iScience, Volume *25*

Supplemental information

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

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

Supplementary Information

Table of contents

Supplementary Figures S1-S14

Supplementary Table S1

Supplementary Figures

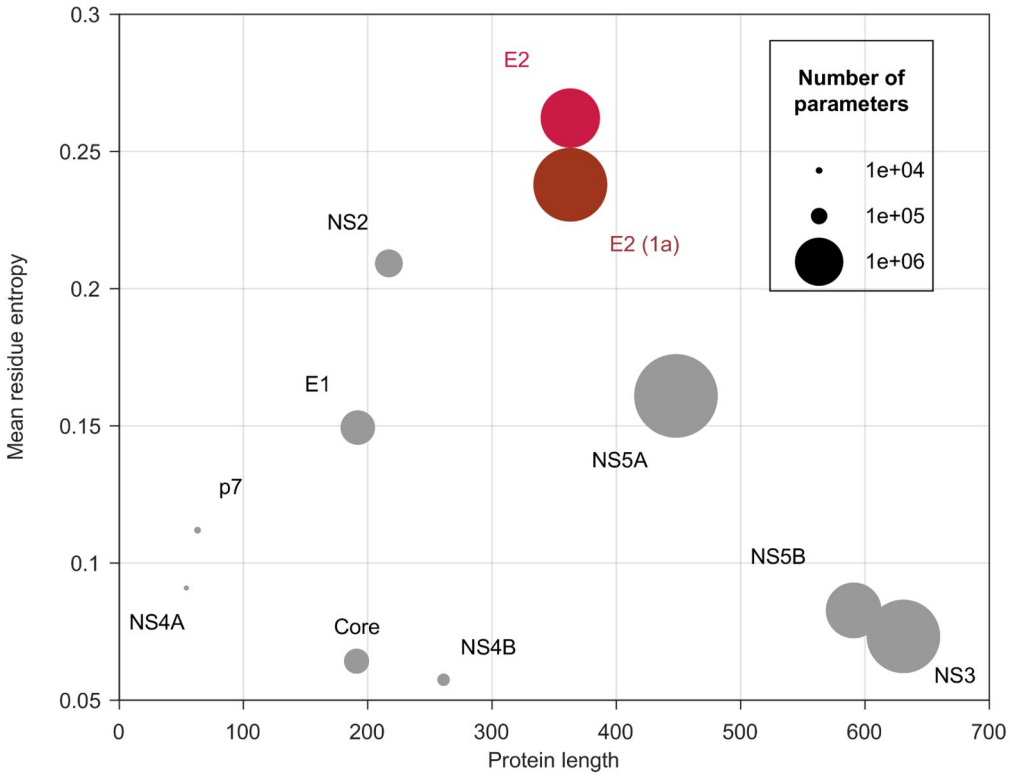

Fig. S1: **Comparison of the number of parameters required to estimate a fitness landscape for different HCV proteins of subtype 1b (related to Fig. 1).** E2 is a long protein with the highest mean residue entropy among HCV proteins. Thus inferring its landscape requires estimating a large number of model parameters. Mean residue entropy H can be calculated as $H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{a=1}^{20} f_i(a)\ln f_i(a)$, where $f_i(a)$ is the frequency of observing amino acid a at residue i, and N is the number of residues in that protein. The number of parameters was calculated by considering all amino acid mutants observed at each residue. The corresponding numbers for E2 subtype 1a have also been included for reference.

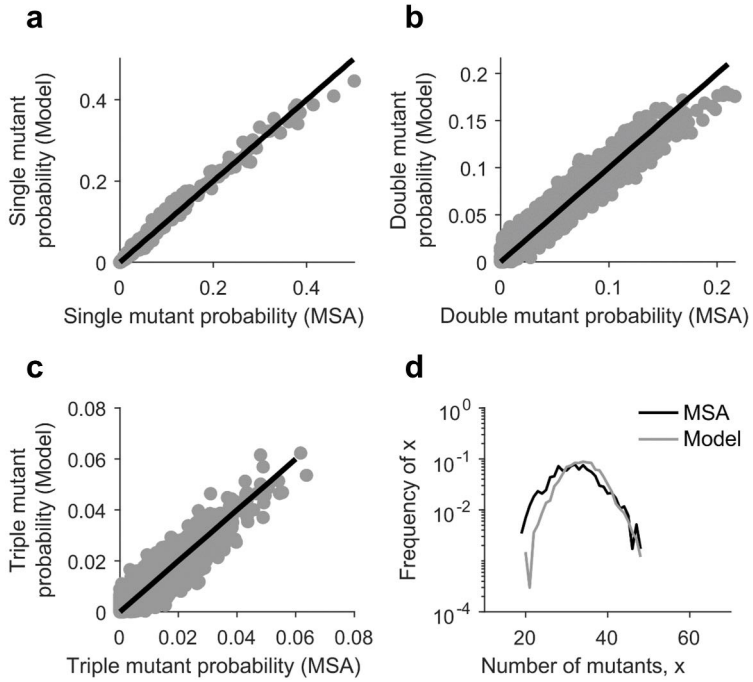

Fig. S2: **Statistical validation of the inferred E2 subtype 1b landscape (related to Fig. 1).** Comparison of the (**a**) single mutant probabilities, (**b**) double mutant probabilities, (**c**) triple mutant probabilities, and (**d**) distribution of the number of mutants per sequence obtained from the MSA and those predicted by the inferred model. Samples were generated from the inferred model using a Markov Chain Monte Carlo (MCMC) procedure (Ferguson et al., 2013).

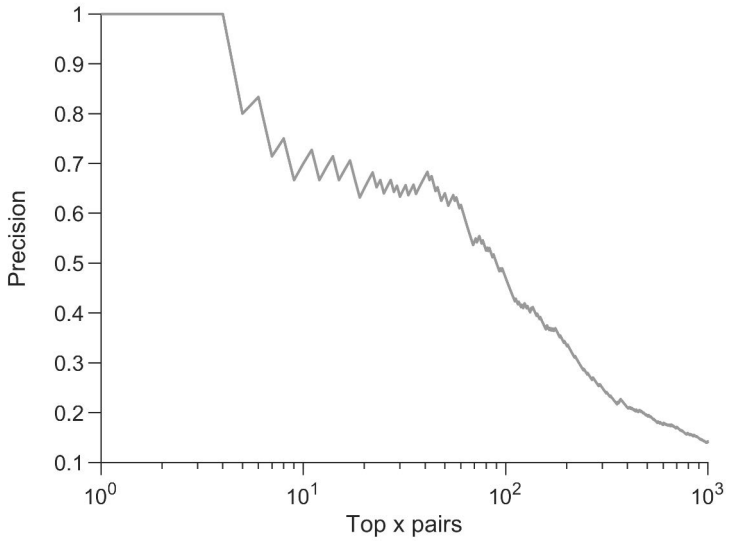

Fig. S3: **Precision of contact predictions vs. the top x pairs obtained using DCA (Morcos et al., 2011) (related to Fig. 1).** Predictions of DCA were generated using the code provided in (Morcos et al., 2011), where the re-weighting procedure was implemented according to Eq. (2) and the pseudocount parameter was set to 0.5. Precision is the proportion of top x pairs that are in contact in the protein tertiary structure. Two residues were assumed to be in contact if their carbon-alpha atoms were $<$8Å apart according to the available crystal structure of E2 1b (PDB ID: 6MEI).

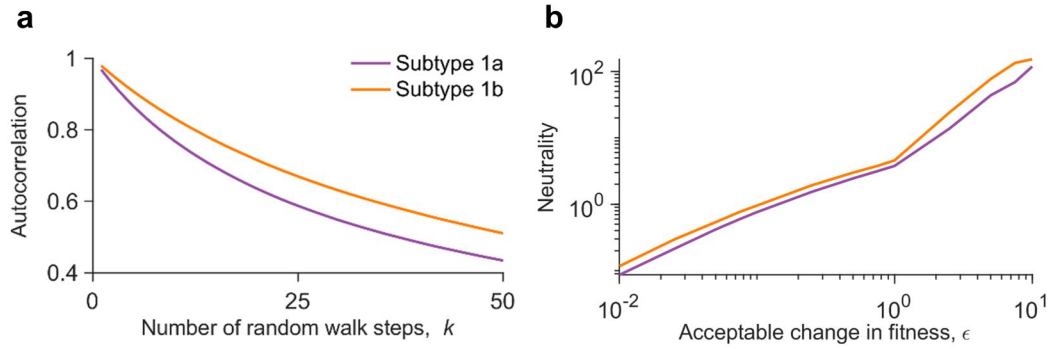

Fig. S4: **Robustness of the autocorrelation and neutrality results to variation of the involved parameters (related to Fig. 2a).** (**a**) Comparison of the autocorrelation of sequence energies of the E2 1a and 1b landscapes, with the starting sequences chosen within a Hamming distance (D) of 30 from the MSA (see STAR Methods for details). (**b**) Comparison of the neutrality of the E2 1a and 1b landscapes. Neutrality was computed for $L = 1000$ steps (see STAR Methods for details). It is observed that the fitness constraints on E2 1b appear lower than E2 1a, independently of whether we use $D = 5$ (Fig. 2a, left panel) or $D = 30$ for computing landscape autocorrelation, and whether we use $L = 500$ (Fig. 2a, right panel) or $L = 1000$ for computing landscape neutrality.

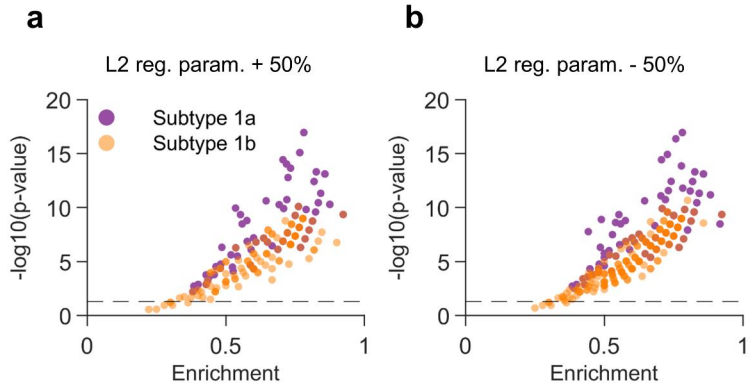

Fig. S5: **Robustness of the statistical enrichment of known escape mutations among landscape peak sequences to changes in landscape inference parameters (related to Fig. 2b left panel).** Almost all peak sequences of each subtype are statistically significantly enriched in escape mutations whether the L_2 regularization parameters used in the landscape inference were (**a**) increased by 50% or (**b**) decreased by 50%.

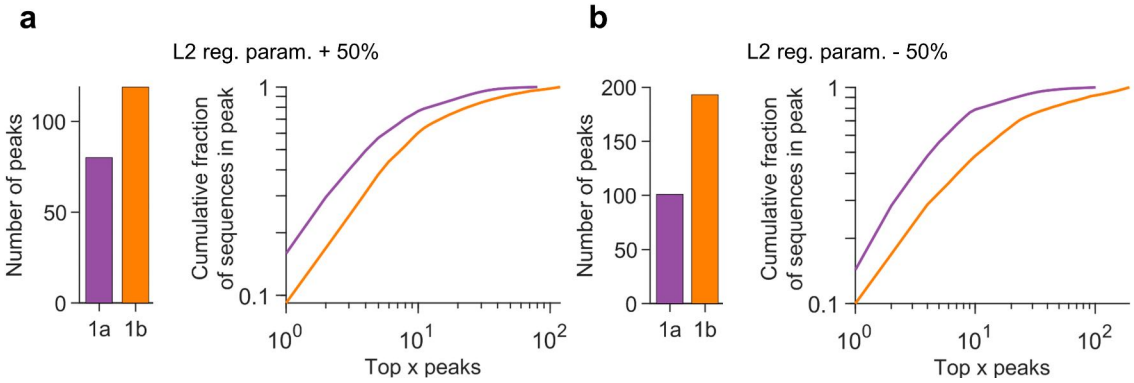

Fig. S6: **Robustness of the statistics of the observed peaks to changes in landscape inference parameters (related to Fig. 2b middle and right panels).** (Left panel) Number of peaks and (right panel) cumulative fraction of sequences in peaks observed in each landscape. More number of peaks are observed in 1b landscape than 1a, and peaks are more spread out across peaks in 1b than 1a whether the L_2 regularization parameters used in the landscape inference were (**a**) increased by 50% or (**b**) decreased by 50%.

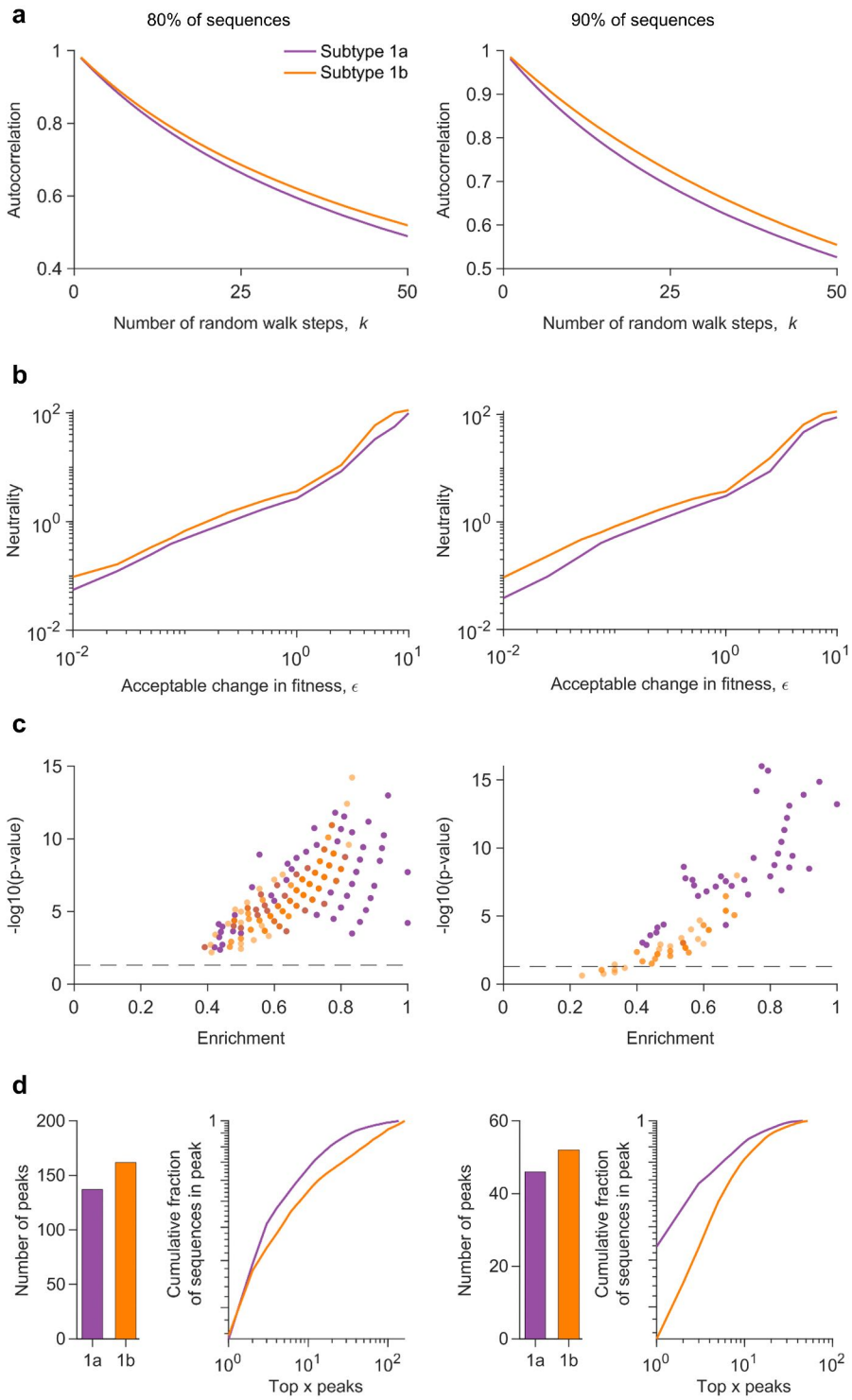

Fig. S7: **Robustness of the comparative analysis of fitness landscapes to changes in the number of sequences used for landscape inference of each subtype (related to Fig. 2).** The results of the comparison between fitness landscapes of subtype 1a and 1b remain qualitatively the same when only 80% (left panel) or 90% (right panel) of the sequence data for each subtype were used for landscape inference. (**a**) Autocorrelation of sequence energies of each landscape, with the starting sequences chosen within a Hamming distance $D = 5$ from the MSA (see STAR Methods for details). (**b**) Comparison of the neutrality of the E2 1a and 1b landscapes, with neutrality computed for $L = 500$ steps (see STAR Methods for details). (**c**) Statistical enrichment of known escape mutations among landscape peak sequences. (**d**) Number of peaks and cumulative fraction of sequences in peaks observed in each landscape.

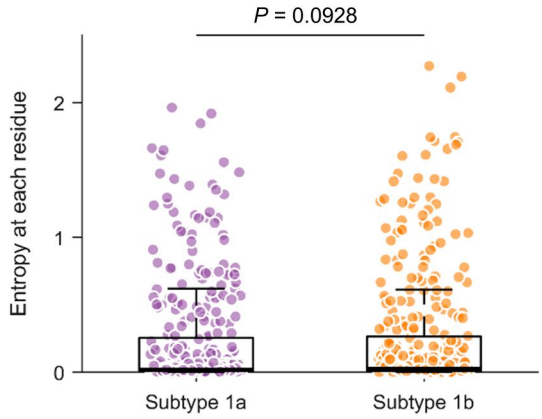

Fig. S8: **Comparison of the residue-wise entropy of E2 subtypes 1a and 1b (related to Fig. 2).** The residue-wise entropy of E2 1a and 1b is not statistically significant different ($P > 0.05$). In each box plot, the middle line indicates the median, the edges of the box represent the first and third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges. The reported p-value was calculated using the one-sided Mann-Whitney test.

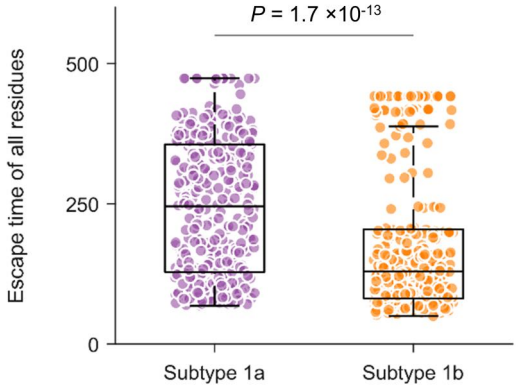

Fig. S9: **Comparison of predicted escape times of all residues of E2 1a and 1b (related to Fig. 4).** Escape times associated with E2 1b are statistically significantly lower than those of E2 1a. In each box plot, the middle line indicates the median, the edges of the box represent the first and third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges. The reported p-value was calculated using the one-sided Mann-Whitney test.

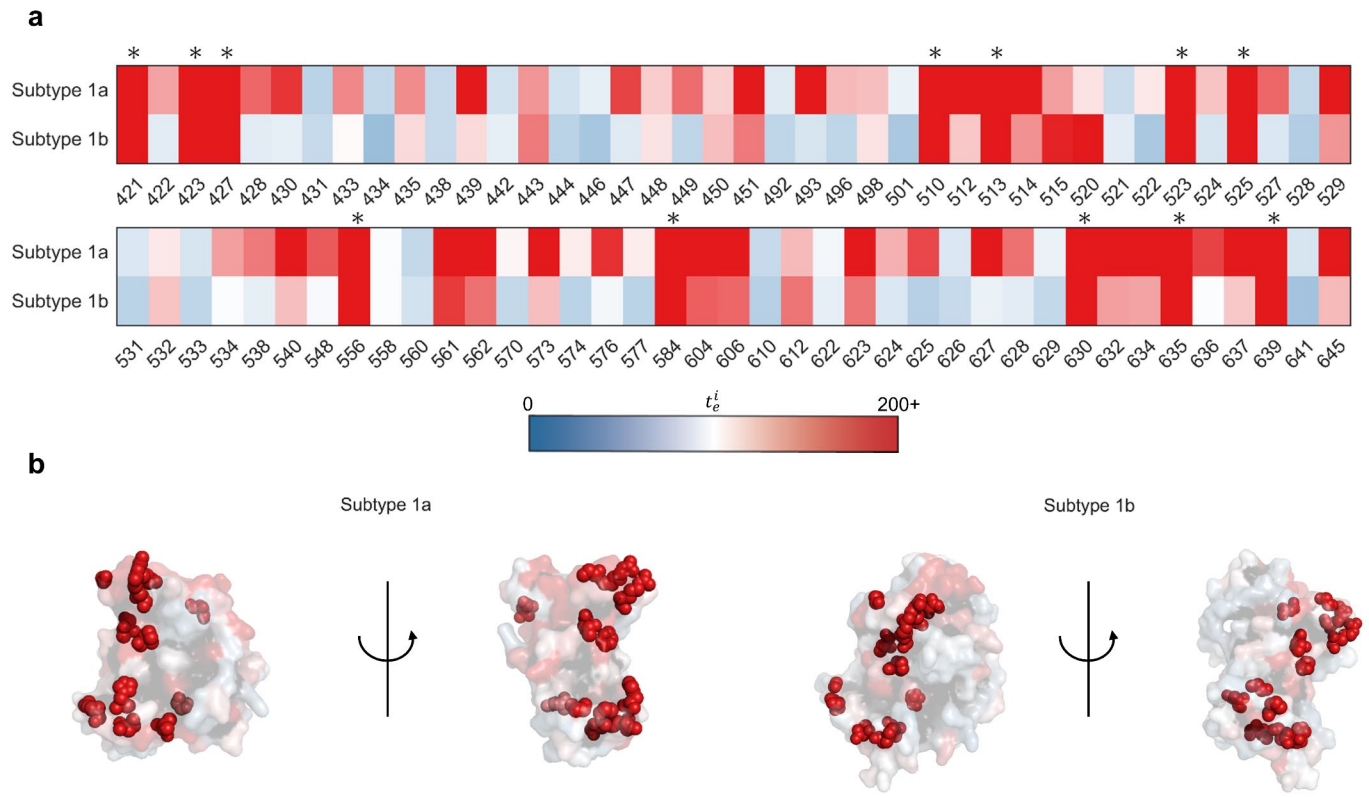

Fig. S10: **Exposed residues identified to be associated with high escape times for both subtypes 1a and 1b (related to Fig. 4).** (**a**) Escape times associated with the E2 residues exposed in both subtypes, with those residues associated high escape times for both subtypes marked with an asterisk on the top of each panel. The x-axis denotes the residue number (according to the H77 reference sequence). While the optimal cut-off value ζ (see STAR Methods for details) to determine whether a residue is difficult to escape or not was 100 generations for 1a and 80 generations for 1b, we selected a conservative minimum escape time of 200 to determine exposed residues associated with high escape time for both subtypes. (**b**) Common exposed residues predicted to be associated with high escape time for both subtypes (in (**a**)) are shown as spheres on the crystal structure of E2 1a (left panel, PDB ID: 4MWF) and E2 1b (right panel, PDB ID: 6MEI).

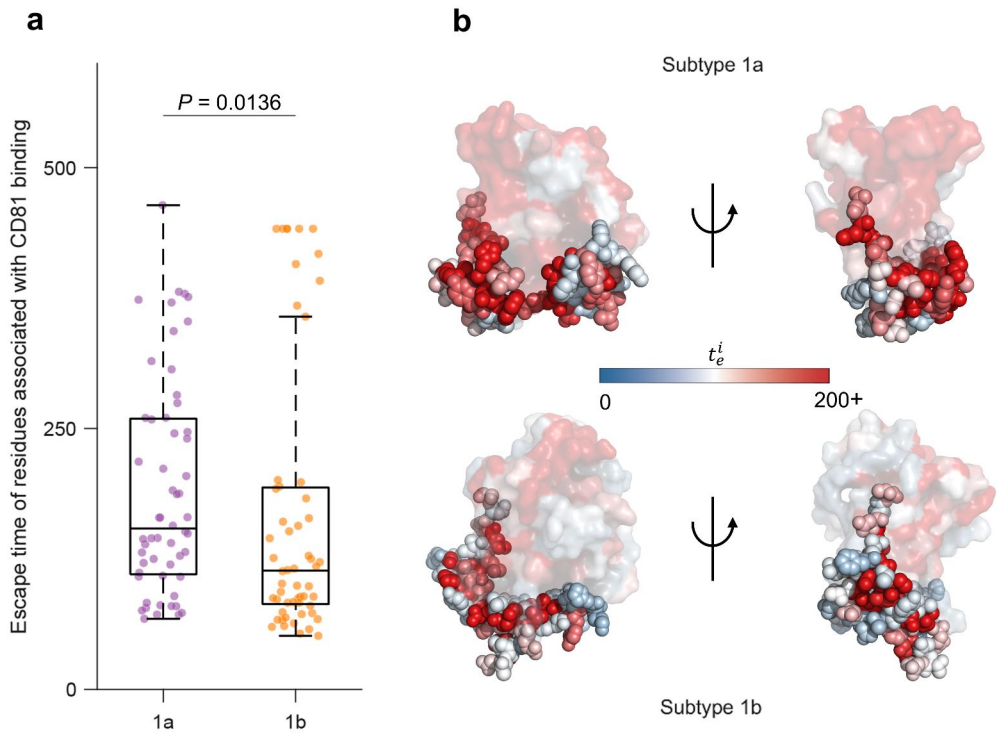

Fig. S11: **Comparison of escape times of CD81 binding residues of subtypes 1a and 1b (related to Fig. 4).** (**a**) CD81 binding residues include a conserved N-terminal region (aa412–423), a front layer region (aa428–446) and an adjacent loop named the CD81-binding loop (aa518–542) (Ströh et al., 2018). The reported p-value was calculated using the one-sided Mann-Whitney test. (**b**) Escape times of CD81 binding residues superimposed on the crystal structure of E2 1a (upper panel, PDB ID: 4MWF) and E2 1b (lower panel, PDB ID: 6MEI).

Fig. S12: **B cell epitope coverage of E2 for subtype 1a and 1b (related to Fig. 4).** The experimentally-identified B cell epitopes cover a similar fraction of E2 along its primary structure for both subtypes. The epitopes for each subtype were obtained from IEDB (https://www.iedb.org) (Vita et al., 2018). Epitopes corresponding to positive B cell assays of HCV were searched, and E2-specific epitopes for which subtype 1a or 1b had been specified were then filtered. This procedure provided 71 unique B cell epitopes for E2 1a and 33 unique ones for 1b. E2 1a epitopes were reported by 26 independent studies while those for 1b were reported by 7 studies, indicating that the high number for E2 1a epitopes was possibly due to this subtype being studied more in the literature. E2 residues covered by the B cell epitopes are shown in green while the remaining ones are shown in white. Residues are numbered according to the H77 reference sequence.

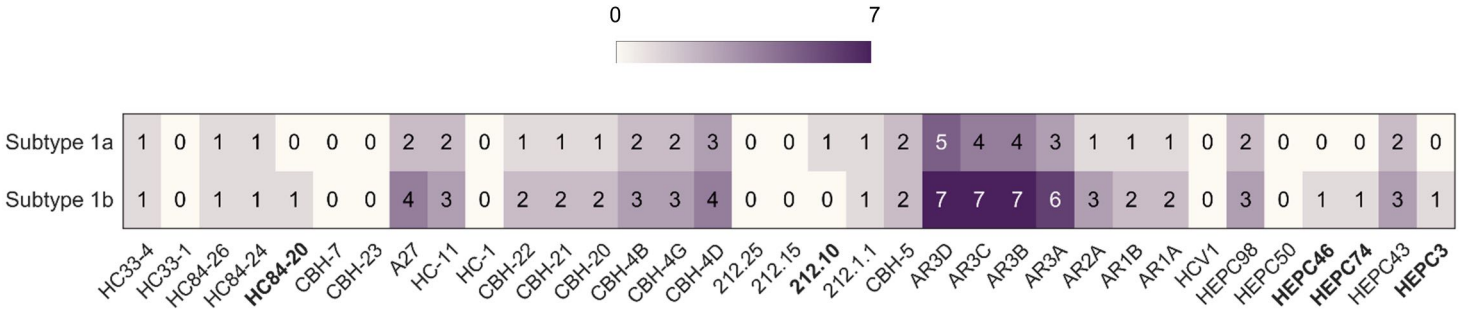

Fig. S13: **Number of residues within the binding sites of each E2-specific HmAb predicted to be associated with escape for subtype 1a and 1b (related to Fig. 4).** The threshold ζ to determine whether or not a residue is potentially associated with escape was set to 100 generations for subtype 1a and 80 generations for subtype 1b (see STAR Methods for details). The specific HmAbs that appear to be relatively easy to escape for exclusively one subtype are shown in bold.

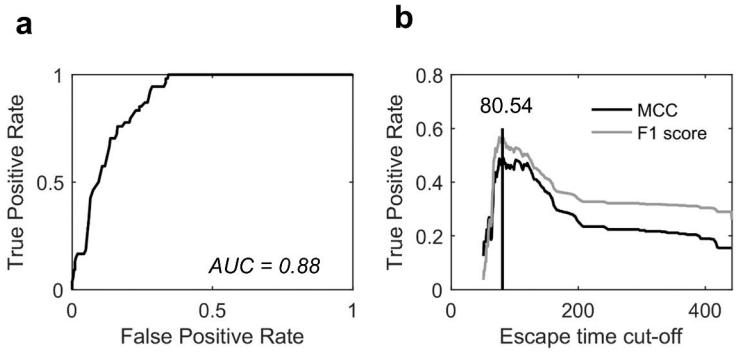

Fig. S14: **Classifier for determining optimal escape time for E2 1b, designed based on knowledge of experimentally or clinically identified escape mutations (related to STAR Methods).** (**a**) Receiver operating characteristic curve for identifying the known escape mutations (listed in Table S1) using the escape time metric. (**b**) Determination of the optimal escape time cut-off based on the F1 score and MCC. In this classification, residues with known escape mutations (listed in Table S1) were considered as true positives while all remaining residues were assumed true negatives.

Supplementary Table

TABLE S1: **List of known escape mutations from E2-specific HmAbs (related to Fig. 3).**

Escape residues	HmAbs	Reference
408	HC33-4	(Keck et al., 2016)
384, 386, 388, 390, 391, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 407, 410	HmAbs targeting HVR1	(Kato et al., 1993)
431	CBH-2	(Keck et al., 2008)
431, 435, 444, 446, 466, 482, 501, 528, 531, 538, 580, 610, 636, 713	CBH-8C, CBH-2, CBH-5, HC-2, HC-11	(Keck et al., 2009)
391, 394, 401, 415, 417, 434, 444, 608	HCV1	(Morin et al., 2012)
416, 422, 424, 431, 433, 438, 442, 446, 453, 456, 461, 475, 482, 520*, 524, 531, 533, 557, 558, 560	CBH-2, CBH-5, HC84-22, HC84-26, AR3A, AR3B, AR3C, AR3D	(Bailey et al., 2015)
431, 438, 442	AR3A	(Velázquez-Moctezuma et al., 2019)

*Excluded from our analysis, as no mutation was observed at this residue in the E2 1b sequence data.