APPENDIX 1: DETAILED METHODS

Laboratory methods

Blood samples were screened for anti-HCV using a third-generation enzyme immunoassay (Abbott Laboratories, Chicago, Ill) and anti-HCV positive specimens were tested again by Murex anti-HCV version 4.0 (Murex Biotech, Kyalami, South Africa) for confirmation. Irrespective of anti-HCV status, all samples were tested for HCV RNA by the COBAS AMPLICOR HCV test version 2.0 (Roche Molecular Systems, Branchburg, NJ). HCV RNA positive blood samples were genotyped by a reverse-phase hybridization line probe assay (LiPA, Versant HCV Genotype Assay, Bayer, Tarrytown, NY) [1]. For molecular studies, amplification was performed using a nested in-house PCR with primers specific to the core region [2] and sequencing was performed on the PCR product using ABI PRISM^{TM} Dye Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems, Foster City, CA).

Identifying genetically distinct (confirmed) HCV reinfections

A change in HCV genotype or subtype was defined as genetically distinct. Where there was no change in HCV genotype of subtype, genetically distinct was defined as follows. All HCV RNA positive bleeds underwent viral sequencing (HCV core region, 331 nucleotides). Viral sequences were compared pairwise, and the maximum composite likelihood distances were calculated. The mean (sd) distance between viral sequences taken from different participants but with the same genotype and subtype was 0.035 (0.013). When consecutive sequences from the same participant were compared, 69.7% of consecutive sequences were identical (distance=0). Participants were defined as having a genetically distinct virus if they changed genotype or subtype or had two consecutive sequences with maximum composite likelihood sequence distance greater than 0.039 (that is, greater than three standard deviations of the distribution of pairwise differences from viral sequences from different participants with the same genotype and subtype). This approach was adapted from that used by Pham and colleagues [3]. A flowchart of HCV reinfection classification is provided in Figure 1.

Computation of transition probabilities

Transition probability matrices were calculated for each testing interval based on the instantaneous rate matrix (main manuscript, section 2.4) using the *expm* package in the R programming environment (*Higham 08b* method) [4]. It is worth noting that although the instantaneous rate of change from the susceptible to the chronic infection states, for example, is zero, the transition probability after a given period of time may be greater than zero. This makes sense from a biological perspective: whilst an uninfected person cannot become chronically infected with hepatitis C instantaneously (they would have to undergo a period of acute infection first), if a number of months had elapsed between tests, the probability of moving from the uninfected (susceptible) state to the chronically infected state may be greater than zero.

Computation of marginal posterior probability distributions

An adaptation of the Metropolis-Hastings method for Markov chain Monte Carlo sampling was used to estimate marginal posterior probability distributions for functions of the parameters [5,6].

In order to ensure that all areas of the parameter-space were sufficiently explored, three sets of disparate starting parameter values were selected; convergence of the runs was assessed using Gelman and Rubin's $\hat{R}$ diagnostic [7]. After convergence, the runs were continued for an additional 30100 steps. Prior to convergence, convergence was assessed every 3010 steps using a burn-in of one quarter of the length of the run. MCMC runs were thinned in order to minimize the computational memory requirements of the analyses: one in every twenty steps was maintained. Plots of the thinned MCMC samples in the order they were drawn and their autocorrelation have been included in Supplementary Figures 3 & 4. The $\hat{R}$ values are printed at the bottom of each page of plots.

Initial parameter values for the runs were chosen as follows. One run began approximately at the point estimates derived using traditional epidemiological methods for reinfection rate, duration of reinfection and probability clearance for the Networks 2 study data. The starting values for the second run were a very short duration of reinfection, a high reinfection rate and a high probability clearance. This second combination was designed so as to be on the other side of the predicted maximum likelihood parameter set from the starting values for the first run, but to still have a reasonably high likelihood because of the combination of short duration of reinfection with high reinfection rate and high probability clearance. In contrast, the starting values for the third run were chosen such that the duration of reinfection, reinfection rate and probability of clearance would be incompatible such that the likelihood of the starting values would be quite low. The values chosen to start each of the three runs are presented in the following table:

	Reinfection rate (per person-month)	Reinfection duration (months)	Probability clearance
1	0.018	4	0.75
2	0.113	1	0.92
3	0.050	3	0.80

Because the three parameters were interdependent and in order to ensure that a reasonable proportion of proposed increments were accepted (>10%), the proposed increments were set to be relatively small. In order to reduce the time to convergence, every thirty steps a jump was proposed, in which new values were proposed for all three parameters simultaneously. Because longer duration of reinfection was associated with a smaller reinfection rate and vice versa, the jumps were designed such that an increase in reinfection rate came with a decrease in reinfection duration and vice versa. Specifically, the jumps were designed such that the product of the reinfection rate and the reinfection duration remained constant (this was achieved by multiplying one by a contstant, *m*, and dividing the other by *m* at each jump, where *m* ranged from 1.3 to 2 and was chosen based on trial and error in order to achieve an acceptance rate for the jump step of around 5-20%). Furthermore, although self-limiting (i.e., spontaneously clearing) reinfections could be missed if the test interval was longer than the reinfection duration, persistent reinfections could not. Therefore, if the proposed rate of persistent reinfection is very different to the previous rate, the likelihood of the data given the proposal may differ markedly from the likelihood of the data under the previous sample. The rate of persistent reinfection is equal to the rate of reinfection minus the rate of reinfections that spontaneously clear – a quantity that depends on the reinfection rate and probability clearance. In order to maximise the probability that the proposal would be accepted, the proposed change in probability clearance was therefore designed such that the rate of persistent reinfection would remain constant given the change to the reinfection rate.

In order to ensure that the MCMC sample is not biased, it is important to maintain detailed balance, meaning that the probability of proposing a given parameter set (θ') from another parameter set θ, is equal to the probability of proposing the first set from the second – that is, $p(\theta \rightarrow \theta')=p(\theta' \rightarrow \theta)$. This was achieved by reversing the direction of the proposal every two jumps. On every second jump, the proposed changes were as follows: multiplying the reinfection rate by *m*, dividing the duration of reinfection by *m* and the following change to the proportion clearance (**γ**): 1-(1- **γ**)/*m*. On alternate jumps, the opposite changes were proposed: dividing the reinfection rate by *m*, multiplying the duration of reinfection by *m*, and the following change to the proportion clearance: 1-*m*(1- **γ**).

Metropolis-Hastings steps

Proposed increments for each parameter were randomly selected from normal distributions with means of 0 and standard deviations of 0.1, 2 and 0.15 for the reinfection rate per person-month, reinfection duration (in months) and probability clearance, respectively. These were selected through a process of trial and error where the objective was to set the values as high as possible while still maintain a reasonably high acceptance rate (>10%) for all of the simulations. At each step, an increment was proposed to only one of the three parameter values such that every three consecutive steps included a proposed change to each of the three parameters.

Methods for simulating data

HCV reinfection and clearance data were simulated for a fixed number of PWID (n=46, the same number as in the Networks 2 study) over time, *t*, using a probabilistic individual-based-model. All PWID in the model were assumed to have cleared their primary HCV infection, and be susceptible to reinfection (state S). After entering the susceptible state, the time to reinfection, whether or not the reinfection was cleared and if so, the time to clearance, were determined stochastically. The average reinfection rate was assumed to be constant (α). Once reinfected, participants moved to the first of two acute reinfection states (I_{A1}), and the average rate of leaving each of these two states was assumed to be constant $\left(\frac{2}{\gamma}\right)$, where γ represents the duration of acute reinfection. While the duration of each of the two acute infection states followed an exponential distribution, the combination of the two member states effectively implements a realistic Gamma-distributed acute infection duration [8]. From the second acute infection state, it was assumed that participants could spontaneously clear (with probability β), thereby returning to the susceptible state, or their infection could progress to chronicity, state I_C.

Thus, the average per-month rate of spontaneous clearance was defined as the product of the spontaneous clearance probability and the per-month rate of leaving the second of these acute infection states $\left(\frac{2\beta}{\gamma}\right)$, whereas the average per-month rate of an infection progressing from the acute infection state to the chronic infection state was defined as the product of the chronic infection probability and the per-month rate of leaving the second of these acute infection states $\left(\frac{2(1-\beta)}{\gamma}\right)$. Those with chronic infection (state I_C) were assumed to remain in that state indefinitely.

Model inputs

Data were simulated for two different scenarios regarding the assumed duration of reinfection (γ), reinfection rate (α) and probability of clearance (β). Because little is known about reinfection duration (γ) and in the model-based estimates it is closely linked to reinfection rate (α, Supplementary Figure 1), we chose two plausible reinfection duration scenarios. Both scenarios involved a Gamma distribution for reinfection duration with mean durations (and variances) of 1 month (0.5 months) and 2 months (2 months), respectively. For each of these two reinfection duration scenarios, we chose reinfection rates and probabilities of spontaneous clearance that were consistent with the Networks 2 study data. Specifically, for each scenario, the associated reinfection rate and probability of clearance were equal to the median posterior for these parameters from the analysis of the empirical study data, when the mean duration of reinfection was limited to the relevant value ± 10%. The values of reinfection rate and probability clearance that were used were:

- Scenario one - mean duration of reinfection equals one month:
  - Mean reinfection rate: 109 per 100 person-years
  - Probability clearance: 0.93
- Scenario two -mean duration of reinfection equals two months:
  - Mean reinfection rate: 72 per 100 person-years
  - Probability clearance: 0.89

Although two scenarios were considered, for brevity, only the results from scenario one are described in detail and results from scenario two are described briefly as a comparator. Three different scenarios regarding the number of participants were also simulated in order to produce data similar to the confirmed reinfections only dataset (containing 16 participants),the confirmed and possible reinfections dataset (containing 46 participants), and a larger dataset containing 100 participants in order to investigate the potential precision gains associated with increasing the number of participants. For each scenario, 100 datasets containing reinfection and spontaneous clearance time points were simulated. Supplementary Table 3 illustrates that for 50-200 replications of each scenario, the number of replications does not affect the simulation results. A range of test intervals (1/2 month, one month, two months and four months) were applied to each simulated reinfection and spontaneous clearance dataset in order to create sets of artificial *observational* data.

References

1. McCaw R, Moaven L, Locarnini SA, Bowden DS (1997) Hepatitis C virus genotypes in Australia. J Viral Hepat 4: 351-357.
2. Dev AT, McCaw R, Sundararajan V, Bowden S, Sievert W (2002) Southeast Asian patients with chronic hepatitis C: the impact of novel genotypes and race on treatment outcome. Hepatology 36: 1259-1265.
3. Pham ST, Bull RA, Bennett JM, Rawlinson WD, Dore GJ, et al. (2010) Frequent multiple hepatitis C virus infections among injection drug users in a prison setting. Hepatology 52: 1564-1572.
4. Goulet V, Dutang C, Maechler M, Firth D, Shapira M, et al. (2011) expm: Matrix exponential. R package version 0.98-5 http://CRAN.R-project.org/package=expm.
5. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21: 1087-1092.
6. Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97-109.
7. Gelman A, Rubin D (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7: 457–511.
8. Lloyd AL (2001) Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. Proc R Soc Lond B Biol Sci 268: 985-993.

Supplementary Table 1: Bayesian terminology

Term	**Abbreviation**	**Description**
Prior probability distribution	Prior	Probability distributions for the parameters that have been assigned by the researchers independent of the observational study data. This represents beliefs about the parameters prior to conducting the current study.
Posterior probability distribution	Posterior	Probability distribution for the parameter set calculated from the prior and the observational study data. This represents updated knowledge about the parameters after conducting the current study.
Marginal posterior probability distribution	Marginal distribution	Posteriors for individual parameters or functions of the parameters.
Credible interval	CrI	Analogous to confidence intervals in frequentist (non-Bayesian) statistics. (see appendix for more detail)
Markov chain Monte Carlo	MCMC	Computational sampling method used to calculate the posterior. (see appendix for more detail)

Supplementary Table 2: Table of equations for the probabilities of possible observed state changes

Observed state change	Transition probability
Susceptible to susceptible	$P_{SS}(t_2 - t_1)$
Susceptible to infected	$P_{SI_{A1}}(t_2 - t_1) + P_{SI_{A2}}(t_2 - t_1) + P_{SI_C}(t_2 - t_1)$
Infected to susceptible	$\frac{S_{I_{A1}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_{A1}S}(t_2 - t_1) + \frac{S_{I_{A2}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_{A2}S}(t_2 - t_1) + \frac{S_{I_C}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_CS}(t_2 - t_1)$
Infected to infected	$\frac{S_{I_{A1}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} \left(P_{I_{A1}I_{A1}}(t_2 - t_1) + P_{I_{A1}I_{A2}}(t_2 - t_1) + P_{I_{A1}I_C}(t_2 - t_1)\right)$ $+ \frac{S_{I_{A2}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} \left(P_{I_{A2}I_{A1}}(t_2 - t_1) + P_{I_{A2}I_{A2}}(t_2 - t_1) + P_{I_{A2}I_C}(t_2 - t_1)\right) + \frac{S_{I_C}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_CI_C}(t_2 - t_1)$
Final two tests: HCV RNA positive followed by HCV RNA negative	$w\left(\frac{S_{I_{A1}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_{A1}S}(t_2 - t_1) + \frac{S_{I_{A2}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_{A2}S}(t_2 - t_1) + \frac{S_{I_C}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_CS}(t_2 - t_1)\right)$ $+ (1 - w)\Bigg(\frac{S_{I_{A1}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} \left(P_{I_{A1}I_{A1}}(t_2 - t_1) + P_{I_{A1}I_{A2}}(t_2 - t_1) + P_{I_{A1}I_C}(t_2 - t_1)\right)$ $+ \frac{S_{2I_{A2}}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} \left(P_{I_{A2}I_{A1}}(t_2 - t_1) + P_{I_{A2}I_{A2}}(t_2 - t_1) + P_{I_{A2}I_C}(t_2 - t_1)\right) + \frac{S_{I_C}(t_1)}{S_{I_{A1}}(t_1) + S_{I_{A2}}(t_1) + S_{I_C}(t_1)} P_{I_CI_C}(t_2 - t_1)\Bigg)$

Table notes: ***P*** is the transition probability matrix; $P_{ij}(t)$ is the probability of transitioning from state *i* to state *j* in time, t. ***S*** is the state probability matrix; $S_j(t)$ is the probability of being in state *j* at time, *t*. The symbols for the states are *S*, I_{A1}, I_{A2} and I_C. *S* is the susceptible state (state 1), I_{A1} is the first acute infection state (state 2), I_{A2} is the second acute infection state (state 3), I_C is the chronic infection state (state 4). t_1 is the time in months since baseline at a given observation and t_2 is the time in months since baseline at the following observation. *w* is the proportion of times in the study in which an HCV RNA positive test followed by an HCV RNA negative test results in spontaneous clearance amongst those with a subsequent study observation.

Supplementary Table 3: The effect of the number of repetitions on simulation results.

Repetitions	Test interval	Reinfection Rate - Simple Epidemiological Estimate	Reinfection Rate - Model Estimate	Duration of Reinfection - Simple Epidemiological Estimate	Duration of Reinfection - Model Estimate	Spontaneous Clearance Probability - Simple Epidemiological Estimate	Spontaneous Clearance Probability - Model Estimate
50	4	19.8 (8.1-25.2)	42.3 (17.0-61.7)	4.0 (4.0-4.6)	1.7 (1.1-2.2)	0.60 (0.22-0.89)	0.82 (0.50-0.98)
100	4	20.3 (9.9-29.2)	42.4 (22.8-73.9)	4.0 (4.0-4.6)	1.7 (1.1-2.4)	0.62 (0.25-0.89)	0.83 (0.56-0.98)
200	4	19.2 (10.8-30.3)	42.0 (17.0-76.3)	4.0 (4.0-4.7)	1.6 (1.0-2.5)	0.62 (0.25-0.88)	0.83 (0.56-0.99)
50	2	33.8 (12.2-46.7)	59.5 (21.7-134.0)	2.1 (2.0-2.3)	1.0 (0.5-2.8)	0.76 (0.50-0.94)	0.90 (0.64-0.99)
100	2	34.0 (22.2-49.4)	62.9 (30.9-134.1)	2.1 (2.0-2.3)	1.0 (0.5-3.1)	0.77 (0.55-0.94)	0.90 (0.73-1.00)
200	2	32.6 (18.2-44.7)	62.1 (30.9-134.1)	2.0 (2.0-2.3)	1.0 (0.5-2.8)	0.76 (0.54-0.96)	0.90 (0.73-1.00)
50	1	54.5 (31.7-71.5)	97.1 (59.7-146.7)	1.1 (1.0-1.2)	0.6 (0.4-1.3)	0.86 (0.66-0.97)	0.94 (0.76-0.99)
100	1	56.2 (42.1-77.7)	98.4 (45.3-149.0)	1.1 (1.0-1.2)	0.6 (0.4-1.5)	0.86 (0.71-0.97)	0.93 (0.84-1.00)
200	1	54.4 (35.8-72.4)	96.6 (55.0-173.8)	1.1 (1.0-1.2)	0.6 (0.4-1.5)	0.85 (0.71-0.95)	0.93 (0.83-1.00)
50	0.5	81.1 (62.3-103.6)	107.9 (73.9-145.5)	0.7 (0.6-0.8)	0.5 (0.3-1.1)	0.90 (0.76-0.96)	0.94 (0.81-0.99)
100	0.5	84.2 (60.8-103.6)	113.2 (73.9-147.8)	0.7 (0.6-0.8)	0.5 (0.3-1.1)	0.90 (0.78-0.96)	0.94 (0.84-0.99)
200	0.5	80.3 (61.5-104.4)	109.0 (73.1-150.5)	0.7 (0.6-0.8)	0.5 (0.3-1.1)	0.90 (0.80-0.97)	0.94 (0.86-0.99)

Notes: All data in this table was simulated using 46 participants.