The Current Hepatitis C Virus Prevalence in China May Have Resulted Mainly from an Officially Encouraged Plasma Campaign in the 1990s: a Coalescence Inference with Genetic Sequences

Ling Lu, Wangxia Tong, Lin Gu, Chunhua Li, Teng Lu, Kok Keng Tee and Guihua Chen

J. Virol. **2013, 87(22):12041. DOI: 10.1128/JVI.01773-13.**
Published Ahead of Print 28 August 2013.

JVI Journals.ASM.org

The Current Hepatitis C Virus Prevalence in China May Have Resulted Mainly from an Officially Encouraged Plasma Campaign in the 1990s: a Coalescence Inference with Genetic Sequences

Ling Lu,[a,b] Wangxia Tong,[a] Lin Gu,[a] Chunhua Li,[b] Teng Lu,[c] Kok Keng Tee,[d] Guihua Chen[a]

Laboratory for Hepatology, Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou, Guangdong, China[a]; Department of Pathology and Laboratory Medicine, University of Kansas Medical Center, Kansas City, Kansas, USA[b]; University of Southern California, Los Angeles, California, USA[c]; Centre of Excellence for Research in AIDS (CERiA), Department of Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia[d]

In this study, we investigated hepatitis C virus (HCV) molecular epidemiology and evolutionary dynamics. Both E1 and NS5B sequences were characterized in 379 of 433 patients in southern China and classified into five major subtypes: 1b in 256 patients, 6a in 67 patients, 2a in 29 patients, 3a in 14 patients, and 3b in 13 patients. Using the E1 sequences obtained, along with those from other studies using samples from China, we inferred the HCV epidemic history by means of coalescence strategies. Five Bayesian skyline plots (BSPs) were estimated for the five subtypes. They concurrently highlighted the rapid growth in the HCV-infected population size from 1993 to 2000, followed by an abrupt slowing. Although flanked on both sides by variable population sizes, the plots showed distinct patterns of rapid HCV growth. Coincidently, 1993 to 2000 was a period when contaminated blood transfusions were common in China due to a procedural error in an officially encouraged plasma campaign. The abrupt slowing in 1998 to 2000 corresponded to the central government outlawing paid blood donations in 1998. Using a parametric model, the HCV population growth rates were estimated during 1993 to 2000. It was revealed that the 6a rate was the highest, followed by those of 1b, 2a, 3b, and 3a. Because these rates differed significantly ($P < 1e{-}9$) from each other, they may help explain why 6a is increasingly prevalent in southern China and 1b is predominant nationwide. These rates are approximately 10-fold higher than those reported elsewhere. These findings suggested that during the plasma campaign, certain barriers to efficient viral transmission were removed, allowing wide HCV dissemination.

Hepatitis C virus (HCV) is classified into six genotypes and >80 subtypes. Different genotypes exhibit distinct geographic distribution patterns. Generally, genotypes 1, 2, and 3 are prevalent worldwide; genotypes 4 and 5 are primarily restricted to Africa (1); and genotype 6 is endemic in Southeast Asia (2, 3). In addition, evidence suggests that different genotypes have spread among different population subsets, and they show different epidemiological patterns (4–6). The routes of HCV transmission (7–9), such as the use of blood products, hemodialysis, blood transfusion, unsafe medical practices, injection drug use, and other forms of parenteral exposure, are related to specific HCV genotypes and subtypes (10, 11). Therefore, it is of considerable importance to elucidate the evolutionary and epidemiological patterns of different HCV genotypes and subtypes, as the knowledge may be required for developing novel antiviral treatments and vaccine strategies. For example, patients infected with genotype 2 have shown better response to alpha interferon treatment and a significantly higher infected-cell death rate than those infected with genotype 1 (12). Based on such differences, new therapies may be designed that specifically target these viral strains.

Recently, the use of BEAST (Bayesian Evolutionary Analysis by Sampling Trees) software has become one of the most common methods to study HCV evolutionary and epidemiological patterns (13, 14). Unlike typical analyses of genetic sequences that use only partial phylogenic information, the BEAST inferences are processed in a stochastic framework based on coalescence theory (15). BEAST reconstructs the ancestral relationship and population history of the contemporarily sampled sequences in the context of a "best-fit" molecular clock. It correlates epidemic events and estimates historical population growth rates. It is powerful for performing retrospective analysis of the changes in effective population size using Bayesian skyline plots (BSPs), and this enables an easy display of the diverse histories from the most recent common ancestor (tMRCA) to the newest descendants (5, 16–18). Therefore, it has been widely used to estimate the past dynamics of HCV in various scenarios. However, different HCV genotypes and subtypes should have evolved at their own evolutionary rates based on the essence of a relaxed molecular clock (13). For this reason, in the present study, we selected the proper prior rates to fit the detected HCV subtypes.

Approximately 170 million people worldwide, or 3% of the global population, are chronically infected with HCV (19). The frequency of infection varies remarkably among populations and geographic regions, with Asia exhibiting levels significantly higher than the global average (20). China is a major Asian country that has >1.3 billion inhabitants, in whom the frequency of HCV infection has been reported to be 3.2% overall and 3.1% in rural areas (21, 22). In other words, >40 million people in China are currently infected with HCV, approximately 95% of which is caused by viruses of five major subtypes: 1b, 2a, 3a, 3b, and 6a (23,

Received 2 July 2013 Accepted 17 August 2013
Published ahead of print 28 August 2013
Address correspondence to Ling Lu, llu@kumc.edu, or Guihua Chen, chgh1955@263.net.
L.L. and W.T. contributed equally to this article.

doi:10.1128/JVI.01773-13

24). We hypothesized that such a high HCV frequency and unique genotype distribution pattern must have resulted from certain founder events and that by using the newly developed algorithms implemented in the BEAST software we might be able to date these events and quantify the historical HCV population growth. To test these hypotheses, we determined the HCV sequences in the samples collected from 423 patients in Guangzhou city, one of the major cities in China, where a considerable portion of the residents are immigrants or migrants from across the country (25). Based on the determined sequences, we performed phylogenetic analyses for a standard HCV classification, followed by subjecting them to BEAST inferences, in combination with the sequences from our previous studies (24–26) and those reported by others (27–30). Because the last two sequence collections were sampled in different provinces, they represent the HCV strains prevalent in the country.

MATERIALS AND METHODS

Subjects and specimens. Serum samples were obtained for routine testing for HCV RNA that was performed from January 2009 to December 2011 on all patients with chronic liver disease who visited the 3rd Affiliated Hospital of Sun Yat-Sen University, Guangzhou City, China. Only those samples positive for HCV RNA were reserved, while those that were negative were excluded, thus providing serum samples from a total of 423 patients. The ethical review committee of the 3rd Affiliated Hospital of Sun Yat-sen University approved this study. Guidelines set by this committee were strictly followed.

Sequence amplification. The E1 and NS5B sequences of HCV were amplified and sequenced as previously described (24, 25). RNA extraction (Qiagen Viral RNA extraction kit; Qiagen, Valencia, CA), cDNA synthesis (RevertAid First Strand cDNA Synthesis kit; Fermentas), and nested PCR (FastStart *Taq* DNA Polymerase, dNTPack; Roche, Indianapolis, IN) were performed following the manufacturers' guidelines. Standard procedures were implemented to avoid possible carryover contamination (31). To obtain the consensus sequences, the amplicons were sequenced directly in both directions using ABI Prism BigDye 3.0 terminators with an appropriate primer on an ABI Prism 3500 genetic analyzer (PE Applied Biosystems, Foster City, CA). Any errors in base calling were corrected using the SeqMan program, and the sequences were edited using the EditSeq program, followed by alignment using the MegAlign program. These programs are parts of the Lasergene 8.1 package (DNAStar Inc., Madison, WI).

Phylogenetic analysis. The HCV isolates were genotyped by means of a phylogenetic analysis of the E1 and/or NS5B sequence with reference sequences representing various genotypes and subtypes. Prior to the tree reconstruction, the best-fitting substitution model was selected using the jModeltest program based on the Akaike information criterion (AIC), by which the GTR + I + Γ (generalized time-reversible model with proportion of invariable sites and shape parameter of the gamma distribution) was found to be the best fit for all the data sets. Under this model, the maximum-likelihood (ML) trees were heuristically searched using the SPR (subtree pruning and regrafting) and NNI (nearest-neighbor interchange) algorithms implemented in PHYML (32). With the tree files generated, the ML tree topology was displayed using Figtree (5).

Evolutionary analysis. To estimate the HCV epidemic history, five sequence data sets were assembled, which comprised the sequences of subtypes 1b, 2a, 3a, 3b, and 6a. The majority of these sequences were from this study and our previous reports (24–26), while a small portion were from the reference literature (27–30). Among the latter, many were from our collaborators, with whom we communicated about the sequencing procedures and sampling dates. For a few sequences, however, this information was obtained from the related papers. We were able to verify that all the included sequences were not from molecular engineering, viral recombinants, or experimental animals and that each represents a single isolate from a single patient and has a known sampling date. To include the sequences from our studies while avoiding sampling bias, an additional three standards were also adhered to: (i) for each isolate, the sequences should be determined in the E1 and the NS5B regions; (ii) the E1 and NS5B sequences of an identical isolate should be positioned consistently in the phylogenetic trees to reflect the consistent genetic relatedness to other isolates; and (iii) no sequences should be identical or nearly identical, or they were removed.

Based on the five sequence data sets, the coalescence analyses were performed using the Bayesian MCMC (Markov chain Monte Carlo) algorithm implemented in the BEAST software (version 1.6.1), for which the best combination (the GTR + I + Γ substitution model, the Bayesian skyline coalescent model, and the uncorrelated exponential clock model) was selected based on its outperforming the other combinations (26). However, the evolutionary prior rates were different and were chosen for their fit with the subtypes or by considering their epidemiological patterns or were estimated directly from the data sets. For the last, a uniform distribution of a prior rate was used, with its lower bound set at 0 and the upper bound at 0.01. Using this uniform prior rate in the MCMC analysis, a posterior distribution of rates was produced and applied as the prior rate of a subsequent MCMC analysis. After these parameters were set, the MCMC analysis was run for 300 million chains, and a tree was output every 10,000 chains. To interpret the MCMC chains and output posterior trees, the tracer v1.5 program was used to estimate tMRCAs and their 95% confidence intervals and to reconstruct the BSP to form a retrospective picture of the epidemic history of HCV. To assess the sampling convergence of the MCMC procedures, the estimated effective sampling sizes (ESS) were evaluated. In this study, when all of the ESS numbers were $\geq$200, sufficient sampling was considered to have been achieved.

Estimation of HCV population growth rates. To identify the differences among subtypes, approximate HCV population growth rates were estimated using regression analysis during the period of rapid growth. The growth rate directly measures the number of new cases that can be caused by an existing patient every year. For this purpose, we first exported a BSP log file to obtain the median number of effective HCV infections each year. We then imported this information into the R software for regression analysis, in which the slope measures the speed of population growth. For an easier comparison, we performed a natural logarithmic transformation to the number of effective infections. After the rapid-growth period was delimited, a simple linear regression was performed on the data within this period. The regression function of each subtype was given, and an approximate growth rate was determined. However, it is not ideal to estimate a growth rate from a BSP, because only the median growth curve is taken into consideration, while the confidence limits are not included.

To obtain more precise growth rates with acceptable confidence limits, we ran BEAST under a "parametric" logistic model that was used by Pybus et al. (33), in which two parameters had to be set manually by editing the initial XML file: growth rate and t_{50}. For the growth rate, the highest posterior density obtained by Pybus et al. (33) was used, which showed a right-skewed shape of distribution. Because the lognormal distribution is also right skewed, we used it as the prior rate distribution with the log mean equaling the log slope and the log standard deviation at 0.3 and used the above-mentioned estimated approximate growth rates as the prior rates. The parameter t_{50} denotes the time at which the viral population size reached 50% of its maximum level, which was before the date on which the most recent descendants appeared. According to the BSP results, we set a normal distribution with its mean at 15 and the standard deviation at 2 as the priors for t_{50}. In addition to these parameters, the other parameters, such as the evolutionary prior rate, the substitution model, and the clock model, remained the same as described above for the evolutionary analysis. After the estimates were given, a two-tailed *t* test was performed to compare the posterior growth rate of one data set with that of another to show their statistical differences.

Nucleotide sequence accession numbers. The nucleotide sequences reported in this study were deposited in GenBank with the following accession numbers: JX676774 to JX677559.

RESULTS

HCV sequencing and phylogenetic analysis. From 257 (60.8%) male and 166 (39.2%) female patients, partial sequences of HCV were determined in the E1 and/or NS5B regions. Among the patients, 417 had the sequences determined in both regions, while 6 patients had the sequences determined only in the NS5B region. After the initial phylogenetic analysis with the reference isolates, we determined that the sequences in 26 patients were identical or nearly identical to another sequence. In one patient, the genotyping was not consistent in the two regions, and in 11 patients, the sequences represented rare subtypes. To meet the standards described in Materials and Methods and for analyzing the five major HCV subtypes, we excluded the sequences from a total of 44 patients and retained those from 379 patients, representing subtypes 1b in 256 patients, 2a in 29 patients, 3a in 14 patients, 3b in 13 patients, and 6a in 67 patients. Each of these 379 patients had sequences determined in both the E1 and NS5B regions, representing a single HCV isolate from a single subject. Phylogenetically, these sequences are distinct from each other.

Figure 1A presents an ML tree reconstructed with the E1 sequences of the 256 1b isolates in this study. The two previously described clusters, A and B, were observed again (24–26). They include 64 (25.0%) and 155 (60.5%) sequences, respectively, and show 88% and 86% bootstrap support. Figure 1B presents the second ML tree reconstructed using the same 256 1b sequences, with the addition of 89 1b sequences that we previously reported (24). The latter were isolated from patients sampled in different provinces in China. The overall topologies of these two trees strongly resemble each other, except that many colored branches were added in Fig. 1B. The latter tree indicates that HCV subtype 1b infection in China is long established and well mixed, so that the sequences sampled in this study from a single hospital can represent those that are prevalent across the country.

Figure 2A shows an ML tree that includes E1 sequences from 97 subtype 2a isolates. In addition to the 29 in this study, 35 were taken from our previous reports (24, 25) and 30 from earlier commercial blood donors (28). At the top of the tree, a large monophyletic cluster includes 44 sequences with 87% bootstrap support. Among these, 26 were from Hebei Province (28), 13 from Guangdong, and five from other provinces. In the middle of the tree, a cluster with 78% bootstrap support includes 14 sequences from Guangdong and 1 from another province. In the rest of the tree, the sequences show a tendency to form clusters, but without significant bootstrap support.

Figure 2B depicts an ML tree including the E1 sequences from 82 subtype 3a isolates. In addition to the 14 in this study, 53 were from our previous reports (25, 26) and 10 were recently described (30). In this tree, five clusters were observed, and four exhibited significant bootstrap support.

Figure 2C shows an ML tree including E1 sequences from 104 subtype 3b isolates. In addition to the 13 in this study, 20 were from our previous studies (24–26) and 71 were from reference reports, and all were from intravenous drug users (IDUs); 44 were from Guangxi Province (27), 24 from Hubei (29), and 3 from Jiangsu (30). Two monophyletic clusters were observed, but they did not show significant bootstrap support.

Figure 3 presents an ML tree that includes E1 sequences from 164 subtype 6a isolates. In addition to the 67 in this study, 97 were taken from our previous reports (24–26). These sequences were assembled to ensure the accuracy of the sampling dates and a sufficient sequence length. The three previously described clusters, I, II, and III, were maintained (25, 26). They contain 74, 24, and 37 sequences, respectively, representing 45.1%, 14.6%, and 22.6% of the 164 isolates, and they show 69%, 64%, and 66% bootstrap support, respectively.

Evolutionary analysis. The first analysis was performed using the 1b sequences shown in Fig. 1B. Initially, the evolutionary rate for 1b was unknown and an estimate was required. For this purpose, a subtype 1b data set was assembled, containing 144 sequences in the same range as that analyzed in this study. These sequences were retrieved from the Los Alamos HCV database with known sampling dates from 1983 to 2008. Phylogenetic and histogram inspection revealed sufficient temporal and phylogenetic structures (34). Based on this data set, we first set a uniform distribution of prior rates for a BEAST analysis, which generated a posterior distribution of rates, $1.21 \times 10^{-3} \pm 2.86 \times 10^{-5}$ substitutions/site/year (Table 1). Using this rate as a prior, a subsequent BEAST analysis was performed based on the E1 sequences shown in Fig. 1B. A flexible nonparametric estimate of the past changes in the HCV-infected population size was illustrated as a BSP (Fig. 4A). It showed a relatively constant population size from 1980 through 1993, followed by a period of rapid population growth until 2000. After that, the growth slowed but has nevertheless continued to increase until the present.

The second analysis was based on the 2a data set shown in Fig. 2A, which was assembled to ensure that there were a sufficient number of sequences with known sampling dates. Initially, the evolutionary rate for 2a was also unknown; therefore, we used the same strategy described above to estimate a rate based on the 2a data set, which generated a posterior distribution rate of $2.92 \times 10^{-3} \pm 9.04 \times 10^{-7}$ substitutions/site/year (Table 1). Using this rate as the prior rate, a subsequent analysis was performed using the same 2a data set, which yielded the BSP shown in Fig. 4B. Compared to 1b, the BSP for 2a showed a lower magnitude but steeper population growth for a shorter period, from 1995 until 2000. However, this growth was followed by an obvious but slow decrease that has persisted to the present.

The third analysis was based on the 3a data set shown in Fig. 2B, but the above-mentioned 1b rate was used as the prior rate. This is because the estimation of the 3a data set did not provide a valid rate, and on the other hand, the epidemiological pattern for 3a is more similar to that of 1b than to those of the others, which are globally distributed rather than geographically restricted. We therefore used the 1b rate as the prior rate. The resulting BSP showed a pattern almost identical to that for 1b, except that the rapid growth had a slightly lower magnitude (Fig. 4C).

The fourth analysis was based on the 3b data set shown in Fig. 2C. For a reason similar to that described above for 3a and because the majority of the 3b sequences were from IDU cohorts, in which 6a infections have been detected more often than any other subtype (10, 27), we selected the 6a rate as the prior rate. This 6a rate of $2.737 \times 10^{-3} \pm 1.60 \times 10^{-7}$ has been recently estimated using a strategy similar to that described above for 1b and 2a (26). A BSP was generated, and it looked to a certain extent similar to that for 2a. However, it exhibited a slow decrease before and a sharp turn

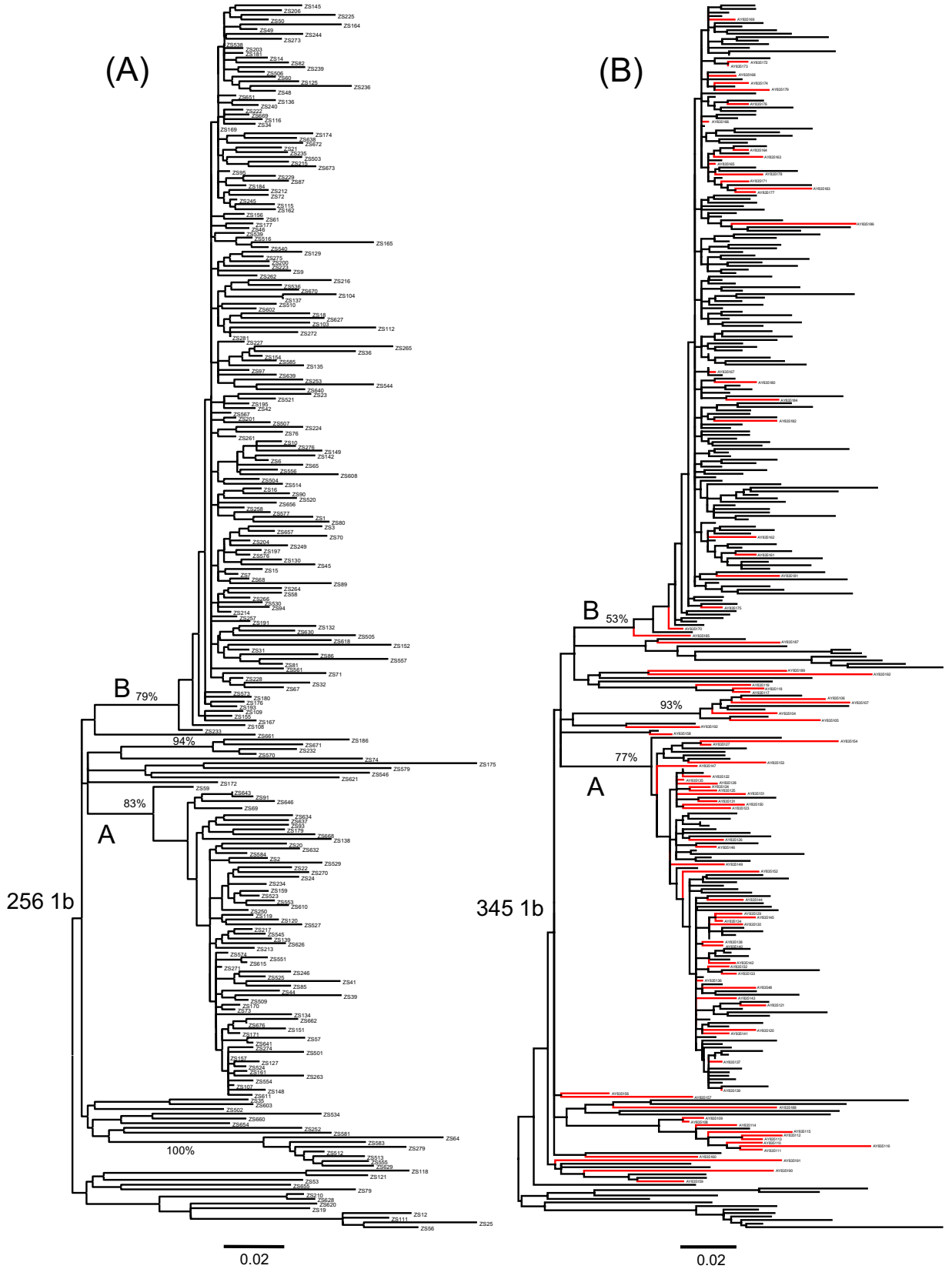

FIG 1 (A) ML tree reconstructed using the E1 sequences of the 256 subtype 1b isolates determined in this study, corresponding to nucleotide positions 869 to 1289 in the H77 genome. Two previously described clusters, A and B (24, 40), are indicated in the tree, and they show significant bootstrap support: 83% and 79%, respectively. The scale bar at the bottom represents 0.02 nucleotide substitution per site. (B) ML tree reconstructed using the same 256 subtype 1b sequences as in panel A, with the addition of 89 subtype 1b sequences we previously reported (24). For differentiation, the branches that lead to the latter 89 sequences are shown in red, and their tips are labeled with GenBank accession numbers. Otherwise, no tips are shown.

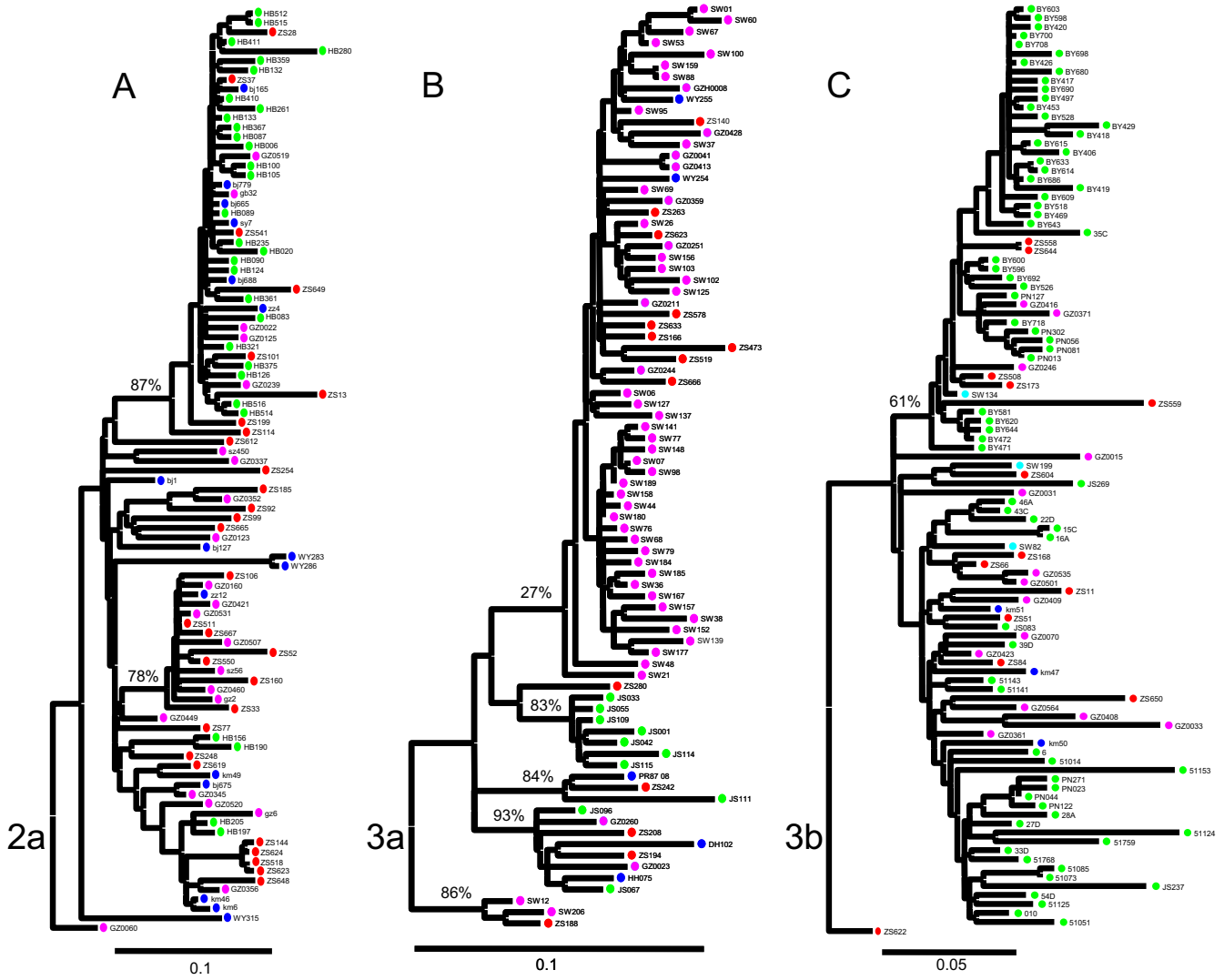

FIG 2 ML trees reconstructed with the E1 sequences from 97 subtype 2a (A), 82 subtype 3a (B), and 104 subtype 3b (C) isolates. (A) In the tree, the red (from this study) or pink (24–26) circles indicate the sequences from Guangdong Province, the blue circles denote the sequences from other provinces (24, 42), and the green circles mark the sequences from earlier commercial blood donors in Hebei Province (28). (B) In the tree, the red circles indicate the sequences from this study, the pink circles denote the sequences from our previous reports, the green circles mark the sequences from IDUs in Jiangsu Province (30), and the blue circles label the sequences retrieved from the HCV database. (C) In the tree, the red circles indicate the sequences from this study, the blue circles denote the sequences from patients, the pink circles mark the sequences from blood donors, and the green circles label the sequences from IDUs. In each tree, the percentages show bootstrap values, and the scale at the bottom indicates the nucleotide substitutions per site.

after the rapid-growth period, followed by a straight median line until the present (Fig. 4D).

The fifth analysis was based on the 6a data set shown in Fig. 3. Although the resulting BSP strongly resembled the one we recently estimated (26), the current one did not show an obvious increase in the size of the HCV-infected population during the most recent decade (Fig. 4E).

In conclusion, the five BSPs highlighted a period from 1993 to 2000 during which rapid growth occurred in the size of the HCV-infected population, followed by an abrupt slowing. Although flanked on both sides by variable population sizes, distinct patterns of rapid growth were shown for different subtypes.

Estimation of the HCV population growth rate. Using a simple linear regression method to analyze the median population sizes exported from the five BSPs, the approximate population growth rates during the period of rapid HCV growth were estimated. The BSPs for 1b and 3a exhibited rapid growth from 1993 to 2000, and their regression slopes were estimated to be 0.8385 and 0.6315, respectively. The BSPs for 2a and 6a exhibited rapid growth from 1995 to 2000, and their respective regression slopes were 0.8145 and 1.043. The BSP for 3b exhibited rapid growth from 1994 to 1999, and its regression slope was 0.898 (Table 2 and Fig. 4F).

Using these approximate data as prior rates to run a constant logistic model in BEAST (33), more precise population growth rates and their 95% confidence limits were determined (Table 2). Consistent with the above regression analysis, the mean growth rate for 6a (1.228) was found to be the highest, while 3a (0.968) was the lowest. Between these two rates were those for 1b, 2a, and 3b, at 1.158, 1.135, and 1.074, respectively. In the regression anal-

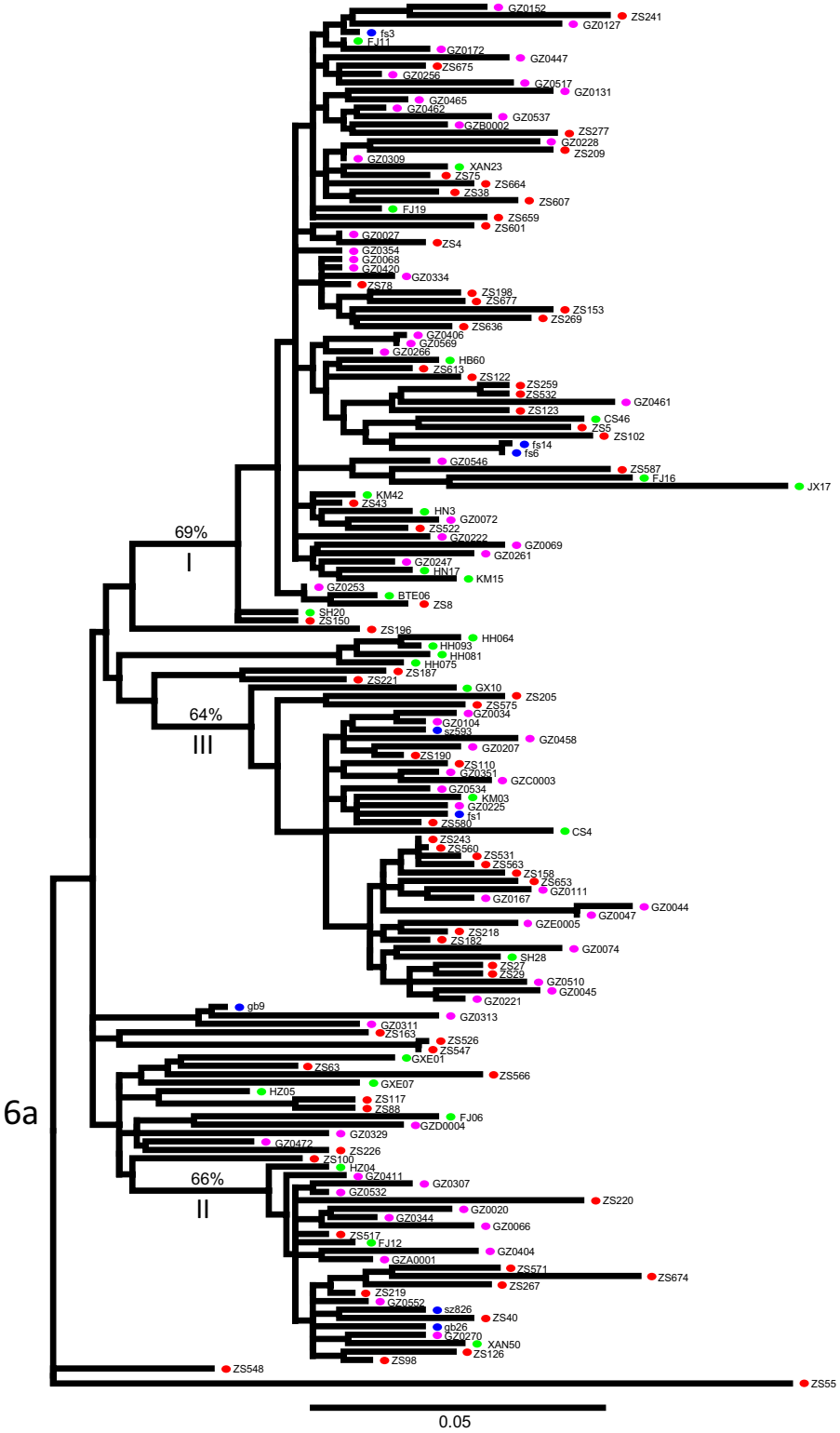

FIG 3 ML tree reconstructed using the E1 sequences of 164 subtype 6a isolates. Except for the 67 sequences (red circles) in this study, all the others are taken from our previous reports: 28 (green circles) from reference 26, 61 (pink circles) from reference 25, and eight (blue circles) from reference 24. The three previously described clusters, I, II, and III, are indicated, with 69%, 66%, and 64% bootstrap support, respectively. The scale bar at the bottom indicates 0.01 nucleotide substitution per site.

ysis, a rate order of 6a > 3b > 1b > 2a > 3a was observed, while under the parametric model, an order of 6a > 1b > 2a > 3b > 3a was found. A possible explanation for the slight difference is that in the regression analysis, the rapid-population-growth period in the BSPs was manually defined by visual inspection, while the confidence limit information was not incorporated. However, under the parametric model, this information was estimated directly using the sequence data sets, which could be more accurate. For pairwise comparison, two-tailed *t* tests were performed, and they showed that these rates are statistically different from each other with *P* values of <1e−9.

DISCUSSION

To estimate the HCV epidemic history in China, Bayesian MCMC analyses were performed using the sequences determined in this study combined with those from our previous studies (24–26) and those reported by others (27–30). These analyses resulted in five BSPs, each representing one subtype. They concurrently highlighted the rapid HCV population growth from 1993 to 2000. This is the period during which HCV infection became explosive in China.

Distinct BSP patterns were characterized for the different HCV subtypes. Compared to 1b, 3a, and 6a, the BSP for 2a exhibited a lower magnitude of rapid population growth for a shorter duration. This is consistent with our previous findings, which showed that in Guangdong Province, the frequency of 2a is now lower than that of 6a, because of the increasing prevalence of 6a, while people infected with 2a were significantly older than those infected with other HCV subtypes (25).

Although the BSP for 3b looked somewhat similar to that for 2a, there was a slow but persistent decrease before the rapid-population-growth period, while after a sharp turn from the rapid growth, the median BSP curve was immediately converted to a straight line. Because the majority of the 3b sequences were from IDUs, such a BSP pattern may not necessarily reflect the general situation.

Using the 6a sequences from a different collection, the resulting BSP looked almost identical to the one we recently reported (26). Such a result robustly confirms that the selected prior rate we used in both studies is appropriate for estimating the 6a epidemic history in China, by which reliable posterior estimates were consistently provided.

In all five BSPs, and particularly in the major one for 1b, a period from 1993 to 2000 was highlighted, during which a rapid increase in the size of the HCV-infected population occurred. This increase coincided with a period when contaminated blood transfusions were common in China, as exemplified by a plasma campaign during 1994 to 1996 in Henan Province that led to >500,000 blood donors being infected with HCV and >300,000 donors being infected with HIV-1 (35–38). Tragically, the majority of the infections were acquired at paid blood donation stations, some of which were operated by local health officials, where many donors' red blood cell samples were mixed, the plasma extracted, and the pooled red blood cells transfused back into the donors. Similar scenarios were also reported in many other provinces (36). To eliminate this risk, the central government in China has outlawed the use of paid blood donations since 1998 (37). This change was reflected in the five BSPs, each with an abrupt slowing in the HCV-infected population growth starting in 1999 to 2000. However, because a huge reservoir of HCV-infected individuals

TABLE 1 Statistics generated from Bayesian MCMC analyses for estimating the five BSPs

Statistic[d]	Value for BSP:				
	1b[a] (*n* = 256)	2a[b] (*n* = 97)	3a[a] (*n* = 82)	3b[c] (*n* = 104)	6a[c] (*n* = 167)
Rate prior	*N*(1.52e−3, 2.8325e−5)	*N*(2.925e−3, 8.28e−5)	*N*(1.52e−3, 2.8325e−5)	*N*(2.737e−3, 1.51e−5)	*N*(2.737e−3, 1.51e−5)
Direct rate	−0.0009	2.2654e−3	−0.0005	5.2106e−4	1.7748e−3
Posterior	−15,436.6 ± 2.3 (−15,530.2, −15,343.8)	−8,226.4 ± 0.8 (−8,273.4, −8,176.3)	−5,398.4 ± 0.64 (−5,444.2, −5,350.0)	−6,922.5 ± 0.91 (−6,970.7,−6,874.4)	−9,762.8 ± 1.19 (−9,830.3,−9,700.4)
Alpha	0.84 ± 1.4e−3 (0.714, 0.98)	1.10 ± 5.9e−5 (1.10, 1.11)	0.811 ± 4.4e−3 (0.48, 1.168)	0.84 ± 3.19e−3 (0.54,1.15)	0.622 ± 1.7e−3 (0.47, 0.77)
pInv	0.42 ± 4.6e−4 (0.375, 0.475)	0.43 ± 1.4e−5 (0.428, 0.431)	0.43 ± 1.03e−3 (0.33, 0.51)	0.34 ± 1.02e−3 (0.24, 0.43)	0.37 ± 7.94e−4 (0.30, 0.43)
Uced.mean (evolutionary rate)	1.21e−3 ± 2.86e−5 (1.16e−3, 1.26e−3)	2.92e−3 ± 9.04e−7 (2.76e−3, 3.1e−3)	1.21e−3 ± 2.86e−5 (1.16e−3, 1.26e−3)	2.737 ± 1.60e−7 (2.71e−3, 2.77e−3)	2.737e−3 ± 1.60e−7 (2.71e−3, 2.77e−3)
COV	0.88 ± 1.50e−3 (0.8158, 0.9444)	0.85 ± 1.08e−3 (0.76, 0.956)	0.84 ± 1.05e−3 (0.73, 0.95)	0.84 ± 7.83e−4 (0.75, 0.92)	0.84 ± 1.0e−3 (0.76, 0.92)
Covariance	8.8e−3 ± 1.23e−3 (−0.08, 0.10)	−0.02 ± 1.46e−3 (−0.15, 0.13)	−0.04 ± 8.46e−4 (−0.19, 0.10)	−0.03 ± 8.17e−4 (−0.16, 0.11)	−0.02 ± 8.26e−4 (−0.13, 8.82e−2)
tMRCA (yr)	57.79 ± 0.83 (31.90,96.22)	43.2 ± 0.18 (32.4, 54.8)	50.8 ± 0.21 (35.4, 67.0)	57.9 ± 0.18 (46.3, 70.7)	30.458 ± 0.198 (24.53, 36.27)
Tree likelihood	−13,746.9 ± 1.9 (−13,796.69, −13,697.2)	−7,668.7 ± 0.44 (−7,696.54, −7,642.55)	−4,872.1 ± 0.25 (−4,895.08, −4,850.38)	−6,304.7 ± 0.49 (−6,329.4, −6,279.2)	−8,737.4 ± 0.78 (−8,772.7, −8,701.5)

[a] Normal rate estimated for 1b in this study.
[b] Normal rate estimated for 2a in this study.
[c] Normal rate estimated for 6a in one of our recent studies (26).
[d] Rate prior, normal distribution (mean, standard deviation). Direct rate, rate directly estimated for the data set using the regression approach in Path-O-Gene, shown here only for comparison. Posterior, mean ± standard error of mean (lower, upper 95% confidence intervals). pInv, proportion of invariable nucleotide sites. Uced.mean, mean rate estimated under the uncorrelated exponential molecular clock model. COV, the coefficient of variation, which measures the rate heterogeneity among branches.

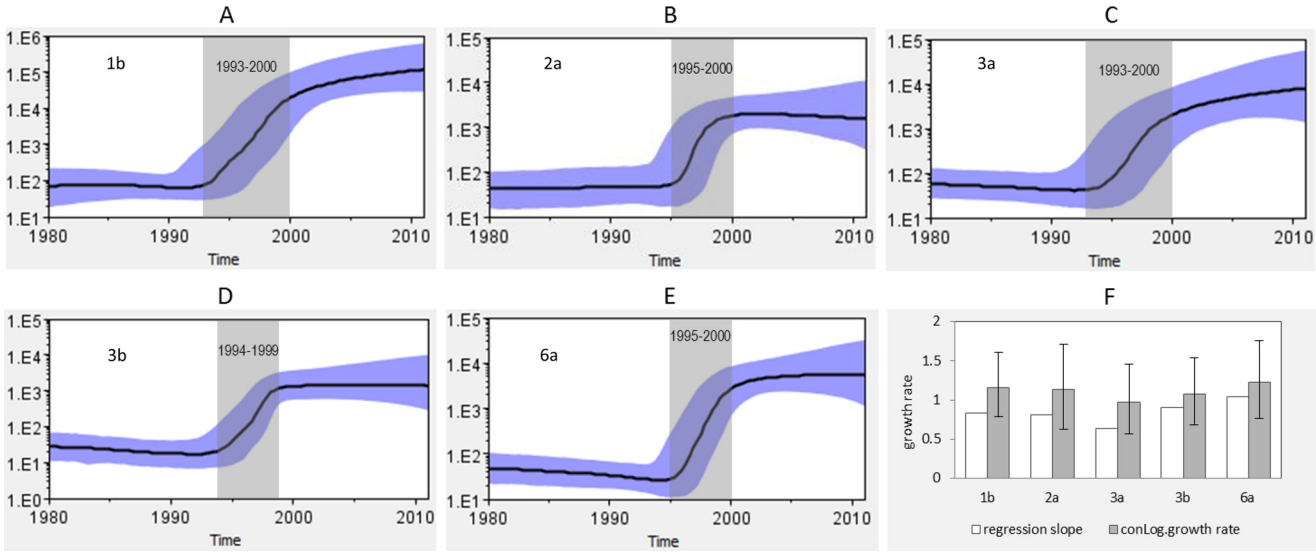

FIG 4 BSPs and HCV population growth rates. The five diagrams show the BSPs estimated for subtypes 1b, 2a, 3a, 3b, and 6a. In each diagram, the solid line represents the effective population size through time. The blue area represents the 95% highest posterior density confidence interval. A vertical ruler on the left measures the effective population size, while a horizontal scale on the bottom measures time. (F) HCV population growth rates. The white bars represent the rates estimated from regression analysis, and the gray bars indicate those estimated under the constant logistic (conLog) model, while the error bars show the 95% confidence intervals.

had been established, they had the opportunity to spread infection in the general population through drug use, unsafe medical practices, sex, and iatrogenic or other parenteral routes. In the major BSP for 1b, and the BSP for 3a as well, this trend was reflected in a slope of slow continuous increase in the HCV-infected population size up to the present.

HCV population growth rates (r) were estimated during the period of rapid growth. These rates were estimated initially by simple linear regression using the information on the median HCV-infected population size exported from the five BSPs and then by running a parametric model that directly used the sequence data sets. They both estimated the 6a rate to be the highest and 3a the lowest. Since the confidence limits and other statistical uncertainties were also included, the order of the rates generated under the parametric model, 6a > 1b > 2a > 3b > 3a, is thought to be more precise. Such an order might help explain why 1b is predominant across the country and 6a is increasingly prevalent in southern China.

TABLE 2 Statistics generated for estimating HCV population growth rates

Subtype	BSP regression slope	ConLog[a].growth rate				
		Mean	95% confidence interval		Median	ConLog.t_{50}[b]
			Lower	Upper		
1b	0.8385	1.158	0.7843	1.6076	1.132	10.596
2a	0.8145	1.135	0.6306	1.7167	1.094	12.390
3a	0.6351	0.968	0.5616	1.4613	0.930	11.738
3b	0.8980	1.074	0.6831	1.5438	1.044	12.713
6a	1.043	1.228	0.7690	1.7609	1.196	12.977

[a] conLog.growth rate, posterior growth rate estimated under the constant logistic model.
[b] Generated under the constant logistic model. t_{50} denotes the time at which the viral population size reached 50% of its maximum level.

The statistical differences in the growth rates for the five HCV subtypes may not necessarily demonstrate a major subtype-specific variation in HCV infectiousness. However, compared to previously published HCV growth rates (0.092 to 0.104) estimated for the IDU population in the United Kingdom, these rates (0.968 to 1.228) are approximately 10-fold higher. In the previous report, the unexpectedly low HCV growth rates were ascribed to a constraint in the growth of the IDU population, and such restricted rates may have underestimated the basic reproductive numbers (R_0) that would be achieved if the virus entered a new, naive population (8). In other words, if the virus is allowed to infect naive people without certain barriers, the growth rates would be much higher. From this point of view, our rates are closer to a theoretically unrestricted rate than those previously reported, since ours were estimated for the virus that became explosively epidemic in China as a result of the plasma campaign. To a certain extent, this suggests that some barriers to efficient HCV transmission were accidentally removed because of the plasma campaign advocated by the local government. However, investigation of this hypothesis would require acquisition of the documented number of blood donors during this period, which is unlikely to occur because of the official prohibition against the provision of such documentation.

The potential uncertainties of the study are discussed below. The first uncertainty concerns the study cohort. Although the samples were leftover material from clinical testing, they represent almost all of the HCV RNA-positive patients who visited the hospital described during the period from January 2009 to December 2011. With special liver programs, the hospital recruits patients not only locally, but also from other provinces. Importantly, the hospital is located in Guangzhou city, the third largest city and one

of the national central cities in China, which has 12.8 million inhabitants (http://en.wikipedia.org/wiki/Guangzhou), a considerable proportion of whom are immigrants or migrants from across the country (25). In addition, we included numerous reference sequences that were sampled in different provinces (24–30). Such a sequence collection can represent the HCV strains prevalent in China, although an ideal representation requires those randomly sampled in the general population in each of the provinces and major cities.

The second uncertainty lies in delimiting the time frames spanning the different phases in the 3a and 3b BSPs during which the viral populations multiplied at different speeds, although these increases were inherently predicted by the founder events. Among the common subtypes in the general population, such time frames may be concurrent, while in particular risk groups, they may vary considerably. According to the relaxed molecular clock rule (5, 14), the epidemic history of a given HCV subtype should be estimated using an evolutionary rate that is specific for that subtype (34). This strategy was used for the 1b, 2a, and 6a data sets, because we have the specific rates estimated in our recent work (26, 34) and in this study, but it was not done for 3a and 3b due to the lack of appropriate information. Instead, we used the 1b rate for 3a and the 6a rate for 3b, because they yielded consistent BSP time frames. However, the strategy may not be appropriate, because 3a and 3b represent HCV strains in China that are often detected in restricted geographic regions and among particular risk groups (26, 27, 29), and therefore, they may present their own unique profiles. Replacing the prior rates with alternative ones for a reanalysis, we observed changes in the 3a and 3b BSP time frames, but not the curve patterns (data not shown). Because we are not confident about which time frames can best reflect the factual scenarios, a question is raised in selecting the correct prior rate for estimating the history of a given HCV subtype that lacks its own specific rate information. Before this study, the majority of the reports used two rates, 7.9×10^{-4} for E1 and 5.0×10^{-4} for NS5B, estimated based on the single-lineage-among-host 1b data sets (7). Although lower than those recently estimated (34, 39), the two rates have been widely used for different HCV genotypes in various scenarios (5, 8, 18, 33, 40, 41), and thus, they must also be applicable to this study. Based on such a premise, we may perform additional BEAST analyses for the 3a and 3b data sets using additional prior rates to obtain more time frames before selecting the most appropriate. Should we select a prior rate only for the desired time frames, regardless of whether the rate is low or high? Perhaps this question is a common uncertainty for the majority of the BEAST analyses that use outer prior rates (i.e., rates estimated from a different HCV subtype), because an appropriate explanation of the estimates requires substantial speculation. However, such speculation may not be supported by data, or we may never find the data for verification due to the complete absence of historically archived samples. This situation may be accurate for 3a and 3b in this study, but it is not accurate for 1b, 2a, and 6a, which represent 93.7% (352/379) of the HCV isolates we characterized.

ACKNOWLEDGMENTS

The study was supported by a grant from NIAID/NIH (5 R01 AI080734-03A). K.K.T. received funding from a Malaysia Ministry of Higher Education High Impact Research (HIR) grant (H-500001-00-A000012-000001).

We have no competing interests.

REFERENCES

1. **Simmonds P, Bukh J, Combet C, Deleage G, Enomoto N, Feinstone S, Halfon P, Inchauspe G, Kuiken C, Maertens G, Mizokami M, Murphy DG, Okamoto H, Pawlotsky JM, Penin F, Sablon E, Shin IT, Stuyver LJ, Thiel HJ, Viazov S, Weiner AJ, Widell A.** 2005. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology **42:**962–973.
2. **Pham VH, Nguyen HD, Ho PT, Banh DV, Pham HL, Pham PH, Lu L, Abe K.** 2011. Very high prevalence of hepatitis C virus genotype 6 variants in southern Vietnam: large-scale survey based on sequence determination. Jpn. J. Infect. Dis. **64:**537–539.
3. **Zhou X, Chan PK, Tam JS, Tang JW.** 2011. A possible geographic origin of endemic hepatitis C virus 6a in Hong Kong: evidences for the association with Vietnamese immigration. PLoS One **6:**e24889. doi:10.1371/journal.pone.0024889.
4. **McOmish F, Yap PL, Dow BC, Follett EA, Seed C, Keller AJ, Cobain TJ, Krusius T, Kolho E, Naukkarinen R.** 1994. Geographical distribution of hepatitis C virus genotypes in blood donors: an international collaborative survey. J. Clin. Microbiol. **32:**884–892.
5. **Pybus OG, Barnes E, Taggart R, Lemey P, Markov PV, Rasachak B, Syhavong B, Phetsouvanah R, Sheridan I, Humphreys IS, Lu L, Newton PN, Klenerman P.** 2009. Genetic history of hepatitis C virus in East Asia. J. Virol. **83:**1071–1082.
6. **Simmonds P.** 2004. Genetic diversity and evolution of hepatitis C virus—15 years on. J. Gen. Virol. **85:**3173–3188.
7. **Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH.** 2001. The epidemic behavior of the hepatitis C virus. Science **292:**2323–2325.
8. **Pybus OG, Cochrane A, Holmes EC, Simmonds P.** 2005. The hepatitis C virus epidemic among injecting drug users. Infect. Genet. Evol. **5:**131–139.
9. **Tanaka Y, Hanada K, Mizokami M, Yeo AE, Shih JW, Gojobori T, Alter HJ.** 2002. A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. Proc. Natl. Acad. Sci. U. S. A. **99:**15584–15589.
10. **Goedert JJ, Chen BE, Preiss L, Aledort LM, Rosenberg PS.** 2007. Reconstruction of the hepatitis C virus epidemic in the US hemophilia population, 1940–1990. Am. J. Epidemiol. **165:**1443–1453.
11. **Hauri AM, Armstrong GL, Hutin YJ.** 2004. The global burden of disease attributable to contaminated injections given in health care settings. Int. J. STD AIDS **15:**7–16.
12. **Soetjipto Handajani R, Lusida MI, Darmadi S, Adi P, Soemarto, Ishido S, Katayama Y, Hotta H.** 1996. Differential prevalence of hepatitis C virus subtypes in healthy blood donors, patients on maintenance hemodialysis, and patients with hepatocellular carcinoma in Surabaya, Indonesia. J. Clin. Microbiol. **34:**2875–2880.
13. **Drummond A, Pybus OG, Rambaut A.** 2003. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitol. **54:**331–358.
14. **Drummond AJ, Rambaut A.** 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7:**214.
15. **Donnelly P, Tavare S.** 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:**401–421.
16. **Ferraro D, Genovese D, Argentini C, Giordano V, Pizzillo P, Stroffolini T, Craxi A, Rapicetta M, Di Stefano R.** 2008. Phylogenetic reconstruction of HCV genotype 1b dissemination in a small city centre: the Camporeale model. J. Med. Virol. **80:**1723–1731.
17. **Markov PV, Pepin J, Frost E, Deslandes S, Labbe AC, Pybus OG.** 2009. Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. J. Gen. Virol. **90:**2086–2096.
18. **Nakano T, Lu L, He Y, Fu Y, Robertson BH, Pybus OG.** 2006. Population genetic history of hepatitis C virus 1b infection in China. J. Gen. Virol. **87:**73–82.
19. **Armstrong GL, Wasley A, Simard EP, McQuillan GM, Kuhnert WL, Alter MJ.** 2006. The prevalence of hepatitis C virus infection in the United States, 1999 through 2002. Ann. Intern. Med. **144:**705–714.
20. **Sievert W, Altraif I, Razavi HA, Abdo A, Ahmed EA, Alomair A, Amarapurkar D, Chen CH, Dou X, El Khayat H, Elshazly M, Esmat G, Guan R, Han KH, Koike K, Largen A, McCaughan G, Mogawer S, Monis A, Nawaz A, Piratvisuth T, Sanai FM, Sharara AI, Sibbel S, Sood A, Suh DJ, Wallace C, Young K, Negro F.** 2011. A systematic review of

hepatitis C virus epidemiology in Asia, Australia and Egypt. Liver Int. **31**(Suppl. 2):61–80.
21. **Lei XZ, Shigeko N, Deng XW, Wang S, Qin S, Liu L, Tang H, Zhao L, Lei B, Yoshihiro A.** 1999. Prevalence of hepatitis C virus infection in the general population and patients with liver disease in China. Hepatol. Res. **14:**135–143.
22. **Xia GL, Liu CB, Cao HL, Li S, Zhan MY, Sub CA, Nan JH, Qi XQ.** 1996. Prevalence of hepatitis B and C virus infections in the general Chinese population: results from a nationwide cross-sectional seroepidemiologic study of hepatitis A, B, C, D, and E virus infections in China, 1992. Int. Hepatol. Commun. **5:**62–73.
23. **Gu L, Tong W, Yuan M, Lu T, Li C, Lu L.** 21 May 2013. An increased diversity of HCV isolates were characterized among 393 patients with liver disease in China representing six genotypes, 12 subtypes, and two novel genotype 6 variants. J. Clin. Virol. **57:**311–317.
24. **Lu L, Nakano T, He Y, Fu Y, Hagedorn CH, Robertson BH.** 2005. Hepatitis C virus genotype distribution in China: predominance of closely related subtype 1b isolates and existence of new genotype 6 variants. J. Med. Virol. **75:**538–549.
25. **Fu Y, Wang Y, Xia W, Pybus OG, Qin W, Lu L, Nelson K.** 2011. New trends of HCV infection in China revealed by genetic analysis of viral sequences determined from first-time volunteer blood donors. J. Viral Hepat. **18:**42–52.
26. **Fu Y, Qin W, Cao H, Xu R, Tan Y, Lu T, Wang H, Tong W, Rong X, Li G, Yuan M, Li C, Abe K, Lu L, Chen G.** 2012. HCV 6a prevalence in Guangdong province had the origin from Vietnam and recent dissemination to other regions of China: phylogeographic analyses. PLoS One **7:**e28006. doi:10.1371/journal.pone.0028006.
27. **Garten RJ, Zhang J, Lai S, Liu W, Chen J, Yu XF.** 2005. Coinfection with HIV and hepatitis C virus among injection drug users in southern China. Clin. Infect. Dis. **41**(Suppl. 1):S18–S24.
28. **Huang CH, Zhou JK, Liu L, Jiang RM, Cao YQ, Mu ZY, Zhang Y.** 2009. Investigating genotype of HCV distribution among residents in a "blood donation" village in Hebei Province. Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi **23:**8–10.
29. **Peng JS, Wang X, Liu MQ, Zhou DJ, Gong J, Xu HM, Chen JP, Zhu HH, Zhou W, Ho WZ.** 2008. Genetic variation of hepatitis C virus in a cohort of injection heroin users in Wuhan, China. Virus Res. **135:**191–196.
30. **Zhang C, Wu N, Liu J, Ge Q, Huang Y, Ren Q, Feng Q, He G.** 2011. HCV subtype characterization among injection drug users: implication for a crucial role of Zhenjiang in HCV transmission in China. PLoS One **6:**e16817. doi:10.1371/journal.pone.0016817.
31. **Kwok S, Higuchi R.** 1989. Avoiding false positives with PCR. Nature **339:**237–238.
32. **Guindon S, Gascuel O.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704.
33. **Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A.** 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol. Biol. Evol. **20:**381–387.
34. **Yuan MQ, Lu T, Li C, Lu L.** 2013. The evolutionary rates of HCV estimated with subtype 1a and 1b sequences over the ORF length and in different genomic regions. PLoS One **8:**e64698. doi:10.1371/journal.pone.0064698.
35. **Ji Y, Ren QH, Zhu ZY, Qu DM, Qiu ZK, Li JF, Mei H, Yan HY.** 1998. Investigation and analysis of HCV infection in Chinese blood donors. Acta Acad. Med. Sinica **203:**240–241.
36. **Kaufman J, Jing J.** 2002. China and AIDS—the time to act is now. Science **296:**2339–2340.
37. **Shan H, Wang JX, Ren FR, Zhang YZ, Zhao HY, Gao GJ, Ji Y, Ness PM.** 2002. Blood banking in China. Lancet **360:**1770–1775.
38. **Shi XL, Ren QH, Zhu ZY, Qu DM, Ji Y, Peng DH, Ni SQ.** 1999. Hepatitis C virus infection in blood donors in the People's Republic of China. Transfusion **39:**913.
39. **Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG.** 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol. Biol. **11:**131.
40. **Njouom R, Frost E, Deslandes S, Mamadou-Yaya F, Labbe AC, Pouillot R, Mbélesso P, Mbadingai S, Rousset D, Pépin J.** 2009. Predominance of hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic. J. Gen. Virol. **90:**2452–2456.
41. **Pouillot R, Lachenal G, Pybus OG, Rousset D, Njouom R.** 2008. Variable epidemic histories of hepatitis C virus genotype 2 infection in West Africa and Cameroon. Infect. Genet. Evol. **8:**676–681.
42. **Xia X, Lu L, Tee KK, Zhao W, Wu J, Yu J, Li X, Lin Y, Mukhtar MM, Hagedorn CH, Takebe Y.** 2008. The unique HCV genotype distribution and the discovery of a novel subtype 6u among IDUs co-infected with HIV-1 in Yunnan, China. J. Med. Virol. **80:**1142–1152.