The Epidemic Behavior of the Hepatitis C Virus

Oliver G. Pybus,* Michael A. Charleston, Sunetra Gupta, Andrew Rambaut, Edward C. Holmes, Paul H. Harvey

Hepatitis C virus (HCV) is a leading worldwide cause of liver disease. Here, we use a new model of HCV spread to investigate the epidemic behavior of the virus and to estimate its basic reproductive number from gene sequence data. We find significant differences in epidemic behavior among HCV subtypes and suggest that these differences are largely the result of subtype-specific transmission patterns. Our model builds a bridge between the disciplines of population genetics and mathematical epidemiology by using pathogen gene sequences to infer the population dynamic history of an infectious disease.

An estimated 170 million people worldwide are at risk of liver cirrhosis and liver cancer due to chronic infection with HCV (*1*). The virus is responsible for 10,000 deaths per year in the United States, and this rate is expected to increase substantially in the next two decades (*2*). HCV is a rapidly evolving single-stranded positive-sense RNA virus that exhibits enormous genetic diversity. It is classified into six types (labeled 1 through 6) and numerous subtypes (labeled 1a, 1b, etc.), which differ in diversity, geographical distribution, and transmission route (*3*). Subtypes appear to differ in treatment response, although their role in variation of disease progression is unclear (*2*, *4*). Any successful HCV vaccination or control strategy, therefore, requires an understanding of the nature and variability of epidemic behavior among subtypes.

HCV was first isolated in 1989, and knowledge of its long-term epidemiology before that date is limited. Highly divergent strains have been found in restricted geographic areas such as West Africa and Southeast Asia, suggesting a long period of infection in these regions. In contrast, several globally prevalent subtypes are much less divergent, indicating a recent worldwide spread of these strains (*5*–*7*).

We investigate HCV epidemiology using coalescent theory, a population genetic model that describes how the demographic history of a population determines the ancestral relationships of individuals sampled from it (*8*, *9*). Phylogenies reconstructed from contemporary HCV gene sequences contain information about past population dynamics and can, therefore, be used to infer viral epidemic behavior (*10*). We also demonstrate one way in which the fundamental epidemiological quantity R_0 (the basic reproductive number of a pathogen) can be estimated from gene sequences. R_0 represents the average number of secondary infections generated by one primary case in a susceptible population and can be used to estimate the level of immunization or behavioral change required to control an epidemic (*11*).

The framework of coalescent theory allows us to estimate $N(t)$, a continuous function that represents the effective number of infections at time t. Time t is zero at the present and increases into the past, hence $N(0)$ is the effective number of infections at the present. $N(t)$ can be considered as the inbreeding effective population size of the viral epidemic (*12*). Previous viral coalescent studies have used simple models for $N(t)$, specifically, constant population size and exponential growth (*13*, *14*). A more appropriate approach, which we use here, is to develop a basic epidemiological model, from which a suitable form for $N(t)$ is obtained. Because there is little protection against HCV reinfection (*15*) and vertical transmission is rare, its epidemic spread can be represented by

$$\frac{dy}{dt} = (1 - y)By - \frac{y}{D} \qquad (1)$$

where y is the proportion of the at-risk population that is infected and D is the average duration of infectiousness. B is a combination of parameters relating the force of infection (the per capita rate of acquisition of infection) to the prevalence of infection. In this model, $R_0 = BD$ and equilibrium prevalence is $1 - (1/R_0)$. A time-reversed version of Eq. 1 was solved for y and then transformed into effective population size using the relation $N(t) = N(0)\,[y(t)/y(0)]$. The resulting demographic model is

$$N(t) = N(0)\frac{1 + c}{1 + ce^{rt}}, \text{ where}$$

$$r = B - D^{-1} \text{ and } c = \frac{re^{-kr}}{B} \qquad (2)$$

r is the growth rate achieved in a wholly susceptible population, c is a logistic shape parameter, and k is the constant of integration. Note that B, D, and k cannot be separated.

Given a molecular phylogeny reconstructed from contemporary viral gene sequences (*16*), it is possible to estimate $N(0)$, r, and c within a maximum likelihood (ML) framework (*17*). Because reconstructed phylogenies represent time in units of nucleotide substitutions per site, some parameters are estimated as functions of the substitution rate (*18*). These parameters can be transformed back into their natural units using the substitution rate of the viral gene concerned. We estimated HCV substitution rates by reanalyzing gene sequences sampled in 1995 from individuals who were infected by a single batch of antibody to rhesus D 17 years earlier (*19*–*21*).

The above methods were used to investi-

Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

*To whom correspondence should be addressed. E-mail: oliver.pybus@zoo.ox.ac.uk

Table 1. Maximum likelihood parameter estimates for each HCV type or subtype. Seq., number of sequences.

HCV	Gene	Seq.	ML parameter estimates (approximate 95% CIs)		
			$N(0)$	r	c
1a	E1	34	13726 (6143, 32615)	0.098 (0.081, 0.114)	∞ (0.28, ∞)
	NS5	59	9858 (5488, 18430)	0.095 (0.079, 0.109)	∞ (0.24, ∞)
1b	E1	76	46484 (11408, 114430)	0.079 (0.068, 0.092)	3.3 (0.12, ∞)
	NS5	53	11800 (3254, 62084)	0.088 (0.068, 0.117)	0.44 (0.015, ∞)
4	E1	22	1817 (1012, 8797)	0.026 (0.008, 0.045)	0.015 (3×10^{-5}, ∞)
	NS5	18	2498 (1328, 5820)	0.043 (0.022, 0.071)	6×10^{-5} (4×10^{-9}, 0.037)
6	E1	23	1579 (948, 3433)	0.012 (0.003, 0.029)	0.028 (2×10^{-6}, ∞)
	NS5	40	2500 (1680, 4532)	0.008 (0.002, 0.018)	0.066 (8×10^{-5}, ∞)

gate four HCV strains. HCV types 6 and 4 are genetically diverse but geographically constrained: type 6 is restricted to Southeast Asia and type 4 is found predominantly in Africa and the Middle East. In contrast, subtypes 1a and 1b are less divergent but are distributed globally (*4*). For each subtype, E1 and NS5 gene sequences were collated from GenBank and aligned by hand (Table 1) (*22*). Phylogenies were estimated from the alignments using a ML approach under the assumption of a constant rate of nucleotide substitution (*23*). In each case, the hypothesis of rate constancy was tested (*24*).

Table 1 reports the ML estimates of $N(0)$, r, and c, with approximate confidence intervals (CIs), obtained from each HCV data set. CIs for estimates of r are considerably smaller than those for $N(0)$ and c (*25*). Figure 1 represents these estimates graphically, and compares them with a nonparametric estimate of $N(t)$ (*26*). For each subtype, the E1 and NS5 results are similar, and our model appears to fit the demographic signal in the data well.

There are significant differences in epidemic history among the HCV strains. Subtypes 1a and 1b seem to have originated about 100 years ago, whereas types 4 and 6 appear to be much older, having arisen about 350 and 700 years ago, respectively (Fig. 1). The growth rates of subtypes 1a and 1b during the last 100 years are considerably greater than those of types 4 and 6, providing confirmation of a recent and rapid spread of subtypes 1a and 1b, in contrast to a long period of localized endemic infection for types 4 and 6 (*5*, *6*).

Types 4 and 6 appear to have reached equilibrium prevalence some time in the past, whereas subtype 1b's growth rate decreased only very recently and subtype 1a is still exponentially growing at the present (Fig. 1). These observations reflect the different modes of transmission that characterize the four strains. The recent and swift global dissemination of subtypes 1a and 1b is largely the result of their effective transmission through modern contact networks, specifically, injecting drug use (IDU) and infected blood products. Subtype 1b transmission is more commonly associated with blood transfusion and hemodialysis, suggesting improved blood screening as the cause of its recent growth rate decrease (*2*), whereas subtype 1a is most strongly linked to IDU (*27–32*). These results corroborate epidemiological surveys, which indicate a decrease in the prevalence of subtype 1b relative to subtype 1a through time (*27–32*). In contrast, type 4 and type 6 HCV infections in less developed regions are often "community acquired" by a variety of undefined social and domestic routes (*2*, *33*, *34*), explaining the earlier spread and lower growth rates observed for these strains.

Estimates of R_0 can be obtained straightforwardly using the relation $R_0 = rD + 1$ (Eq. 2). Figure 2 displays estimates of R_0 for each subtype under a range of plausible D (*2*, *35*). For each D, the R_0 values of subtypes 1a and 1b are significantly higher than those of types 4 and 6. Integrating uniformly across this range and averaging the E1 and NS5 results, we obtain the following point estimates of R_0: 2.93 for subtype 1a, 2.67 for subtype 1b, 1.68 for type 4, and 1.21 for type 6. Because there is little reason to believe D varies substantially among strains, the observed differences in R_0 probably result from differences in the transmission parameters that collectively define B. These differences most likely arise from the association of subtypes with specific transmission routes. This conclusion is strengthened by two observations: (i) types 4 and 6 can spread quickly if they enter efficient contact networks (*36–39*) and (ii) in the absence of such networks, HCV type 1 in West Africa shows evidence of long-term endemic infection (*7*). However, the possibility of viral genetic variability in infectiousness among subtypes should not yet be discounted entirely (*40*).

Extrapolating our estimates into the near future, it is clear that, in terms of new infections, subtype 1a poses the greatest threat to public health. Subtype 3a, which was not included in our analysis, is also strongly linked to IDU and may pose a similar risk (*27–32*). Furthermore, we can use our estimates of R_0 to predict that the eventual equilibrium prevalence of subtype 1a will be ~65%. This value is not unrealistically high because it represents prevalence within the subtype 1a risk group rather than within the

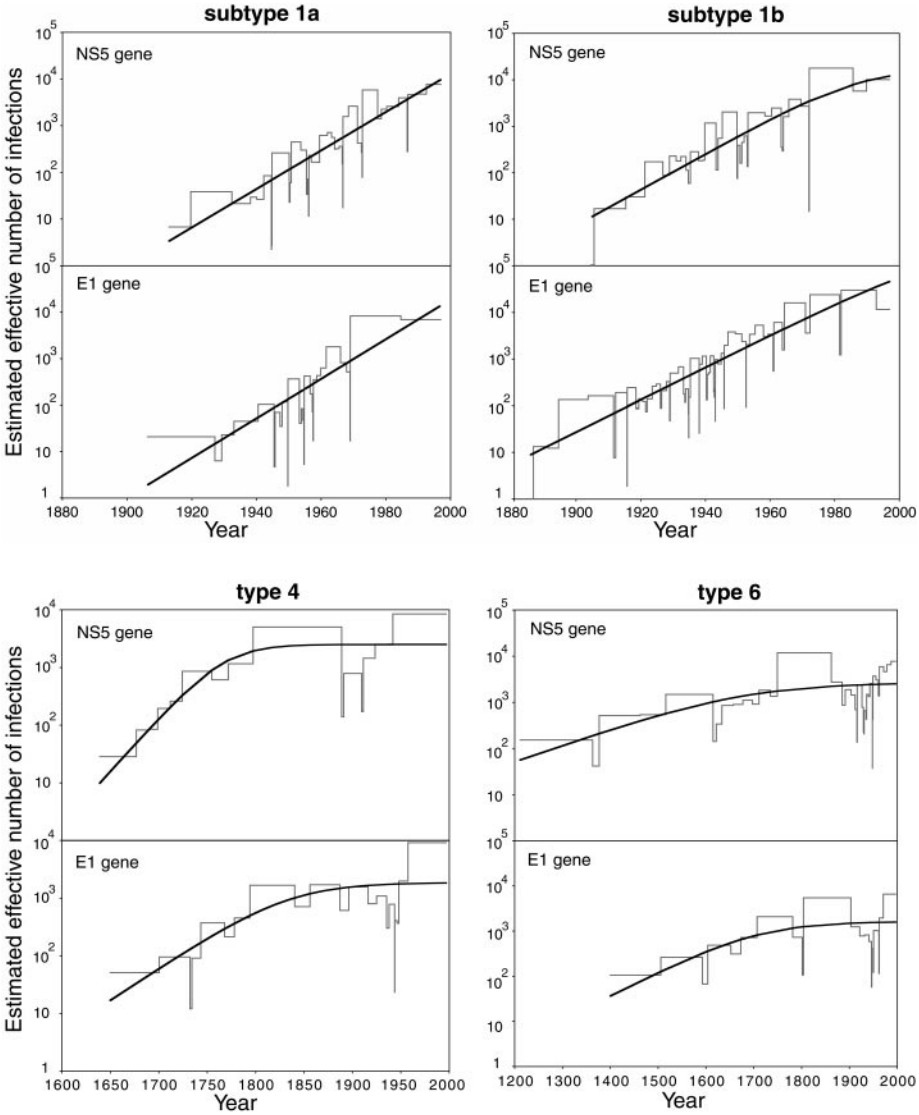

Fig. 1. Maximum likelihood estimates of $N(t)$, the effective number of infections through time, for each HCV data set (black curves) (*17*). The gray, stepwise plots represent corresponding nonparametric estimates of $N(t)$ (*26*). Genetic distances were transformed into a time scale of years using estimates of E1 and NS5 nucleotide substitution rates (*20*). These plots are point estimates of $N(t)$ and, thus, contain no information about uncertainty in $N(0)$, r and c (Table 1) or μ (*20*).

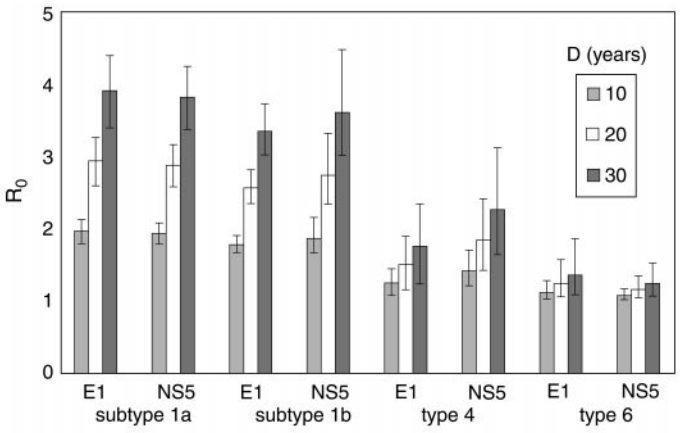

Fig. 2. Estimates of R_0 for each HCV data set, when the average duration of infectiousness, *D*, is 10, 20, and 30 years.

general population. HCV prevalence among injecting drug users is already at this level (70 to 80%) (*2*) and most new HCV infections occur soon after initiation of IDU (*41*), suggesting that the sustained exponential increase in subtype 1a infections (Fig. 1) is at least partly driven by non-IDU transmission and continual recruitment into the IDU risk group.

Our analysis has been aided by three factors: the abundance and variability of HCV sequence data, the absence of observable recombination in HCV (*42*), and the existence of independent substitution rate estimates. However, we recognize that selection, uncertainty in phylogeny estimation, and variable substitution rates are probably present in our data and may confound the interpretation of our results. Yet the consistency of our results among structural and nonstructural genes that are under different selective constraints (*43*, *44*) and their concordance with current epidemiological data (*27–32*) suggest that our conclusions are at least qualitatively robust. Importantly, no confounding factor appears to vary in a subtype-specific manner so as to produce the results we observe.

The methods introduced here demonstrate that viral gene sequences constitute a potentially significant source of information about epidemiological processes. These methods are especially suitable for rapidly evolving viruses that do not induce lifelong immunity, because the R_0 values of such infections cannot be estimated from the average age at first infection (*45*). We hope that other HCV subtypes will be similarly investigated as more sequence data becomes available. However, analysis of other viruses may require more complex epidemiological models than that used here, and it is possible that coalescent-based approaches will be less effective when applied to pathogens, such as influenza, that exhibit strong cyclical population dynamics, due to the loss of genetic information during population bottlenecks.

References and Notes

1. World Heath Organization, *Wkly. Epidemiol. Rec.* **72**, 65 (1997).
2. Centers for Disease Control and Prevention, *Morb. Mortal. Wkly. Rep.* **47**, RR-19 (1998).
3. P. Simmonds *et al.*, *J. Gen. Virol.* **74**, 2391 (1993).
4. X. Forns, J. Bukh, *Viral Hepatitis Rev.* **4**, 1 (1998).
5. J. Mellor *et al.*, *J. Gen. Virol.* **76**, 2493 (1995).
6. D. B. Smith *et al.*, *J. Gen. Virol.* **78**, 321 (1997).
7. D. Jeannel *et al.*, *J. Med. Virol.* **55**, 92 (1998).
8. J. F. C. Kingman, *J. Appl. Probab.* **19A**, 27 (1982).
9. R. C. Griffiths, S. Tavaré, *Philos. Trans. R. Soc. London Ser. B* **344**, 403 (1994).
10. S. Nee, E. C. Holmes, A. Rambaut, P. H. Harvey, *Philos. Trans. R. Soc. London Ser. B* **349**, 25 (1995).
11. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and control*. (Oxford Univ. Press, Oxford, 1991).
12. The relation between the true number of infections at time t, $I(t)$, and the effective number of infections, is $N(t) = I(x)/\sigma^2$, where σ^2 is the variance in reproductive success among infections (*8*, *9*), and is assumed to be constant through time.
13. O. G. Pybus, A. Rambaut, P. H. Harvey, *Genetics* **155**, 1429 (2000).
14. O. G. Pybus, E. C. Holmes, P. H. Harvey, *Mol. Biol. Evol.* **16**, 953 (1999).
15. P. Farci *et al.*, *Science* **258**, 135 (1992).
16. In mathematical terms, this is a rooted connected acyclic graph whose tips represent contemporary gene sequences and whose internal nodes are dated according to a given time scale.
17. If P is a viral phylogeny (*16*) and φ is a vector representing the parameters of model $N(t)$, then it is possible to calculate $l[\varphi|P]$, the log-likelihood of φ given P (*9*). ML estimates of φ are found by numerically optimizing $l[\varphi|P]$. Approximate 95% CIs for these estimates are obtained using the likelihood ratio statistic (*13*). Software to perform these analyses and details of the optimization algorithms used are available at http://evolve.zoo.ox.ac.uk/software/.
18. Specifically, we estimate $N(0)\mu$ and r/μ, where μ is the substitution rate in substitutions per site per year. Parameter c is unaffected by linear scaling of time.
19. J. P. Power *et al.*, *Lancet* **345**, 1211 (1995).
20. HCV substitution rates were obtained using E1 and NS5 gene sequences sampled in 1995 from individuals who were infected by a single batch of antibody to rhesus D in 1978 (*19*). There was little variation in the infected antibody to rhesus D batch, so the phylogeny of the sequences was assumed to be a star with each branch representing 17 years of time. For each gene, a single constant rate of substitution was estimated using a ML method (*21*). A HKY85 substitution model with codon-position rate heterogeneity was used. The estimates obtained were $\mu = 7.9 \times 10^{-4}$ (6.1×10^{-4}, 9.9×10^{-4}) for the E1 gene and $\mu = 5.0 \times 10^{-4}$ (3.6×10^{-4}, 6.7×10^{-4}) for the NS5 gene.
21. A. Rambaut, *Bioinformatics* **16**, 395 (2000).
22. To reduce nonrandom sampling, sequences were excluded if they came from the same patient or if the infections were related by direct transmission (information obtained from primary sources). We used gene regions that matched the alignments used to estimate E1 and NS5 substitution rates (*20*). Subtypes 6a and 4a were removed because they are unlikely to be representative of types 6 and 4 as a whole. The high prevalence of subtype 4a in Egypt is probably the result of its widespread iatrogenic transmission during antischistosomal treatment campaigns (*36*). Additionally, it appears that subtype 6a is spreading via IDU in Hong Kong (*37*, *38*). Subtypes 4a and 6a should therefore be analyzed as separate populations (like subtypes 1a and 1b) when enough sequences are available. Alignments are available on request.
23. ML phylogenies were estimated using the HKY85 substitution model and a codon-position model of rate heterogeneity. The final tree for each alignment was obtained by reestimating the branches of the ML topology using the REV substitution model. Phylogenies were estimated using PAUP* 4.0d65 (D. L. Swofford, Sinauer Associates, Sunderland, MA) and are available on request.
24. The molecular clock was tested by maximizing the likelihood of each phylogeny with and without the restriction of a clock (*46*). Type 6 was the only strain for which the clock was rejected in both genes. For the remaining types, the clock was rejected in one of the two genes. There was no relation between significance and sample size. Because our estimates of epidemic history are consistent among genes, it appears that the level of rate variation within HCV subtypes is not large enough to systematically bias demographic inferences.
25. We tested the robustness of our r estimates to changes in c by reestimating r whilst constraining c at a value equal to either the upper or lower CI of c. In every case, the new estimates of r fell within the CIs of r reported in Table 1.
26. The nonparametric estimate is the skyline plot. See (*13*).
27. S. Pol *et al.*, *Gastroenterology* **108**, 581 (1995).
28. J. M. Pawlotsky *et al.*, *J. Infect. Dis.* **171**, 1607 (1995).
29. H. Rosen, S. Chou, A. Sasaki, D. Gretch, *Am. J. Gastroenterol.* **94**, 3015 (1999).
30. L. B. Seeff *et al.*, *Ann. Intern. Med.* **132**, 105 (2000).
31. E. Silini *et al.*, *J. Hepatol.* **22**, 691 (1995).
32. F. Dubois, J. Desenclos, N. Mariotte, A. Goudeau, *Hepatology* **25**, 1490 (1997).
33. J. H. Kao, D. S. Chen, *J. Gastroenterol. Hepatol.* **15**, E91 (2000).
34. O. Shobokshi, F. Serebour, L. Skakni, Y. Al-Saffy, M. Ahdal, *J. Med. Virol.* **58**, 44 (1999).
35. S. Sherlock, in *Viral Hepatitis*, A. Zuckermann, C. Thomas, Eds. (Churchill Livingstone, New York, 1993), pp. 1–17.
36. C. Frank *et al.*, *Lancet* **355**, 887 (2000).
37. L. E. Prescott *et al.*, *J. Med. Virol.* **50**, 168 (1996).
38. D. Wong, L. Tong, W. Lim, *Eur. J. Epidemiol.* **14**, 421 (1998).
39. Y. Morice *et al.*, *J. Gen. Virol.* **82**, 1001 (2001).
40. Y. Kimura, K. Hayashida, H. Ishibashi, Y. Niho, Y. Yanagi, *J. Med. Virol.* **61**, 37 (2000).
41. R. S. Garfein, D. Vlahov, N. Galai, M. Doherty, K. Nelson, *Am. J. Public Health* **86**, 655 (1996).
42. L. E. Prescott *et al.*, *J. Med. Virol.* **53**, 237 (1997).
43. R. H. Miller, R. H. Purcell, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2057 (1990).
44. D. B. Smith, P. Simmonds, *J. Mol. Evol.* **45**, 238 (1997).
45. S. Gupta, K. Trenholme, R. M. Anderson, K. P. Day, *Science* **263**, 961 (1994).
46. J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981).

47. We thank P. Donnelly, B. Griffiths, M. Worobey, and the anonymous referees for their helpful comments. Supported by the Wellcome Trust (50275), the BBSRC (BIF05332), and the Royal Society.

15 December 2000; accepted 15 May 2001