Supplementary Information for

Superinfection and Cure of Infected Cells as Mechanisms For Hepatitis C Virus adaptation and Persistence

Ruian Ke[1], Hui Li[1], Shuyi Wang, Wenge Ding, Ruy M. Ribeiro, Elena E. Giorgi, Tanmoy Bhattacharya, Richard J.O. Barnard, Beatrice H. Hahn, George M. Shaw, Alan S. Perelson

[1]R.K. and H.L. contributed equally to this work.

Corresponding authors: Beatrice H. Hahn (bhahn@pennmedicine.upenn.edu) and Alan S. Perelson (asp@lanl.gov)

This PDF file includes:

SI Methods and Results
Figs. S1 to S17
Tables S1 to S9
References for SI reference citations

SI Methods and Results

www.pnas.org/cgi/doi/10.1073/pnas.1805267115

Table of contents

Parameter fitting, sensitivity analysis and model selection

For each participant, we fitted the viral load and the sequence data simultaneously. To do this, we calculated the sum of squared residuals (SSR) between model simulations and both the viral load data and the sequence data as follows. We first normalized the viral load data and the simulated viral loads by their initial values (values before treatment), respectively, and then calculated the squared difference between the normalized viral load data and the simulation results (on a $\log_{10}$ scale) for the time points where sequence data are not available. For the time points where sequence data are available, we derived the normalized viral loads for each strain considered in the model as a product of the total normalized viral load and the frequency of each strain in the sequence data, and then calculated the squared difference between the normalized viral loads of each strain derived from sequence data and the normalized viral load from the model simulation (on a $\log_{10}$ scale). The final SSR is calculated by summing all the squared differences calculated with equal weighting. The expression for SSR is:

$$SSR = \sum_{i} (\log_{10}(\tilde{V}_i) - \log_{10}(V_i))^2 + \sum_{j} \sum_{k} (\log_{10}(\widetilde{W}_{j,k}) - \log_{10}(W_{j,k}))^2$$

where $\tilde{V}_i$ and V_i are the normalized viral loads (from data and simulation, respectively) at the ith time point at which sequence data is not available. $\widetilde{W}_{j,k}$ and $W_{j,k}$ are the normalized viral loads (from data and simulation, respectively) for the kth mutant in a subject at the jth time point at which sequence data is available.

Note that, in cases where the viral load is below the limit of quantification or a viral strain is below the limit of sequencing quantification, i.e. no sequence belonging to the viral strain is detected in a sample, we calculate the SSR as follows. If the viral load is below the lower limit of quantification in the data, the SSR for that data point is set to 0 if the simulated viral load is also below the quantification threshold; otherwise, if the simulated viral load is above the quantification threshold, the SSR is calculated as the squared difference between the simulated viral load and the quantification threshold. In cases where no sequence belonging to a viral strain is detected in a sample, we first calculate the lower limit of quantification for the viral load of that strain based on the total number of sequences in the sample, *m*, and the total viral load at that time point. Specifically, we assume binomial sampling for the sequence data, and calculate the threshold below which there is a 95% chance that the viral strain is not sampled if we sample *m* sequences. If the simulated viral load for the strain is higher than the quantification threshold, we calculate the squared difference between the simulated viral load and the threshold; otherwise, the squared difference is set to 0.

We fitted each model to the viral load data and the sequence data sampled for each subject simultaneously. A total of 10,000 optimization runs were performed for each model using the Nelder-Mead algorithm (1), with each run starting with a parameter set randomly drawn from the range of plausible parameters. The parameters that are fitted and the best-fit parameter values are shown in Tables S3-S7 for all 5 subjects. Note that, the EC_{50} values for a number of drug sensitive/baseline viruses were measured previously using *in vitro* systems (shown in Table S2) (2, 3). In situations where both the *in vitro* measurements of the EC_{50} values for the baseline virus ($EC_{50,1}$) and the *i*th mutant virus

($EC_{50,i}$) are available, we fit the EC_{50} value for the baseline virus ($EC_{50,1}$), and calculate the EC_{50} values for the resistant mutants according to their reported fold-resistance relative to the wild-type (i.e. $EC_{50,i}/EC_{50,1}$). In this way, we account for the potential differences in the EC_{50} values between *in vitro* and *in vivo* systems, and the differences between the plasma and tissue drug concentrations (4) in our fitting.

To compare multiple models, we calculated the corrected Akaike Information Criterion (AICc) scores, which corrects for small sample size (5):

$$AICc = n \cdot \log\left(\frac{SSR}{n}\right) + 2 \cdot k \cdot \frac{n}{n-k-1}$$

where k is the number of fitted parameters (between 8 and 16 in our models) and n is the number of observations/data points (between 31 and 42). As the data from each participant was fitted independently, we also calculated the total AICc score, which represents an overall score for how well a model variation performs in 5 independent 'tests'.

To evaluate the uncertainties in the parameter estimation, we performed likelihood profiling (6) for each estimated parameter in the best model in each subject. Specifically, for each parameter, we fixed its value above or below its best-fit value, and fitted the values of other parameters. Repeating the fitting by fixing the parameter of interest at different values, we determine how the likelihood changes with changes in the parameter of interest. We determine the confidence intervals of the parameter of interest by choosing a 2-log likelihood cutoff, i.e. the value of the parameter above and below its fitted value that give a 2-log decrease in the likelihood, which corresponds to approximately the 95% confidence interval for the estimated parameter (6).

Inferring the most closely related strain to a strain of interest using sequence data

We inferred the evolutionary relationship between different strains considered in the ODE model at different sampling time points. Each strain in the model consists of a group of sequences. Inferring evolutionary relationship between groups of sequences represents a serious methodological challenge in general. Here, we implement a simple method to infer the evolutionary relationship by calculating the nucleotide differences between two groups of sequences. This method is sufficient for the particular problem we deal with, because of the short time period of evolution and the temporal information of the sequences collected.

For each strain/group at each time point of sampling, we derive the most closely related group to the group of interest at that time point and previous time points. Specifically, we first calculate the nucleotide Hamming distances between each sequence belonging to the group of interest and all sequences belonging to groups that were sampled at that time point or at previous time points. We then determine the sequence that has the smallest nucleotide distance to each sequence in the group of interest and to which group this sequence belongs to (see Figs. S15-S17 for results for Subjects 1, 3 and 4, where the mutant strains show the most complicated dynamics). This calculation allows us to infer from which groups the sequences in the group of interests are derived. We then infer the group to which the group of interest is most closely related to as the group from which most of the sequences are derived.

Single genome sequence (SGS) analysis of data from subjects 2-5

Subjects 2 and 3 (Figs. S1 and S2) showed the appearance of similar patterns of resistance mutations on a Q80K mutant background that was present prior to treatment. At day 34, multiple sets of distinct low diversity lineages were detected (Q80K/D168E in subject 2 and Q80K/Y56H in subject 3). This mutational pattern remained unchanged at day 62 in subject 2 but shifted substantially in subject 3 to a combination of Q80K/Y56H, Q80K/D168E and Q80K/Y56H/R62K mutations.

Viral sequences from subject 4 (Fig. S3) revealed phylogenetic and resistance patterns similar to subject 1. At baseline, no known drug resistance mutations (DRMs) were identified in any of 102 HCV sequences examined. In this subject, we also analyzed a plasma sample from the first day on treatment. A total of 144 day 1 sequences revealed no DRMs (Fig. S3B) concomitant with a five log reduction in viral load from 9,287,191 to 1926 IU/ml (Fig. S3A). At day 14, 7 days after treatment cessation, all 43 sequences obtained carried an A156T or A156V drug resistance mutation. As was the case for subject 1 (Fig. 1), these mutant sequences were well separated in the phylogenetic tree of pre-treatment sequences (Fig. S3B), suggesting that they represented the progeny of many different productively infected hepatocytes each of which contained a virus that carried an A156T/V mutation in a distinct sequence background prior to or near the time of treatment initiation. At day 34, 116 sequences were obtained. Discrete low diversity lineages of double mutants Y56H/D168V, Y56H/D168A, and D168E/F169L comprised about 34% of the sequences and then were rapidly replaced by D168E/A/Y mutants (52%) and wildtype virus (41%) at day 60, when 109 sequences were analyzed.

For subject 5, the first time point after treatment cessation with sequencing data was day 34 (*i.e.*, 27 days after treatment cessation). Single amino acid resistance mutations were found in 12.6% (D168E/A) and 1.5% (R155K) of the sequences (obtained from two independent samplings; Fig. S4). At day 62, 93% were wildtype, and these sequences were well distributed throughout the phylogenetic tree of pre-treatment sequences.

Testing alternative mathematical models

1. A model with a DAA independent cure

We tested whether a DAA independent cure of infected cells could explain the rapid turnover of dominant viral variants seen in subjects 1 and 4 after treatment cessation (without the assumption of superinfection). For a DAA independent cure, the cure process occurs constantly irrespective of the presence of DAAs.

We constructed two models to test the role of a DAA independent cure, one with DAA dependent cure and one without. In the first model (which we term the ‘DAA independent cure' model), the DAA-dependent cure or the superinfection is not included. The ODEs are:

$$\frac{dT}{dt} = \rho_T \cdot T \cdot (1 - \frac{T + \sum_{i=1}^{n} I_i + N}{T_{max}}) - d \cdot T - \sum_{i=1}^{n} \beta \cdot T \cdot V_i + \sum_{i=1}^{n} k_{cure,ind} \cdot I_i$$

$$\frac{dI_i}{dt} = \beta \cdot T \cdot V_i - \delta \cdot I_i - k_{cure,ind} \cdot I_i$$

$$\frac{dV_i}{dt} = (1 - \varepsilon_i) \cdot r_i \cdot p \cdot I_i - c \cdot V_i$$

$$\varepsilon_i = \frac{D \cdot \exp(-w \cdot \max(t - 7{,}0))}{EC_{50,i} + D \cdot \exp(-w \cdot \max(t - 7{,}0))}$$

where $k_{cure,ind}$ is the per capita rate of cure of infected cells (independent of DAAs). This model is constructed by adding a DAA independent cure term, $\sum_{i=1}^{n} k_{cure,ind} \cdot I_i$, to the baseline model.

In the second model (which we term the 'DAA dependent and independent cure' model), we included the DAA dependent cure (as in the cure model) in addition to the DAA independent cure. The ODEs are:

$$\frac{dT}{dt} = \rho_T \cdot T \cdot (1 - \frac{T + \sum_{i=1}^{n} I_i + N}{T_{max}}) - d \cdot T - \sum_{i=1}^{n} \beta \cdot T \cdot V_i + \sum_{i=1}^{n} (k_{cure} \cdot (-log_{10}(1 - \varepsilon_i)) + k_{cure,ind}) \cdot I_i$$

$$\frac{dI_i}{dt} = \beta \cdot T \cdot V_i - \delta \cdot I_i - (k_{cure} \cdot (-log_{10}(1 - \varepsilon_i)) + k_{cure,ind}) \cdot I_i$$

$$\frac{dV_i}{dt} = (1 - \varepsilon_i) \cdot r_i \cdot p \cdot I_i - c \cdot V_i$$

$$\varepsilon_i = \frac{D \cdot \exp(-w \cdot \max(t - 7{,}0))}{EC_{50,i} + D \cdot \exp(-w \cdot \max(t - 7{,}0))}$$

Table A1 summarizes the AICc scores for the two new models as well as the cure model and the full model in the main text. In general, the full model is the best model in terms of the overall AICc score. It is also significantly better than all other models in explaining data from subjects 1 and 4 where rapid turnovers of dominant viral variants were observed after treatment cessation. These results strongly support the conclusion that intracellular competition through superinfection is a crucial process to explain the pattern in the data from subjects 1 and 4.

We further examined the estimated rate of DAA independent cure needed to explain the rapid turnover of dominant mutants seen in the data. From the two models tested, we estimated this rate to be around 0.4 day^{-1} (Table A2). This means that under non-DAA therapy, such as the interferon therapy, the loss rate of infected cells, $k_{cure,ind} + \delta$, is (0.4+0.14=) 0.54 day^{-1}. This is inconsistent with previous clinical data from patients treated with interferon, where the loss rate of infected cells is estimated to be 0.14 day^{-1}

(7). This again suggests against the importance of the role of DAA independent cure in explaining the data.

Table A1. Summary of the model characteristics and the fitting results, i.e. sums of squared residuals (SSR) and the AICc scores, of each model for each subject. Bold AICc scores denote the best model fit among all models for the 5 subjects.

Model Characteristics								
Parameter	**Cure model**		**Full model**		**DAA independent cure model**		**DAA dependent and independent cure model**	
k_{cure}	Fitted		Fitted		-		Fitted	
k_{super}	-		Fitted		-		-	
$k_{cure,ind}$	-		-		Fitted		Fitted	
Fitting Results								
Subject	**SSR**	**AICc**	**SSR**	**AICc**	**SSR**	**AICc**	**SSR**	**AICc**
1	5.9	-32.8	2.4	**-63.8**	2.8	-59.0	2.7	-55.2
2	2.1	-53.2	1.8	-51.8	1.3	**-64.1**	1.3	-58.7
3	2.3	-47.6	2.1	-44.5	1.5	**-58.0**	1.5	-51.4
4	6.7	-42.6	1.9	**-90.6**	3.1	-72.1	2.8	-71.6
5	0.9	**-86.5**	0.9	-83.6	1.1	-77.6	0.9	-83.6
Total AICc	17.9	-262.7	9.1	**-334.3**	9.9	-330.8	9.3	-320.5

Table A2. Summary of the estimated rates of the DAA independent cure ($k_{cure,ind}$). The last line shows the predicted rate of 2nd phase V.L. decline under effective IFN therapy, where the rate is driven by both cell death (δ=0.14/day) and DAA independent cure ($k_{cure,ind}$).

Subjects	**Model**			
	Cure model	**Full model**	**DAA independent cure model**	**DAA dependent and independent cure model**
Pt1	0.00	0.00	0.34	0.23
Pt2	0.00	0.00	0.19	0.19
Pt3	0.00	0.00	0.22	0.22
Pt4	0.00	0.00	0.41	0.35
Pt5	0.00	0.00	0.23	0.23
Mean	**0.00**	**0.00**	**0.28**	**0.24**
Predicted rate of 2nd phase V.L. decline under IFN therapy ($k_{cure,ind}$+ δ)	**0.14 /day**	**0.14 /day**	**0.42 /day**	**0.38 /day**

Overall, our analysis here suggests that viral competition through superinfection is the best hypothesis (among all hypotheses) to explain the rapid turnover of dominant viral variants observed in the data.

2. A model with a constant rate of target cell generation

In the analyses in the main text, we assumed density dependent target cells proliferation (as in Rong *et al.* (8)), so that new target cells become available rapidly when infected cells are lost. To test the role of this rapid proliferation assumed in the model in the main text and to test if our conclusions about superinfection and cure of infected cells are robust to variations in these assumptions, we modified the models using a constant rate of target cell generation as in Neumann *et al.* (7), instead of the density-dependent proliferation. The ODE for the rate of change in the target cell concentration (*T*) becomes

$$\frac{dT}{dt} = \lambda - d \cdot T - \sum_{i=1}^{n} \beta \cdot T \cdot V_i + \sum_{i=1}^{n} k_{cure} \cdot (-log_{10}(1 - \varepsilon_i)) \cdot I_i$$

where λ is the constant rate of target cell generation and it is set to be 6.5 x 10^4 ml^{-1} day^1, such that the target cell concentration at the infection free equilibrium is 6.5 x 10^6 cells ml^{-1}, to be consistent with our study and the study in Ref. (8).

Fitting results (Table A3) show that in subjects 1 and 4, the modified models do not fit the data as well as models assuming density-dependent target cell proliferation (see the large differences between the AICc scores in Table A3 below and in Table 1 in the main text). Rapid target cell proliferation is necessary to explain the fast viral load rebound and selection of resistant mutants seen by day 14 in subjects 1 and 4 (but not in the other subjects). This suggests that in addition to superinfection and cure of infected cells, rapid target cell proliferation is another mechanism that drives rapid population expansion and selection of resistance seen in subjects 1 and 4. Note that, target cell proliferation is not needed to explain the data from subjects 2, 3 and 5. This could be due to the sparse sampling of data in these patients, which makes it impossible to accurately estimate the viral load rebound kinetics.

Table A3. Summary of the fitting results for the model assuming a constant target cell generation. Values are AICc scores. Bolded AICc scores denote the best model fit among all models for the 5 subjects.

Subject	Baseline model with δ=0.14 day^{-1}	Cure model	Superinfection model	Full model
1	5.2	6.1	9.5	**2.3**
2	-2.2	**-53.0**	-17.3	-53.3
3	4.2	**-46.0**	3.8	-42.1
4	17.2	-1.4	21.2	**-26.9**
5	-24.3	**-85.7**	-20.6	-81.9
Total AICc	0.1	-180.0	-3.4	**-201.9**

Overall, although the model assuming a constant rate of target cell generation fits the data worse than the model in the main text, similar patterns arise from the two model variations: the full model is the best model to explain data from subjects 1 and 4 and the cure model is the best model for the other subjects. This again strongly suggests that the conclusions about the role of superinfection and cure of infected cells are robust to the assumptions of how rapidly target cells are generated.

Evolutionary survival of the drug-resistant strains

We observed that drug-sensitive forms appeared at late time points in subjects 1, 4 and 5. To characterize the nature and persistence of the drug sensitive forms, we tested whether the reappearance of drug sensitive forms at later time points was due to reversion of the resistant form or latent survival of the original forms. If the reappearance of drug sensitive forms is due to reversion, then we expect to see these variants would emerge from within the drug resistant clades that dominate the time-points immediately after treatment. Contrary to this expectation, every drug-sensitive form that reappeared was basal to the drug-resistant clades, separated from it by further mutations at other positions. This suggests that the drug-sensitive viruses seen at the last time point are originated from drug-sensitive viruses before treatment.

On the other hand, one could expect that if these forms survived from the earliest time point, they would descend equally from any branch in the phylogenetic trees of the first time points. To see if this were the case, we devised a Fisher exact test as follows: first, we divided the phylogenetic tree composed of all sequences from all time points into phylogenetic clades defined by long parental branches; next, for each clade, we compared the number of drug-sensitive sequences at the first time point with the ones at the last time point. In subjects 4 and 5, there was a significant difference ($p=0.0021$ and $p < 2\times10^{-12}$, respectively) between the distributions of the drug-sensitive forms in the first and last time points: almost all such sequences from the last time point belonged to a single cluster. In subject 1 we only defined two clusters, and the distribution of the drug-sensitive forms was not significantly different ($p=0.24$) at the first and last time points.

These surviving forms, however, do not appear uniformly in the phylogenetic tree of pre-treatment sequences; rather, they are close to the forms that develop drug-resistant mutations. This could indicate that either there are significant fitness differences between the various clades in the pre-treatment viral quasi-species, or replication complexes containing drug-resistance may rarely be able to rescue neighboring drug-sensitive forms.

SI Figures

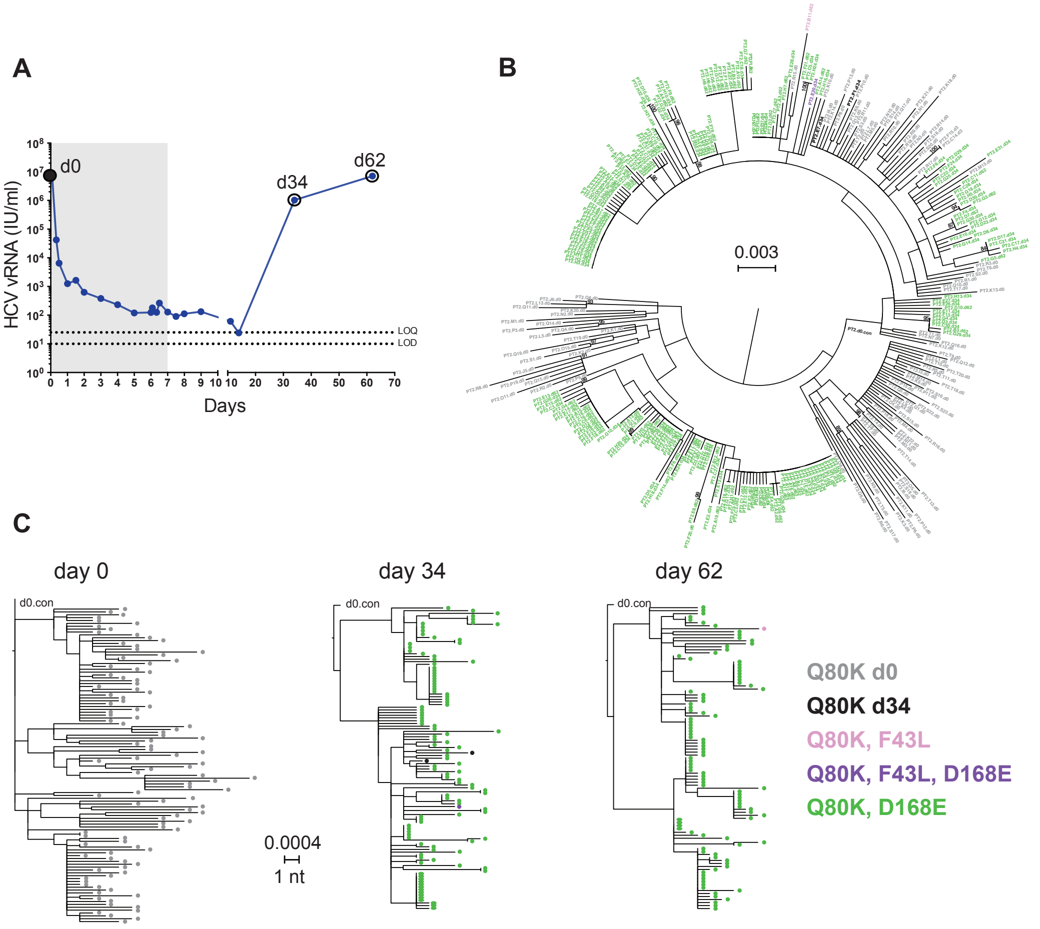

Figure S1. Sequential plasma virus load and sequences from subject 2. (A) Time course of treatment with MK-5172 (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0, 34 and 62). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 2 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 2 at each time point.

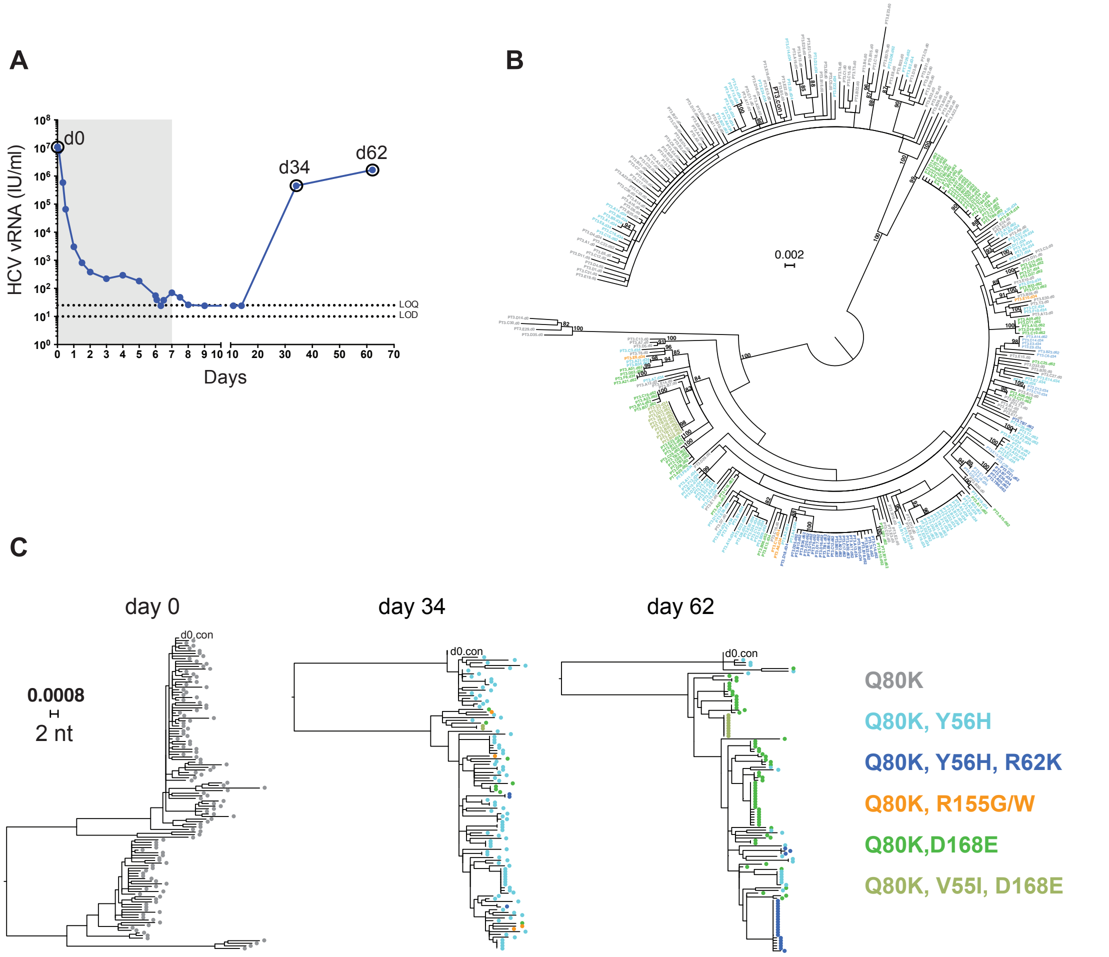

Figure S2. Sequential plasma virus load and sequences from subject 3. **(A)** Time course of treatment with MK-5172 (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0, 34 and 62). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 3 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 3 at each time point.

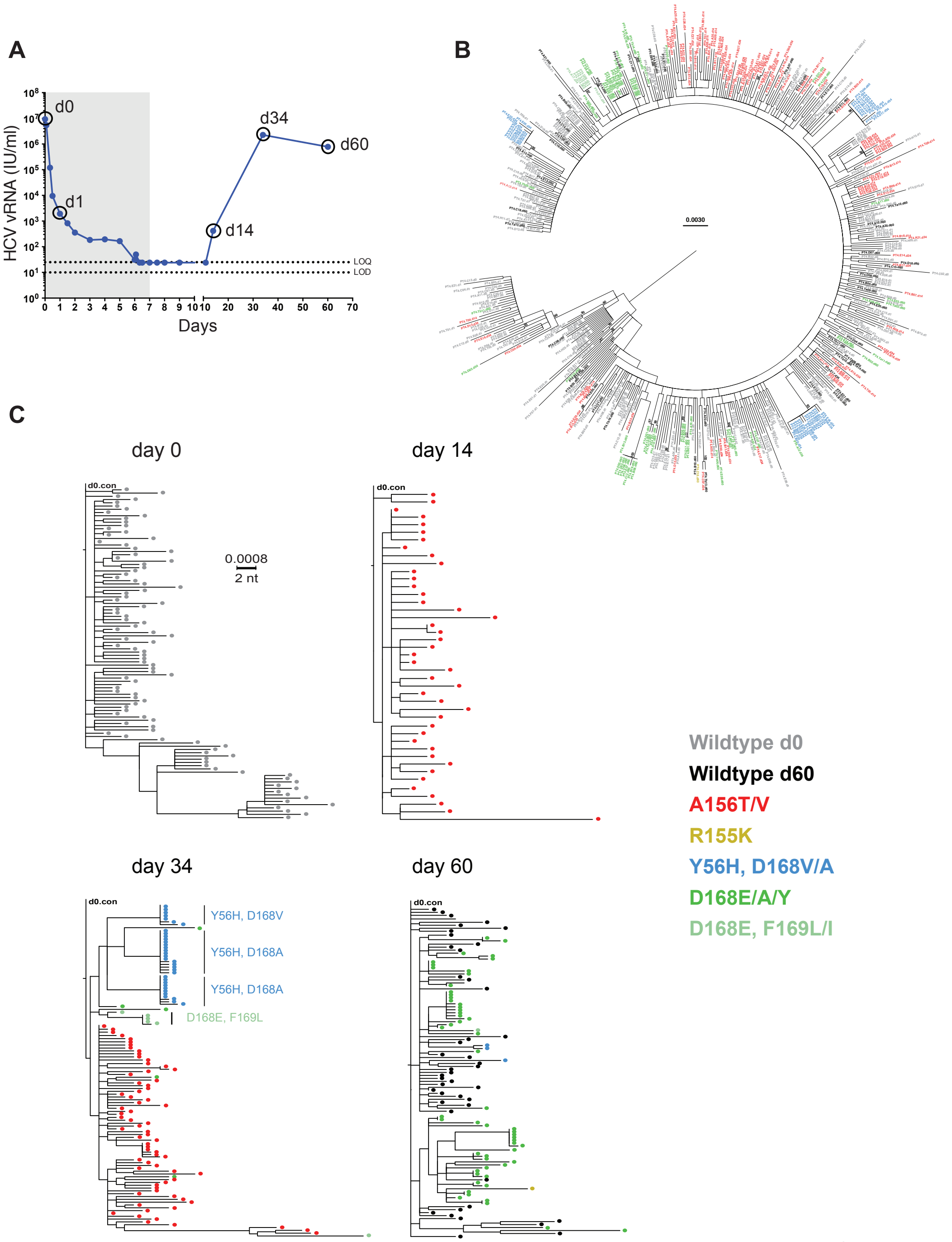

Figure S3. Sequential plasma virus load and sequences from subject 4. (A) Time course of treatment with MK-5172 (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0, 14, 34 and 60). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 4 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 4 at each time point.

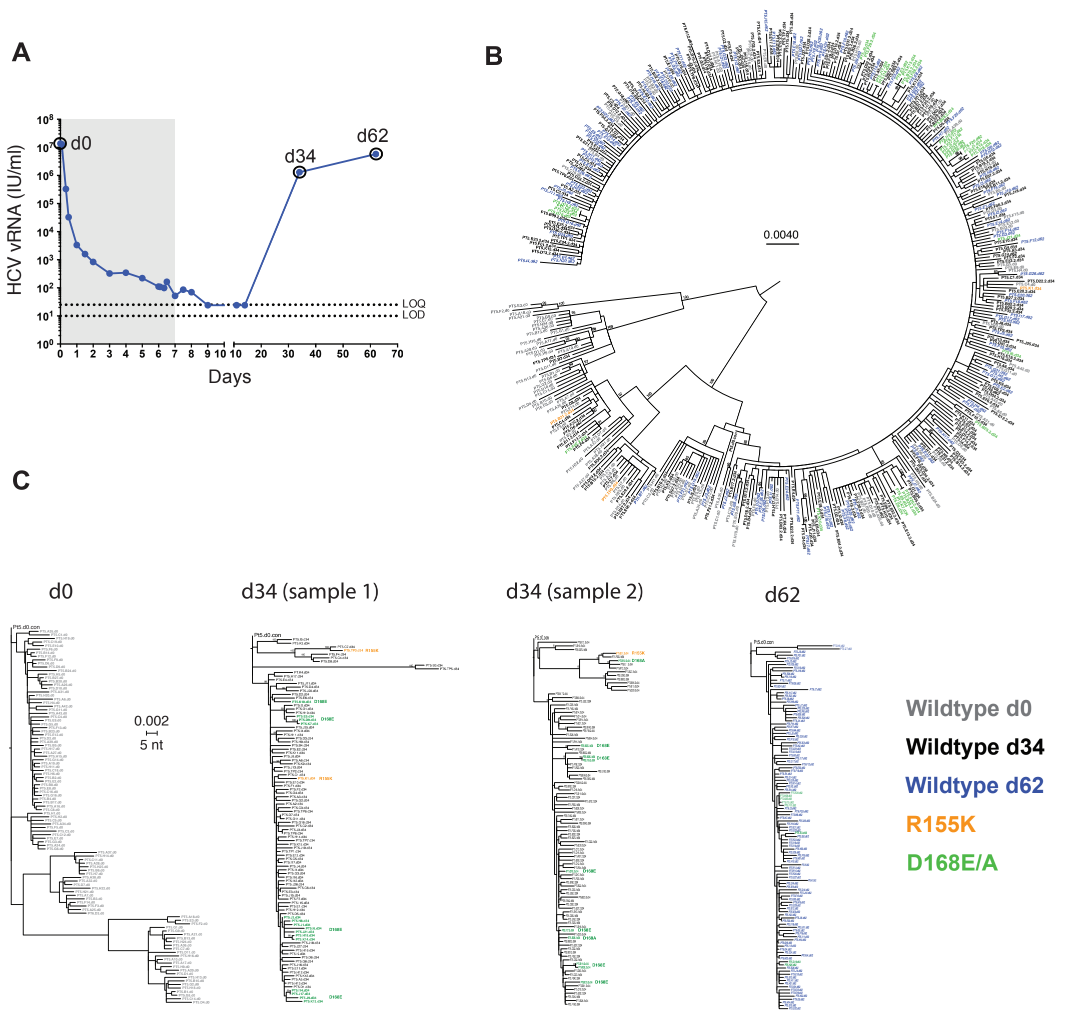

Figure S4. Sequential plasma virus load and sequences from subject 5. (A) Time course of treatment with MK-5172 (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0, 34 and 62). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 5 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 5 at each time point (including two samples at day 34, which shows similar patterns).

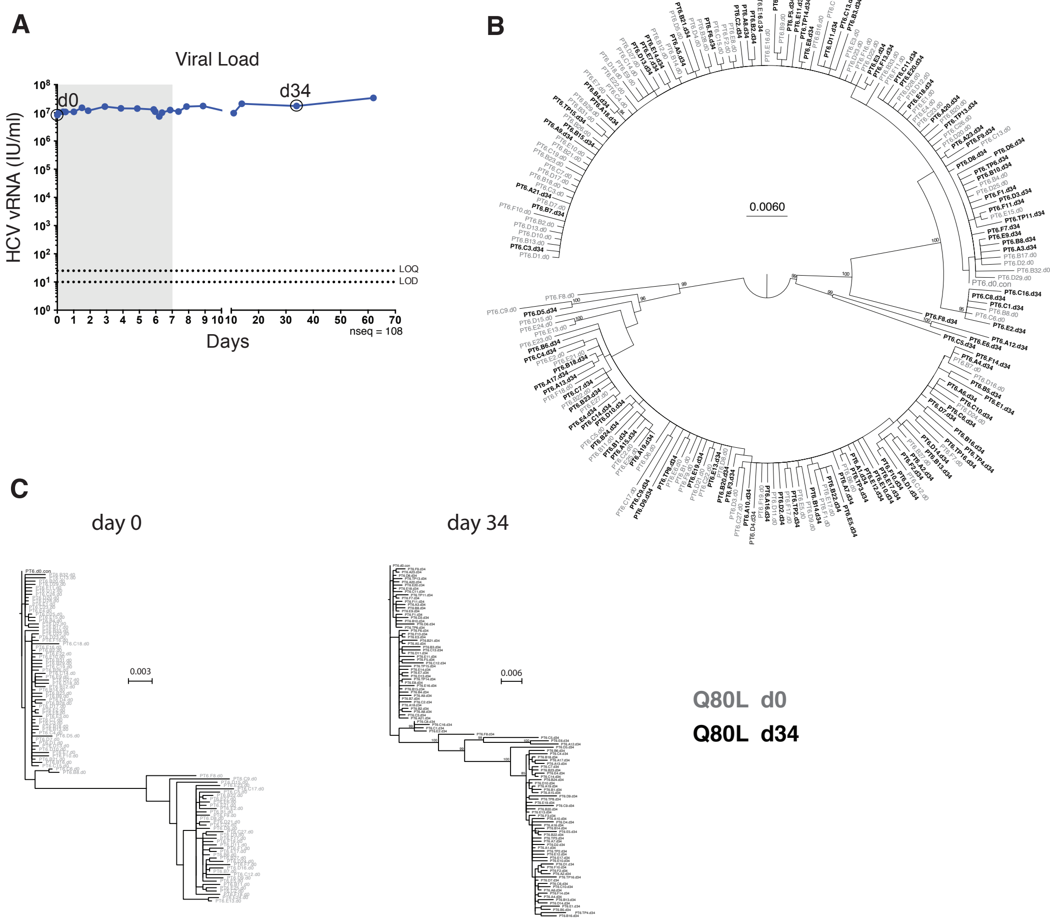

Figure S5. Sequential plasma virus load and sequences from subject 6. (A) Time course of treatment with placebo (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0 and 34). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 6 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 6 at each time point.

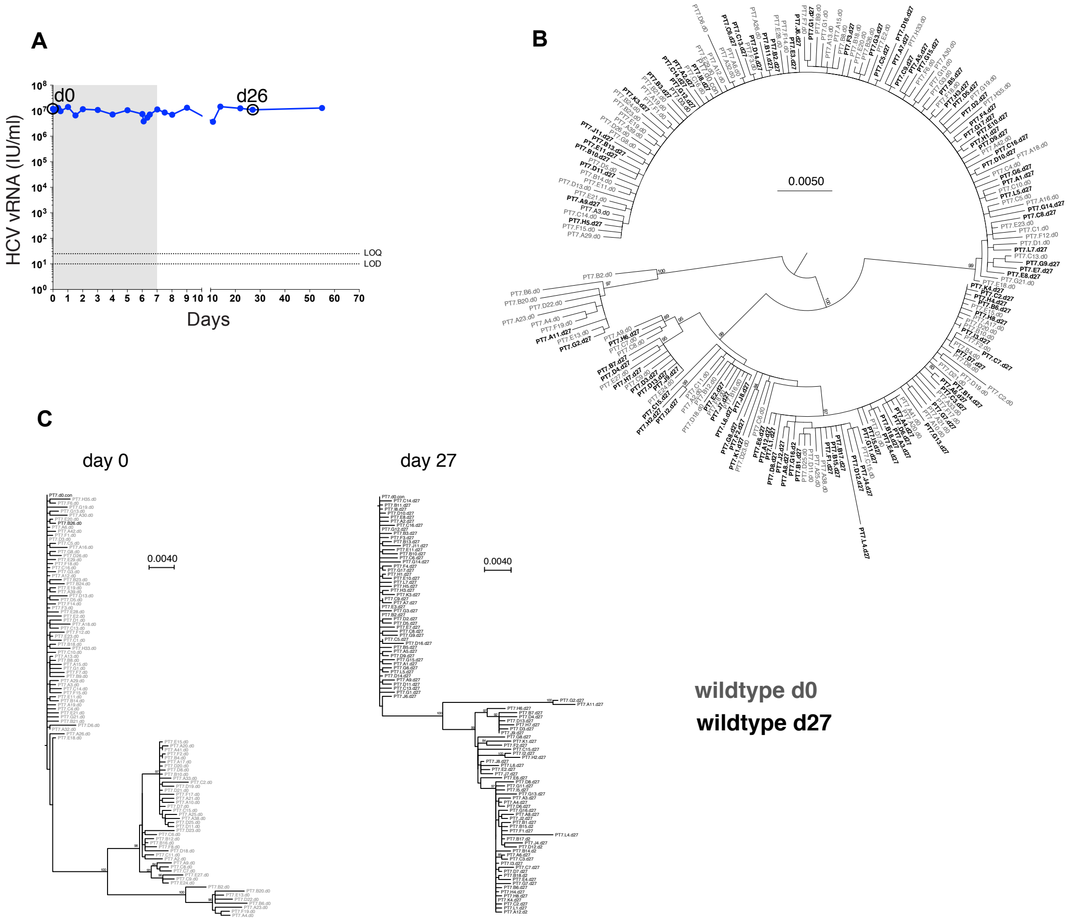

Figure S6. Sequential plasma virus load and sequences from subject 7. (A) Time course of treatment with placebo (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0 and 27). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 7 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 7 at each time point.

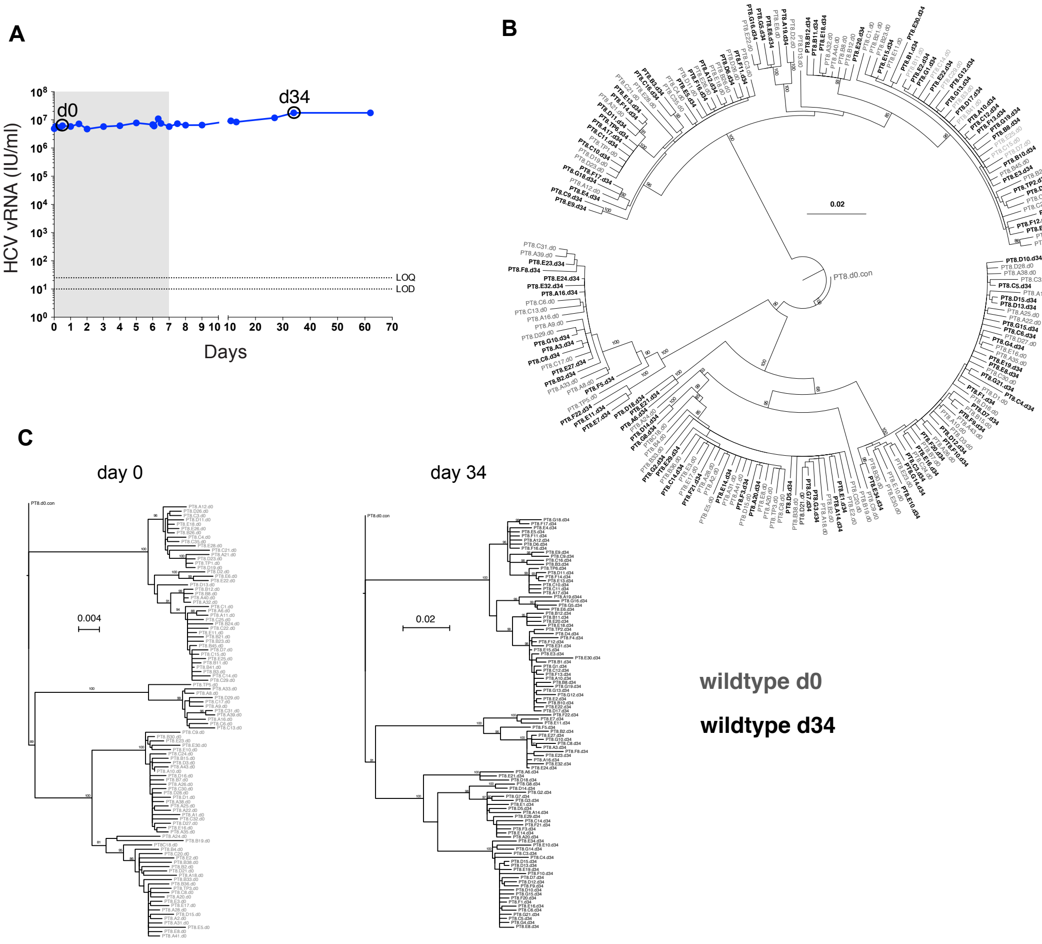

Figure S7. Sequential plasma virus load and sequences from subject 8. (A) Time course of treatment with placebo (shaded area, days 1-7), viral load determinations (blue solid dots), and viral sequence analyses (open circles at days 0 and 34). **(B)** A maximum likelihood (ML) phylogenetic tree of all viral sequences sampled from subject 8 from all time points. **(C)** ML phylogenetic trees of viral sequences sampled from subject 8 at each time point.

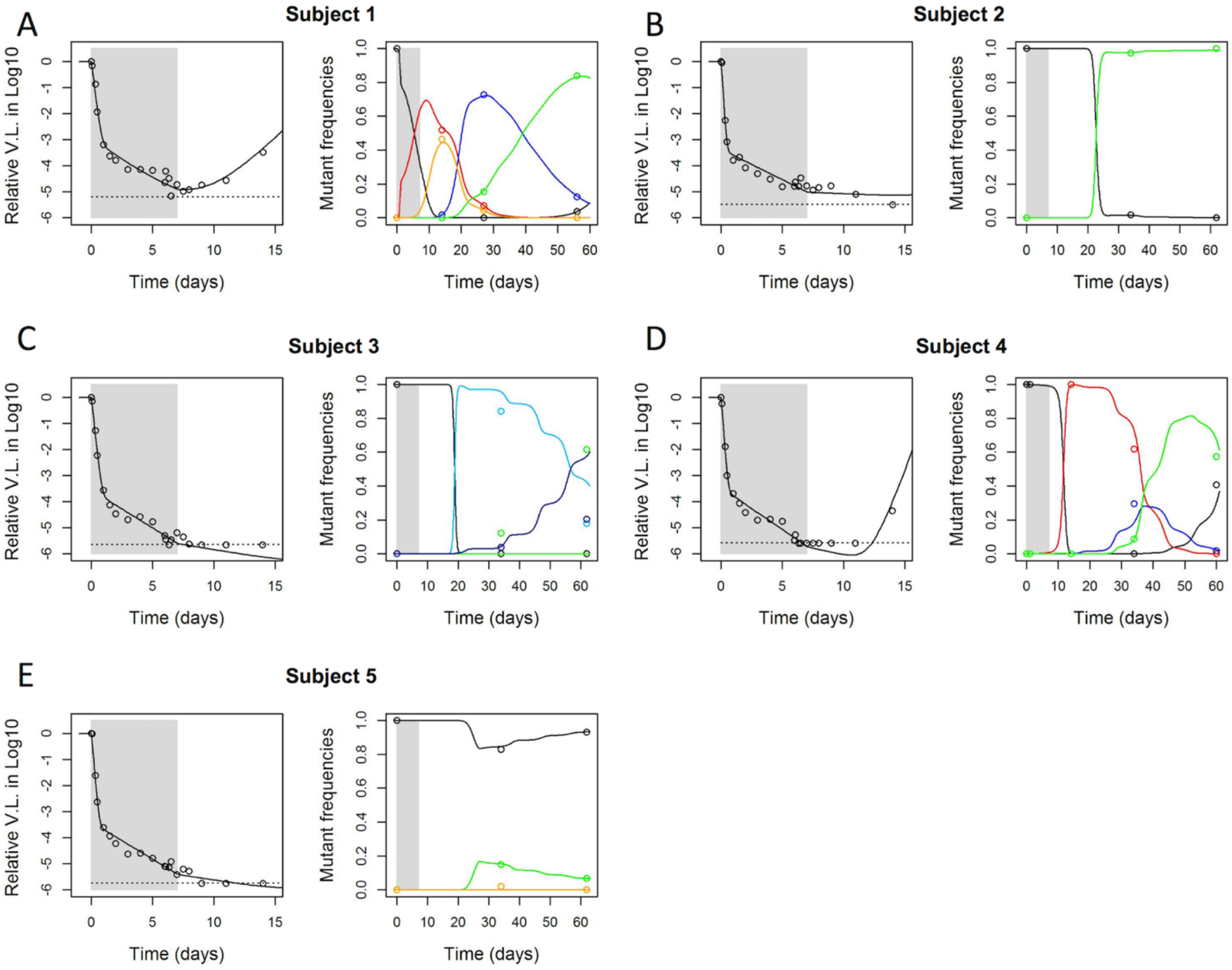

Figure S8. A baseline HCV multi-strain model with unrealistically high death rates of infected cells describes the clinical data well in subjects treated with MK-5172. The five panels show data points (in open circles) and simulation trajectories using best-fit parameters (solid lines) for the 5 subjects. In each panel, the data and simulation results for viral loads are shown in open circles and black lines, respectively, on the left; the data and simulation results for mutant frequencies are shown in colored open circles and lines, respectively, on the right. Note that the x-axis of the viral load plot is scaled between 0-15 days to show the agreement between the data and the model fit in the early time points. The mutants are color coded according to Table S2.

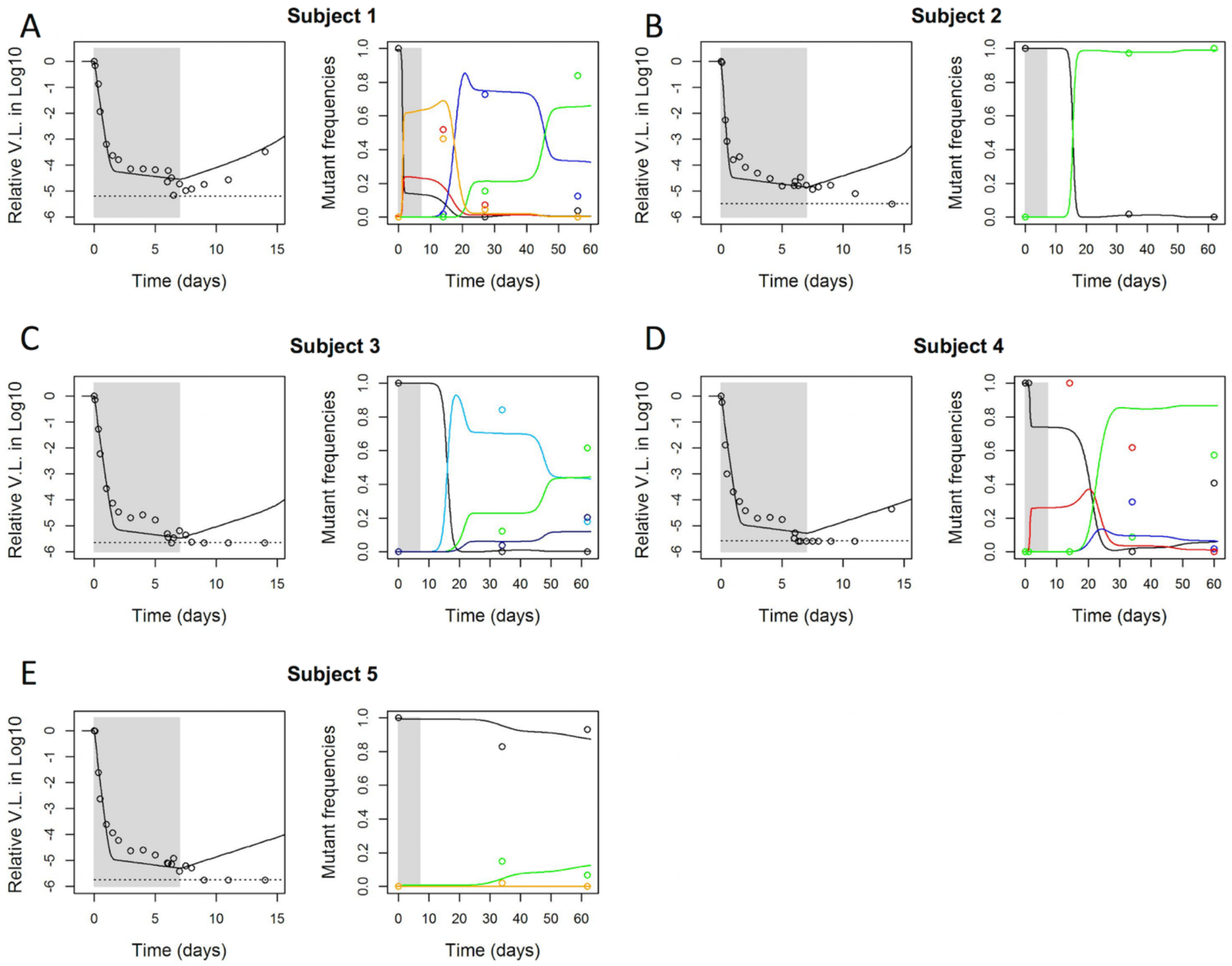

Figure S9. A baseline HCV multi-strain model with the death rate of infected cells fixed at 0.14 day^{-1} does not describe the clinical data well in 5 subjects treated with MK-5172. The five panels show data points (in open circles) and simulation trajectories using best-fit parameters (solid lines) for the 5 subjects. In each panel, the data and simulation results for viral loads are shown in open circles and black lines, respectively, on the left; the data and simulation results for mutant frequencies are shown in colored open circles and lines, respectively, on the right. Note that the x-axis of the viral load plot is scaled between 0-15 days to show the discrepancy between the data and the model fit in the early time points. The mutants are color coded according to Table S2.

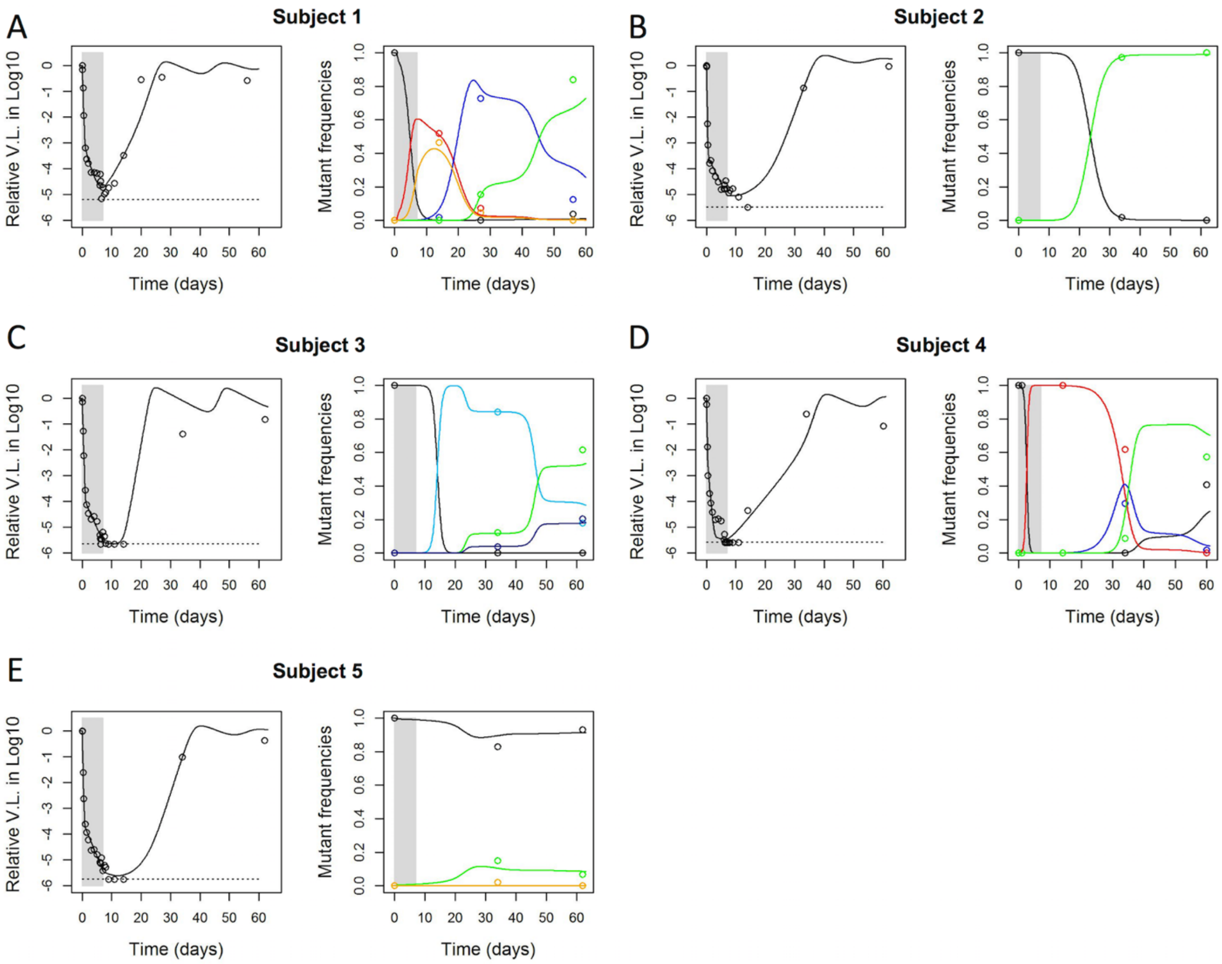

Figure S10. Fitting results of the 'cure' model (lines) to the clinical data (circles) in subjects treated with MK-5172. In each panel, the data and simulation results for viral loads are shown in open circles and black lines, respectively, on the left; the data and simulation results for mutant frequencies are shown in colored open circles and lines, respectively, on the right. The mutants are color coded according to Table S2.

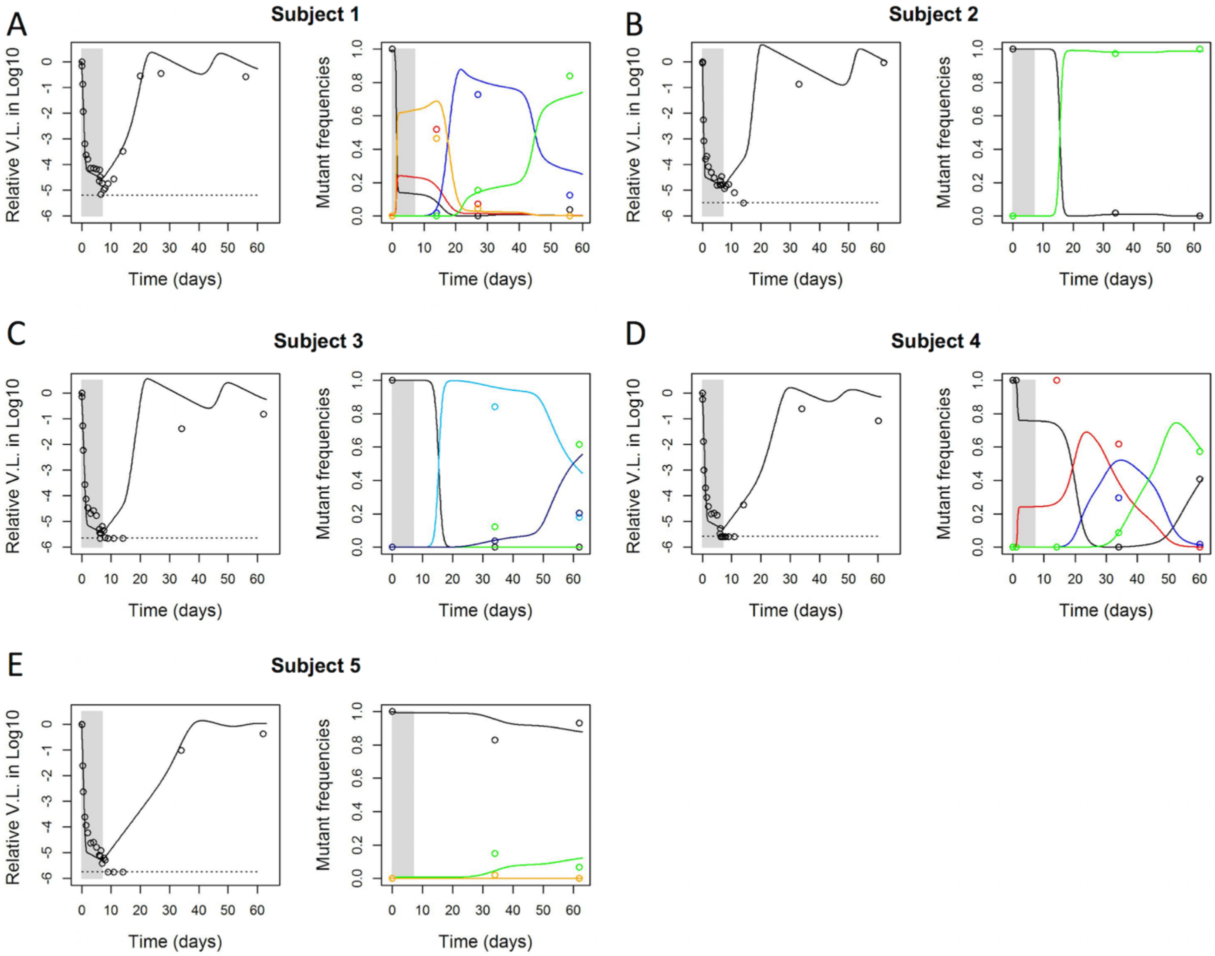

Figure S11. Fitting results of the 'superinfection' model (lines) to the clinical data (circles) in subjects treated with MK-5172. In each panel, the data and simulation results for viral loads are shown in open circles and black lines, respectively, on the left; the data and simulation results for mutant frequencies are shown in colored open circles and lines, respectively, on the right. The mutants are color coded according to Table S2.

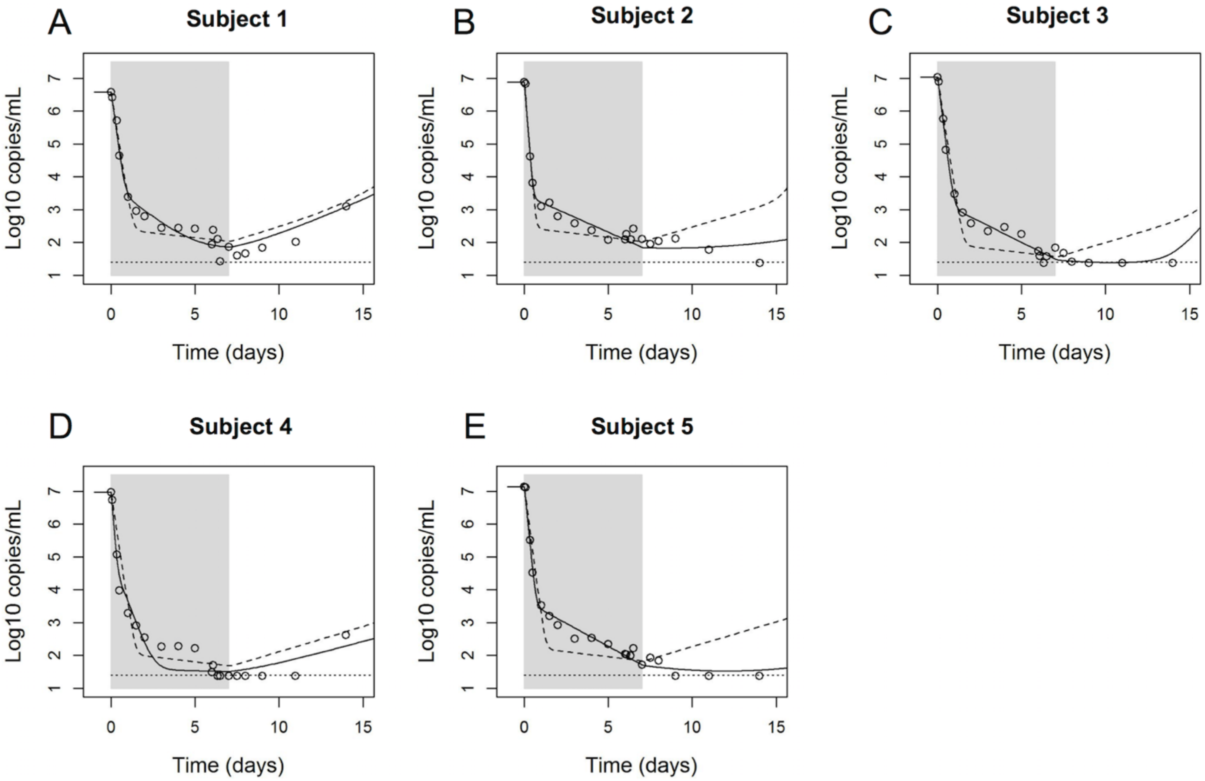

Figure S12. Cure of infected cells is important in explaining the rapid second phase decline in all 5 subjects. Shown are comparisons of viral loads between fitting results of the model without cure, the 'basic' model (dashed lines), and the model with cure, the 'cure' model (solid lines), in 5 subjects (panels A-E). Viral load data are shown as open circles.

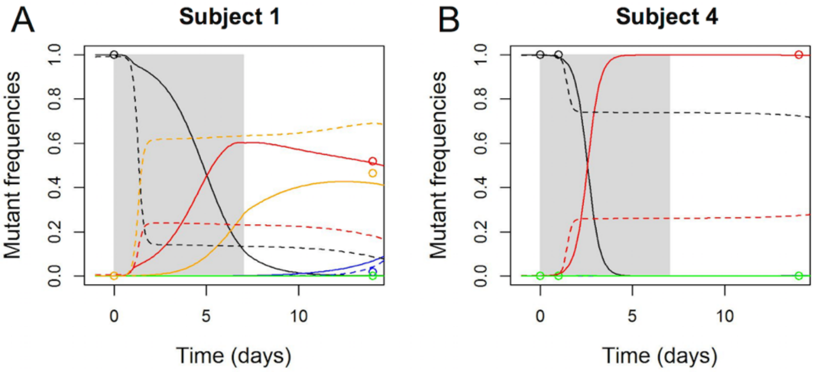

Figure S13. Cure of infected cells is important in explaining the rapid selection of early resistant mutants seen in subjects 1 and 4. Shown are comparisons of mutant frequencies between fitting results of a model without cure, the null model (dashed lines), and a model with cure, the 'cure' model (solid lines). Mutant frequencies derived from sequence data are shown as open circles, which are color coded according to mutants shown in Table S2.

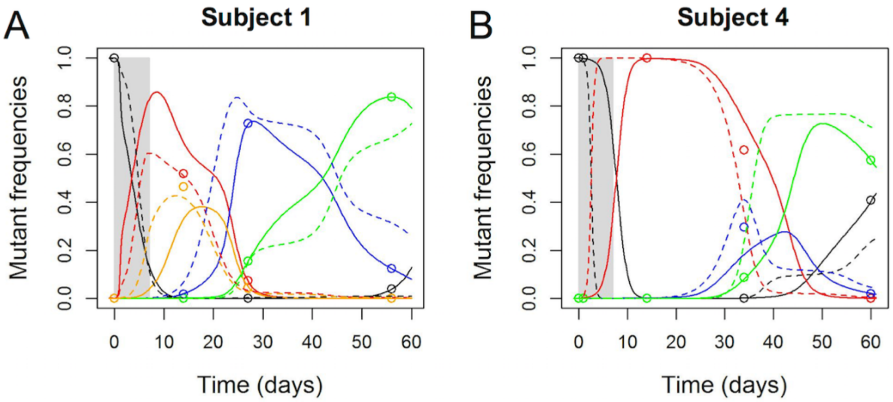

Figure S14. Superinfection of infected cells is important in explaining the rapid turnover of resistant mutants seen after Day 20 in subjects 1 and 4. Panels A and B show comparisons of mutant frequencies between fitting results of the 'full' model, model with superinfection (solid lines), and the 'cure' model, i.e. model without superinfection (dashed lines), for Subjects 1 and 4, respectively. Mutant frequencies derived from sequence data are shown as open circles, which are color coded according to mutants shown in Table S2.

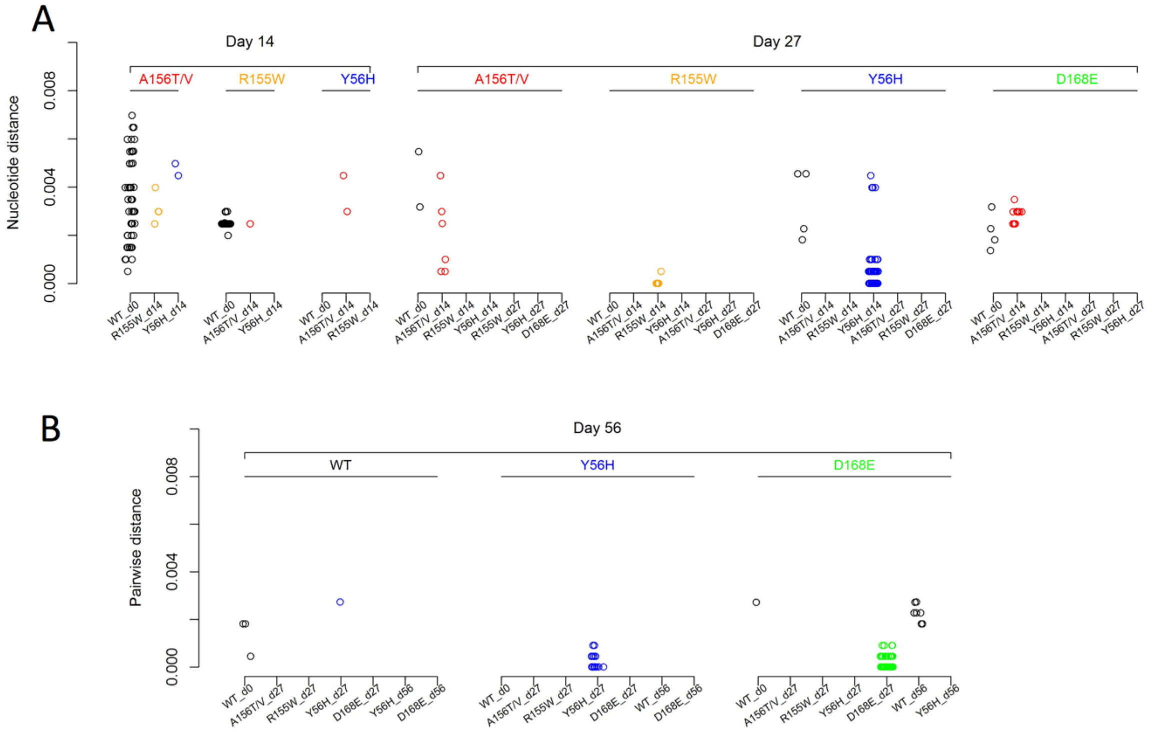

Figure S15. Derivation of the most closely related mutant strains to a mutant strain of interest at a particular time point for subject 1. We calculated the pairwise distances between sequences belonging to one mutant strain of interest at a particular time point and sequences belonging to other mutant strains at that time point and previous time points. The name of the mutant strain of interest and day of sampling are shown at the top of the plots. For each sequence belonging to the mutant strain of interests, there exists a sequence belonging to other mutant strains that are the mostly closely related to it, i.e. has the shortest pairwise distance. We plot the shortest pairwise distance (y-axis) for each sequence (colored dots). The ticks on the x-axis show the mutant strain and the day of sampling to which the sequence is most closely related. The dots are also color coded according to the mutant strain shown as the tick on the x-axis. **(A)** The shortest pairwise distances calculated for sequences sampled at day 14 and 27. **(B)** The shortest pairwise distances calculated for sequences sampled at day 56.

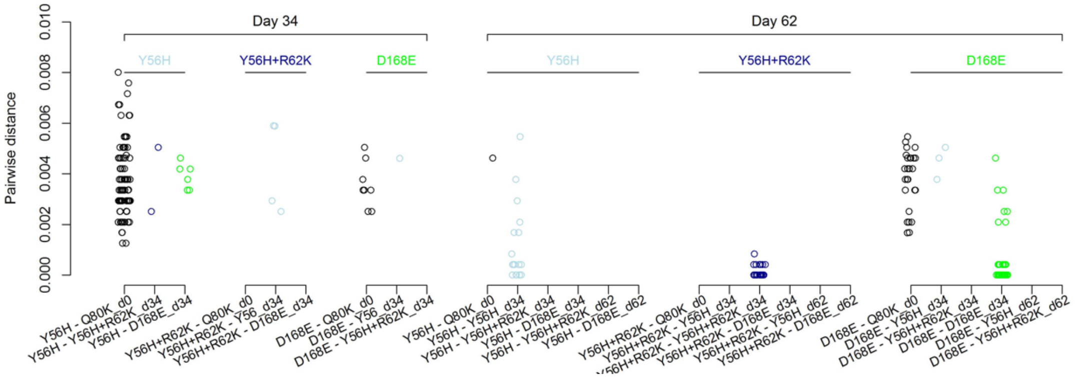

Figure S16. Derivation of the most closely related mutant strains to a mutant strain of interest at a particular time point for subject 3. The same plot as Fig. S15, except using sequence data from subject 3.

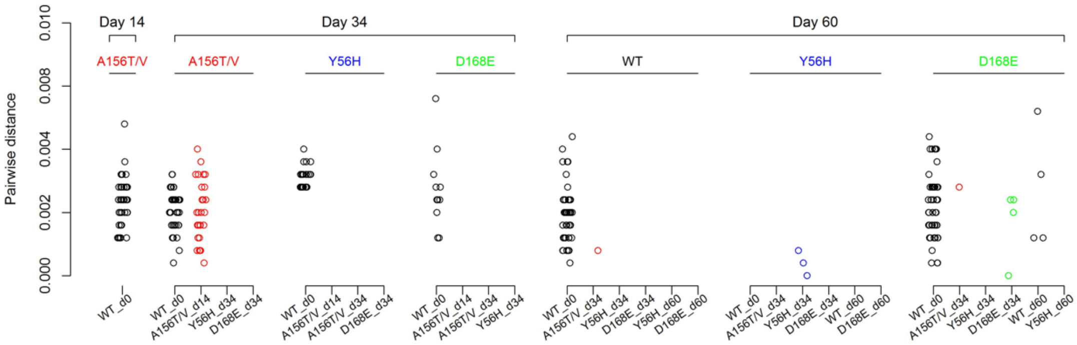

Figure S17. Derivation of the most closely related mutant strains to a mutant strain of interest at a particular time point for subject 4. The same plot as Fig. S15, except using sequence data from subject 4.

SI Tables

Table S1. Sequence diversity analyses before and after MK5172 treatment.

	Genotype	Dose (mg)	Days post treatment	Number of sequence	Sequence diversity %			
					Maximum	**Minimum**	**Median**	**Mean**
Pt 1	1b	50	0	103	2.44	0.05	1.22	1.12
			14	111	2.17	0.00	0.86	0.89
			27	111	1.90	0.00	0.50	0.54
			56	106	1.13	0.00	0.77	0.53
Pt 2	1a	800	0	110	1.17	0.00	0.67	0.62
			34	113	1.00	0.00	0.42	0.42
			62	101	1.00	0.00	0.46	0.44
Pt 3	1a	800	0	120	4.34	0.08	2.15	1.85
			34	107	3.88	0.00	0.92	1.59
			62	112	4.22	0.00	0.88	1.05
Pt 4	1a	800	0	102	1.29	0.04	0.42	0.53
			1	144	1.63	0.00	0.46	0.58
			14	43	1.38	0.00	0.38	0.42
			34	116	1.42	0.00	0.46	0.46
			34	111	1.46	0.00	0.46	0.44
			60	109	1.34	0.00	0.42	0.45
Pt 5	1a	800	0	105	4.80	0.17	2.13	2.37
			34	100	4.26	0.00	0.71	0.94
			34	99	2.84	0.00	0.75	1.02
			62	117	2.55	0.00	0.58	0.67
Pt 6	1a	placebo	0	100	2.72	0.00	0.92	1.25
			34	108	2.72	0.00	1.96	1.38
Pt 7	1b	placebo	0	104	3.35	0.05	1.72	1.37
			27	102	3.75	0.00	1.81	1.33
Pt 8	1a	placebo	0	100	5.64	0.17	4.47	3.45
			34	101	5.72	0.13	4.51	3.50

Table S2. The strains considered in the models and the colors chosen to represent them in the plots. The values of EC_{50} measured in the replicon system are shown if available; theses values are taken from Refs. (2, 3).

Subject (Genotype)	Strain # (*i*)	Resistant mutations	Color in the figure	EC_{50}
1 (1b)	1	Wild-type	Black	0.5 nM
	2	A156T/V	Red	140 nM
	3	R155G R155W+A156G	Orange	1540 nM
	4	Y56H+D168N/V	Blue	
	5	D168T D168E+F169I/L	Green	
2 (1a)	1	Q80K	Black	
	2	Q80K+D168E	Green	
3 (1a)	1	Q80K	Black	
	2	Q80K+Y56H	Light blue	
	3	Q80K+Y56H+YR62K	Dark blue	
	4	Q80K+D168E, Q80K+V55I+D168E	Green	
4 (1a)	1	Wild-type	Black	0.35 nM
	2	A156T/V	Red	108 nM
	3	Y56H+D168V	Blue	
	4	D168E/A/Y, D168E+F169L	Green	29 nM
5 (1a)	1	Wild-type	Black	0.35 nM
	2	D168E	Green	1.6 nM

Table S3. Best fit parameter values and least square residuals of 5 models fitted to data collected from subject 1.

Parameter	Parameter variations (Lower, higher bound)	Baseline model	Cure Model	Superinfection Model	Full Model (confidence intervals)
p (days^{-1})	*	11.5	30.8	27.1	28.1 (20.4, 28.1)
Log$_{10}$ ($I_{0,2}/I_{0,1}$)	-20,-1	-2.48	-3.05	-2.16	-2.58 (-4.2, -1.9)
Log$_{10}$ ($I_{0,3}/I_{0,1}$)	-20,-1	-12.20	-7.32	-7.74	-10.29 (-12.6, -2.8)
Log$_{10}$ ($I_{0,4}/I_{0,1}$)	-20,-1	-12.93	-2.94	-14.72	-7.88 (-16.1, -2.9)
Log$_{10}$ ($I_{0,5}/I_{0,1}$)	-20,-1	-6.03	-5.24	-2.49	-6.17 (-7.2, -5.1)
r_2 (red)	0,2	0.45	0.29	0.79	0.67 (0.51,0.72)
r_3 (blue)	0,2	0.65	0.51	0.93	0.79 (0.68, 0.86)
r_4(green)	0,2	0.76	0.90	1.70	0.87 (0.72, 0.91)
r_5 (orange)	0,2	0.43	0.31	0.50	0.62 (0.51, 0.72)
Log$_{10}$ (1-ε_1)**	-2,5	3.12	3.22	5.00	3.64 (3.15, 3.76)
Log$_{10}$ (1-ε_2)**	Calculated***	-	-	-	-
Log$_{10}$ (1-ε_3)**	-2,5	-0.99	-0.02	0.31	-0.29 (-2.0,-0.20)
Log$_{10}$ (1-ε_4)**	-2,5	-0.09	2.25	0.45	0.63 (0.13,0.70)
Log$_{10}$ (1-ε_5)**	Calculated***	-	-	-	-
c (days^{-1})	0,30	8.17	9.72	7.45	7.74 (6.7, 9.1)
δ (days^{-1})	0,2	0.77	N/A (=0.14)	N/A (=0.14)	N/A (=0.14)
k_{cure} (days^{-1})	0,5	N/A (=0.0)	0.28	N/A (=0.0)	0.17 (0.13, 0.32)
k_{super}	0,5	N/A (=0.0)	N/A (=0.0)	0.04	3.62 (2.5, 4.6)
Sum of squared residuals (SSR)		1.28	5.94	10.14	2.37

* The range of variation of p is constrained by the range of variation of R_0 (which we set as 5-15 (9)), where $R_0 = \frac{\beta\cdot p}{\delta\cdot c}\cdot T_0$.

** The value of ε_i is calculated as $D/(EC_{50,i}+D)$, where D is the drug concentration.

*** These values are calculated based on in vitro measurement of values of EC_{50}.

Table S4. Best fit parameter values and least square residuals of 5 models fitted to data collected from subject 2.

Parameter	Parameter variations (Lower, higher bound)	Baseline model	Cure Model (confidence intervals)	Superinfection Model	Full Model
p (days^{-1})	*	27.4	13.0 (12.5, 14.3)	47.1	14.3
Log$_{10}$ (I$_{0,2}$/I$_{0,1}$)	-20,-1	-12.03	-8.84 (-9.8, -5.2)	-13.96	-4.8
r$_2$ (green)	0,2	1.19	1.94 (1.0,1.99)	1.82	1.55
Log$_{10}$ (1-ε$_1$)**	-2,5	3.39	3.00 (3.2,3.7)	4.43	3.14
Log$_{10}$ (1-ε$_2$)**	-2,5	1.74	0.32 (0.05, 0.81)	-0.06	1.91
c (days^{-1})	0,30	14.89	12.13 (14.1, 15.5)	14.19	12.75
δ (days^{-1})	0,2	0.54	N/A (=0.14)	N/A (=0.14)	N/A (=0.14)
k$_{cure}$ (days^{-1})	0,5	N/A (=0.0)	0.14 (0.09, 0.12)	N/A (=0.0)	0.12
k$_{super}$	0,5	N/A (=0.0)	N/A (=0.0)	N/A (=0.0)	2.47
Sum of squared residuals (SSR)		1.13	2.05	6.42	1.83

* The range of variation of p is constrained by the range of variation of R_0 (which we set as 5-15 (9)), where $R_0 = \frac{\beta \cdot p}{\delta \cdot c} \cdot T_0$.

** The value of ε_i is calculated as $D/(EC_{50,i}+D)$, where D is the drug concentration.

Table S5. Best fit parameter values and least square residuals of 5 models fitted to data collected from subject 3.

Parameter	Parameter variations (Lower, higher bound)	Baseline model	Cure Model (confidence intervals)	Superinfection Model	Full Model
p (days^{-1})	*	20.2	33.3 (26.3, 33.8)	20.4	33.4
Log$_{10}$ (I$_{0,2}$/I$_{0,1}$)	-20,-1	-10.65	-9.40 (-11.1,-6.0)	-12.98	-9.18
Log$_{10}$ (I$_{0,3}$/I$_{0,1}$)	-20,-1	-7.52	-14.21 (-15.2, -9.1)	-8.16	-14.45
Log$_{10}$ (I$_{0,4}$/I$_{0,1}$)	-20,-1	-12.04	-14.17 (-14.6, -11.9)	-12.57	-14.24
r$_2$ (light blue)	0,2	1.11	0.84 (0.65, 0.97)	1.63	0.96
r$_3$ (green)	0,2	1.32	1.98 (1.11, 2.0)	1.72	1.99
r$_4$(dark blue)	0,2	1.33	1.97 (1.08,2.0)	1.72	1.97
Log$_{10}$ (1-ε$_1$)**	-2,5	3.61	3.79 (3.3, 4.0)	5.04	3.76
Log$_{10}$ (1-ε$_2$)**	-2,5	1.65	0.23 (0.08,0.30)	0.07	0.32
Log$_{10}$ (1-ε$_3$)**	-2,5	2.30	1.17 (0.8,1.18)	1.41	1.18
Log$_{10}$ (1-ε$_4$)**	-2,5	1.93	1.25 (0.9,1.36)	0.68	1.24
c (days^{-1})	0, 30	9.50	9.27 (9.0,9.5)	7.63	9.23
δ (days^{-1})	0, 2	0.67	N/A (=0.14)	N/A (=0.14)	N/A (=0.14)
k$_{cure}$ (days^{-1})	0, 5	N/A (=0.0)	0.12 (0.08, 0.13)	N/A (=0.0)	0.12
k$_{super}$	0, 5	N/A (=0.0)	N/A (=0.0)	2.00	0.03
Sum of squared residuals (SSR)		0.87	2.28	7.01	2.10

* The range of variation of p is constrained by the range of variation of R_0 (which we set as 5-15 (9)), where $R_0 = \frac{\beta\cdot p}{\delta\cdot c}\cdot T_0$.

** The value of $\boldsymbol{\varepsilon}_i$ is calculated as $D/(EC_{50,\mathrm{i}}+D)$, where D is the drug concentration.

Table S6. Best fit parameter values and least square residuals of 5 models fitted to data collected from subject 4.

Parameter	Parameter variations (Lower, higher bound)	Baseline model	Cure Model	Superinfection Model	Full Model (confidence intervals)
p (days^{-1})	*	47.5	52.6	27.8	50.5 (41.1, 50.5)
Log$_{10}$ ($I_{0,2}/I_{0,1}$)	-20,-1	-4.88	-4.31	-2.62	-5.14 (-5.5, -4.7)
Log$_{10}$ ($I_{0,3}/I_{0,1}$)	-20,-1	-7.32	-7.44	-12.29	-6.79 (-7.5, -6.2)
Log$_{10}$ ($I_{0,4}/I_{0,1}$)	-20,-1	-4.38	-4.26	-7.79	-7.96 (-9.2, -6.7)
r_2 (red)	0,2	0.47	0.22	0.67	0.74 (0.68, 0.76)
r_3 (blue)	0,2	0.63	0.40	0.73	0.80 (0.78, 0.81)
r_4(green)	0,2	0.79	0.69	0.89	0.92 (0.87,0.95)
Log$_{10}$ (1-ε_1)**	-2,5	3.51	2.05	5.00	3.12 (3.0, 3.18)
Log$_{10}$ (1-ε_2)**	Calculated***	-	-	-	-
Log$_{10}$ (1-ε_3)**	-2,5	1.30	0.03	-0.83	0.84 (0.75, 0.89)
Log$_{10}$ (1-ε_4)**	-2,5	1.91	1.15	2.05	0.94 (0.83, 0.98)
c (days^{-1})	0, 30	13.07	13.52	7.87	13.89 (11.67, 14.1)
δ (days^{-1})	0, 2	0.74	N/A (=0.14)	N/A (=0.14)	N/A (=0.14)
k_{cure} (days^{-1})	0, 5	N/A (=0.0)	1.37	N/A (=0.0)	0.28 (0.25, 0.29)
k_{super}	0, 5	N/A (=0.0)	N/A (=0.0)	2.00	2.25 (2.01, 2.31)
Sum of squared residuals (SSR)		1.80	6.66	16.25	1.92

* The range of variation of p is constrained by the range of variation of R_0 (which we set as 5-15 (9)), where $R_0 = \frac{\beta \cdot p}{\delta \cdot c} \cdot T_0$.

** The value of $\boldsymbol{\varepsilon}_i$ is calculated as $D/(EC_{50,\mathrm{i}}+D)$, where D is the drug concentration.

*** These values are calculated based on in vitro measurement of values of EC_{50}.

Table S7. Best fit parameter values and least square residuals of 5 models fitted to data collected from subject 5.

Parameter	Parameter variations (Lower, higher bound)	Baseline model	Cure Model (confidence intervals)	Superinfection Model	Full Model
p (days^{-1})	*	26.0	20.9 (18.8, 23.3)	12.8	20.8
Log$_{10}$ ($I_{0,2}$/$I_{0,1}$)	-20,-1	-5.81	-2.93 (-3.9, -2.6)	-2.81	-2.95
r_2	0,2	0.95	0.96 (0.9, 1.1)	1.19	0.99
Log$_{10}$ (1-ε_1)**	-2,5	3.45	3.49 (3.3, 3.7)	4.89	3.49
Log$_{10}$ (1-ε_2)**	Calculated***	-	-	-	-
c (days^{-1})	0,30	11.72	11.60 (10.2, 13.0)	8.96	11.59
δ (days^{-1})	0, 2	0.66	N/A (=0.14)	N/A (=0.14)	N/A (=0.14)
k_{cure} (days^{-1})	0, 5	N/A (=0.0)	0.14 (0.12, 0.17)	N/A (=0.0)	0.14
k_{super}	0, 5	N/A (=0.0)	N/A (=0.0)	0.00	4.80
Sum of squared residuals (SSR)		0.69	0.92	7.56	0.89

* The range of variation of p is constrained by the range of variation of R_0 (which we set as 5-15 (9)), where $R_0 = \frac{\beta \cdot p}{\delta \cdot c} \cdot T_0$.

** The value of $\boldsymbol{\varepsilon}_i$ is calculated as *D*/($EC_{50,\mathrm{i}}$+*D*), where *D* is the drug concentration.

*** These values are calculated based on in vitro measurement of values of EC_{50}.

Table S8. Estimated drug efficacy (against non-resistant virus), rate of cure and rate constant for superinfection for 5 subjects treated with MK-5172.

Subject ID – Genotype (viral type at baseline)	ε_1	$k_{cure}\times(-\log_{10}(1-\varepsilon_1))$[a] / corresponding average time to cure a cell	k_{super} [b]
1 - G1b (WT)	0.9997	0.59 day^{-1} / 1.7 days	3.62
2 - G1a (Q80K)	0.9996	0.38 day^{-1} / 2.6 days	
3 - G1a (Q80K)	0.9998	0.45 day^{-1} / 2.2 days	
4 - G1a (WT)	0.9992	0.88 day^{-1} / 1.1 days	2.25
5 - G1a (WT)	0.9997	0.50 day^{-1} / 2.0 days	
Mean	**0.9996**	**0.56 day^{-1} / 1.9 days**	
STD	**0.0002**	**0.20 / 0.56**	

[a] This quantity is the estimated cure rate of cells infected by the baseline virus in each subject.

[b] Note that the actual rate of superinfection is a product of k_{super} and the fitness differences between two strains.

Table S9. Parameter descriptions and values in the HCV models.

Parameter	**Description**	**Baseline value**	**Unit**	**Reference**
ρ_T	Logistic proliferation constant	2.0	day^{-1}	(8)
T_{max}	Carrying capacity of hepatocytes	1.3e+7	cells ml^{-1} day^{-1}	(8)
N	The concentration of liver cells that are not target cells	6.5e+6	cells ml^{-1}	(8)
d	Death rate of target cells	0.01	day^{-1}	(8)
β	HCV infectivity rate constant	8.88e-8	mL day^{-1}	(8)
δ	Death rate of infected hepatocytes	0.14	day^{-1}	(7)
i	Drug effectiveness for viral strain i	Calculated according to the expression shown in the ODEs		
$EC_{50,i}$	The value of EC50 for viral strain i	Fitted		
w	Elimination rate of the drug	0.49	day^{-1}	(10)
r_i	Fitness of viral strain i relative to the wild-type	Fitted		
p	Viral production rate	Fitted	day^{-1}	
c	Viral clearance rate	Fitted	day^{-1}	
k_{cure}	Rate constant for the cure of infected cells	Fitted (or set to 0 in the 'superinfection' model)	day^{-1}	
k_{super}	Dimensionless constant for the superinfection of infected cells	Fitted (or set to 0 in the 'cure' model)		

References:

1. Nelder JA & Mead R (1965) A simplex-method for function minimization. *Comput J* 7(4):308-313.
2. Romano KP*, et al.* (2012) The molecular basis of drug resistance against hepatitis C virus NS3/4A protease inhibitors. *PLoS pathogens* 8(7):e1002832.
3. Summa V*, et al.* (2012) MK-5172, a selective inhibitor of hepatitis C virus NS3/4a protease with broad activity across genotypes and resistant variants. *Antimicrob Agents Chemother* 56(8):4161-4167.
4. Ke RA*, et al.* (2014) Modelling clinical data shows active tissue concentration of daclatasvir is 10-fold lower than its plasma concentration. *J Antimicrob Chemoth* 69(3):724-727.
5. Burnham KP & Anderson DR (2002) *Model selection and multimodel inference: a practical information-theoretic approach* (Springer, New York) 2nd Ed p 488.
6. Bolker BM (2008) *Ecological models and data in R.* (Princeton University Press, Princeton) p 408.
7. Neumann AU*, et al.* (1998) Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science* 282(5386):103-107.
8. Rong L, Dahari H, Ribeiro RM, & Perelson AS (2010) Rapid emergence of protease inhibitor resistance in hepatitis C virus. *Science translational medicine* 2(30):30ra32.
9. Snoeck E*, et al.* (2010) A comprehensive hepatitis C viral kinetic model explaining cure. *Clinical pharmacology and therapeutics* 87(6):706-713.
10. Brainard DM*, et al.* (2010) Safety and antiviral activity of MK-5172, a novel HCV NS3/4a protease inhibitor with potent activity against known resistance mutants, in genotype 1 and 3 HCV-infected patients. *Hepatology* 52:706A-707A.