SUPPLEMENTARY INFORMATION

Empirical fitness models for hepatitis C virus immunogen design

Gregory R. Hart[1] and Andrew L. Ferguson[2,3,*]

[1]*Department of Physics,* [2]*Department of Materials Science and Engineering,*

and [3]*Department of Chemical and Biomolecular Engineering,*

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

**To whom correspondence should be addressed: Email:*

alf@illinois.edu. Phone: (217) 300-2354. Fax: (217) 333-2736.

I. AVERAGE IMMUNE PRESSURE ESTIMATE

We previously showed that the effects of host immune pressure on strain distribution is averaged out if the MSA samples a genetically diverse host population [1–3]. In other words, the fact that hosts with different haplotypes target very different regions of the viral proteome means that if the number of sequences in the ensemble used to fit the model are sufficiently numerous and drawn from hosts with diverse immunological haplotypes, no single position in the viral proteome is subjected to a disproportionate mutational pressure when averaged over the sequence ensemble. The sequences where collected in multiple locales in nine countries on three continents [4–6], and while the majority of patients were Caucasian, there exists a strong representation of Africans as well as Hispanics, Asians, and other ethnic groups. This geographic and ethnic diversity suggests that the MSA contains sufficient genetic diversity to eliminate signatures of adaptive immunity. To quantify this assertion we estimated the frequency with which each amino acid position in NS5B is expected to be subject to immune pressure using the approach detailed in Ref. [2]. We compiled from the Immune Epitope Database (`http://www.iedb.org`) the 24 CTL NS5B epitopes that were both exactly defined and the HLA association known [7], and determined the frequency with which each HLA occurs within the North American population as a representative group (`http://www.ncbi.nlm.nih.gov/projects/gv/mhc`) [8]. Finally, we estimated the probability that persons possessing these HLA types recognize and target the cognate CTL as the mean of the reported non-zero recognition frequencies of CTL epitopes across the HCV proteome [9]. We observe that this provides a conservative estimate higher than any value reported for a NS5B epitope. Using these values, we estimated no position within the NS5B protein to be targeted by more than 8% of the population. This percentage is substantially lower than the values of 17% and 23% estimated

for the HIV-1 proteins p17 and p24, for which we have previously computed fitness landscapes and validated in extensive comparisons against experimental and clinical data to demonstrate that they are not contaminated with signatures of adaptive immune pressure [1, 2]. In sum, our targeting frequency estimate for NS5B, prior empirical, numerical, and theoretical studies [2], and direct comparisons against clinical data and experimental measurements described in the main text, all provide support that our inferred NS5B fitness landscape does not contain footprints of adaptive immunity.

II. MODEL AUGMENTATION

The Potts model fitted to the MSA data detailed in the main text contained parameters describing the fitness impact of each amino acid residue in each single position, and each pair of amino acids in each pair of positions. In comparing our model predictions against clinical and experimental data, we twice encountered a situation in which the experimentally reported viral strains contained amino acid residues absent in our MSA and therefore not contained within our model. Rather than simply discarding these sequences from our comparisons, we constructed two separate augmented models containing parameters for amino acid variants that were unobserved in the MSA.

The first augmentation was required in order to fully compare our model predictions with the measured *in vitro* fitness data in Section 3.1. Of the 31 *in vitro* measurements we collated from the literature, 30 of them – all from the same lab [10, 11] – used the H77 sequence (GenBank Accession No. M67463) as their wild type baseline sequence, rather than the more commonly used H77S.3 sequence (GenBank Accession No. AF011751). Our fitted Potts model contained all amino acids in H77S.3, but did not contain parameters for six residues in H77: K2469, A2512, L2637, R2703, R2715, and W2925. To assign energies to all strains considered, we augmented our model with parameters for these six unobserved residues to generate Augmented Model I.

The second augmentation was required to assign energies to all of the clinically observed escape, and compensatory, mutations that we analyzed in Section 3.2, and the sequences in the longitudinal studies considered in Section 3.4. The sequences from the longitudinal studies contained 456 residues in particular positions not observed in our MSA. The clinical escape mutations contained two amino acids that were not contained in our MSA, and which were coincident with two of the 456 unobserved amino acids within the longitudinal data. Accordingly, in order to assign energies to all longitudinal strains and escape mutations we augmented our model with parameters for the 456 unobserved residues to generate Augmented Model II.

To perform model augmentation it is necessary to incorporate h_i and J_{ij} parameters for each unobserved residue in position i. To estimate values of these model parameters, we specified the probability with which the unobserved amino acids appear within the MSA to be non-zero using pseudo-counts [12, 13]. This procedure adds a (possibly non-integer) number of fictitious observations of the amino acid within the MSA to reflect the belief that the probability of observing this amino acid in this position is not precisely zero, but rather a low-probability event that is not observed within the finite number of strains within the MSA. From a Bayesian perspective, the use of pseudo-counts may be considered the incorporation of prior knowledge into the model inference procedure [12], in this case the prior knowledge that the probability that these amino acids exist within an HCV strain should be non-zero. In this work, we specify the pseudo-count modified probability of observing amino acid A in position i, $P_i(A)$, as,

$$P_i(A) = \frac{1}{\lambda + N}\left(\frac{\lambda}{q_i} + \sum_{k=1}^{N} \delta_{A,z_i^k}\right), \tag{1}$$

where N is the number of sequences in our MSA, q_i is the number of distinct amino acids (including this unobserved amino acids added to the model) at position i, z_i^k is the identity of the amino acid in position i in sequence k of the MSA, δ_{A,z_i^k} is an indicator function that is unity when $A = z_i^k$ and zero otherwise, and λ is a pseudo-count [13, 14]. At positions where we supplement our model with unobserved amino acids, we choose $\lambda = \frac{q_i N}{N+1-q_i}$ such that for an unobserved amino acid A at that position, $P_i(A) = \frac{1}{N+1}$. This quantity may be interpreted as an estimated upper bound on the frequency with which amino acid A is observed in position i corresponding to supplementing the MSA with one (hypothetical) additional sequence containing amino acid A in position i. We then iteratively rescaled the two-position target probabilities, $P_{ij}(A, B)$, such that the marginal probabilities over i and j are consistent with the one-position target probabilities, $P_i(A)$ [1]. At positions where no unobserved amino acids were added, no pseudo-counts were added (i.e., $\lambda = 0$).

Having specified the pseudo-count modified probabilities, we re-fitted the parameters of the Potts model using the procedure detailed in Section 2.2. Constituting a relatively small perturbation to the target probabilities for the fitting procedure, the model parameters changed very little from their unaugmented values, and by initializing the $\{h_i\}$ and $\{J_{ij}\}$ parameters to their unaugmented values, the fitting procedure quickly converged. The energy predictions of the augmented and unaugmented models are in close agreement, as illustrated by the correspondence of strain energies in Figure 2 (augmented model) and Figure S3 (unaugmented model).

III. PREDICTED FITNESS COSTS OF CLINICAL ESCAPE MUTATIONS

In Section 3.2 we looked at the energy cost (fitness cost) of documented escape mutations and where these costs fall in the spectrum of possible mutations. Three of the single mutations, K2471R, Q2467K, and R2937S, fall in the 36^{th}, 74^{th} and 96^{th} percentiles, respectively. Two of the double mutations, R2937G/I2940T and Q2467K/K2471R fall in the 33^{rd} and 61^{st} percentiles, respectively. The relatively high energy (fitness) costs of these mutations can be rationalized by the fact that they are almost always observed in concert with compensatory mutations that place them far lower on the energy (fitness) cost spectrum. The details of those compensatory mutations follow.

K2471R (36^{th} percentile) and Q2467K (74^{th} percentile) are typically observed as the double mutant, Q2467K/K2471R (61^{st} percentile). Furthermore, they are almost always seen in connection with another mutation H2453Y which is compensatory and mediates further immune escape [15]. H2453Y is an escape mutation in a nearby epitope presented by the same HLA molecule. H2453Y/K2471R falls in the 20^{th} percentile of all double mutants with $\Delta E = 12.50$, and H2453Y/Q2467K falls in the 34^{th} percentile of all double mutants with $\Delta E = 14.05$. The triple mutant H2453Y/Q2467K/K2471R falls in the 35^{th} percentile of all triple mutants with $\Delta E = 21.42$.

We observe that the Q2467K/K2471R double mutant may represent a temporary "metastable" escape, since K2471R was observed to revert back to wild type after Q2467K was replaced by the Q2467L polymorphism. This alternative escape mutation has been reported to be less effective at mediating CTL escape, but is of much higher fitness (lower energy), falling in the 4^{th} percentile of the energy spectrum [15].

R2937S (96^{th} percentile) is very rare and is always observed to be accompanied by E2875K and P2881Q [10]. This E2875K/P2881Q/R2937S triple mutant falls in the 4^{th} percentile of all triple mutants and has an energy cost $\Delta E = 13.26$. R2937G/I2940T (33^{rd} percentile) is also very rare and always observed with E2875K and P2881Q [10].

IV. LONGITUDINAL CLONAL SEQUENCING STUDY

In Section 3.4 we analyzed longitudinal sequencing data of HCV progression in Patient M003 and the two children to whom she gave birth during the study and vertically transmitted HCV – Patients C003 and D003 – as reported in Ref. [15]. For concision, we considered in the main text only the average energy assigned by our model to the strains at each time point, $\bar{E}$, in our predictions of the fitness of the viral ensemble over the course of the study. Here we present a more detailed analysis of the clonal sequencing data for M003 (and her children C003 and D003), for whom multiple

sequences are available for each time point.

Table S5 presents for each viral strain identified at each time point the energy of the strain assigned by our model, and the sequence of the six epitopes B*15-LLRHHNMVY$_{2450-2458}$, B*15-SQRQKKVTF$_{2466-2474}$, A*02-RLIVFPDLGV$_{2578-2587}$, A*02-ALYDVVSKL$_{2594-2602}$, A*02-GLQDCTMVL$_{2727-2735}$, and A*31-VGIYLLPNR$_{3003-3011}$ for which M003 possesses the cognate HLA molecules, and position 2510 which is associated with an A*31 associated polymorphism (S2510N) [16]. As our reference sequence for this analysis we adopt the consensus sequence at the first time point of the study at 0.0 months.

M003 presented with acute HCV during her pregnancy with C003 at 0.0 months. Consistent with a suppressed immune system due to maternofatel immune tolerance, our model assigns high fitness (low energy) to the sequences retrieved at this time and at the time of delivery (1.3 months). The sequences reveal scattered mutations within epitopes, but all are transient and do not appear to be correlated with host immune pressure.

After delivery of C003 at 7.2 months our model predicts a sharp decrease in fitness (increase in energy) in all sequences reported in M003, that appears to be correlated with an increase in host immune pressure due to disappearance of the maternofetal immune tolerance mechanism after delivery of the infant. In particular, the sequence ensemble contains four immune related polymorphisms: H2453Y, Q2467X, K2471R, and S2510N. The H2453Y mutation appears in epitope B*15-LLRHHNLVY$_{2450-2458}$ and is known to abrogate T-cell recognition [15]. The mutations Q2467X and K2471R appear in epitope B*15-SQRQKKVTF$_{2466-2474}$. Although the specific information on all polymorphisms present is not available, Q2476L is reported to decrease T-cell recognition and Q2467K/K2471R to abolish it [15]. Our model predicts the energy cost of Q2467K ($\Delta E = 9.0$) to be more than three times that of Q2467L ($\Delta E = 2.8$). The elimination of Q2467L in favor of Q2467K by month 10.7 is consistent with a scenario of mounting immune pressure on this epitope, and the fixation of a higher fitness cost escape mutation that more effectively abrogates T-cell recognition. The A*31 associated polymorphism S2510N also becomes fixed in M003 after delivery, consistent with her HLA haplotype.

As M003 enters a second pregnancy with D003 at month 8.3, our model again predicts an increase in the fitness of the viral ensemble (decrease in energy) consistent with suppression of specific HCV host immune pressure by the maternofetal immune tolerance mechanism. In addition to an increase in fitness (decrease in energy) of most sequences, by month 16.8, K2471R has reverted to wild type, and Q2467K/H to Q2467L in all but one sequence, consistent with a decrease in host immune pressure mediating the reversion of high fitness cost escape mutations. These reversions

persisted for several months post-delivery of D003, but at 4.3 months after delivery (21.5 months) the rebounding of M003 host immune pressure induced the more costly, but more effective, escape mutations to arise again, with Q2467K/K2471R appearing in half of the clonal population.

The next ensemble of viral sequences from M003 are reported more than a year later, at 36.9 months. At this time all the sequences possess the wild type amino acid at position 2451 and a histidine residue at position 2467. Our model predicts the cost of Q2467H ($\Delta E = 4.0$) to be intermediate to that of Q2467L and Q2467K/K2451R. We suggest that Q2467H offers some immune escape and hence is tolerated over the more fit Q2467L. The final sequence data from M003 come after another year, at 49.0 months. Continuing to evolve under immune pressure a new polymorphism arises in half the sequences, Q2467T. In all cases Q2467T appears with K2571R. Our model predicts that Q2467T/K2451R ($\Delta E = 18.2$) is less fit than Q2467K/K2451R ($\Delta E = 16.3$); however we see a increase in the fitness of the strains with Q2467T/K2451R indicating that compensatory mutations arose to make it more fit.

None of the sequences vertically transmitted to either child C003 and D003 show significant differences from the maternal sequences at (or near) the time of birth. Sequences within C003 25 weeks after birth (at 7.2 months) contain none of the escape mutations that arose within in the mother after delivery, consistent with transmission of a fit HCV strain from the immunosuppressed mother "outrunning" the nascent immune system of the newborn [15] . Sequences within D003 12 weeks after birth (at 20.1 months) contain the S2510N polymorphism and H2453Y and Q2467L polymorphisms within the HLA-B*15 associated epitopes $\text{LLRHHNLVY}_{2450-2458}$ and $\text{SQRQKKVTF}_{2466-2474}$ present in the mother before delivery, and constitute a slightly more fit population ($\overline{E} = 52.2$) of viral strains than those of the mother at time of delivery. D003 inherited the HLA-B*1501 class I molecule from the mother, and that reversion of the polymorphisms within these unfit strains is not observed is consistent with continued immune pressure at these epitopes by the child's immune system.

V. FIGURES

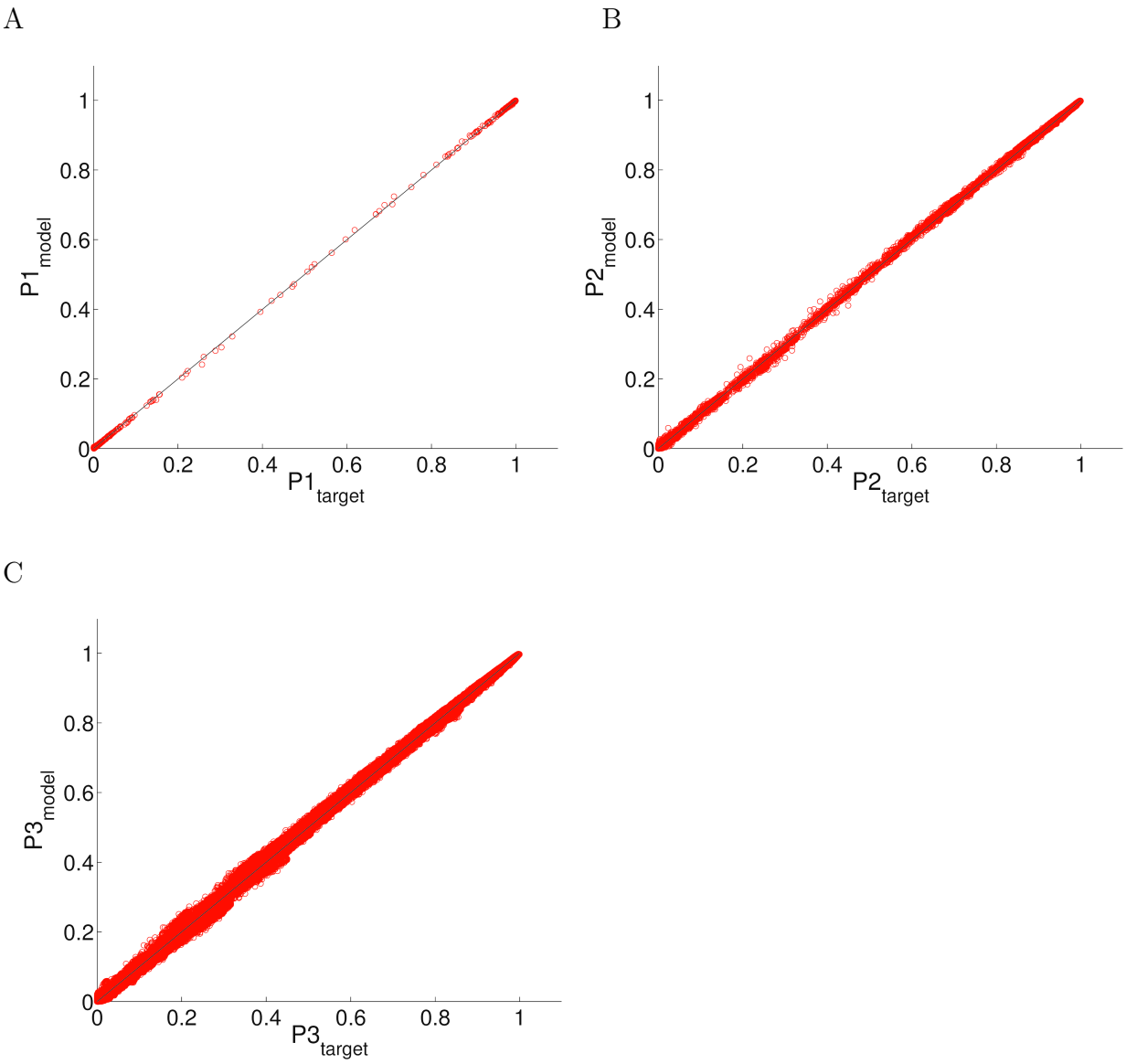


Figure S1. Comparison for each amino acid, A_i, at each position, i, the (A) one-position, $P1(A_i)$, (B) two-position, $P2(A_i, A_j)$, and (C) three-position, $P3(A_i, A_j, A_k)$, amino acid frequencies observed within the MSA, $P1_{target}$, to those computed by the fitted Potts model, $P1_{model}$, by performing 99,990 rounds of Monte-Carlo sampling from the model (cf. Ref. [1]). The parameters of the model were explicitly fitted to reproduce the one and two-position frequencies and so are expected to reproduce the observed mutational frequencies. That the model also predicts the three-position amino acid frequencies observed within the MSA demonstrates that our model *predicts* higher order mutational correlations within its effective one and two-position interaction parameters.

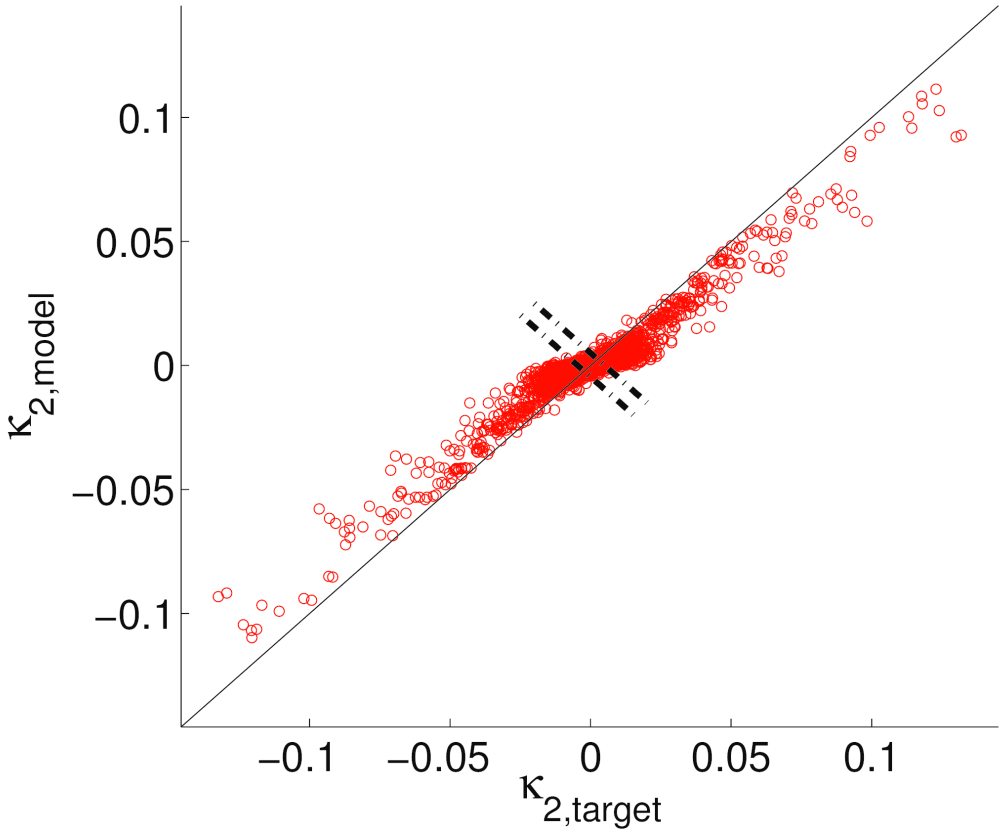


Figure S2. Comparison of the second order cumulants for each pair of amino acids, A_i and A_j, at each pair of positions, $\kappa_2(A_i, A_j) = P2(A_i, A_j) - P1(A_i)P1(A_j)$ observed within the MSA, $\kappa_{2,\,target}$, to those computed by the fitted Potts model by performing 99,990 rounds of Monte-Carlo sampling from the model, $\kappa_{2,\,model}$, (cf. Ref. [1]). κ_2 measures the difference between actual two-position probability of observing a particular pair of amino acids at a particular pair of positions and the two-position probability that would be expected if the two positions were mutationally uncoupled. $\kappa_2 \in [-0.25, 0.25]$, where $\kappa_2 > 0$ indicates that the mutations are correlated, $\kappa_2 = 0$ uncorrelated, and $\kappa_2 < 0$ anti-correlated. To define a statistically-significant correlation, we performed 10 independent scrambles of the columns of the MSA to randomize the amino acids located in each position of the protein and artificially break mutational correlations. The dashed lines in the plot indicate the 0.5^{th} and 99.5^{th} percentiles of the observed distribution of κ_2 under this permutation test – $\kappa^{0.5\%}_{2,\,model} = -2.3 \times 10^{-3}$ and $\kappa^{99.5\%}_{2,\,model} = 2.5 \times 10^{-3}$ – presenting an empirical measure of the expected range of κ_2 in the absence of mutational correlations and defining a 1% significance level for measured values of κ_2. The distribution of $\kappa_{2,\,target}$ indicates that while most mutational pairs are relatively uncorrelated, there are a significant number of strongly correlated and anti-correlated mutations, reflecting the presence of important epistatic effects within the protein. Furthermore, the clustering of the data around the diagonal indicates that our model captures these epistatic effects.

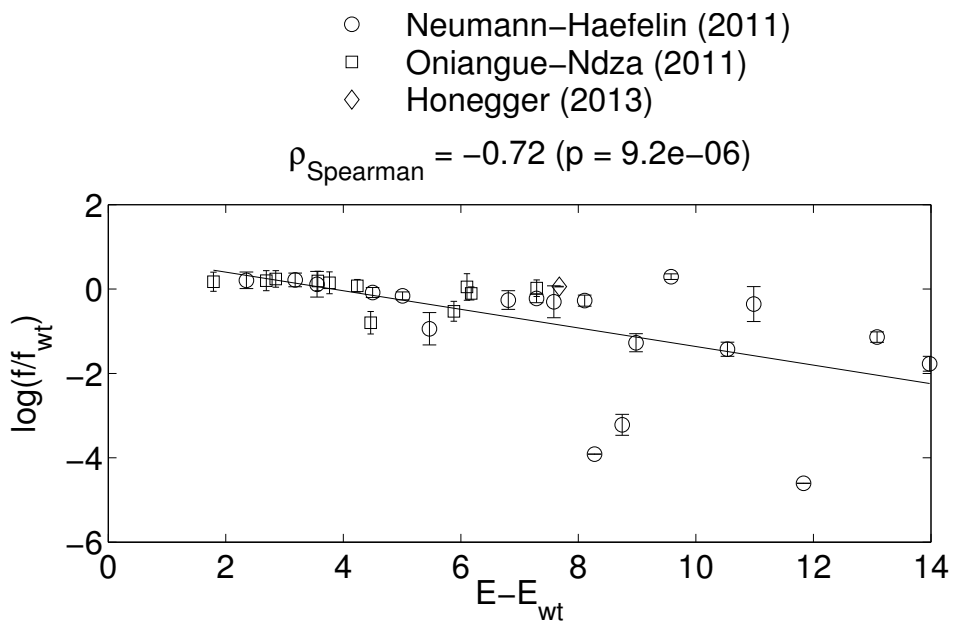
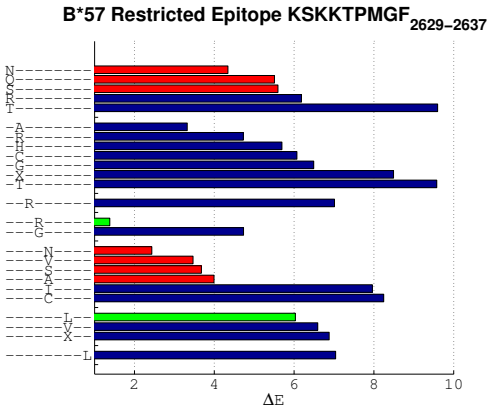
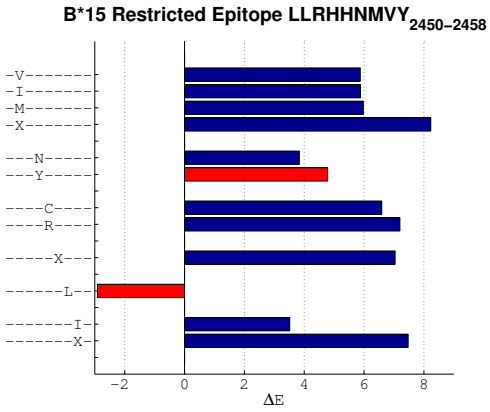


Figure S3. Comparison of the *in vitro* replicative fitness relative to wild type, f/f_{wt}, measured for 31 engineered NS5B mutants containing up to four polymorphisms [10, 11, 15] against the energy relative to H77S.3 reference sequence, $(E - E_{wt})$, of each strain predicted by our unaugmented model. A strong and statistically significant negative correlation, $\rho_{Spearman} = -0.72$ ($p = 9.2 \times 10^{-6}$), validates our fitted model as a good predictor of intrinsic viral fitness. A linear least-squares fit is provided to guide the eye, and error bars delineate estimated uncertainties in the measured relative fitness.

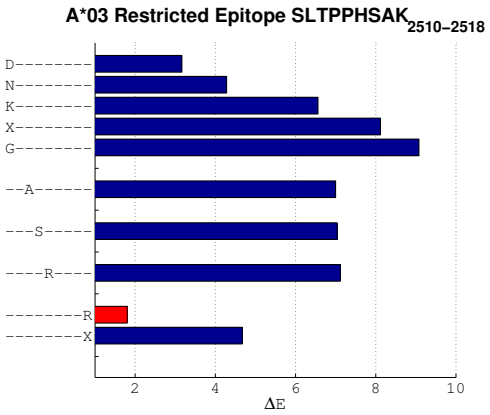
A


B*57 Restricted Epitope KSKKTPMGF2629–2637
ΔE

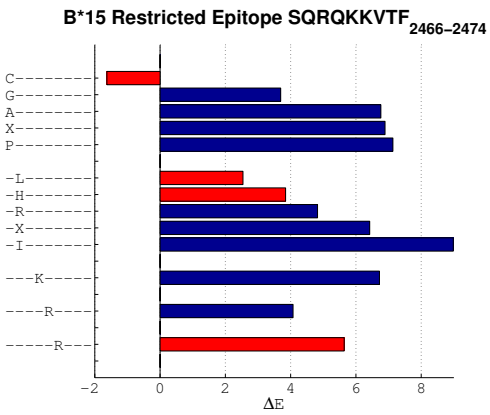
B


B*15 Restricted Epitope LLRHHNMVY2450–2458
ΔE

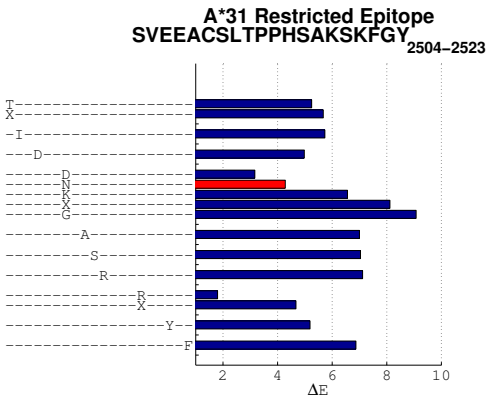
C


A*03 Restricted Epitope SLTPPHSAK2510–2518
ΔE

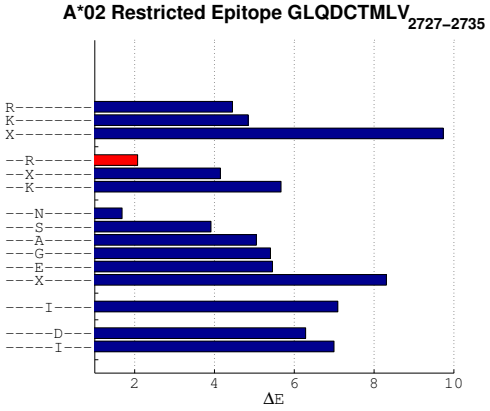
D


B*15 Restricted Epitope SQRQKKVTF2466–2474
ΔE

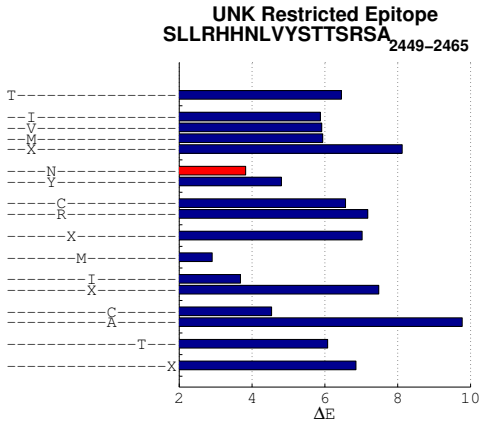
E


A*31 Restricted Epitope
SVEEACSLTPPHSAKSKFGY2504–2523
ΔE

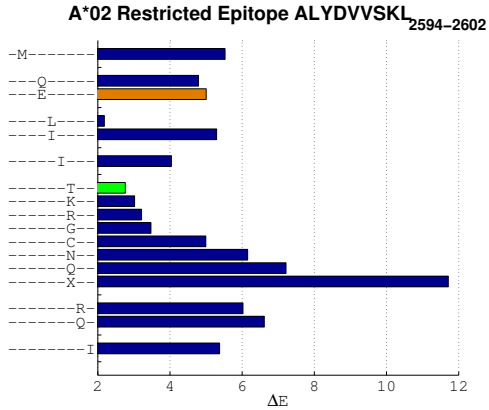
F


A*02 Restricted Epitope GLQDCTMLV2727–2735
ΔE

G



H



I

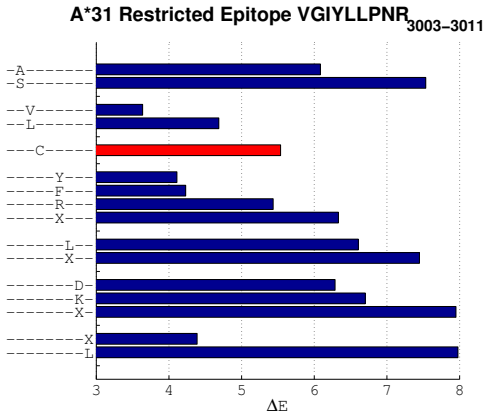
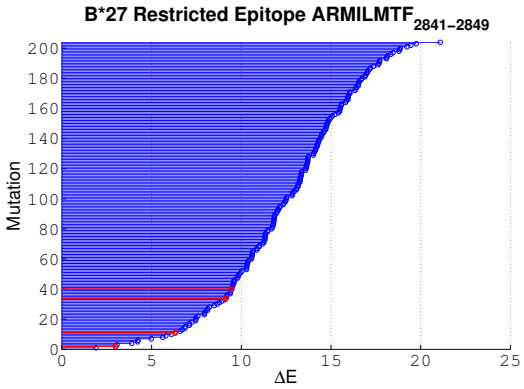


Figure S4. Comparison of the energy costs (fitness penalties) relative to the H77 wild type reference sequence predicted by our model for all polymorphisms observed within our MSA occurring within the nine indicated CTL epitopes. All nine epitopes possess single amino acid mutations known to confer CTL escape. The energy cost associated with each single mutation, ΔE, is along the abscissa, and the mutations are shown along the ordinate. Dashes indicate unmutated positions, and letters the mutant amino acid residue. The letter X indicates an unknown amino acid type that was inconclusively identified by experimental sequencing within the ensemble sequences constituting the MSA used to fit our model. The greater the energy cost, the higher the fitness penalty. Polymorphisms possessing negative ΔE values are those predicted to *elevate* fitness relative to the H77 reference sequence. Red bars denote documented escape mutations that abrogate CTL recognition, green bars denote cross-reactive mutations that do not mediate escape, brown bars denote polymorphisms that have been reported both as escapes and as cross-reactive, blue bars denote mutations for which no specific clinical information is available. In panels A-G, one or more of the first, second, or third least costly polymorphisms within the epitope corresponds to a documented escape mutation. In panel H the escape mutation has the ninth lowest cost, although we note that it remains disputed as to whether this polymorphism conveys escape or is cross reactive [17–22]. In panel I the escape mutation is the seventh lowest cost polymorphism.

A



B

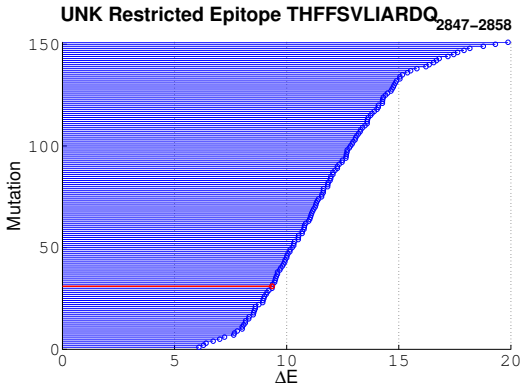


Figure S5. Plots of the energy (fitness) cost of all double mutations observed within our MSA within the CTL epitopes (a) $\mathrm{ARMILMTHF}_{2841-2849}$ and (b) $\mathrm{THFFSVLIARDQ}_{2847-2858}$. These two epitopes require at least two mutations in order to escape CTL pressure. The energy cost associated with each double mutation, ΔE, is along the abscissa, and the double mutations – indexed according to their energy cost – are shown along the ordinate; the greater the energy cost, the higher the fitness penalty. Red bars denote documented escape mutations. In panel A, the second least costly of all double mutants predicted by our model corresponds to a clinically documented escape mutation. In panel B, the documented escape mutation lies in the bottom fifth of the energy spectrum of all double mutations, ranked as the 31/151 lowest energy double mutant.

VI. TABLES

EU781776	EU482846	EU781802	EU781779	EU781748	EU781781	EU781747	EU781771	EU256056	EU781749
EU781797	EU256044	EU781796	EU256055	EU781798	EU482840	EU256053	EU781788	EU482841	EU781793
EU781799	GQ149768	EU781751	EU781800	EU781787	EU781794	EU781795	EU781786	EU781782	EU781804
EU155276	EU781789	EU781790	EU781760	EU155277	EU781810	EU781780	EU781803	EU781752	EU781823
EU781783	EU781750	EU781768	EU781777	EU781774	EU482837	EU482847	EU260396	EU256040	EU482878
EU256039	EU862828	EU155247	EU256041	EU482843	EU482848	EU256058	EU256043	EU256051	EU255940
EU569723	EU155265	EU155270	EU862831	EU482882	EU155271	EU482844	EU155278	EU155275	EU155282
EU155248	EU482845	EU155249	EU256067	EU482842	EU155266	EU256052	EU155251	EU256060	EU482838
EU155267	EU256046	EU155252	EU256047	EU155272	EU256057	EU256048	EU256049	EU255941	EU256068
EU155268	EU155273	EU155283	EU155274	EU256050	EU255939	EU155269	EU255942	EU781770	EU781753
EU781772	EU781767	EU781763	EU781773	EU781785	EU862823	EU482831	EU155378	EU482873	EU155379
EU781809	EU256106	EU482832	EU155380	EU256107	EU862834	EF032890	EU255963	EU255964	EU255965
EU155338	EU255966	EU155339	EU255968	EU255969	EU862839	EU255970	EU234063	EU482889	EU255973
EU255974	EU234064	EU255975	EU255976	EU155340	EU482850	EU255977	EU255978	EU255979	EU255980
EU255981	EU155341	EU255982	EU255983	EU155342	EU255984	EU255985	EU255986	EU255987	EU255988
EU255989	EU255990	EU255991	EU234065	EU255992	EU155343	EU569722	EU155344	EU482852	EU862827
EU781821	EU781815	EU781820	EU781812	EU781817	EU781807	EU781814	EU781808	EU781824	EU781822
EU781818	EU781811	EU781764	EU781792	EU781765	EU781754	EU781758	EU781762	EU687193	EU687194
EU687195	EU781775	EU781757	EU781761	EU781766	EU781759	EU781756	EU781791	EU781784	EU781801
EU781778	EU781769	EU781755	EU155311	EU239713	EU155284	EU155312	EU155285	EU155286	EU155287
EU256096	EU155288	EU155289	EU256097	EU155321	EU256105	EU155322	EU155290	EU482834	EU155291
EU595697	EU482836	EU155292	EU250017	EU155319	EU155293	EU155320	EU155323	EU482876	EU155295
EU239716	EU239715	EU482835	EU155296	EU155297	EU155298	EU155309	EU155299	EU660387	EU155313
EU155310	EU155314	EU256104	EU255927	EU255943	EU255944	EU255945	EU255946	EU482853	EU255947
EU256069	EU255948	EU255928	EU256070	EU256071	EU255949	EU155346	EU255930	EU256072	EU155347
EU155348	EU482854	EU255951	EU155349	EU255952	EU255953	EU482855	EU482856	EU529681	EU255954
EU255955	EU155350	EU155351	EU660385	EU255956	EU255957	EU482872	EU482857	EU256074	EU255958
EU155353	EU255959	EU155354	EU155355	EU482858	EU482884	EU529676	EU256002	EU595698	EU256004
EU482865	EU256008	EU862841	EU781816	EU781819	EU256009	EU256010	EU256011	EU529677	EU256013
EU155239	EU660384	EU256014	EU256015	EU256017	EU256018	EU256019	EU256020	EU529678	EU256021
EU256022	EU482887	EU256023	EU155240	EU482868	EU482869	EU256024	EU155242	EU256094	EU482870
EU529679	EU482871	EU595699	EU256025	EU529680	EU256095	EU781813	EU256026	EU256027	EU256028
EU256029	EU155243	EU256030	EU255938	EU256031	EU155244	EU155245	EU256032	EU155246	EU256034
EU256035	EU256036	EU256037	EU256038	EU660383	EU155213	EU255993	EU255994	EU155214	EU155215
EU255995	EU255996	EU255931	EU255997	EU256086	EU255932	EU255998	EU256087	EU155216	EU255999
EU256003	EU155233	EU482861	EU256012	EU155236	EU256005	EU256006	EU255934	EU482862	EU255936
EU482863	EU155237	EU255937	EU482864	EU482866	EU482867	EU255935	EU256007	EU155238	FJ024087
FJ181999	FJ024274	FJ024275	FJ205867	FJ024276	FJ182000	FJ024278	FJ390399	FJ205868	FJ182001
FJ024280	FJ024281	FJ024282	FJ390394	FJ410172	FJ390395				

Table S1. Los Alamos National Laboratory HCV database (http://www.hcv.lanl.gov) accession numbers of 386 of 412 sequences shared with us by Dr. Todd Allen (Harvard Medical School) which are identical to publicly available sequences that now appear in this database.

Mutations	Epitope	ΔE	%	References
D2597E	A*02 ALYDVVSKL$_{2594-2602}$	5.9	22.2	[22]
Q2729R	A*02 GLQDCTMLV$_{2727-2735}$	2.7	4.5	[23]
K2518R	A*03 SLTPPHSAK$_{2510-2518}$	1.7	2.5	[24]
S2510N	A*31 epitope incompletely defined	4.7	12.4	[15, 16]
Y3006C	A*31 VGIYLLPNR$_{3003-3011}$	6.5	27.8	[15]
H2453Y	B*15 LLRHHNMVY$_{2450-2458}$	5.2	15.6	[15, 25]
M2456L	B*15 LLRHHNMVY$_{2450-2458}$	0.0	0.1	[25]
S2466C	B*15 SQRQKKVTF$_{2466-2474}$	0.0	0.1	[25–27]
K2471R	B*15 SQRQKKVTF$_{2466-2474}$	7.2	36.1	[15]
Q2467H	B*15 SQRQKKVTF$_{2466-2474}$	4.0	7.8	[15]
Q2467K	B*15 SQRQKKVTF$_{2466-2474}$	9.0	74.4	[15]
Q2467L	B*15 SQRQKKVTF$_{2466-2474}$	2.8	4.6	[15, 25]
R2937S	B*27 GRAAICGKY$_{2936-2944}$	10.2	96.6	[10]
R2937K	B*27 GRAAICGKY$_{2936-2944}$	0.0	0.1	[10]
I2940T	B*27 GRAAICGKY$_{2936-2944}$	5.5	18.4	[10]
K2943R	B*27 GRAAICGKY$_{2936-2944}$	3.6	7.0	[10]
K2629N	B*57 KSKKTPMGF$_{2629-2637}$	4.6	11.2	[11]
K2629Q	B*57 KSKKTPMGF$_{2629-2637}$	6.8	31.6	[11]
K2629S	B*57 KSKKTPMGF$_{2629-2637}$	5.8	21.6	[11]
T2633A	B*57 KSKKTPMGF$_{2629-2637}$	4.8	13.0	[11]
T2633S	B*57 KSKKTPMGF$_{2629-2637}$	4.4	9.8	[11]
T2633V	B*57 KSKKTPMGF$_{2629-2637}$	4.2	9.0	[11]
T2633N	B*57 KSKKTPMGF$_{2629-2637}$	3.1	5.1	[11, 28]
H2453N	unknown SLLRHHNLVYSTTSRSA$_{2449-2465}$	4.6	10.9	[21]
Q2467K/K2471R	B*15 SQRQKKVTF$_{2466-2474}$	16.3	61.4	[15]
A2841V/I2844V	B*27 ARMILMTHF$_{2841-2849}$	7.3	3.7	[29]
A2841V/M2846L	B*27 ARMILMTHF$_{2841-2849}$	4.0	2.1	[29]
M2846L/T2847P	B*27 ARMILMTHF$_{2841-2849}$	10.5	10.1	[29]
I2844V/T2847P	B*27 ARMILMTHF$_{2841-2849}$	10.8	11.1	[29]
R2937K/I2940V	B*27 GRAAICGKY$_{2936-2944}$	5.2	2.5	[10]
R2937G/I2940T	B*27 GRAAICGKY$_{2936-2944}$	14.0	33.5	[10]
F2849L/I2854M	unknown THFFSVLIARDQ$_{2847-2858}$	9.9	8.2	[21]
A2841V/I2844V/M2846L	B*27 ARMILMTHF$_{2841-2849}$	7.8	1.9	[29]
A2841V/I2844V/T2847S	B*27 ARMILMTHF$_{2841-2849}$	16.5	9.3	[29]
A2841V/M2846L/T2847P	B*27 ARMILMTHF$_{2841-2849}$	10.3	3.0	[29]

Table S2. List of the 35 escape mutations analyzed in Section 3.2 of the main text, the associated epitope and HLA allele, energy of the mutant relative to the H77 wild type $\Delta E = (E - E_{wt})$, and the percentile within which the mutant is located on the energy spectrum of all possible mutants of the same order.

Index	Epitope	Position	HLA	$\Delta < E >$	Dominant	Protective	References
1	QPEKGGRKPA	2568-2577	B*55	7.06	N	N	[17, 18, 22]
2	KSKKTPMGF*	2629-2637	B*57	6.45	Y	Y	[11, 17, 22, 28, 30]
3	SPGEINRVAA	2898-2907	B*55	6.38	N	N	[17, 18]
4	RVCEKMALY	2588-2596	A*03	5.28	N	N	[30–32]
5	CYSIEPLDL	2871-2879	A*24:02	4.97	N	N	[33]
6	LGVPPLRAWR	2912-2921	B*57	4.71	N	Y	[17, 21]
7	VGIYLLPNR	3003-3011	A*31	4.51	N	N	[30]
8	RMILMTHFF	2842-2850	A*24:02	4.50	N	N	[33]
9	TARHTPVNSW*	2819-2828	A*25	3.88	N	N	[17, 34]
10	ARHTPVNSW	2820-2828	B*27, B*27:02	3.85	N	Y	[35, 36]
11	ARMILMTHF	2841-2849	B*27, B*27:01, B*27:02	3.85	Y	Y	[19, 22, 29, 35, 36]
12	APTLWARMVL*	2836-2845	B*07	3.67	N	N	[22, 37]
13	HDGAGKRVYYL	2794-2804	B*38	3.47	N	N	[30]
14	HDGAGKRVY	2794-2802	A*03	3.47	N	N	[32]
15	GRAAICGKY	2936-2944	B*27, B*27:02	3.24	N	Y	[10, 35, 36]
16	ALYDVVTKL*	2594-2602	A*02, A*02:01	3.16	Y	N	[17–22]
17	RLIVFPDLGV	2578-2587	A*02	2.89	N	N	[38, 39]
18	GLQDCTMLV	2727-2735	A*02, A*02:01	1.15	N	N	[18, 23, 32, 39]
19	SVRARLLSR	2926-2934	A*03	1.11	N	N	[22]
20	SLTPPHSAK	2510-2518	A*03	0.83	N	N	[30]
21	VYSTTSRSASL	2457-2467	A*24:02	-0.06	N	N	[40]
22	SQRQKKVTF	2466-2474	B*15	-0.07	N	N	[15, 25, 37]
23	LLRHHNMVY	2450-2458	B*15	-0.09	N	N	[15, 25, 37]
24	SYTWTGALI	2423-2431	A*24:02	-0.12	N	N	[41]

Table S3. List of the 24 NS5B CTL epitopes discussed in Section 3.5 of the main text which are precisely mapped and for which the restricting allele is known. The index reported in the first column corresponds to the indices reported in the "Epitopes" column in Table S4. The asterisk character in the second column indicates those epitopes for which some mutations are known to be cross reactive. In calculating the cost of escape for these epitopes we eliminated under our simulated CTL targeting procedure detailed in Section 3.5 all strains reported to be antigenic. We also report the change in the average energy of a strain in the population, $\Delta\langle E\rangle$, upon eliminating those strains with wild type epitopes, whether the epitope is reported to be immunodominant, and if the associated HLA allele is correlated with protection.

Index	# Components	$\overline{\Delta\langle E\rangle}$	Population Coverage	Epitopes
1	1	1.1	35.0%	16
2	2	2.1	35.0%	16 17
3	2	1.9	50.1%	4 16
4	3	2.9	50.1%	4 16 17
5	3	2.1	54.4%	4 9 16
6	3	2.6	52.9%	4 12 16
7	3	2.3	54.4%	2 4 16
8	4	3.0	54.4%	4 9 16 17
9	4	3.6	52.9%	4 12 16 17
10	4	3.3	54.4%	2 4 16 17
11	4	2.8	57.2%	4 9 12 16
12	4	2.5	58.7%	2 4 9 16
13	4	3.0	57.2%	2 4 12 16
14	5	3.8	57.2%	4 9 12 16 17
15	5	3.5	58.7%	2 4 9 16 17
16	5	4.4	52.9%	4 12 14 16 17
17	5	4.0	57.2%	2 4 12 16 17
18	5	3.2	61.4%	2 4 9 12 16
19	5	2.6	61.9%	2 4 9 13 16
20	6	4.8	52.9%	4 12 14 16 17 18
21	6	4.5	57.2%	4 9 12 14 16 17
22	6	4.2	61.4%	2 4 9 12 16 17
23	6	3.6	61.9%	2 4 9 13 16 17
24	6	4.8	57.2%	2 4 12 14 16 17
25	6	3.3	64.7%	2 4 9 12 13 16
26	7	5.0	57.2%	4 9 12 14 16 17 18
27	7	5.2	57.2%	2 4 12 14 16 17 18
28	7	3.4	66.7%	2 4 5 9 12 13 16
29	7	4.9	61.4%	2 4 9 12 14 16 17
30	7	4.3	64.7%	2 4 9 12 13 16 17
31	7	3.4	67.1%	2 4 7 9 12 13 16
32	8	5.4	61.4%	2 4 9 12 14 16 17 18
33	8	5.6	57.2%	2 4 6 12 14 16 17 18
34	8	4.4	66.7%	2 4 5 9 12 13 16 17
35	8	3.5	69.2%	2 4 5 7 9 12 13 16
36	8	5.0	64.7%	2 4 9 12 13 14 16 17
37	8	4.4	67.1%	2 4 7 9 12 13 16 17
38	9	5.5	64.7%	2 4 9 12 13 14 16 17 18
39	9	5.8	61.4%	2 4 6 9 12 14 16 17 18
40	9	5.9	57.2%	2 4 6 12 14 16 17 18 19
41	9	5.2	66.7%	2 4 5 9 12 13 14 16 17
42	9	4.5	69.2%	2 4 5 7 9 12 13 16 17
43	9	3.5	71.0%	2 4 5 7 9 12 13 16 22

Continued on next page

44	9	3.5	69.3%	1 2 4 5 7 9 12 13 16
45	9	5.2	67.1%	2 4 7 9 12 13 14 16 17
46	10	5.6	66.7%	2 4 5 9 12 13 14 16 17 18
47	10	6.0	61.4%	2 4 6 9 12 14 16 17 18 19
48	10	5.6	67.1%	2 4 7 9 12 13 14 16 17 18
49	10	5.9	64.7%	2 4 6 9 12 13 14 16 17 18
50	10	5.3	69.2%	2 4 5 7 9 12 13 14 16 17
51	10	4.5	71.0%	2 4 5 7 9 12 13 16 17 22
52	10	4.6	69.3%	1 2 4 5 7 9 12 13 16 17
53	10	3.5	71.2%	1 2 4 5 7 9 12 13 16 22
54	11	5.7	69.2%	2 4 5 7 9 12 13 14 16 17 18
55	11	6.1	66.7%	2 4 5 6 9 12 13 14 16 17 18
56	11	6.2	64.7%	2 4 6 9 12 13 14 16 17 18 19
57	11	6.0	67.1%	2 4 6 7 9 12 13 14 16 17 18
58	11	5.3	71.0%	2 4 5 7 9 12 13 14 16 17 22
59	11	5.3	69.3%	1 2 4 5 7 9 12 13 14 16 17
60	11	4.6	71.2%	1 2 4 5 7 9 12 13 16 17 22
61	12	6.3	66.7%	2 4 5 6 9 12 13 14 16 17 18 19
62	12	5.7	71.0%	2 4 5 7 9 12 13 14 16 17 18 22
63	12	5.7	69.3%	1 2 4 5 7 9 12 13 14 16 17 18
64	12	6.2	69.2%	2 4 5 6 7 9 12 13 14 16 17 18
65	12	6.3	67.1%	2 4 6 7 9 12 13 14 16 17 18 19
66	12	5.3	71.2%	1 2 4 5 7 9 12 13 14 16 17 22
67	13	6.4	69.2%	2 4 5 6 7 9 12 13 14 16 17 18 19
68	13	5.7	71.2%	1 2 4 5 7 9 12 13 14 16 17 18 22
69	13	6.2	71.0%	2 4 5 6 7 9 12 13 14 16 17 18 22
70	13	6.2	69.3%	1 2 4 5 6 7 9 12 13 14 16 17 18
71	14	6.5	69.2%	2 4 5 6 7 8 9 12 13 14 16 17 18 19
72	14	6.4	71.0%	2 4 5 6 7 9 12 13 14 16 17 18 19 22
73	14	6.4	69.3%	1 2 4 5 6 7 9 12 13 14 16 17 18 19
74	14	6.2	71.2%	1 2 4 5 6 7 9 12 13 14 16 17 18 22
75	15	6.5	69.2%	2 4 5 6 7 8 9 10 12 13 14 16 17 18 19
76	15	6.5	71.0%	2 4 5 6 7 8 9 12 13 14 16 17 18 19 22
77	15	6.5	69.3%	1 2 4 5 6 7 8 9 12 13 14 16 17 18 19
78	15	6.4	71.2%	1 2 4 5 6 7 9 12 13 14 16 17 18 19 22
79	16	6.5	71.0%	2 4 5 6 7 8 9 10 12 13 14 16 17 18 19 22
80	16	6.5	71.2%	1 2 4 5 6 7 8 9 12 13 14 16 17 18 19 22
81	16	6.5	69.3%	1 2 3 4 5 6 7 8 9 12 13 14 16 17 18 19
82	17	6.5	69.3%	1 2 3 4 5 6 7 8 9 10 12 13 14 16 17 18 19
83	17	6.5	71.2%	1 2 3 4 5 6 7 8 9 12 13 14 16 17 18 19 22
84	18	6.5	69.3%	1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17 18 19
85	18	6.5	71.2%	1 2 3 4 5 6 7 8 9 10 12 13 14 16 17 18 19 22
86	19	6.5	71.2%	1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17 18 19 22

Continued on next page

Table S4: List of the 86 optimal immunogen candidates residing on the Pareto frontier of Figure 7. The number of components corresponds to the number of epitopes in the immunogen candidate, the population coverage is the fraction of the target population of the the top 66 haplotypes of North Americans who respond to at least one epitope in the immunogen candidate, and $\overline{\Delta\langle E\rangle}$ is the weighted averaged impact of the immunogen upon the fitness of the viral ensemble within the target population as defined in Section 3.6. The particular epitopes in the immunogen candidate reported in the last column correspond to the indices in the "Index" column in Table S3.

Months	E	2450 2458 LLRHHNLVY	···	2466 2474 SQRQKKVTF	···	2510 S	···	2578 2587 RLIVFPDLGV	···	2594 2602 ALYDVVSKL	···	2727 2735 GLQDCTMLV	···	3003 3011 VGIXXXXXX
0.0	10.6	.....		.....		.		.....		.....		.....		.....
	10.6	.....		.....		.		.....		.....		.....		.....
	15.7	.....		.....		.		.....		.....		.....		.....
	18.9	.....		.....		.		.....		.....		.....		.....
	22.1	.....		.....		.		.....		.....		.....		.....
	23.1	.....		.....		N		.....		.....		.....		.....
	24.9	.....		.....		.		.....		.....		.....		.....
	25.3	.....		.....		.		.....		.....		.....		.....
	27.4	.....		.....		.		.....		.....		.....		.....
	33.8	.....		.....		.		.....		.....		.....		.....
	36.8	.....		.....		.		.....		..H.....		.....		.....
	39.9	.....		.....		.		.....		.....		.....		.....
	40.5	.....		....R....		.		.....		.....		.....		.....
	49.7	.....		.....		.		.....		.....		.....		.....
	53.4	..P.....		.......S.		.		.....		.....		.....		.....
1.3	10.9	.....		.....		.		.....		.....		.....		.....
	19.5	.....		.....		.		.....H....		.....		.....		.....
	19.7	.....		.....		.		.....		.....		.....		.....
	23.7	.....		.....		.		.....		.....		.....		.....
	27.5	.....		.....		.		.....		.....		.....		.....
	35.5	.....		.....		.		.....		.....		.....		.....
	43.1	.....		.....		.		....S.....		.....		.....		.....
	52.0	..P.....		.....		.		.....		.....		.....		.....
	58.3	........H		.....		.		.....		......G.P		.....		.....
7.2	51.4	...Y.....		.H...R...		N		.....		.....		.....		.....
	57.0	...Y.....		.L...R...		.		.....		.....		.....		.....
	63.1	...Y.....		.K...R...		N		.....		.....		.......P.		.....
	65.1	...Y.....		.H...R...		N		.....		.....		.....		.....
	68.9	...Y.....		.L...R..L		N		.....		.....		.....		.....
	104.6	...Y.....		.K...R...		N		.....		.....		.....		.....
10.7	69.7	...Y.....		.K...R...		N		.....		.....		.....		.....
	77.7	...Y.....		.K...R...		N		.....		.....		.....		.....
	84.0	...Y.....		.K..ER...		N		.....		.....		.....		.....
	84.0	...Y.....		.K..ER...		N		.....		.....		.....		.....
	87.4	...Y.....		.H...R...		N		.....		.....		.....		E.....
	101.2	...Y.....		.H...R...		N		.....		.....		.....		..V.....
	115.9	...Y.....		.H...R...		N		.......P..		.....		......T..		.....
	120.0	...Y.....		.K...R...		N		.....		.....		.....		.....
	133.5	...Y.....		.K...R...		N		.....		.....		.....		..V.....

Continued on next page

Months	E	2450 2458 LLRHHNLVY	···	2466 2474 SQRQKKVTF	···	2510 S	···	2578 2587 RLIVFPDLGV	···	2594 2602 ALYDVVSKL	···	2727 2735 GLQDCTMLV	···	3003 3011 VGIXXXXXX
14.5	62.0	...Y.....		.K...R...		N		....		....		....		....
	65.6	...Y.....		.H...R...		N		....		....		....		....
	65.6	...Y.....		.K...R...		N		....		....		....		....
	69.3	...Y.....		.K...R...		N		....		....		....		....
	70.0	...Y....H		.K...R...		N		....		....		....		....
	70.9	...Y.....		.R...R...		N		....		....		....		....
	73.4	...Y.....		.H...R...		N		....		....		....		....
	80.1	...Y.....		.K...R...		N		....		....		....		....
16.8	31.2	...Y.....		.L.....		N		....		....		....		....
	38.0	...Y.....		.L.....		N		....		....		....		....
	46.6	...Y.....		.L.....		N		...A.....		.P.....		....		....
	46.6	...Y.....		.L.....		N		...A.....		.P.....		....		....
	55.1	...Y.....		.L.....		N		....		....		....		....
	57.1	...Y.....		.L..E....		N		....		....		....		....
	61.1	...Y.....		.K...R...		N		..L.....		....		....		....
	70.7	...Y.....		.L.....		N		....		....		....		..V.....
	87.4	...Y.....		.L.....		N		....		....		....		....
	99.1	...Y.....		.L.....		N		....		....		....		....
18.2	29.1	...Y.....		.L.....		N		....		....		....		....
	36.0	...Y.....		.K...R...		N		....		....		....		....
	45.7	...Y.....		.L.....		N		....		.......R.		....		....
	58.4	...Y.....		.L.....		N		....		....		....		....
	66.2	...Y.....		.L.....		N		....		....		....		....
	80.6	...Y.....		.L.....		N		....		..H.A....		....		....
	81.1	...Y.....		.L.R...A.		N		....		....		....		....
	82.5	...Y.....		.L.....		N		....		....		....		....
	84.4	...Y.....		.L.....		N		....		....		....		....
	93.7	...Y.....		.L.....		N		....		....A....		....		....
21.5	28.0	...Y.....		.H.....		N		....		....		....		....
	31.1	...Y.....		.R.....		N		....		....		....		....
	31.8	...Y.....		.H.....		N		....		....		....		....
	47.2	...Y.....		.H.....		N		....		....		....		....
	48.0	...Y.....		.K...R...		N		....		....		....		....
	64.6	...Y.....		.K...R...		N		....		....		....		....
	71.2	...Y....H		.K...R...		N		....		....		....		....
	76.3	...Y.....		.R.....		N		....		....		....		....
	97.9	...Y.....		.KG..R...		N		....		....		....		....
36.9	76.6	...Y.....		.H.....		N		....V.....		....		....		....
	83.7	...Y.....		.H.....		N		....		....		....		....
	102.2	...Y.....		.H.....		N		....		..H.....		.P.....		....
	102.6	...Y.....		.H.....		N		....		....		....		....
	122.1	...Y.....		.H.....		N		....		....		....		....
49.0	36.6	...Y.....		.H.....		N		....		....		....		....

Continued on next page

Months	E	2450 2458 LLRHHNLVY	···	2466 2474 SQRQKKVTF	···	2510 S	···	2578 2587 RLIVFPDLGV	···	2594 2602 ALYDVVSKL	···	2727 2735 GLQDCTMLV	···	3003 3011 VGIXXXXXX
	46.0	...Y.....		.T...R...		N		....		....		....		....
	46.5	...Y.....		.H.....		N		....		....		....		....
	46.6	...Y.....		.H.....		N		....		....		....		....
	46.7	...Y.....		.T...R...		N		....		....		....		....
	54.5	...Y.....		.T...R...		N		....		....		....		....
	54.5	...Y.....		.T...R...		N		....		....		....		I.....
Child C003														
7.2	10.6	....		....		.		....		....		....		....
	10.6	....		....		.		....		....		....		....
	10.6	....		....		.		....		....		....		....
	19.4	....		....		.		.P.....		....		....		....
	23.8	....		....		.		....		....		....		....
	32.0	....		....		.		....		....		.....A		....
	34.2	....		....		.		....		....		....		....
	44.6	....		....		.		....		....		....		....
Child D003														
20.1	31.3	...Y.....		.L.....		N		....		....		....		....
	39.2	...Y.....		.L.....		N		....		....		....		....
	78.3	...Y.....		.L.....		N		....V.....		....		....		.V.....
	85.3	...Y.....		.L.....		N		....		....		....		....
	98.1	...Y.....		.L.....		N		....		....		....		....
	112.2	...Y.....		.L.....		N		....		....		....		.X.....
	153.3	...Y.....		.L.....		N		....		....		....		.V.....

Table S5. The amino acid sequence of six epitopes (B*15-LLRHHNMVY$_{2450-2458}$, B*15-SQRQKKVTF$_{2466-2474}$, A*02-RLIVFPDLGV$_{2578-2587}$, A*02-ALYDVVSKL$_{2594-2602}$, A*02-GLQDCTMVL$_{2727-2735}$, and A*31-VGIYLLPNR$_{3003-3011}$) and one HLA associated polymorphism (S2510N) from M003 and her children (C003 and D003) with the energies assigned to the complete NS5B sequence by our model. The shaded regions of the table indicate periods of time during which M003 was pregnant with C003 and, subsequently, D003.

[1] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, and A. K. Chakraborty, Immunity **38**, 606 (2013).

[2] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, Physical Review E **88**, 062705 (2013).

[3] P. C. Matthews, A. J. Leslie, A. Katzourakis, H. Crawford, R. Payne, A. Prendergast, K. Power, A. D. Kelleher, P. Klenerman, J. Carlson, D. Heckerman, T. Ndung'u, B. D. Walker, T. M. Allen, O. G. Pybus, and P. J. R. Goulder, Journal of Virology **83**, 4605 (2009).

[4] M. J. Donlin, N. A. Cannon, E. Yao, J. Li, A. Wahed, M. W. Taylor, S. H. Belle, A. M. Di Bisceglie, R. Aurora, and J. E. Tavis, Journal of Virology **81**, 8211 (2007).

[5] T. Kuntzen, J. Timm, A. Berical, N. Lennon, A. M. Berlin, S. K. Young, B. Lee, D. Heckerman, J. Carlson, L. L. Reyor, M. Kleyman, C. M. McMhon, C. Birch, J. Schulze zur Wiesch, T. Ledlie, M. Koehrsen, G. M. Lauer, H. R. Rosen, F. Bihl, A. Cerny, U. Spengler, Z. Liu, A. Y. Kim, Y. Xing, A. Schneidwind, J. F. Madey, Margaret A. Fleckenstein, V. M. Park, J. E. Galagan, C. Nusbaum, B. D. Walker, E. S. Lake-Bakaar, Gerond V. Daar, I. M. Jacobson, B. R. Gomperts, Edward D. Edlin, S. M. Donfield, R. T. Chung, A. H. Talal, T. Marion, B. W. Birren, M. R. Henn, and T. M. Allen, Hepatology **48**, 1769 (2008).

[6] D. J. Bartels, J. C. Sullivan, E. Z. Zhang, A. M. Tigges, J. L. Borrian, S. De Meyer, D. Takemoto, E. Dondero, A. D. Kwong, G. Picchio, and T. L. Kieffer, Journal of Virology **87**, 1544 (2013).

[7] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette, and B. Peters, Nucleic Acids Research **38**, D854 (2010).

[8] K. Cao and M. Fernández-Viña, in *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference*, Vol. 1, edited by J. A. Hansen (IHWG Press, Seattle, WA, 2007) pp. 648–651.

[9] K. P. Burke, S. Munshaw, W. O. Osburn, J. Levine, L. Liu, J. Sidney, A. Sette, S. C. Ray, and A. L. Cox, Journal of Immunology **188**, 5177 (2012).

[10] C. Neumann-Haefelin, C. Oniangue-Ndza, T. Kuntzen, J. Schmidt, K. Nitschke, J. Sidney, C. Caillet-Saguy, M. Binder, N. Kersting, M. W. Kemper, K. A. Power, S. Ingber, L. L. Reyor, K. Hills-Evans, A. Y. Kim, G. M. Lauer, V. Lohmann, A. Sette, M. R. Henn, S. Bressanelli, R. Thimme, and T. M. Allen, Hepatology **54**, 1157 (2011).

[11] C. Oniague-Ndza, T. Kuntzen, M. Kemper, A. Berical, Y. E. Wang, C. Neumann-Haefelin, P. K. Foote, K. Hills-Evans, L. L. Reyor, K. Kane, A. D. Gladden, A. K. Bloom, K. A. Power, R. Thimme, G. M. Lauer, M. R. Henn, A. Y. Kim, and T. M. Allen, Journal of Virology **85**, 11883 (2011).

[12] D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*, 2nd ed. (Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, 2006).

[13] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic,

T. Hwa, and M. Weigt, Proceedings of the National Academy of Sciences **108**, E1293 (2011).

[14] R. Durbin, S. R. Eddy, A. Kroph, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, 1st ed. (Cambridge University Press, 1998).

[15] J. R. Honegger, S. Kim, A. A. Price, J. A. Kohout, K. L. McKnight, M. R. Prasad, S. M. Lemon, A. Grakoui, and C. M. Walker, Nature Medicine **19**, 1529 (2013).

[16] A. Rauch, I. James, K. Pfafferott, D. Nolan, P. Klenerman, W. Cheng, L. Mollison, G. McCaughan, N. Shackel, G. P. Jeffrey, R. Baker, E. Freitas, I. Humphreys, H. Furrer, H. F. Günthard, B. Hirschel, S. Mallal, M. John, M. Lucas, E. Barnes, and S. Gaudieri, Hepatology **50**, 1017 (2009).

[17] G. M. Lauer, E. Barnes, M. Lucas, J. Timm, K. Ouchi, A. Y. Kim, C. L. Day, G. K. Robbins, D. R. Casson, M. Reiser, G. Dusheiko, T. M. Allen, R. T. Chung, B. D. Walker, and P. Klenerman, Gastroenterology **127**, 924 (2004).

[18] A. L. Cox, T. Mosbruger, G. M. Lauer, D. Pardoll, D. L. Thomas, and S. C. Ray, Hepatology **42**, 104 (2005).

[19] J. Schmidt, A. K. N. Iversen, S. Tenzer, E. Gostick, D. A. Price, V. Lohmann, U. Distler, P. Bowness, H. Schild, H. E. Blum, P. Klenerman, C. Neumann-Haefelin, and R. Thimme, PLOS Pathogens **8**, E1003042 (2012).

[20] H. C. Spangenberg, S. Viazov, N. Kersting, C. Neumann-Haefelin, D. McKinney, M. Roggendorf, F. von Weizsäcker, H. E. Blum, and R. Thimme, Hepatology **42**, 828 (2005).

[21] T. Kuntzen, J. Timm, A. Berical, L. L. Lewis-Ximenez, A. Jones, B. Nolan, J. S. zur Wiesch, B. Li, A. Schneidwind, A. Y. Kim, R. T. Chung, G. M. Lauer, and T. M. Allen, Journal of Virology **81**, 11658 (2007).

[22] C. Neumann-Haefelin, J. Timm, H. C. Spangenberg, N. Wischniowski, N. Nazarova, N. Kersting, M. Roggendorf, T. M. Allen, H. E. Blum, and R. Thimme, Hepatology **47**, 1824 (2008).

[23] K. M. Chang, B. Rehermann, J. G. McHutchison, C. Pasquinelli, S. Southwood, A. Sette, and F. V. Chisari, The Journal of Clinical Investigation **100**, 2376 (1997).

[24] S. Merani, D. Petrovic, I. James, A. Chopra, D. Cooper, E. Freitas, A. Rauch, J. di Iulio, M. John, M. Lucas, K. Fitzmaurice, S. McKiernan, S. Norris, D. Kelleher, P. Klenerman, and S. Gaudieri, Hepatology **53**, 396 (2011).

[25] M. Ruhl, P. Chhatwal, H. Strathmann, T. Kuntzen, D. Bankwitz, K. Skibbe, A. Walker, F. M. Heinemann, P. A. Horn, D. Allen, Todd M. Hoffmann, T. Pietschmann, and J. Timm, Journal of Virology **86**, 991 (2012).

[26] I. Tester, S. Smyk-Pearson, P. Wang, A. Wertheimer, E. Yao, D. M. Lewinsohn, J. E. Tavis, and H. R. Rosen, Journal of Experimental Medicine **201**, 1725 (2005).

[27] J. S. zur Wiesch, G. M. Lauer, C. L. Day, A. Y. Kim, K. Ouchi, J. E. Duncan, A. G. Wurcel, J. Timm, A. M. Jones, B. Mothe, T. M. Allen, B. McGovern, L. Lewis-Ximenez, J. Sidney, A. Sette, R. T. Chung, and B. D. Walker, Journal of Immunology **175**, 3603 (2005).

[28] A. Y. Kim, T. Kuntzen, J. Timm, B. E. Nolan, M. A. Baca, L. L. Reyor, A. C. Berical, A. J. Feller,

K. L. Johnson, J. S. Z. Wiesch, G. K. Robbins, R. T. Chung, W. B. D., M. Carrington, T. M. Allen, and G. M. Lauer, Gastroenterology **140**, 686 (2011).

[29] E. Dazert, C. Neumann-Haefelin, S. Bressanelli, K. Fitzmaurice, J. Kort, J. Timm, S. McKiernan, D. Kelleher, N. Gruener, J. E. Tavis, H. R. Rosen, J. Shaw, P. Bowness, H. E. Blum, P. Klenerman, R. Bartenschlager, and R. Thimme, Journal of Clinical Investigation **119**, 376 (2009).

[30] D. K. Wong, D. D. Dudley, N. H. Afdhal, J. Dienstag, C. M. Rice, L. Wang, M. Houghton, B. D. Walker, and M. J. Koziel, Journal of Immunology **160**, 1479 (1998).

[31] M. J. Koziel, D. Dudley, N. Afdhal, A. Grakoui, C. M. Rice, Q. L. Choo, M. Houghton, and B. D. Walker, Journal of Clinical Investigation **96**, 2311 (1995).

[32] N. H. Gruener, M.-C. Jung, A. Ulsenheimer, J. T. Gerlach, R. Zachoval, H. M. Diepolder, G. Baretton, R. Schauer, G. R. Pape, and C. A. Schirren, Liver Transplantation **10**, 1487 (2004).

[33] Y. Nakamoto, S. Kaneko, H. Takizawa, Y. Kikumoto, M. Takano, Y. Himeda, and K. Kobayashi, Journal of Medical Virology **70**, 51 (2003).

[34] C. Kuiken, K. Yusim, L. Boykin, and R. Richardson, Bioinformatics **21**, 376 (2005).

[35] C. Neumann-Haefelin, S. McKiernan, S. Ward, S. Viazov, H. C. Spangenberg, T. Killinger, T. F. Baumert, N. Nazarova, I. Sheridan, O. Pybus, F. von Weizsäcker, M. Roggendorf, D. Kelleher, P. Klenerman, H. E. Blum, and R. Thimme, Hepatology **43**, 563 (2006).

[36] K. Nitschke, A. Barriga, J. Schmidt, J. Timm, S. Viazov, T. Kuntzen, A. Y. Kim, G. M. Lauer, T. M. Allen, S. Gaudieri, A. Rauch, C. M. Lange, C. Sarrazin, T. Eiermann, J. Sidney, A. Sette, R. Thimme, D. López, and C. Neumann-Haefelin, Journal of Hepatology **60**, 22 (2014).

[37] M. Ruhl, T. Knuschke, K. Schewior, L. Glavinic, C. Neumann-Haefelin, D.-I. Chang, M. Klein, F. M. Heinemann, H. Tenckhoff, M. Wiese, P. A. Horn, S. Viazov, U. Spengler, M. Roggendorf, N. Scherbaum, J. Nattermann, D. Hoffmann, and J. Timm, Gastroenterology **140**, 2064 (2011).

[38] P. Scognamiglio, D. Accapezzato, M. A. Casciaro, A. Cacciani, M. Artini, G. Bruno, M. L. Chircu, J. Sidney, S. Southwood, S. Abrignani, A. Sette, and V. Barnaba, Journal of Immunology **162**, 6681 (1999).

[39] N. H. Grüner, T. J. Gerlach, M. C. Jung, H. M. Diepolder, C. A. Schirren, W. W. Schraut, R. Hoffmann, R. Zachoval, T. Santantonio, M. Cucchiarini, A. Cerny, and G. R. Pape, Journal of Infectious Diseases **181**, 1528 (2000).

[40] T. Hakamada, K. Funatsuki, H. Morita, T. Ugajin, I. Nakamura, H. Ishiko, Y. Matsuzaki, N. Tanaka, and M. Imawari, Journal of General Virology **85**, 1521 (2004).

[41] T. Mashiba, K. Udaka, Y. Hirachi, Y. Hiasa, T. Miyakawa, Y. Satta, T. Osoda, S. Kataoka, M. Kohara, and M. Onji, Immunogenetics **59**, 197 (2007).