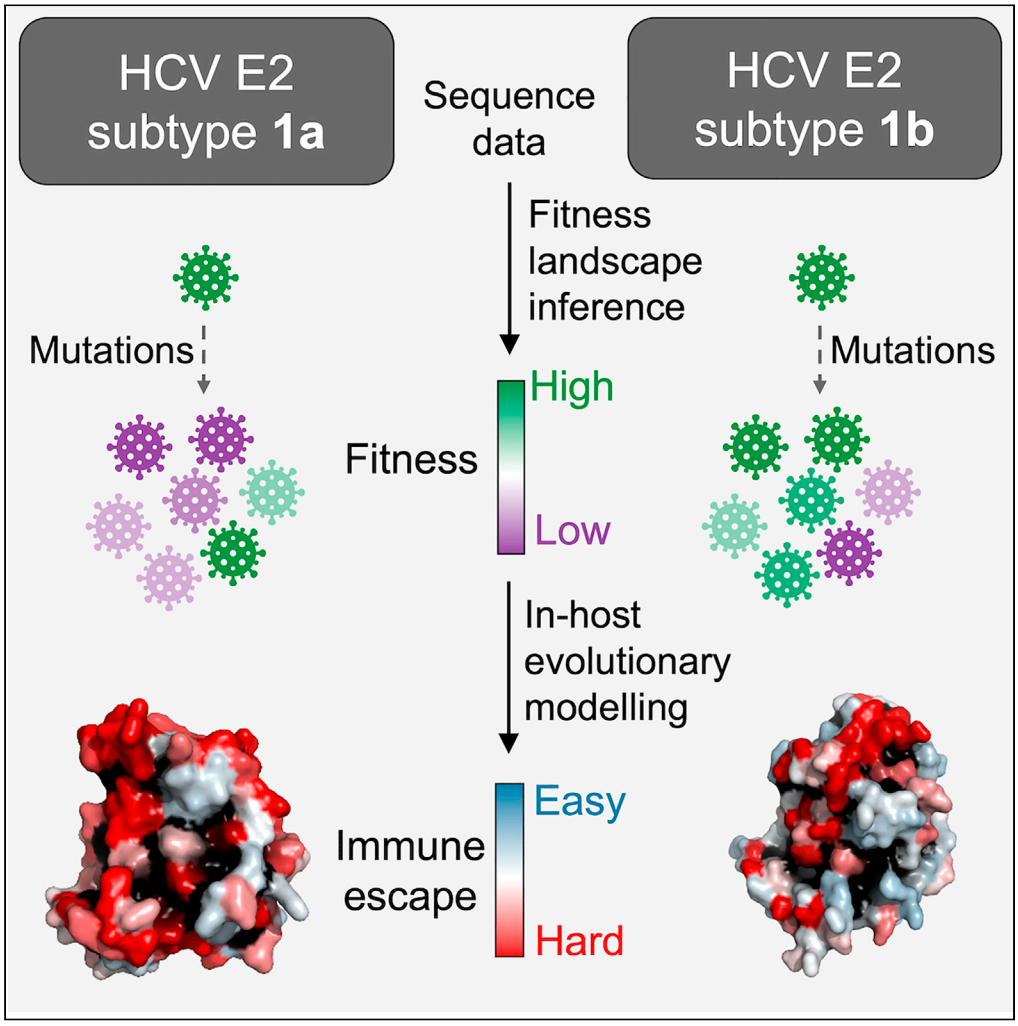
iScience

CellPress
OPEN ACCESS

Article

Evolutionary modeling reveals enhanced mutational flexibility of HCV subtype 1b compared with 1a



Hang Zhang,
Ahmed A.
Quadeer,
Matthew R. McKay

eeaaquadeer@ust.hk (A.A.Q.)
matthew.mckay@unimelb.
edu.au (M.R.M.)

Highlights

Comparative analysis of the fitness landscapes of HCV subtypes 1a and 1b

Subtype 1b evolution is subject to less constraints than 1a

Subtype 1b appears to evade antibodies more easily compared with 1a

Antibodies are identified that are difficult to escape for both subtypes 1a and 1b

Zhang et al., iScience *25*, 103569
January 21, 2022
https://doi.org/10.1016/j.isci.2021.103569

iScience

CellPress
OPEN ACCESS

Article

Evolutionary modeling reveals enhanced mutational flexibility of HCV subtype 1b compared with 1a

Hang Zhang,[1,5] Ahmed A. Quadeer,[1,5,*] and Matthew R. McKay[1,2,3,4,6,*]

SUMMARY

Hepatitis C virus (HCV) is a leading cause of liver-associated disease and liver cancer. Of the major HCV subtypes, patients infected with subtype 1b have been associated with having a higher risk of developing chronic infection and hepatocellular carcinoma. However, underlying reasons for this increased disease severity remain unknown. Here, we provide an evolutionary rationale, based on a comparative study of fitness landscape and in-host evolutionary models of the E2 glycoprotein of HCV subtypes 1a and 1b. Our analysis demonstrates that a higher chronicity rate of 1b may be attributed to lower fitness constraints, enabling 1b viruses to more easily escape antibody responses. More generally, our results suggest that differences in evolutionary constraints between HCV subtypes may be an important factor in mediating distinct disease outcomes. Our analysis also identifies antibodies that appear escape-resistant against both subtypes 1a and 1b, providing directions for designing HCV vaccines having cross-subtype protection.

INTRODUCTION

HCV is a highly mutable single-stranded RNA virus (Rosen, 2011). Approximately 15%–25% of infected people clear the virus spontaneously, while the remaining develop chronic liver disease, commonly leading to cirrhosis and hepatocellular carcinoma (HCC) (Centers for Disease Control and Prevention, 2018). HCV is estimated to cause chronic infection in 58 million people worldwide (World Health Organization, 2021). Although therapeutic treatments for chronic HCV infection have improved since the introduction of direct-acting antivirals (DAAs), these drugs are only administered to a limited number of infected individuals because of high cost (Rosenthal and Graham, 2016) and low diagnostic rate (World Health Organization, 2021). The effectiveness of DAAs also has limitations due to the inability to prevent reinfection (Rossi et al., 2018) and the emergence of drug-resistant variants (Wyles and Luetkemeyer, 2017). Thus, for complete eradication of HCV, developing an effective vaccine is essential.

HCV is classified into eight major genotypes and 90 subtypes (ICTV, 2019). Among the major genotypes, genotype 1 is the most prevalent, representing roughly half of all HCV infections worldwide (Petruzziello et al., 2016). Of the genotype 1 infections for which subtypes have been specified, 99% are either subtype 1a or 1b (Messina et al., 2014), with subtype 1b known to be the most prevalent subtype worldwide (Gower et al., 2014). The dominant HCV subtype however varies across geographical regions, with subtype 1a more prevalent in North America, Tropical Latin America, Northwestern Europe, and Australia, and subtype 1b more prevalent in Asia, Western and Eastern Europe, Southern and Central Latin America, and North Africa (Messina et al., 2014). The most prominent difference between these two highly prevalent HCV subtypes is in terms of disease outcome. Multiple studies have shown that the chronicity rate of subtype 1b infections is about two times higher than that of 1a and other subtypes (Amoroso et al., 1998; Cho et al., 2014; Hwang et al., 2001). Moreover, subtype 1b infections have been reported to increase the risk of developing cirrhosis (Osella et al., 2001), and present almost double the risk of developing HCC compared with 1a and other subtypes (Bruno et al., 2007; Lee et al., 2014; Raimondi et al., 2009; Silini et al., 1996).

Despite these observations, it remains unclear as to why subtype 1b infections may lead to more severe disease outcomes. This may be attributed to various reasons, including differences in viral phenotype, host-specific immune factors, demographics, transmission route (injectable drug use in the case of 1a

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, SAR, China

[2]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, SAR, China

[3]Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, VIC, Australia

[4]Department of Microbiology and Immunology, University of Melbourne, The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

[5]These authors contributed equally

[6]Lead contact

*Correspondence: eeaaquadeer@ust.hk (A.A.Q.), matthew.mckay@unimelb.edu.au (M.R.M.)

https://doi.org/10.1016/j.isci.2021.103569

and blood transfusion in the case of 1b (Magiorkinis et al., 2009)), etc. One plausible explanation is that subtype 1b may evolve under less rigid fitness constraints than other subtypes, allowing it to escape immune responses more easily and to more effectively propagate infection. Evidence of a similar phenomenon has been presented for HIV, for which the fitness of the infecting subtype has been reported to be a factor associated with disease outcome (Venner et al., 2016). For HCV, host-specific factors such as patient's ethnicity, gene composition (e.g., IL28B polymorphisms), alcohol consumption, and co-infection with HIV have been shown to contribute to the disparate disease outcome of infections (Chen and Morgan, 2006; Missiha et al., 2008; Yan and Wang, 2017). However, it is not yet known whether different HCV subtypes exhibit markedly different fitness properties, and if so, whether this could be a significant factor in determining disease outcome.

We aim to address these questions by using computational modeling, leveraging sequence data as well as available clinical and experimental data. Our approach is to first infer a model for the fitness landscape of HCV E2 (the primary target of neutralizing antibodies) for subtype 1b, which we validate against in-vitro fitness measurements, and compare it with an analogous model for subtype 1a that we had inferred previously (Quadeer et al., 2019a). Our comparative analysis suggests E2 1b to be more tolerant to mutations compared with E2 1a, indicating that it may be easier for subtype 1b viruses to evade immune responses. This is further corroborated by an analysis of the average time it takes for each subtype to escape from antibody responses, which we quantify by using population-genetics-based models of in-host viral evolution (Quadeer et al., 2019a). Our analysis, in general, points to significant differences in viral evolutionary constraints experienced by HCV subtypes 1a and 1b, which may contribute to observed subtype-specific differences in disease outcomes. We additionally employ the evolutionary models to identify potentially escape-resistant human monoclonal antibodies (HmAbs) against both subtypes 1a and 1b, which may assist in the rational design of a vaccine that is effective against the most prevalent HCV subtypes worldwide.

RESULTS

Inference and validation of the HCV E2 1b fitness landscape

We inferred a computational model for the fitness landscape of E2 1b using the sequence data available at the HCV-GLUE database (Singer et al., 2018, 2019) (see STAR Methods for details). This involved obtaining a maximum entropy (least-biased) probabilistic model—a "prevalence landscape" that captures the probability of observing a virus with a particular E2 protein sequence in circulation. In this model, the probability of any sequence $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ can be expressed as

$$P_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \frac{e^{-E_{\mathbf{h},\mathbf{J}}(\mathbf{x})}}{Z}, \text{ where } E_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(x_i, x_j) + \sum_{i=1}^{N} h_i(x_i), \qquad \text{(Equation 1)}$$

where $\mathbf{h}$ is the set of all fields that represent the effect of mutations at a single residue, $\mathbf{J}$ is the set of all couplings that represent the effect of interactions between mutations at two different residues, and $Z = \sum_{\mathbf{x}} e^{-E_{\mathbf{h},\mathbf{J}}(\mathbf{x})}$ is a normalization factor. The quantity $E_{\mathbf{h},\mathbf{J}}(\mathbf{x})$ represents the energy of sequence $\mathbf{x}$, which is inversely related to its prevalence. Inferring a maximum entropy model involves choosing the model parameters (fields and couplings) such that the single and double mutant probabilities obtained from the model match with those of the multiple sequence alignment (MSA). Maximum entropy models have been used for inferring the fitness landscape of HCV polymerase protein (Hart and Ferguson, 2015) and of several proteins of HIV (Barton et al., 2016; Ferguson et al., 2013; Flynn et al., 2017; Louie et al., 2018; Mann et al., 2014), and for designing a T cell-based HIV vaccine candidate, shown to be immunogenic in rhesus macaques (Murakowski et al., 2021). The maximum entropy model inference for E2 protein is challenging as it involves estimating a large number of model parameters, a consequence of the high mutational diversity of E2 compared to other HCV proteins (Figure S1). To tackle this problem, we used an efficient computational approach introduced in (Louie et al., 2018) to infer fitness landscapes of the HIV envelope protein. The method was also applied to infer a fitness landscape for the E2 protein of HCV subtype 1a (Quadeer et al., 2019a).

The single and double mutant probabilities obtained from our inferred model for HCV E2 subtype 1b aligned well with those of the MSA (Figures S2A and S2B). Other statistics, such as triple mutant probabilities and the distribution of the number of mutations computed from our inferred model, although not explicitly included while training, also matched well with those of the MSA (Figures S2C and S2D). These

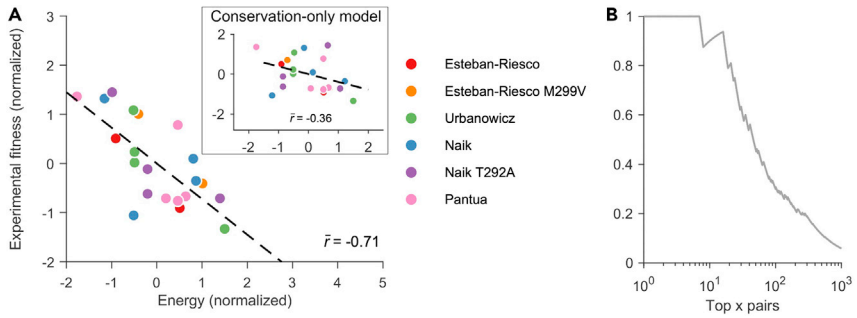


Figure 1. Validation of the inferred E2 1b fitness landscape

(A) Normalized experimental fitness measurements correlate strongly with the energy computed from the inferred landscape. In contrast, a much lower correlation is observed for the conservation-only model (Quadeer et al., 2019a) that does not take couplings into account (inset). References for fitness (infectivity) measurements are shown in the legend. Normalization of fitness measurements and predicted model energies were performed by subtracting the mean from each dataset and dividing by its standard deviation. In (Esteban-Riesco et al., 2013) and (Naik et al., 2017), E2 fitness measurements were reported in two different E1 backgrounds. In (Esteban-Riesco et al., 2013), one background involved methionine (M) at residue 299, while the other involved valine (V). Similarly, in (Naik et al., 2017), one background involved threonine (T) at residue 292, while the other involved alanine (A).

(B) Precision of contact predictions vs. the top x pairs according to the inferred model couplings. Precision is the proportion of top x pairs that are truly in contact. Two residues were assumed to be in contact if their carbon-alpha atoms were less than 8Å apart according to the available E2 1b crystal structure (PDB ID: 6MEI). See also Figure S3.

results indicate that our inferred prevalence landscape model accurately captures the statistical variations in the observed E2 1b sequence data.

We curated in-vitro infectivity measurements of subtype 1b strains from four experimental studies (Esteban-Riesco et al., 2013; Naik et al., 2017; Pantua et al., 2013; Urbanowicz et al., 2015) (see Data S1 for details). Comparing fitness predictions from the inferred model with these in-vitro infectivity measurements demonstrated that the inferred landscape for E2 1b is a reasonably good representative of the underlying protein fitness landscape (Figure 1A). Specifically, a strong negative Spearman correlation ($\bar{r} = -0.71$; see STAR Methods for details) was observed between the model-predicted energies (inversely related to prevalence; see Equation (1)) and measured fitness values. This accuracy is commensurate with fitness predictions reported by studies of other proteins (Ferguson et al., 2013; Flynn et al., 2017; Hart and Ferguson, 2015; Louie et al., 2018; Mann et al., 2014; Quadeer et al., 2019a).

In addition to mutations at individual residues, interactions between mutations at different residues have been shown to be important contributors to HCV fitness (Parera and Martinez, 2014; Quadeer et al., 2019a). Thus, we compared the predictions of our model, which takes into account residue interactions via couplings, with those of a model based only on amino acid conservation (or single mutant probabilities (Quadeer et al., 2019a)). Compared to our model, the conservation-only model provided a much lower correlation ($\bar{r} = -0.36$) between the model-predicted energy and in-vitro infectivity measurements (Figure 1A, inset), corroborating the importance of incorporating residue interactions in determining fitness. This suggests that the fitness effect of individual mutations in HCV also depends on the sequence background in which they are introduced, a phenomenon commonly called "epistasis" in genetics. Moreover, the strong couplings of maximum entropy models are known to be informative of contacts in the protein tertiary structure (Weigt et al., 2008; Dunn et al., 2007; Morcos et al., 2011). The inferred E2 1b model couplings were also found to be good predictors of contacts in the E2 1b protein structure [PDB ID: 6MEI] (Figure 1B; see STAR Methods for details). The contact prediction from the model couplings was comparable with direct coupling analysis (DCA) (Morcos et al., 2011), a standard contact prediction method (Figure S3).

Fitness models indicate subtype 1b to be less evolutionarily constrained than subtype 1a

We compared the fitness landscape model of HCV E2 1b with an analogous model developed previously for E2 1a (Quadeer et al., 2019a) to investigate whether the observed differences in disease outcome of the two subtypes' infections may be related to the associated fitness constraints. We first compared the fitness landscapes based on two standard metrics used for quantifying landscape ruggedness: autocorrelation

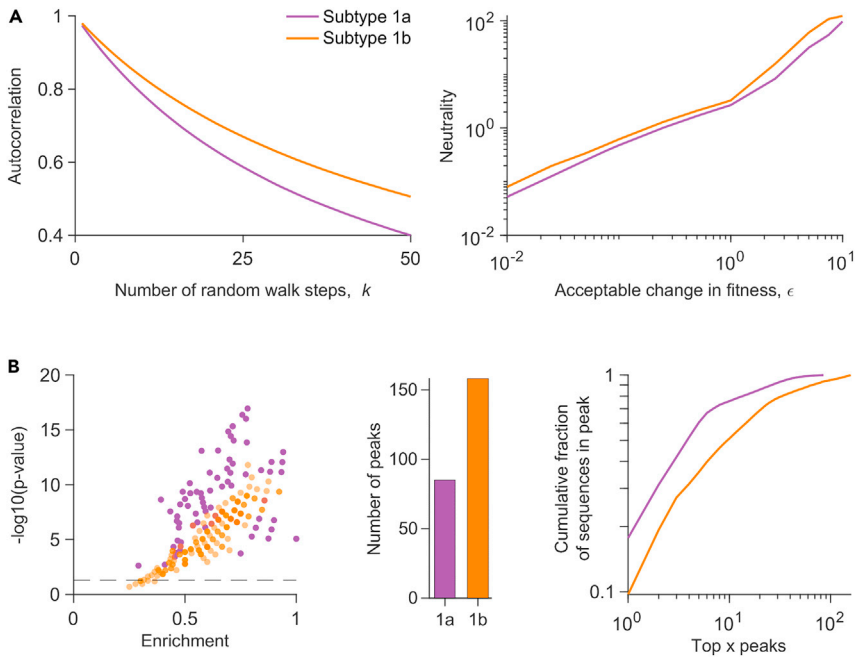


Figure 2. Comparison of the fitness landscapes of E2 1a and 1b

(A) (Left panel) Comparison of the autocorrelation of the sequence energies of each subtype's landscape. The x axis denotes the number of random walk steps (K) for which the autocorrelation was computed. The starting sequences were chosen within a Hamming distance (D) of 5 from the MSA (see STAR Method for details). (Right panel) Comparison of the neutrality of the landscape of each subtype. Neutrality was computed for $L = 500$ random walk steps. A random step in the walk was accepted only when the change in fitness from the sequence at the previous step was within a small value, ϵ (shown on the xaxis) (see STAR Methodfor details).

(B) (Left panel) Statistically significant enrichment of almost all peak sequences in known escape mutations (listed in Table S1), for each subtype. Enrichment is the fraction of known escape mutations within all mutations in a peak sequence, and the *p* value measures the probability of observing by random chance at least the observed number of escape mutations among all observed mutations in a peak sequence (see STAR Methods for details). The dashed horizontal line is the cut-off for statistically significant results (*p*value <0.05), with all peak sequences above this line being classed as statistically significant. (Middle panel) Number of peaks observed in each landscape. (Right panel) The cumulative fraction of sequences associated with the peaks versus the top x peaks in each landscape, plotted on a log-log scale. See also Figures S4–S7.

and neutrality (Kouyos et al., 2012; Quadeer et al., 2020; Vassilev et al., 2003). Autocorrelation quantifies the average change in fitness as one moves randomly along the landscape (see STAR Methods for details). We observed that the decay of autocorrelation of the E2 1b landscape was slower than that of 1a (Figure 2A, left panel), suggesting that the average change in fitness while moving along the landscape of 1b is less than that of 1a. The neutrality associated with each landscape quantifies the maximum number of mutation steps one can take on average without much change in fitness while moving randomly along the landscape (see STAR Methods for details). Our results showed that the E2 1b landscape was more neutral than 1a (Figure 2A, right panel), further suggesting that the average fitness cost upon mutation in 1b is lower than that of 1a. These results were qualitatively robust to the choice of the involved parameters (Figure S4).

We also compared the two landscapes using a low-dimensional representation that characterizes each landscape by its local "peaks", representing fitness maxima. Such local peaks have been shown to be informative of immune-escape pathways employed by HIV (Barton et al., 2015), and of the evolution of poliovirus under vaccine-induced and natural selective pressures (Quadeer et al., 2020). Peaks were obtained by following a steepest ascent walk from each sequence in the MSA. In the previous HIV study (Barton et al., 2015), sequences representing each peak in the landscape, referred to as "peak sequences", were found to be strongly enriched in HLA-associated mutations driven by host immune responses. Investigating the peak sequences of each subtype for HCV E2 revealed that almost all of the peak sequences of both subtypes were statistically significantly enriched in known escape mutations from E2-specific

HmAbs (Keck et al., 2008, 2009, 2016; Kato et al., 1993; Morin et al., 2012; Bailey et al., 2015; Velázquez-Moctezuma et al., 2019) (Figure 2B left panel; see STAR Methods for details). This suggests that peaks in the fitness landscapes of E2 may be representative of immune-escape pathways employed by HCV for both subtypes. This result was also qualitatively robust to the change of parameter values involved in the landscape inference (Figure S5). We observed that there were far more peaks in the landscape of 1b than 1a (Figure 2B, middle panel), suggesting that more escape pathways may be available for subtype 1b than 1a. Rank-ordering the peaks according to the fraction of MSA sequences represented by each peak revealed that the same number of top peaks in 1b comprised a smaller fraction of sequences than 1a (Figure 2B, right panel). This indicates that the sequences are more spread out across peaks in 1b, which is suggestive of relatively diverse escape pathways utilized by subtype 1b in comparison with 1a. These qualitative results were also robust to the change of parameter values involved in the landscape inference (Figure S6). We additionally confirmed, using a procedure similar to (Barton et al., 2015; Quadeer et al., 2020), that the peaks observed in each landscape are not an artifact of finite sampling, but arise owing to the interplay between mutations at different residues in each subtype (Quadeer et al., 2019a). Generally, the comparative analysis of the 1a and 1b fitness landscapes (Figure 2) was robust to finite sampling, returning qualitatively similar results even when fewer sequences were used for inferring the landscapes (Figure S7).

Overall, this comparison of fitness landscapes demonstrates that intrinsic fitness constraints governing the evolution of HCV subtypes may differ significantly. Our results are suggestive of subtype 1b to be under less fitness constraints compared with 1a, potentially making it easier for 1b to escape from antibody-mediated immune pressure. Such a difference between the two subtypes is not apparent by comparing simple sequence level statistics such as residue-wise entropies (Figure S8).

Evolutionary modeling and escape time prediction for subtype 1b

To further investigate whether the ease of escaping immune responses is subtype-dependent, we quantified the time it takes for E2 1b to escape immune responses using an in-host viral evolutionary model similar to the one we had employed previously for E2 1a (Quadeer et al., 2019a). This evolutionary model takes into account stochastic dynamics during in-host viral evolution, such as host-viral interactions, competition within the viral population, and multiple pathways that may be employed by the virus to escape from immune pressure. This was accomplished by incorporating the inferred E2 1b fitness landscape into a Wright–Fisher-like population genetics model (Ewens, 2004). The parameters involved in this model, such as mutation rate and effective size of the viral population, were set according to known values for HCV (Bull et al., 2011; Cuevas et al., 2009; Sanjuan et al., 2010), while the sequence survival probability in the population from one generation to the next was determined by the inferred E2 1b fitness landscape (see STAR Methods for details).

For each E2 1b residue, we used the evolutionary model to predict the time for an escape mutation to reach a majority in the population. Specifically, for calculating the escape time associated with a particular residue, we started the evolutionary simulation by first initializing the population with duplicates of a sequence randomly picked from the MSA having the consensus amino acid at the selected residue. Immune pressure was modeled as a fixed reduction in fitness of sequences having the consensus amino acid at the residue, thereby allowing a selective advantage to sequences that incur a mutation at that residue. We continued the simulation until the sequences having a mutation at the selected residue reached a majority in the population. We considered this as a representative marker of viral escape from the corresponding immune pressure. The number of generations for escape was recorded, and this procedure was repeated multiple times with the same initial sequence, as well as with multiple distinct initial sequences. The mean number of generations over all these simulation runs, termed "escape time", was computed and recorded for every E2 1b residue (see STAR Methods for details).

We validated the predicted escape times for E2 1b residues using experimental and clinical data. First, we assessed the ability of our model to predict known escape mutations from multiple E2-specific HmAbs (Keck et al., 2008, 2009, 2016; Kato et al., 1993; Morin et al., 2012; Bailey et al., 2015) (listed in Table S1). Such mutations would be expected to be associated with lower escape times compared to mutations at other residues, enabling the virus to escape the associated antibody pressure. Our results demonstrated that this was indeed the case ($P = 1.4 \times 10^{-19}$, Mann-Whitney test, Figure 3A). We also validated the escape time metric by predicting the escape times associated with mutating the buried and exposed (surface)

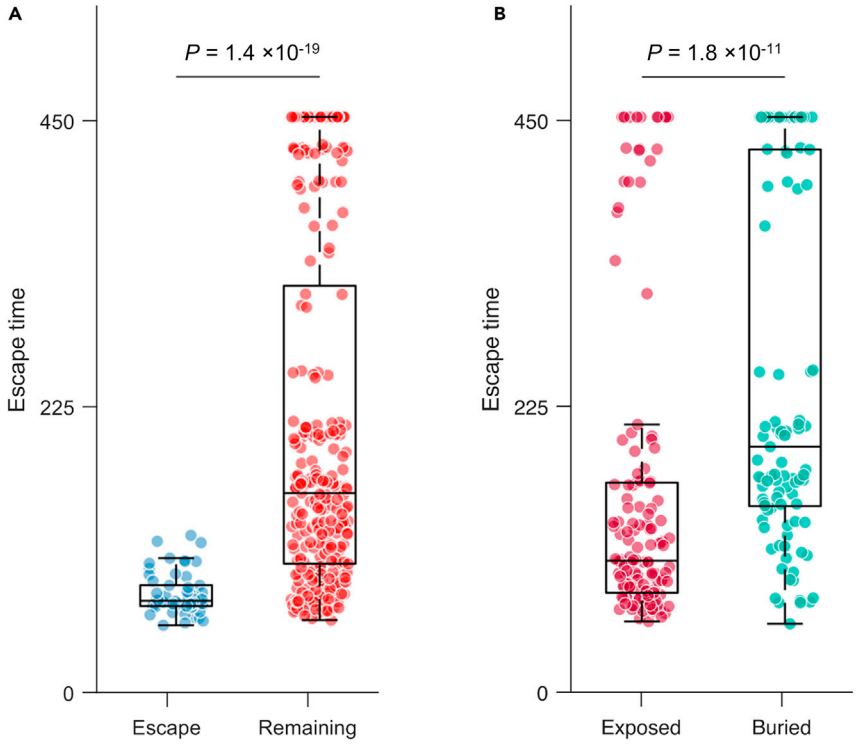


Figure 3. Validation of E2 1b escape times predicted using the in-host evolutionary model

(A) Comparison of escape times associated with the residues at which mutation is known to assist in escape from HmAbs (listed in Table S1) and those of the remaining E2 residues.

(B) Comparison of escape times associated with the mutations at exposed and buried residues. Residues of the available crystal structure of E2 1b (PDB ID: 6MEI) (Flyak et al., 2018) were classified as exposed or buried according to the standard RSA metric (see STAR methods for details). For each box plot, the middle line indicates the median, the edges of the box represent the first and third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges. All reported *P* values were calculated using the one-sided Mann-Whitney test.

residues in E2 1b, respectively. Buried residues that form the protein core are more likely to be crucial for stability (Chen and Zhou, 2005), implying that mutations at these residues would be expected to be associated with higher escape times than mutations at the exposed ones. Residues of the available crystal structure of E2 1b (PDB ID: 6MEI) (Flyak et al., 2018) were classified as exposed or buried according to the standard relative solvent accessibility (RSA) metric (see STAR Methods for details). Comparing the escape times associated with these two sets of residues revealed that the mutations at buried residues were predicted to have higher escape times ($P = 1.8 \times 10^{-11}$, Mann-Whitney test; Figure 3B), as expected. Overall, these tests provide confidence in the capability of the evolutionary model to distinguish E2 1b residues associated with low and high escape times.

Escape time comparison indicates subtype 1b may evade HmAbs more easily than subtype 1a

We compared the predicted escape times of the exposed residues of E2 1b (which are more likely to be targeted by HmAbs) with those previously predicted for E2 1a (Quadeer et al., 2019a). All relevant model parameters were matched to perform a fair comparison between the two subtypes (see STAR Methods for details). We found that the escape times of exposed residues of E2 1b were generally lower than those of 1a ($P = 9.5 \times 10^{-9}$, Mann-Whitney test; Figure 4A), suggesting that it may be easier for subtype 1b to escape from antibody responses than 1a. This was also true if all residues of both subtypes were considered ($P = 1.7 \times 10^{-13}$, Mann-Whitney test; Figure S9). To visualize the escape times associated with exposed residues, we superimposed them as a heatmap on the resolved crystal structure of E2 for each subtype (Figure 4B). Here, as in Figure 4A, the low concentration of residues with high escape time (colored in red) in subtype 1b compared with 1a is also evident. Moreover, we observed that few common exposed residues were associated with very high escape times for both subtypes (Figure S10). Interestingly, an epitope 412–425 comprising such residues (423 and 425) was shown in a recent study to be capable of inducing potent

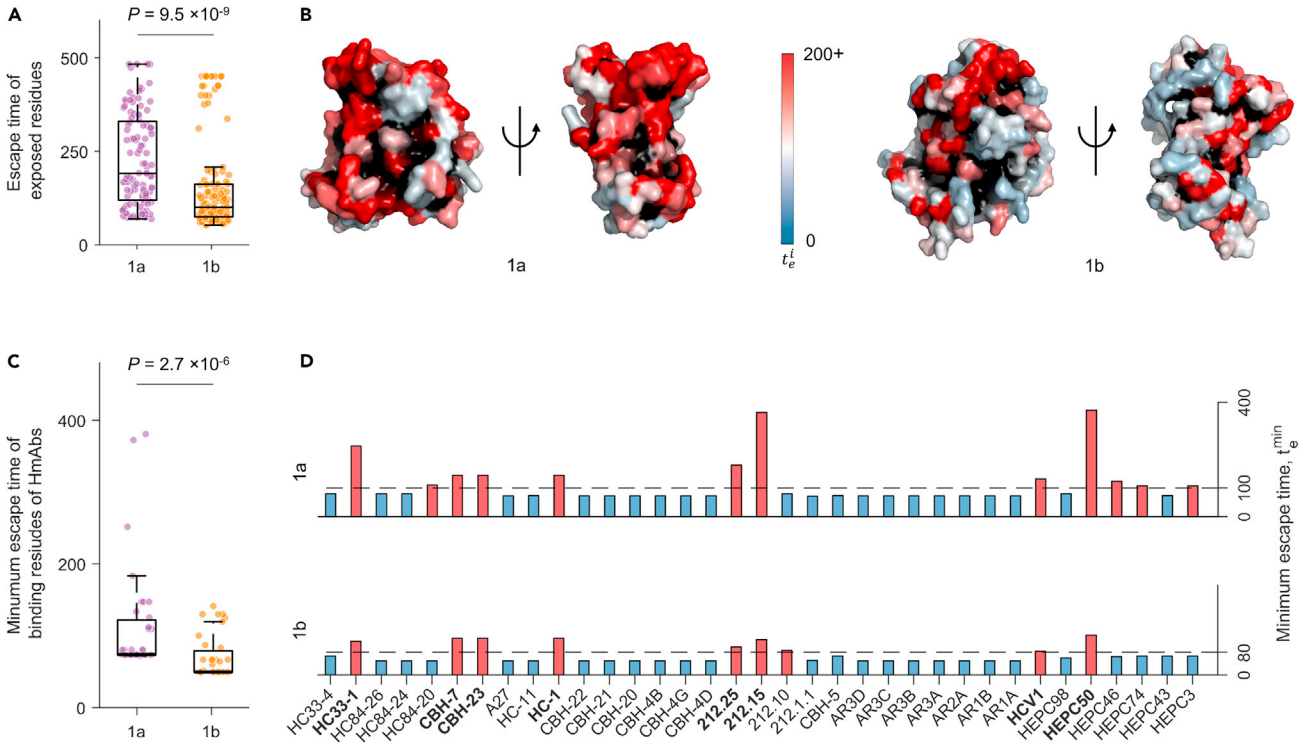


Figure 4. Escape time comparison of E2 subtype 1a and 1b

(A) Comparison of escape times associated with exposed residues of E2 of each subtype.

(B) Escape times of exposed residues superimposed on the crystal structure of E2 1a (left panel, PDB ID: 4MWF (Kong et al., 2013)) and E2 1b (right panel, PDB ID: 6MEI (Flyak et al., 2018)). Buried residues are shown in black.

(C and D) Comparison of the minimum escape time associated with mutating the binding residues of known E2-specific HmAbs for each subtype (C) taken together and (D) individually. Binding residues of an antibody were determined by global alanine scanning mutagenesis (Pierce et al., 2016; Gopal et al., 2017; Keck et al., 2019; Bailey et al., 2017), where each residue of the wild-type sequence was substituted by alanine (or glycine/serine if the residue in the wild-type was alanine). For the majority of antibodies (Pierce et al., 2016; Keck et al., 2019; Gopal et al., 2017), the fraction of the mutant sequence binding with respect to the wild-type sequence, called relative binding (RB), was reported for each residue, and we defined binding residues of any one of these antibodies as residues with RB less than or equal to 20%. For some specific antibodies (HEPC98, HEPC50, HEPC46, HEPC74, HEPC43, and HEPC3), critical binding residues as reported in (Bailey et al., 2017) were used. The dashed line for each subtype denotes the optimal cut-off value ζ (see STAR Methods for details). HmAbs predicted to be escape-resistant for any subtype are colored in red, while the remaining ones are colored in blue. The HmAbs predicted to be escape-resistant for both subtypes are shown in bold. In each box plot in (A and C), the middle line indicates the median, the edges of the box represent the first and third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges. All reported P values were calculated using the one-sided Mann-Whitney test.

See also Figures S9–S11.

neutralizing antibodies against multiple HCV genotypes including 1a and 1b (Czarnota et al., 2020). Hence, targeting these residues may assist in eliciting an effective immune response against both subtypes.

As the majority of the neutralizing antibodies block HCV entry (Harman et al., 2015), we also investigated the escape times of residues of each subtype that are involved in binding to CD81, the main HCV receptor (Ströh et al., 2018). We found that the escape times of these residues for subtype 1b were lower than those of 1a, suggesting that the entry of subtype 1b is more flexible and thereby making it potentially easier for 1b to escape CD81-targeting antibodies without significantly compromising viral entry (Figure S11).

We further employed our model to compare the two subtypes based on the escape resistance of known HmAbs. Specifically, we focused on HmAbs with binding residues determined using global alanine scanning experiments (Pierce et al., 2016; Keck et al., 2019; Gopal et al., 2017; Bailey et al., 2017). We compared the minimum escape time (t_e^{min}, see STAR Method) predicted by each model for the binding residues of 35 antibodies. Similar to Figure 4A, this analysis also showed that the minimum escape times predicted for each antibody for subtype 1b were statistically significantly lower than those predicted for 1a ($P = 2.7 \times 10^{-6}$, Mann-Whitney test; Figure 4C).

In addition, we considered a linear classifier (Quadeer et al., 2019a) to compute an optimal escape time cut-off value ζ for each subtype to determine whether a residue is relatively escape-resistant or not. This binary classifier was designed by taking the residues with known escape mutations (listed in Table S1) as true positives and all remaining residues as true negatives (see STAR Methods for details). We then assessed each antibody by comparing the minimum escape time predicted by each model with the corresponding optimal cut-off value ζ. Based on this analysis, we observed that fewer HmAbs were predicted to be difficult to escape for 1b than 1a (colored in red in Figure 4D). All the HmAbs predicted to be difficult to escape for 1b were also predicted to be the same for 1a, except for HmAb 212.10. However, the opposite was not true, further suggesting the potentially greater ability for subtype 1b to escape from HmAbs. Nonetheless, we identified eight HmAbs, HC33-1, CBH-7, CBH-23, HC-1, HEPC50, 212.25, 212.15, and HCV1, predicted to be difficult to escape for both subtypes (shown in bold). This result is consistent with information available for these HmAbs in the literature. Specifically, multiple studies have reported HC33-1 as a potentially escape-resistant antibody for multiple genotypes (Keck et al., 2014; Pierce et al., 2016) and HCV1 as a potent cross-reactive antibody (Kong et al., 2012; Broering et al., 2009). Other studies have isolated HmAbs HEPC50, 212.25, and 212.15 from patients who have spontaneously cleared HCV (subtypes 1a, 1b, or 3a), and these antibodies have also been shown to be cross-neutralizing (Keck et al., 2019; Bailey et al., 2017). Additionally, among the HmAbs that we predicted to be relatively easy to escape for both subtypes, HmAbs CBH-4D, CBH-4G, CBH-4B, CBH-20, CBH-21, CBH-22, AR1A, AR1B, and AR2A have been reported to be non-neutralizing or isolate-specific antibodies in previous studies (Keck et al., 2005; Kong et al., 2016; Law et al., 2008). As for HmAbs AR3A, AR3B, AR3C, and AR3D, there are contradicting results in the literature. Certain studies have reported them as potent broadly neutralizing antibodies (Law et al., 2008; Merat et al., 2019), whereas other studies have found escape mutations against these antibodies (Bailey et al., 2015; Velázquez-Moctezuma et al., 2019). Our results are consistent with the latter set of studies, as both 1a and 1b models predicted these antibodies to be relatively easy to escape.

DISCUSSION

Subtypes 1a and 1b are the two prevalent subtypes of HCV, together representing roughly half of all HCV infections worldwide. Clinical outcomes of subtype 1b infections have been shown to be considerably more severe than those of 1a, showing higher chronicity rate and higher risk of developing cirrhosis and HCC. The underlying reasons for these differences are not well understood. We reasoned that this may be due to subtype 1b being under relatively less fitness constraints than 1a, potentially enabling it to evade immune pressure more easily than 1a. We developed computational models to validate this hypothesis. First, by inferring a fitness landscape of HCV E2 1b and comparing it with an analogous model for 1a obtained previously, our analysis suggested that HCV 1b is subject to less fitness constraints. This result, taken more generally, showed that different subtypes of HCV may possess markedly distinct evolutionary properties. We incorporated the inferred fitness landscape of E2 1b into a population-genetics evolutionary model to predict the time it takes HCV to escape immune responses targeting any specific residue of E2 1b, and compared with the escape times predicted for E2 1a in our previous study (Quadeer et al., 2019a). The escape times of E2 1b were generally lower than those of 1a, suggesting that it may be easier for 1b to escape antibody responses.

The significance of incorporating interactions between mutations at different residues has been demonstrated in multiple previous studies for inferring protein fitness landscapes (Sohail et al., 2021; Hart and Ferguson, 2015; Ferguson et al., 2013; Barton et al., 2016; Flynn et al., 2017; Louie et al., 2018; Quadeer et al., 2019a, 2020) and for determining networks of residues important for mediating protein structure and function (Dahirel et al., 2011; Quadeer et al., 2014, 2018; Ahmed et al., 2019; Gaiha et al., 2019). However, residue–residue interactions, represented in the maximum entropy fitness landscape model by the coupling parameters (Equation 3 in STAR Methods), seem to be especially important for HCV E2 1b. This is evident from the significantly higher correspondence between our model predictions and in-vitro fitness (infectivity) measurements ($\bar{r} = -0.71$), compared with that obtained with a model without couplings ($\bar{r} = -0.36$) (Figure 1A). This difference is more apparent than observed previously for models of HIV proteins (Ferguson et al., 2013; Mann et al., 2014; Flynn et al., 2017; Louie et al., 2018) or other HCV proteins (Hart and Ferguson, 2015) (as well as E2 subtype 1a (Quadeer et al., 2019a)), suggesting that collective mutational effects, including compensatory mutations, may play a particularly dominant role in shaping the evolution of E2 1b and in providing coordinated evolutionary pathways to facilitate immune escape. Importantly, the incorporation of residue–residue interactions exposed evolutionary differences between the HCV subtypes 1a and 1b (Figure 2) that were not evident with site-independent models

(the single-site entropy of both subtypes appeared noticeably similar; Figure S8). Hence, the incorporation of collective mutational effects appears essential to elucidate the distinct evolutionary properties of the different HCV subtypes, which may in turn help to explain the different associated chronicity rates.

The differences in the fitness constraints of the two HCV subtypes that we report here do not seem to stem from differences in the quality of the respective model fit. This is evident from the similar statistical accuracy of the inferred model for each subtype (1a: Figure 1 in (Quadeer et al., 2019a) and 1b: Figure S2) and from nearly the same correspondence between model predictions and experimental fitness measurements (1a: −0.72 (Quadeer et al., 2019a) and 1b: −0.71 (Figure 1)). Additionally, the qualitative results related to the comparison of the two landscapes appeared robust to changes in the number of sequences as well as the specific parameter values used in inferring the landscapes (Figures S5–57).

Our results suggest that distinct fitness constraints of HCV subtype 1a and 1b may contribute to the observed differences in disease outcomes of these subtypes. However, other factors may also contribute to disparate disease outcomes, including host-specific factors such as patient's ethnicity, alcohol consumption, and co-infection with HIV (Chen and Morgan, 2006; Missiha et al., 2008; Yan and Wang, 2017). Immunogenicity of the viral subtypes may also play a role because if HCV subtype 1b was found to be less immunogenic than 1a, it could contribute to higher rates of chronic disease. However, there is little evidence to support this, with experimental studies in fact suggesting that the two subtypes have similar immunogenicity (Rodríguez-López et al., 1999). This was demonstrated by stimulating the sera from HCV-infected patients with peptides selected from regions of the HCV proteome having high inter-subtype variability. Similar immunogenicity of both subtypes is also consistent with experimentally identified HCV E2-specific B cell epitope data available from the Immune Epitope Database (IEDB; https://www.iedb.org) (Vita et al., 2018), which shows that the coverage of epitopes along the E2 primary structure is similar for both 1a and 1b (Figure S12). Based on current information, it appears unlikely that differences in immunogenicity is a key factor responsible for the disparate disease outcomes of HCV subtypes 1a and 1b.

While to our knowledge possible association between viral fitness constraints and disease outcome of HCV subtypes has not been elucidated previously, related results have been reported for HIV. Like HCV, HIV subtypes have also been shown to associate with different disease outcomes, defined based on the rate of $CD4^+$ T cell decline and speed of progression to AIDS (Easterbrook et al., 2010; Amornkul et al., 2013; Ssemwanga et al., 2013; Kiwanuka et al., 2013; Venner et al., 2016). Clinical studies on HIV have pointed to replicative fitness of the infecting subtype as an important factor associated with disease outcome (Venner et al., 2016; Claiborne et al., 2015). For instance, in (Venner et al., 2016), by following the decay of $CD4^+$ T cell counts in untreated patients infected with HIV subtype A, C, or D, patients infected with subtype D were reported to be associated with faster disease progression. Further, by performing dual HIV competition assays of strains isolated from patients infected with each subtype against the reference HIV-1 subtype B isolates, it was shown that the relative replicative fitness of subtype D strains was higher than that of the other two subtypes. Overall, these studies are supportive of our findings for HCV by providing evidence that associations between the viral subtype and disease progression may be explained, at least in part, by differences in subtype fitness constraints. Our study motivates targetedHCV experimental studies for further investigating the effects of subtype fitness constraints on disease outcome. These may include longitudinal studies of patients infected with HCV subtypes 1a and 1b, which compare the relative fitness (or infectivity) of the infecting strains as well as their association with chronicity or clearance.

Our escape time analysis of residues known to be important for CD81 binding showed that viral entry of subtype 1b may be more flexible than that of subtype 1a (Figure S11). That is, subtype 1b may be more capable of mutating CD81 binding residues while remaining fit. This would make it easier for subtype 1b to escape CD81-targeting antibodies, which form the majority of neutralizing antibodies (Harman et al., 2015). Thus, the potential ability of 1b to escape neutralizing antibodies while maintaining binding to CD81 may lead to more chronic infections compared with 1a. This predicted subtype-specific disparity in viral entry can be confirmed via targeted experiments such as performing site-directed mutagenesis of CD81 binding residues for subtypes 1a and 1b and measuring the difference in viral entry using CD81 binding assays.

The escape times predicted by our in-host evolutionary models identified five E2-specific HmAbs (HEPC3, HEPC74, HEPC46, HC84-20, and 212.10) that appear to be relatively easy to escape for exclusively one

subtype. Investigating the escape times associated with the binding residues of these HmAbs for both subtypes revealed that this distinction was due to a single residue in the footprint of each HmAb, for which mutational escape for one subtype was predicted to be relatively easy, but not for the other subtype (Figure S13). Specifically, residue 437 for HEPC3 and HEPC74, residue 546 for HEPC46, residue 608 for HC84-20, and residue 442 for 212.10 appear to be relatively easy to escape for one subtype only. Of these, residues 437, 546, and 608 are orthologous in E2 subtypes 1a and 1b. This suggests that the distinctive association of mutations at these residues with immune escape in each subtype may be due to the different amino acids preferred at these positions along with their corresponding mutational interactions with other residues in the protein. Thus, our study provides motivation for further experiments to systematically explore whether the difference in the ability to escape neutralizing antibodies is HCV genotype/subtype-specific. For instance, the sequences representing the top peaks in the fitness landscape of each subtype—which are locally most fit and enriched in antibody escape mutations (Figure 2B)—can serve as a good representative panel of HCV E2 sequences to test the neutralization patterns of antibodies.

We also identified eight HmAbs (HC33-1, CBH-7, CBH-23, HC-1, HEPC50, 212.25, 212.15, and HCV1) that appear relatively difficult to escape for both subtypes, and also identified specific exposed residues in E2 that are associated with high escape times for both subtypes (Figure S10). Incorporating the epitopes targeted by the identified escape-resistant HmAbs into rationally designed vaccines (Bailey et al., 2018; Sohail et al., 2020) may aid in eliciting an effective immune response against both HCV subtypes 1a and 1b, and eventually help to curb their global spread. Targeting the exposed yet difficult-to-escape E2 residues, e.g., by synthetic antibodies (Muyldermans, 2013; Sormanni et al., 2015; Zimmermann et al., 2018), may also present an effective therapeutic strategy. Our study motivates further investigation of the identified escape-resistant antibodies and difficult-to-escape exposed E2 residues in the context of HCV vaccines and therapies via targeted experiments.

Limitations of the study

There are multiple limitations of our study. We used the available HCV subtype 1b sequence data and *in vitro* fitness measurements for the inference and validation of our fitness landscape model, respectively (Figure 1). This available data for subtype 1b is limited compared with that of 1a. Thus, further sequencing efforts and fitness studies specific to HCV subtype 1b would assist in improving its computational modeling. Moreover, while our study suggests that subtype 1b is evolving under lower fitness constraints compared with 1a, it is difficult to resolve specific underlying mechanisms that lead to the observed disparity between the outcomes of the two subtypes. This is because fitness is a broad phenomenon which can be attributed to multiple factors, including the ability of the virus to infect host cells, bind to host receptor, transmit to a new host, etc. Our analysis suggests that compared with 1a, subtype 1b is more flexible in binding to the host receptor (Figure S11). However, the two subtypes may also differ in other mechanisms. Further, targeted experimental studies are needed to investigate this aspect.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data preprocessing
  - Inference of HCV E2 1b fitness landscape
  - Fitness verification
  - Residue-residue contact prediction
  - Metrics for comparing the ruggedness of fitness landscapes
  - Comparing the fitness landscapes based on the local peaks
  - In-host evolutionary model
  - Relative solvent accessibility
  - Estimating escape resistance of HmAbs
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103569.

ACKNOWLEDGMENTS

The authors were supported by the General Research Fund of the Hong Kong Research Grants Council (RGC) [Grant No. 16204519].

AUTHOR CONTRIBUTIONS

A.A.Q. and M.R.M. conceptualized the idea and designed the research. H.Z. performed the research and generated the figures and tables. All authors analyzed the data and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 2, 2021
Revised: November 19, 2021
Accepted: December 2, 2021
Published: January 21, 2022

REFERENCES

Ahmed, S.F., Quadeer, A.A., Morales-Jimenez, D., and McKay, M.R. (2019). Sub-dominant principal components inform new vaccine targets for HIV Gag. Bioinformatics *35*, 3884–3889.

Amornkul, P.N., Karita, E., Kamali, A., Rida, W.N., Sanders, E.J., Lakhi, S., Price, M.A., Kilembe, W., Cormier, E., Anzala, O., et al. (2013). Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-saharan africa. AIDS *27*, 2775–2786.

Amoroso, P., Rapicetta, M., Tosti, M.E., Mele, A., Spada, E., Buonocore, S., Lettieri, G., Pierri, P., Chionne, P., Ciccaglione, A.R., et al. (1998). Correlation between virus genotype and chronicity rate in acute hepatitis C. Hepatology *28*, 939–944.

Bailey, J.R., Barnes, E., and Cox, A.L. (2018). Approaches, progress, and challenges to hepatitis C vaccine development. Gastroenterology *156*, 418–430.

Bailey, J.R., Flyak, A.I., Cohen, V.J., Li, H., Wasilewski, L.N., Snider, A.E., Wang, S., Learn, G.H., Kose, N., Loerinc, L., et al. (2017). Broadly neutralizing antibodies with few somatic mutations and hepatitis C virus clearance. JCI Insight *2*, e92872.

Bailey, J.R., Wasilewski, L.N., Snider, A.E., El-Diwany, R., Osburn, W.O., Keck, Z., Foung, S.K., and Ray, S.C. (2015). Naturally selected hepatitis C virus polymorphisms confer broad neutralizing antibody resistance. J. Clin. Invest. *125*, 437–447.

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

Barton, J.P., Kardar, M., and Chakraborty, A.K. (2015). Scaling laws describe memories of host-pathogen riposte in the HIV population. Proc. Natl. Acad. Sci. U S A *112*, 1965–1970.

Broering, T.J., Garrity, K.A., Boatright, N.K., Sloan, S.E., Sandor, F., Thomas, W.D., Szabo, G., Finberg, R.W., Ambrosino, D.M., Babcock, G.J., et al. (2009). Identification and characterization of broadly neutralizing human monoclonal antibodies directed against the E2 envelope glycoprotein of hepatitis C virus. Virology *83*, 12473–12482.

Bruno, S., Crosignani, A., Maisonneuve, P., Rossi, S., Silini, E., and Mondelli, M.U. (2007). Hepatitis C virus genotype 1b as a major risk factor associated with hepatocellular carcinoma in patients with cirrhosis: a seventeen-year prospective cohort study. Hepatology *46*, 1350–1356.

Bull, R.A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S.T., Chopra, A., Cameron, B., Maher, L., Dore, G.J., White, P.A., et al. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog. *7*, 1–14.

Centers for Disease Control and Prevention (2018). Hepatitis C questions and answers for the public. https://www.cdc.gov/hepatitis/HCV/cfaq.htm.

Chen, H., and Zhou, H.-X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res. *33*, 3193–3199.

Chen, S.L., and Morgan, T.R. (2006). The natural history of hepatitis C virus (HCV) infection. Int. J. Med. Sci. *3*, 47–52.

Cho, Y.-K., Kim, Y.N., and Song, B.-C. (2014). Predictors of spontaneous viral clearance and outcomes of acute hepatitis C infection. Clin. Mol. Hepatol. *20*, 368–375.

Claiborne, D.T., Prince, J.L., Scully, E., Macharia, G., Micci, L., Lawson, B., Kopycinski, J., Deymier, M.J., Vanderford, T.H., Nganou-Makamdop, K., et al. (2015). Replicative fitness of transmitted HIV-1 drives acute immune activation, proviral load in memory CD4+ T cells, and disease progression. Proc. Natl. Acad. Sci. U S A *112*, E1480–E1489.

Cuevas, J.M., Gonzalez-Candelas, F., Moya, A., and Sanjuan, R. (2009). Effect of ribavirin on the mutation rate and spectrum of hepatitis C virus in vivo. Virology *83*, 5760–5764.

Czarnota, A., Offersgaard, A., Pihl, A.F., Prentoe, J., Bukh, J., Gottwein, J.M., Bieńkowska-Szewczyk, K., and Grzyb, K. (2020). Specific antibodies induced by immunization with hepatitis B virus-like particles carrying hepatitis C virus envelope glycoprotein 2 epitopes show differential neutralization efficiency. Vaccines *8*, 1–19.

Dahirel, V., Shekhar, K., Pereyra, F., Miura, T., Artyomov, M., Talsania, S., Allen, T.M., Altfeld, M., Carrington, M., Irvine, D.J., et al. (2011). Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc. Natl. Acad. Sci. U S A *108*, 11530–11535.

Dunn, S., Wahl, L., and Gloor, G. (2007). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics *24*, 333–340.

Easterbrook, P.J., Smith, M., Mullen, J., O'shea, S., Chrystie, I., Ruiter, A., Tatt, I.D., Geretti, A.M., and Zuckerman, M. (2010). Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. J. Int. AIDS Soc. *13*, 4.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. Phys. Rev. E *87*, 012707.

Esteban-Riesco, L., Depaulis, F., Moreau, A., Bacq, Y., Dubois, F., Goudeau, A., and Gaudy-Graffin, C. (2013). Rapid and sustained autologous neutralizing response leading to early spontaneous recovery after HCV infection. Virology *444*, 90–99.

Ewens, W.J. (2004). Mathematical Population Genetics (Springer).

Ferguson, A.L., Mann, J.K., Omarjee, S., Ndung'U, T., Walker, B.D., and Chakraborty, A.K. (2013). Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity *38*, 606–617.

Flyak, A.I., Ruiz, S., Colbert, M.D., Luong, T., Crowe, J.E., Bailey, J.R., and Bjorkman, P.J. (2018). HCV broadly neutralizing antibodies use a CDRH3 disulfide motif to recognize an E2 glycoprotein site that can be targeted for vaccine design. Cell Host Microbe *24*, 703–716.e3.

Flynn, W.F., Haldane, A., Torbett, B.E., and Levy, R.M. (2017). Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. Mol. Biol. Evol. *34*, 1291–1306.

Gaiha, G.D., Rossin, E.J., Urbach, J., Landeros, C., Collins, D.R., Nwonu, C., Muzhingi, I., Anahtar, M.N., Waring, O.M., Piechocka-Trocha, A., et al. (2019). Structural topology defines protective CD8+ T cell epitopes in the HIV proteome. Science *364*, 480–484.

Gopal, R., Jackson, K., Tzarum, N., Kong, L., Ettenger, A., Guest, J., Pfaff, J.M., Barnes, T., Honda, A., Giang, E., et al. (2017). Probing the antigenicity of hepatitis C virus envelope glycoprotein complex by high-throughput mutagenesis. PLoS Pathog. *13*, e1006735.

Gower, E., Estes, C., Blach, S., Razavi-Shearer, K., and Razavi, H. (2014). Global epidemiology and genotype distribution of the hepatitis C virus infection. J. Hepatol. *61*, S45–S57.

Harman, C., Zhong, L., Ma, L., Liu, P., Deng, L., Zhao, Z., Yan, H., Struble, E., Virata-Theimer, M.L., Zhang, P., et al. (2015). A view of the E2-CD81 interface at the binding site of a neutralizing antibody against hepatitis C virus. J. Virol. *89*, 492–501.

Hart, G.R., and Ferguson, A.L. (2015). Empirical fitness models for hepatitis C virus immunogen design. Phys. Biol. *12*, 066006.

Hwang, S.-J., Lee, S.-D., Lu, R.-H., Chu, C.-W., Wu, J.-C., Lai, S.-T., and Chang, F.-Y. (2001). Hepatitis C viral genotype influences the clinical outcome of patients with acute posttransfusion hepatitis C. J. Med. Virol. *65*, 505–509.

ICTV (2019). Table 1 - Confirmed HCV Genotypes/subtypes (May 2019). https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/634/table-1—confirmed-hcv-genotypes-subtypes-may-2019.

Jardine, J.G., Sok, D., Julien, J.-P., Briney, B., Sarkar, A., Liang, C.-H., Scherer, E.A., Henry Dunand, C.J., Adachi, Y., Diwanji, D., et al. (2016). Minimally mutated HIV-1 broadly neutralizing antibodies to guide reductionist vaccine design. PLoS Pathog. *12*, 1–33.

Kato, N., Sekiya, H., Ootsuyama, Y., Nakazawa, T., Hijikata, M., Ohkoshi, S., and Shimotohno, K. (1993). Humoral immune response to hypervariable region 1 of the putative envelope glycoprotein (gp70) of hepatitis C virus. Virology *67*, 3923–3930.

Keck, Z.-Y., Angus, A.G.N., Wang, W., Lau, P., Wang, Y., Gatherer, D., Patel, A.H., and Foung, S.K.H. (2014). Non-random escape pathways from a broadly neutralizing human monoclonal antibody map to a highly conserved region on the hepatitis C virus E2 glycoprotein encompassing amino acids 412-423. PLoS Pathog. *10*, 1–13.

Keck, Z.-Y., Girard-Blanc, C., Wang, W., Lau, P., Zuiani, A., Rey, F.A., Krey, T., Diamond, M.S., and Foung, S.K.H. (2016). Antibody response to hypervariable region 1 interferes with broadly neutralizing antibodies to hepatitis C virus. Virology *90*, 3112–3122.

Keck, Z.-Y., Li, S.H., Xia, J., von Hahn, T., Balfe, P., McKeating, J.A., Witteveldt, J., Patel, A.H., Alter, H., Rice, C.M., et al. (2009). Mutations in hepatitis C virus E2 located outside the CD81 binding sites lead to escape from broadly neutralizing antibodies but compromise virus infectivity. Virology *83*, 6149–6160.

Keck, Z.-Y., Li, T.-K., Xia, J., Bartosch, B., Cosset, F.-L., Dubuisson, J., and Foung, S.K.H. (2005). Analysis of a highly flexible conformational immunogenic domain a in hepatitis C virus E2. Virology *79*, 13199–13208.

Keck, Z.-Y., Olson, O., Gal-Tanamy, M., Xia, J., Patel, A.H., Dreux, M., Cosset, F.-L., Lemon, S.M., and Foung, S.K.H. (2008). A point mutation leading to hepatitis C virus escape from neutralization by a monoclonal antibody to a conserved conformational epitope. Virology *82*, 6067–6072.

Keck, Z.-Y., Pierce, B.G., Lau, P., Lu, J., Wang, Y., Underwood, A., Bull, R.A., Prentoe, J., Velázquez-Moctezuma, R., Walker, M.R., et al. (2019). Broadly neutralizing antibodies from an individual that naturally cleared multiple hepatitis C virus infections uncover molecular determinants for E2 targeting and vaccine design. PLoS Pathog. *15*, e1007772.

Kiwanuka, N., Ssetaala, A., Mpendo, J., Wambuzi, M., Nanvubya, A., Sigirenda, S., Nalutaaya, A., Kato, P., Nielsen, L., Kaleebu, P., et al. (2013). High HIV-1 prevalence, risk behaviours, and willingness to participate in HIV vaccine trials in fishing communities on Lake Victoria, Uganda. J. Int. AIDS Soc. *16*, 18621.

Kong, L., Giang, E., Nieusma, T., Kadam, R.U., Cogburn, K.E., Hua, Y., Dai, X., Stanfield, R.L., Burton, D.R., Ward, A.B., et al. (2013). Hepatitis C virus E2 envelope glycoprotein core structure. Science *342*, 1090–1094.

Kong, L., Giang, E., Robbins, J.B., Stanfield, R.L., Burton, D.R., Wilson, I.A., and Law, M. (2012). Structural basis of hepatitis C virus neutralization by broadly neutralizing antibody HCV1. Proc. Natl. Acad. Sci. U S A *109*, 9499–9504.

Kong, L., Lee, D.E., Kadam, R.U., Liu, T., Giang, E., Nieusma, T., Garces, F., Tzarum, N., Woods, V.L., Ward, A.B., et al. (2016). Structural flexibility at a major conserved antibody target on hepatitis C virus E2 antigen. Proc. Natl. Acad. Sci. U S A *113*, 12768–12773.

Kouyos, R.D., Leventhal, G.E., Hinkley, T., Haddad, M., Whitcomb, J.M., Petropoulos, C.J., and Bonhoeffer, S. (2012). Exploring the complexity of the HIV-1 fitness landscape. PLoS Genet. *8*, e1002551.

Law, M., Maruyama, T., Lewis, J., Giang, E., Tarr, A.W., Stamataki, Z., Gastaminza, P., Chisari, F.V., Jones, I.M., Fox, R.I., et al. (2008). Broadly neutralizing antibodies protect against hepatitis C virus quasispecies challenge. Nat. Med. *14*, 25–27.

Lee, M.-H., Yang, H.-I., Lu, S.-N., Jen, C.-L., You, S.-L., Wang, L.-Y., L'Italien, G., Chen, C.-J., Yuan, Y., and the REVEAL-HCV Study Group. (2014). Hepatitis C virus genotype 1b increases cumulative lifetime risk of hepatocellular carcinoma. Int. J. Cancer *135*, 1119–1126.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Social Psychol. *49*, 764–766.

Louie, R.H.Y., Kaczorowski, K.J., Barton, J.P., Chakraborty, A.K., and Mckay, M.R. (2018). Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. Proc. Natl. Acad. Sci. U S A *115*, E564–E573.

Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S.Y.W., Shapiro, B., Pybus, O.G., Allain, J.-P., and Hatzakis, A. (2009). The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. PLoS Med. *6*, 1–12.

Mann, J.K., Barton, J.P., Ferguson, A.L., Omarjee, S., Walker, B.D., Chakraborty, A., and Ndungu, T. (2014). The fitness landscape of HIV-1 Gag: Advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput. Biol. *10*, e1003776.

Merat, S.J., Bru, C., Berg, D.V.D., Molenkamp, R., Tarr, A.W., Koekkoek, S., Kootstra, N.A., Prins, M., Ball, J.K., Bakker, A.Q., et al. (2019). Cross-genotype AR3-specific neutralizing antibodies confer long-term protection in injecting drug users after HCV clearance. Hepatology *71*, 14–24.

Messina, J.P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G.S., Pybus, O.G., and Barnes, E. (2014). Global distribution and prevalence of hepatitis C virus genotypes. Hepatology *61*, 77–87.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. (1987). Interior and surface of monomeric proteins. J. Mol. Biol. *196*, 641–656.

Missiha, S.B., Ostrowski, M., and Heathcote, E.J. (2008). Disease progression in chronic hepatitis C: modifiable and nonmodifiable factors. Gastroenterology *134*, 1699–1714.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U S A *108*, E1293–E1301.

Morin, T.J., Broering, T.J., Leav, B.A., Blair, B.M., Rowley, K.J., Boucher, E.N., Wang, Y., Cheslock, P.S., Knauber, M., Olsen, D.B., et al. (2012). Human monoclonal antibody HCV1 effectively prevents and treats HCV infection in chimpanzees. PLoS Pathog. *8*, e1002895.

Murakowski, D.K., Barton, J.P., Peter, L., Chandrashekar, A., Bondzie, E., Gao, A., Barouch, D.H., and Chakraborty, A.K. (2021). Adenovirus-

vectored vaccine containing multidimensionally conserved parts of the HIV proteome is immunogenic in rhesus macaques. Proc. Natl. Acad. Sci. U S A *118*, e2022496118.

Muyldermans, S. (2013). Nanobodies: natural single-domain antibodies. Annu. Rev. Biochem. *82*, 775–797.

Naik, A.S., Owsianka, A., Palmer, B.A., O'Halloran, C.J., Walsh, N., Crosbie, O., Kenny-Walsh, E., Patel, A.H., and Fanning, L.J. (2017). Reverse epitope mapping of the E2 glycoprotein in antibody associated hepatitis C virus. PLoS One *12*, 1–20.

Osella, A.R., Misciagna, G., Guerra, V., Elba, S., Buongiorno, G., Cavallini, A., Di Leo, A., Sonzogni, L., Mondelli, M.U., and Silini, E.M. (2001). Hepatitis C virus genotypes and risk of cirrhosis in southern Italy. Clin. Infect. Dis. *33*, 70–75.

Pantua, H., Diao, J., Ultsch, M., Hazen, M., Mathieu, M., McCutcheon, K., Takeda, K., Date, S., Cheung, T.K., Phung, Q., et al. (2013). Glycan shifting on hepatitis C virus (HCV) E2 glycoprotein is a mechanism for escape from broadly neutralizing antibodies. J. Mol. Biol. *425*, 1899–1914.

Parera, M., and Martinez, M.A. (2014). Strong epistatic interactions within a single protein. Mol. Biol. Evol. *31*, 1546–1553.

Petruzziello, A., Marigliano, S., Loquercio, G., Cozzolino, A., and Cacciapuoti, C. (2016). Global epidemiology of hepatitis C virus infection: an up-date of the distribution and circulation of hepatitis C virus genotypes. World J. Gastroenterol. *22*, 7824.

Pierce, B.G., Keck, Z.-Y., Lau, P., Fauvelle, C., Gowthaman, R., Baumert, T.F., Fuerst, T.R., Mariuzza, R.A., and Foung, S.K.H. (2016). Global mapping of antibody recognition of the hepatitis C virus E2 glycoprotein: implications for vaccine design. Proc. Natl. Acad. Sci. U S A *113*, E6946–E6954.

Quadeer, A.A., Barton, J.P., Chakraborty, A.K., and Mckay, M.R. (2020). Deconvolving mutational patterns of poliovirus outbreaks reveals its intrinsic fitness landscape. Nat. Commun. *11*, 377.

Quadeer, A.A., Louie, R.H.Y., and Mckay, M.R. (2019a). Identifying immunologically-vulnerable regions of the HCV E2 glycoprotein and broadly neutralizing antibodies that target them. Nat. Commun. *10*, 2073.

Quadeer, A.A., Louie, R.H.Y., Shekhar, K., Chakraborty, A.K., Hsing, I.-M., and McKay, M.R. (2014). Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design. J. Virol. *88*, 7628–7644.

Quadeer, A.A., McKay, M.R., Barton, J.P., and Louie, R.H.Y. (2019b). MPF-BML: a standalone GUI-based package for maximum entropy model inference. Bioinformatics *36*, 2278–2279.

Quadeer, A.A., Morales-Jimenez, D., and McKay, M.R. (2018). Co-evolution networks of HIV/HCV are modular with direct association to structure and function. PLoS Comput. Biol. *14*, e1006409.

Raimondi, S., Bruno, S., Mondelli, M.U., and Maisonneuve, P. (2009). Hepatitis C virus genotype 1b as a risk factor for hepatocellular carcinoma development: a meta-analysis. Hepatology *50*, 1142–1154.

Rodríguez-López, M., Riezu-Boj, J.I., Ruiz, M., Berasain, C., Civeira, M.P., Prieto, J., and Borrás-Cuesta, F. (1999). Immunogenicity of variable regions of hepatitis C virus proteins: selection and modification of peptide epitopes to assess hepatitis C virus genotypes by ELISA. J. Gen. Virol. *80*, 727–738.

Rosen, H.R. (2011). Clinical practice. Chronic hepatitis C infection. N. Engl. J. Med. *364*, 2429–2438.

Rosenthal, E.S., and Graham, C.S. (2016). Price and affordability of direct-acting antiviral regimens for hepatitis C virus in the United States. Infect. Agents Cancer *11*, 24.

Rossi, C., Butt, Z.A., Wong, S., Buxton, J.A., Islam, N., Yu, A., Darvishian, M., Gilbert, M., Wong, J., Chapinal, N., et al. (2018). Hepatitis C virus reinfection after successful treatment with direct-acting antiviral therapy in a large population-based cohort. Hepatology *69*, 1007–1014.

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., and Belshaw, R. (2010). Viral mutation rates. Virology *84*, 9733–9748.

Silini, E., Bottelli, R., Asti, M., Bruno, S., Candusso, M., Brambilla, S., Bono, F., Iamoni, G., Tinelli, C., Mondelli, M., et al. (1996). Hepatitis C virus genotypes and risk of hepatocellular carcinoma in cirrhosis: a case-control study. Gastroenterology *111*, 199–205.

Singer, J., Thomson, E., Hughes, J., Aranday-Cortes, E., Mclauchlan, J., Filipe, A.D.S., Tong, L., Manso, C., Gifford, R., Robertson, D., et al. (2019). Interpreting viral deep sequencing data with GLUE. Viruses *11*, 323.

Singer, J.B., Thomson, E.C., McLauchlan, J., Hughes, J., and Gifford, R.J. (2018). GLUE: a flexible software system for virus sequence data. BMC Bioinformatics *19*, 532.

Sohail, M.S., Louie, R.H.Y., McKay, M.R., and Barton, J.P. (2021). MPL resolves genetic linkage in fitness inference from complex evolutionary histories. Nat. Biotechnol. *39*, 472–479.

Sohail, M.S., Quadeer, A.A., and McKay, M.R. (2020). How genetic sequence data can guide vaccine design. IEEE Potentials *39*, 31–37.

Sormanni, P., Aprile, F.A., and Vendruscolo, M. (2015). Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. Proc. Natl. Acad. Sci. U S A *112*, 9902–9907.

Ssemwanga, D., Nsubuga, R.N., Mayanja, B.N., Lyagoba, F., Magambo, B., Yirrell, D., Van der Paal, L., Grosskurth, H., and Kaleebu, P. (2013). Effect of HIV-1 subtypes on disease progression in rural Uganda: a prospective clinical cohort study. PLoS One *8*, 1–7.

Strimmer, K., and Haeseler, A.V. (2009). Genetic distances and nucleotide substitution models. In The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny, P. Lemey, M. Salemi, and A.-M. Vandamme, eds. (Cambridge University Press), pp. 112–113.

Ströh, L.J., Nagarathinam, K., and Krey, T. (2018). Conformational flexibility in the CD81-binding site of the hepatitis C virus glycoprotein E2. Front. Immunol. *9*, 1396.

Urbanowicz, R.A., Mcclure, C.P., Brown, R.J.P., Tsoleridis, T., Persson, M.A.A., Krey, T., Irving, W.L., Ball, J.K., and Tarr, A.W. (2015). A diverse panel of hepatitis C virus glycoproteins for use in vaccine research reveals extremes of monoclonal antibody neutralization resistance. Virology *90*, 3288–3301.

Vassilev, V.K., Fogarty, T.C., and Miller, J.F. (2003). Smoothness, ruggedness and neutrality of fitness landscapes: from theory to application. Nat. Comput. Ser. Adv. Evol. Comput. 3–44.

Velázquez-Moctezuma, R., Galli, A., Law, M., Bukh, J., and Prentoe, J. (2019). Hepatitis C virus escape studies of human antibody AR3A reveal a high barrier to resistance and novel insights on viral antibody evasion mechanisms. Virology *93*, e01909–e01918.

Venner, C.M., Nankya, I., Kyeyune, F., Demers, K., Kwok, C., Chen, P.-L., Rwambuya, S., Munjoma, M., Chipato, T., Byamugisha, J., et al. (2016). Infecting HIV-1 subtype predicts disease progression in women of sub-saharan africa. EBioMedicine *13*, 305–314.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2018). The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2008). Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. U S A *106*, 67–72.

World Health Organization (2021). Hepatitis C, fact sheet. https://www.who.int/news-room/fact-sheets/detail/hepatitis-c.

Wyles, D.L., and Luetkemeyer, A.F. (2017). Understanding hepatitis C virus drug resistance: clinical implications for current and future regimens. Top. Antivir. Med. *25*, 103–109.

Yan, Z., and Wang, Y. (2017). Viral and host factors associated with outcomes of hepatitis C virus infection. Mol. Med. Rep. *15*, 2909–2924.

Zimmermann, I., Egloff, P., Hutter, C.A., Arnold, F.M., Stohler, P., Bocquet, N., Hug, M.N., Huber, S., Siegrist, M., Hetemann, L., et al. (2018). Synthetic single domain antibodies for the conformational trapping of membrane proteins. eLife *7*, e34317.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MPF-BML standalone package	Louie et al. (2018); Quadeer et al. (2020)	https://github.com/ahmedaq/MPF-BML-GUI
DCA	Morcos et al. (2011)	http://dca.rice.edu/portal/dca/home
PyMOL	Schrödinger, Inc.	https://www.pymol.org
Other		
Fitness landscape and the mean escape time predicted by the in-host evolutionary model for each residue of E2 subtype 1a	Quadeer et al. (2019a)	https://github.com/ahmedaq/HCV-E2
The E2 subtype 1b infectivity measurements, used for validating the fitness landscape model	This manuscript	Data S1
Accession numbers of E2 subtypes 1b sequences used for inferring the model	This manuscript	Data S2
The mean escape time predicted by the in-host evolutionary model for each residue of E2 subtype 1b	This manuscript	Data S3
Data and scripts for reproducing the results reported in the manuscript	This manuscript	https://github.com/hangzhangust/HCVE21a1b

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Matthew R. McKay (matthew.mckay@unimelb.edu.au).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data used in this work has been provided in the supplemental information and is publicly available as of the date of publication. The E2 1b infectivity measurements, used for validating the fitness landscape model, are included in Data S1. Accession numbers of E2 1b sequences used for inferring the model are listed in Data S2. The mean escape time predicted by the in-host evolutionary model for each residue of E2 1b is provided in Data S3.
- Data and scripts for reproducing the results are available at GitHub: https://github.com/hangzhangust/HCVE21a1b_Hang. The GUI-based software implementation of the MPF-BML method (Louie et al., 2018), used for inferring the fitness landscape parameters, is available at GitHub: https://github.com/ahmedaq/MPF-BML-GUI (Quadeer et al., 2019b). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data preprocessing

We downloaded 1,559 aligned HCV E2 subtype 1b amino acid sequences (genome coverage $\geq$ 99%) from the HCV-GLUE database (http://hcv.glue.cvr.ac.uk) (Singer et al., 2018, 2019). To exclude any outlying sequences, we constructed a pairwise similarity matrix (1559 $\times$ 1559) of the sequences (Strimmer and Haeseler, 2009), with each (i, j)th entry representing the fraction of residues at which the sequence i and

sequence j are identical. By investigating the first two principal components (PCs) of this matrix, we excluded 260 outlying sequences which appeared at more than 3 scaled median absolute deviations (Leys et al., 2013) away from the median of either the first or second PC. We also excluded 16 sequences from chimpanzees and 162 sequences having no patient information. This filtering procedure resulted in a total of $M = 1121$ sequences from $W = 579$ patients. In addition, to control residue quality, we excluded 43 residues that were fully conserved. Thus, the final processed MSA comprised $M = 1121$ sequences and $N = 320$ residues.

Inference of HCV E2 1b fitness landscape

We inferred a maximum entropy model, i.e., the "prevalence landscape", for E2 1b to serve as a representative of its underlying fitness landscape. The maximum entropy model is a least-biased model that can reproduce the single and double mutant probabilities of the MSA, defined as

$$f_i(a) = \frac{1}{W}\sum_{k=1}^{M} w_k \delta\left(x_i^{(k)}, a\right)$$
$$f_{ij}(a,b) = \frac{1}{W}\sum_{k=1}^{M} w_k \delta\left(x_i^{(k)}, a\right)\delta\left(x_j^{(k)}, b\right). \quad \text{(Equation 2)}$$

Here, $x_i^{(k)}$ is the ith residue of the kth sequence from the MSA which takes on a value from either the consensus amino acid ($x_i^{(k)} = 0$) or the mth most frequently observed mutant ($x_i^{(k)} = m$) for $m = 1, \ldots, q_i$, where q_i denotes the number of mutants at residue i. δ is the Kronecker delta function, $\delta(a,b) = 1$ if $a = b$ and 0 otherwise, and w_k is one divided by the number of sequences contributed by the patient from which sequence k was obtained. For a given sequence $\mathbf{x} = [x_1, x_2, \ldots, x_N]$, the maximum entropy model assigns the probability

$$P_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \frac{e^{-E_{\mathbf{h},\mathbf{J}}(\mathbf{x})}}{Z}, \text{ where } E_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(x_i, x_j) + \sum_{i=1}^{N} h_i(x_i), \quad \text{(Equation 3)}$$

where $\mathbf{h}$ is the set of all fields that represent the effect of mutations at a single residue, and $\mathbf{J}$ is the set of all couplings that represent the effect of interactions between mutations at two different residues. $Z = \sum_{\mathbf{x}} e^{-E_{\mathbf{h},\mathbf{J}}(\mathbf{x})}$ is a normalization factor, and $E_{\mathbf{h},\mathbf{J}}(\mathbf{x})$ represents the energy of sequence $\mathbf{x}$. The fields $\mathbf{h}$ and couplings $\mathbf{J}$ are chosen such that the single and double mutant probabilities obtained from the model match the single and double mutant probabilities of the MSA, i.e.,

$$f_i(a) = \sum_{\mathbf{x}} \delta(x_i, a) P_{\mathbf{h},\mathbf{J}}(\mathbf{x})$$
$$f_{ij}(a,b) = \sum_{\mathbf{x}} \delta(x_i, a)\delta(x_j, b) P_{\mathbf{h},\mathbf{J}}(\mathbf{x}). \quad \text{(Equation 4)}$$

The problem of inferring the model parameters can be cast as the following convex optimization problem

$$(\mathbf{h}^*, \mathbf{J}^*) = \arg\min_{\mathbf{h},\mathbf{J}} \mathrm{KL}(P_0 \,||\, P_{\mathbf{h},\mathbf{J}}) = \arg\min_{\mathbf{h},\mathbf{J}} \sum_{\mathbf{x}} P_0(\mathbf{x}) \ln \frac{P_0(\mathbf{x})}{P_{\mathbf{h},\mathbf{J}}(\mathbf{x})}, \quad \text{(Equation 5)}$$

where $\mathrm{KL}(\bullet \,||\, \bullet)$ denotes the Kullback-Leibler divergence between probability distributions, and $P_0(\mathbf{x}) = \frac{1}{W}\sum_{k=1}^{M} w_k \delta(\mathbf{x}^{(k)}, \mathbf{x})$ is the patient-weighted probability of observing strain $\mathbf{x}$ in the MSA.

As E2 1b is a long protein and has a high mean residue entropy, the total number of parameters to estimate is very large (Figure S1). To solve this problem, we considered the inference framework, MPF-BML, introduced in (Louie et al., 2018). Specifically, we inferred the model parameters using the GUI-based software implementation of MPF-BML (Quadeer et al., 2019b). The parameters that we used for model inference are as follows: (i) The sample weights were set according to w_k in Equation (2); (ii) both L_1 and L_2 regularization parameters were set to 30 for couplings and 10^{-3} for fields, respectively; (iii) the termination condition was set to $\epsilon_1 \leq 2.5$ and $0.7 \leq \epsilon_2 \leq 1.3$; and (iv) all other parameters were set to their default values. The statistics of the E2 1b model inferred using these parameters aligned well with those obtained from the MSA (Figure S2).

Fitness verification

We used in-vitro experimental fitness measurements compiled from the literature (Esteban-Riesco et al., 2013; Urbanowicz et al., 2015; Naik et al., 2017; Pantua et al., 2013) to validate that our inferred prevalence

landscape is a good proxy of the E2 1b fitness landscape. The detailed selection procedure of the specific fitness measurements (listed in Data S1) from each study is presented below.

In (Esteban-Riesco et al., 2013), the authors compared in Figure 3A infectivity (in RLU) of subtype 1b HCVpps, subtype 1a reference strain, positive and negative controls. We selected all four measurements for subtype 1b and divided them into two groups, with the same E1 background for each group. In (Urbanowicz et al., 2015), the authors compared in Figure 2 infectivity (in RLU) of subtype 1b HCVpps. We selected all eight measurements and divided them into two groups, with the same E1 background for each group. In (Naik et al., 2017), the authors compared in Figure 1B infectivity (in RLU) of HCVpps from multiple genotypes. We selected four subtype 1b measurements, the only ones that have the same E1 background. In (Pantua et al., 2013), the authors compared in Figure 1F infectivity (in RLU) of HCVpps from multiple genotypes. We selected all five subtype 1b measurements.

For each study, normalization of fitness measurements and predicted model energies was performed by subtracting the mean from the dataset and dividing by its standard deviation. To further normalize for potential experimental biases, we considered the weighted average of Spearman correlation coefficients from different experiments. This is given by

$$\bar{r} = \frac{\sum_{i=1}^{s_{exp}} S_i r_i}{\sum_{i=1}^{s_{exp}} S_i},$$

where S_i is the number of measurements and r_i the Spearman correlation coefficient for experiment i, and s_{exp} is the total number of experiments.

Residue-residue contact prediction

To investigate if the inferred maximum entropy model couplings can predict residue-residue tertiary contacts in the E2 1b protein, we considered the Frobenius norm F_{ij} of the corresponding coupling matrix J_{ij} between residue i and j, a common measure of the strength of residue-residue connections (Weigt et al., 2008; Ekeberg et al., 2013). We then applied an average product correction (APC) to suppress phylogenetic bias and finite sampling effects (Dunn et al., 2007; Ekeberg et al., 2013). These are given by:

$$F_{ij}^{APC} = F_{ij} - \frac{F_i F_j}{F}, \qquad \text{(Equation 6)}$$

where

$$F_{ij} = \sqrt{\sum_{a,b} J_{ij}(a,b)^2},$$

and

$$F_i = \frac{1}{N-1}\sum_{j,j\neq i}^{N} F_{ij}$$

$$F = \frac{1}{N^2-N}\sum_{i,j,i\neq j}^{N} F_{ij}.$$

As residues in contact are likely to be strongly coupled, we expected the quantity calculated in Equation (6) between these residue pairs to be large. The true contacts were determined from the available E2 1b crystal structure (PDB ID: 6MEI) (Flyak et al., 2018), where two residues were assumed to be in contact if their carbon-alpha atoms were less than 8Å apart.

Metrics for comparing the ruggedness of fitness landscapes

We adopted two metrics to compare the fitness landscape of E2 1b with E2 1a (Kouyos et al., 2012).

Autocorrelation. Autocorrelation is used to quantify the average change in fitness as one moves randomly along the landscape. Starting with $M_c = 10^6$ random sequences chosen within a Hamming distance $D = \{5, 30\}$ from the MSA, we performed a 50-step random walk along the landscape starting from each sequence. The autocorrelation of the sequence energies at the kth step was calculated as

$$a_k = \frac{\langle E(\boldsymbol{x}^0)E(\boldsymbol{x}^k)\rangle_{M_c} - \langle E(\boldsymbol{x}^0)\rangle_{M_c}\langle E(\boldsymbol{x}^k)\rangle_{M_c}}{\sqrt{\langle E(\boldsymbol{x}^0)^2\rangle_{M_c} - \langle E(\boldsymbol{x}^0)\rangle_{M_c}^2}\sqrt{\langle E(\boldsymbol{x}^k)^2\rangle_{M_c} - \langle E(\boldsymbol{x}^k)\rangle_{M_c}^2}}, \quad k = 1, 2, \ldots, 50,$$

where $\boldsymbol{x}^k$ is the sequence at the kth step, and $\langle \bullet \rangle_{M_c}$ is the average over M_c sequences.

Neutrality. Neutrality quantifies the maximum number of steps (measured in terms of mutations) that one can take in a landscape without much change in fitness, i.e., the resulting difference in fitness remains within a small value ϵ. We started with $M_c = 10^6$ random sequences chosen within a Hamming distance $D = 5$ from the MSA. For each starting sequence $\boldsymbol{x}^0$, we performed an $L = \{500, 1000\}$ random walk along the landscape. At each step k, we accepted the new sequence $\boldsymbol{x}^{k'}$ generated by the random walk if the difference in fitness between the new sequence and the current sequence was less than ϵ. Otherwise, we kept the current sequence, i.e., $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k$. We calculated the Hamming distance d^k between the sequence $\boldsymbol{x}^k$ and the starting sequence $\boldsymbol{x}^0$. The neutrality was then calculated as the average over the maximum Hamming distances obtained from all M_c random sequences for a particular ϵ, i.e., $\langle \max_k d^k \rangle_{M_c}$.

Comparing the fitness landscapes based on the local peaks

As the local fitness maxima of the landscape, known as local "peaks", have been shown to be representative of immune escape pathways of HIV (Barton et al., 2015), we studied the peak structure of the inferred landscapes of subtype 1a and 1b. For each sequence present in the MSA, we obtained the corresponding local peak using the following procedure: For a given sequence of the MSA, we compared the energies of all its neighboring sequences (defined as one Hamming distance away from the sequence), and then chose the most fit sequence (i.e., the sequence with the lowest energy). We repeated the above procedure until the "peak sequence", i.e., the sequence which has higher fitness than all of its neighbors, was reached. The obtained unique number of local maxima represented the number of local peaks in the specific landscape.

We calculated the statistical significance of enrichment of known escape mutations (listed in Table S1) in the peak sequences using a P value. The enrichment of a peak sequence is defined as the fraction of escape mutations among all mutations defining that peak sequence. The P value corresponds to the probability of observing at least i mutations out of j escape mutations in a peak sequence, where there are n total mutations defining the peak sequence out of 363 total residues of the E2 protein. Mathematically, this can be written as

$$P = \sum_{q=i}^{\min(j,n)} \frac{\binom{j}{q}\binom{363-j}{n-q}}{\binom{363}{n}}.$$

We tested the null hypothesis that these i escape mutations were observed in a peak sequence by a random chance, and it was rejected if P<0.05.

In-host evolutionary model

We considered a population genetics viral evolutionary model similar to (Quadeer et al., 2019a) (which drew upon an earlier work (Barton et al., 2016)) for quantifying the ease of escape from antibody responses for each residue in E2 1b. This was accomplished using the "escape time" metric, which represents the number of generations required for mutations at a residue under immune pressure to reach a frequency >0.5 in a fixed-sized virus population.

Specifically, we employed a well-known population genetics Wright-Fisher model (Ewens, 2004). In this model, sequences in the population undergo mutation, selection, and random sampling steps in each generation. The virus population size was fixed at $M_e = 2000$, in line with the known HCV effective population size in in-host evolution (Bull et al., 2011). For a given E2 1b residue i, we started the simulation with a homogeneous population comprising copies of a sequence randomly selected from the MSA sequences having the consensus amino acid at residue i. In the mutation step, each nucleotide in the sequences was randomly mutated to another nucleotide with a fixed probability $\mu = 10^{-4}$, in line with the HCV mutation rate reported in (Cuevas et al., 2009; Sanjuan et al., 2010). In the selection step, the survival probability of each sequence in the population was calculated based on its fitness predicted from the inferred landscape (see Quadeer et al., 2019a for details). To model the immune pressure at residue i, the fitness of

all sequences having the consensus amino acid at residue i was decreased by a fixed value b, thereby providing a selective advantage to the sequences having a mutation at this residue. Similar to (Quadeer et al., 2019a), b was set according to the largest value of the field parameter in the inferred landscape. In the sampling step, the new generation of the population was generated through a standard multinomial sampling process parameterized by the survival probabilities calculated in the previous step and M_e. These three steps (mutation, selection and random sampling) were repeated until the number of sequences having a mutation at residue i reached a majority in the population and the number of generations was recorded. This number was considered a representative of the time (generations) taken by the virus to escape the immune pressure at residue i. We repeated this procedure multiple times using the same initial sequence, as well as multiple distinct initial sequences. The escape time t_e^i associated with residue i was calculated by averaging all number of generations over all these simulation runs.

For a fair comparison of the predicted escape times of E2 1b with those of E2 1a, we used the same parameter values in running the evolutionary simulations for both subtypes. Specifically, we used the same value for the parameter involved in mapping the predicted fitness of a sequence to its survival probability in the population for both 1a and 1b subtypes (see Quadeer et al., 2019a). Moreover, simulations were run for both subtypes for the same number of generations (500), the same number of distinct transmitted/founder (T/F) sequences (25) for each residue, and the same number of simulation runs (100) for each T/F sequence.

Relative solvent accessibility

To determine whether a residue in the crystal structure of each subtype is exposed or buried, we used the get_area( ) function in the PyMOL software (www.pymol.org) with a 1.4 solvent radius parameter to assign each residue in the crystal structure of a subtype (subtype 1a: PDB ID: 4MWF (Kong et al., 2013); subtype 1b: PDB ID: 6MEI (Flyak et al., 2018)) with a solvent accessible surface area (SASA). We then obtained the RSA values of each residue by normalizing the respective SASA values per residue in a Gly-X-Gly tripeptide construct (Miller et al., 1987). As suggested in (Jardine et al., 2016), residues with RSA >0.2 were considered as exposed, while the remaining residues were considered buried.

Estimating escape resistance of HmAbs

To compare the escape resistance of known E2-specific HmAbs against each subtype (Figures 4C and 4D), we computed the minimum escape time associated with their binding residues (determined using global alanine scanning experiments (Pierce et al., 2016; Keck et al., 2019; Gopal et al., 2017; Bailey et al., 2017)). The minimum escape time $t_e^{\min}$ associated with an antibody was defined as

$$t_e^{\min} = \min_i \; t_e^i,$$

where i is selected from the set of binding residues of that antibody. To distinguish whether an E2 residue is relatively escape-resistant or not based on its predicted escape time, we designed a binary classifier using the information of known E2-specific escape mutations (listed in Table S1). Specifically, we considered a classifier that takes the residues with known escape mutations as true positives and all remaining residues as true negatives. The classifier for subtype 1b achieved an area under the receiver operating characteristic curve (AUC) of 0.88 (Figure S14A). We chose the optimal cut-off value $\zeta \sim 80$ (Figure S14B) based on the maximum F1 score and the maximum Matthews correlation coefficient (MCC)—two commonly used metrics to evaluate the performance of a binary classifier with different thresholds. For subtype 1a, ζ was determined to be 100 using a similar statistical analysis in (Quadeer et al., 2019a). The HmAbs were classified as relatively escape-resistant for a subtype if their corresponding $t_e^{\min}$ was greater than ζ for that subtype, and vice versa.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using MATLAB (R2019b). Fisher's exact test was used to compute the statistical significance associated with enrichment of escape mutations among peaks in fitness landscapes of subtypes 1a and 1b. All other P values in each figure were calculated by performing one-sided Mann-Whitney tests. A P value less than 0.05 was considered as statistically significant.