Biol. 24: 2643–51.

442 Peris J. B., Davis P., Cuevas J. M., Nebot M. R., Sanjuán R., 2010 Distribution of fitness effects caused by
443 single-nucleotide substitutions in bacteriophage f1. Genetics 185: 603–9.

444 Puchta O., Cseke B., Czaja H., Tollervey D., Sanguinetti G., *et al.*, 2015 Network of epistatic interactions
445 within a yeast snoRNA. Science 352: 840–844.

446 Qi H., Olson C. A., Wu N. C., Ke R., Loverdo C., *et al.*, 2014 A quantitative high-resolution genetic profile
447 rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. PLoS Pathog.
448 10: e1004064.

449 Ramsey D. C., Scherrer M. P., Zhou T., Wilke C. O., 2011 The Relationship Between Relative Solvent
450 Accessibility and Evolutionary Rate in Protein Evolution. Genetics 188: 479–488.

451 Rihn S. J., Wilson S. J., Loman N. J., Alim M., Bakker S. E., *et al.*, 2013 Extreme genetic fragility of the HIV-

1 capsid. PLoS Pathog. 9: e1003461.

Robinson M., Tian Y., Delaney W. E., Greenstein A. E., 2011 Preexisting drug-resistance mutations reveal unique barriers to resistance for distinct antivirals. Proc. Natl. Acad. Sci. U. S. A. 108: 10290–5.

Rokyta D., Joyce P., Caudle S., Wichman H., 2005 An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat. Genet. 37: 441–444.

Rosenbloom D. I. S., Hill A. L., Rabi S. A., Siliciano R. F., Nowak M. A., 2012 Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. Nat. Med. 18: 1378–85.

Sanjuan R., Moya A., Elena S. F., 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. 101: 8396–8401.

Shapovalov M. V, Dunbrack R. L., 2011 A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19: 844–58.

Silander O. K., Tenaillon O., Chao L., 2007 Understanding the Evolutionary Fate of Finite Populations: The Dynamics of Mutational Effects. PLoS Biol. 5: e94.

Soskine M., Tawfik D. S., 2010 Mutational effects and the evolution of new protein functions. Nat. Rev. Genet. 11: 572–82.

Stiffler M. A., Hekstra D. R., Ranganathan R., 2015 Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell 160: 882–892.

Tellinghuisen T. L., Marcotrigiano J., Rice C. M., 2005 Structure of the zinc-binding domain of an essential component of the hepatitis C virus replicase. Nature 435: 374–9.

Thyagarajan B., Bloom J. D., 2014 The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife 3: e03300.

Tien M. Z., Meyer A. G., Sydykova D. K., Spielman S. J., Wilke C. O., 2013 Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8: e80635.

Turner P. E., Elena S. F., 2000 Cost of Host Radiation in an RNA Virus. Genetics 156: 1465–1470.

Visher E., Whitefield S. E., McCrone J. T., Fitzsimmons W., Lauring A. S., 2016 The Mutational Robustness of Influenza A Virus. PLOS Pathog. 12: e1005856.

Visser J. A. G. M., Hermisson J., Wagner G. P., Meyers L. A., Bagheri-Chaichian H., *et al.*, 2003 Perspective: Evolution and detection of genetic robustness. Evolution 57: 1959–1972.

Visser J. A. G. M. de, Krug J., 2014 Empirical fitness landscapes and the predictability of evolution. Nat. Rev. Genet. 15: 480–90.

Wakita T., Pietschmann T., Kato T., Date T., Miyamoto M., *et al.*, 2005 Production of infectious hepatitis C

virus in tissue culture from a cloned viral genome. Nat. Med. 11: 791–796.

Wright S., 1932 The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. 1: 356–366.

Wu N. C., Young A. P., Dandekar S., Wijersuriya H., Al-Mawsawi L. Q., *et al.*, 2013 Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. J. Virol. 87: 1193–9.

Wu N. C., Olson C. A., Du Y., Le S., Tran K., *et al.*, 2015 Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLOS Genet. 11: e1005310.

Wu N. C., Dai L., Olson C. A., Lloyd-Smith J. O., Sun R., 2016 Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife 5: e16965.

Wylie C. S., Shakhnovich E. I., 2011 A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc. Natl. Acad. Sci. 108: 9916–9921.

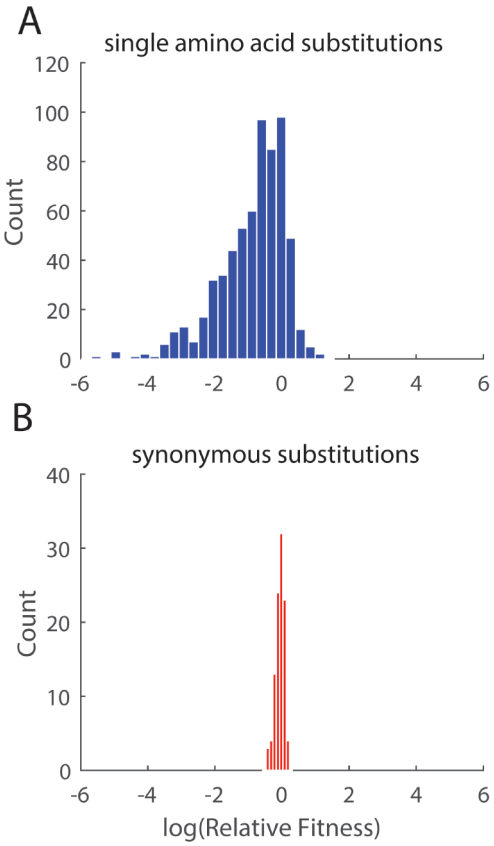

Figure 1. Distribution of fitness effects (DFE) of single amino acid substitutions in domain IA of HCV NS5A protein without drug selection. DFE of single amino acid substitutions (A) and synonymous substitutions (B). Lethal mutations are not shown in the histogram.

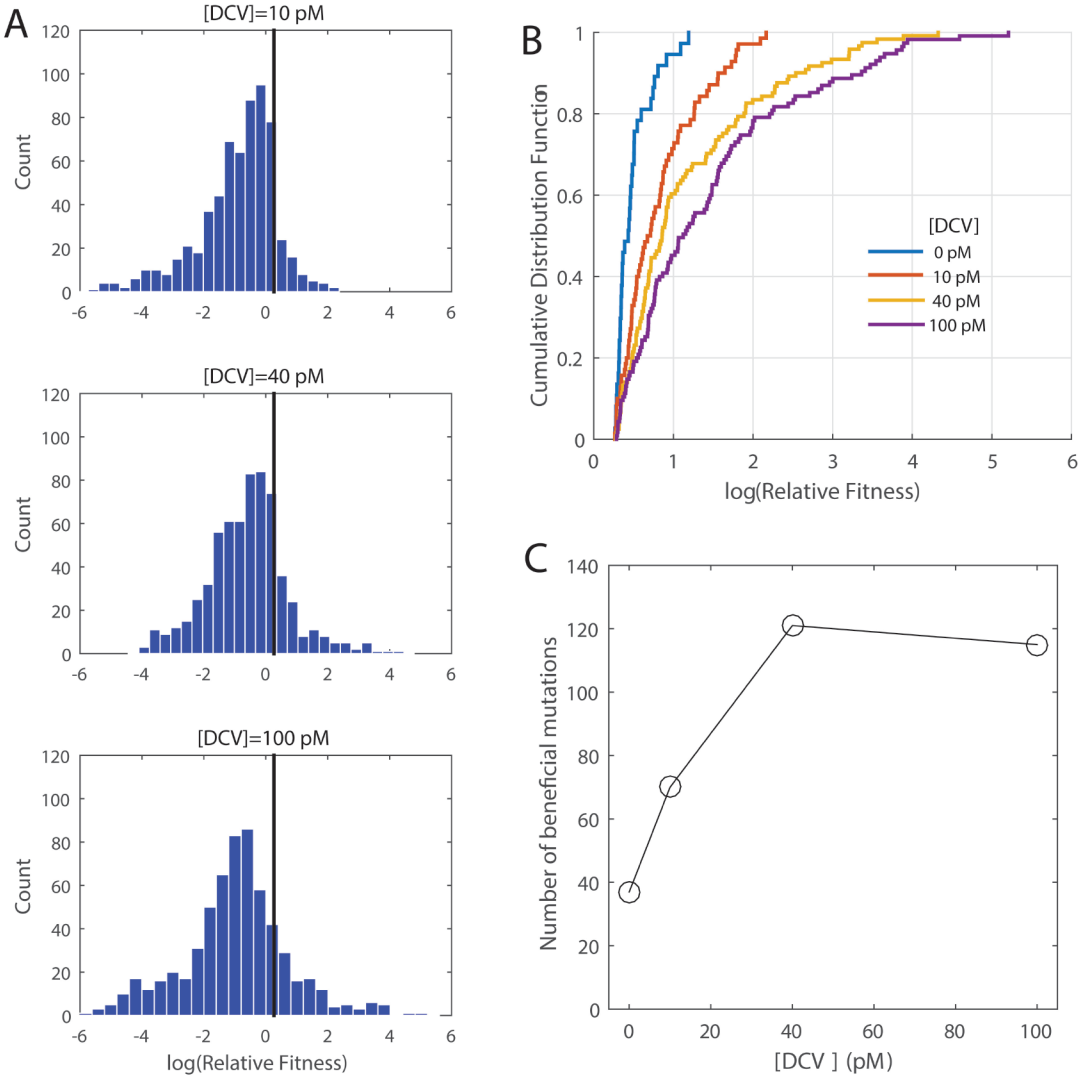

Figure 2. The spectrum of beneficial mutations shifts under increasing positive selection imposed by the antiviral drug Daclatasvir. (A) DFE of single amino acid substitutions in domain IA of HCV NS5A protein under increasing positive selection by Daclatasvir. The black line indicates the threshold used for classifying beneficial mutations (Methods). (B) The cumulative distribution function of the fitness effect of beneficial mutations. (C) The number of beneficial mutations as a function of positive selection imposed by Daclatasvir.

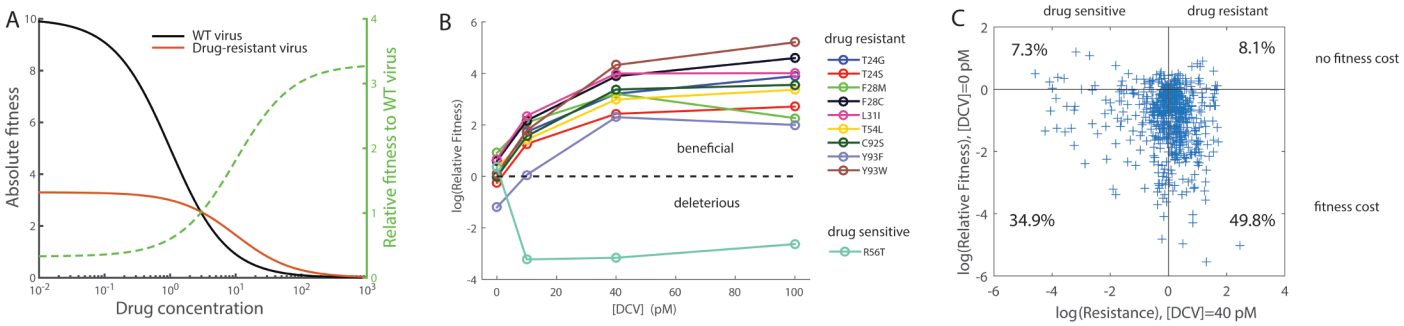

Figure 3. The adaptive potential under drug selection is determined by the effects of mutations on replication fitness and drug resistance. (A) Hypothetical dose response curves of the wild-type virus and a drug-resistant mutant virus. Relative fitness of the drug-resistant mutant is expected to increase with drug concentration. (B) Relative fitness of validated drug-resistant and drug-sensitive mutants (Supplementary Figure 4) as a function of [DCV]. (C) The effects of mutations on replication fitness (i.e. fitness without drug) and drug resistance score W at [DCV]=40 pM (Methods).

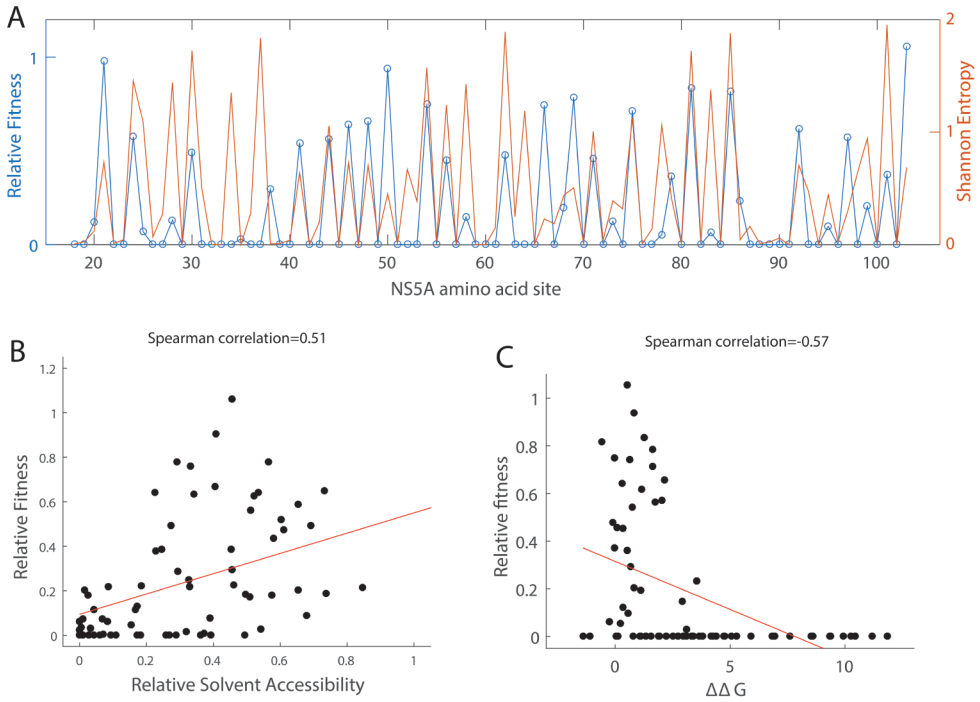

Figure 4. Mutations with deleterious fitness effects reveal constraints of protein evolution. (A) The pattern of sequence conservation observed in patient sequences is highly correlated to the replication fitness measured in cell culture. (B) Mutations at amino acid sites with lower solvent accessibility tend to incur larger fitness costs. (C) Mutations at amino acid sites with larger effects on destabilizing protein stability (ΔΔG>0) tend to reduce the viral replication fitness. Changes in folding free energy ΔΔG (Rosetta Energy Unit) of NS5A monomer were predicted by PyRosetta. The median ΔΔG at each amino acid site is shown. In (B) and (C), the median fitness of observed mutants at each amino acid site is shown. Red lines represent the fits by linear regression and are only used to guide the eye.