Origin and Evolution of the Unique Hepatitis C Virus Circulating Recombinant Form 2k/1b

Jayna Raghwani, Xiomara V. Thomas, Sylvie M. Koekkoek, Janke Schinkel, Richard Molenkamp, Thijs J. van de Laar, Yutaka Takebe, Yasuhito Tanaka, Masashi Mizokami, Andrew Rambaut and Oliver G. Pybus


Updated information and services can be found at:
http://jvi.asm.org/content/86/4/2212

**SUPPLEMENTAL MATERIAL**

These include:

Supplemental material

**REFERENCES**

This article cites 67 articles, 28 of which can be accessed free at:  http://jvi.asm.org/content/86/4/2212#ref-list-1

**CONTENT ALERTS**

Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article),  more»
Origin and Evolution of the Unique Hepatitis C Virus Circulating Recombinant Form 2k/1b

Jayna Raghwani,a Xiomara V. Thomas,b Sylvie M. Koekkoek,b Janke Schinkel,b Richard Molenkamp,b Thijss J. van de Laar,c Yutaka Takebe,c Yasuhiro Tanaka,c Masashi Mizokami,c Andrew Rambaut,a,b and Oliver G. Pybusb

Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh, United Kingdoma; Academic Medical Center, Department of Medical Microbiology, Section of Clinical Virology, Amsterdam, The Netherlandsb; VU University Medical Centre, Department of Medical Microbiology and Infection Control, Amsterdam, The Netherlandsb; AIDS Research Center, National Institute of Infectious Diseases, Tokyo, Japanb; Department of Virology and Liver Unit, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japanb; The Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Ichikawa, Japanb; Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USAa; and Department of Zoology, University of Oxford, Oxford, United Kingdoma

Since its initial identification in St. Petersburg, Russia, the recombinant hepatitis C virus (HCV) 2k/1b has been isolated from several countries throughout Eurasia. The 2k/1b strain is the only recombinant HCV to have spread widely, raising questions about the epidemiological background in which it first appeared. In order to further understand the circumstances by which HCV recombinants might be formed and spread, we estimated the date of the recombination event that generated the 2k/1b strain using a Bayesian phylogenetic approach. Our study incorporates newly isolated 2k/1b strains from Amsterdam, The Netherlands, and has employed a hierarchical Bayesian framework to combine information from different genomic regions. We estimate that 2k/1b originated sometime between 1923 and 1956, substantially before the first detection of the strain in 1999. The timescale and the geographic spread of 2k/1b suggest that it originated in the former Soviet Union at about the time that the world’s first centralized national blood transfusion and storage service was being established. We also reconstructed the epidemic history of 2k/1b using coalescent theory–based methods, matching patterns previously reported for other epidemic HCV subtypes. This study demonstrates the practicality of jointly estimating dates of recombination from flanking regions of the breakpoint and further illustrates that rare genetic-exchange events can be particularly informative about the underlying epidemiological processes.

Recombination is thought to play a comparatively minor role in shaping the genetic diversity of HCV; however, an increasing number of reports suggests that it is not entirely insignificant in HCV evolution. Most notable of these was the initial discovery of a natural recombinant form of HCV circulating in injecting drug users resident in St. Petersburg, Russia (20). This recombinant, labeled 2k/1b, has a 5’ genome region that is most closely related to subtype 2k and a 3’ genome region that is most closely related to the global epidemic subtype 1b, with a single recombination breakpoint located at genomic position 3175 or 3176 in the NS2 gene (20). Since the discovery of 2k/1b, several other studies have reported both inter- and intragenotypic HCV recombinants in natural populations, although the evidence presented for recombination varies in strength; the weakest studies report only discordant genotyping results between genome regions (which could also result from coinfection), whereas the most convincing studies repeatedly sequence the same recombination breakpoint from independent extractions (thereby excluding the possibility of in vitro genetic...
exchange). Thus far there have been nine descriptions of HCV recombinant forms, although only in six cases have the breakpoints been sequenced (6–8, 19, 28, 29, 42).

Inspection of the recombination breakpoint positions within the HCV genome reveals a difference between inter- and intragenotypic recombinants. Breakpoints in the intrasubtypic recombinants (1a/1c and 1b/1a) are located in the E1/E2 region, while in the intergenotypic recombinants (including 2k/1b), the breakpoints are consistently found in the NS2-NS3 region (8, 19, 28, 29, 39, 42). Interestingly, naturally occurring intergenotypic HCV recombinants have more often than not involved genotype 2 in the 5’ genome region (19, 20, 28, 29, 42). This may reflect some inherent yet unknown biological or ecological properties of this genotype to produce viable recombinant viruses.

Nevertheless, the low rate of discovery of novel recombinant forms suggests that although recombination does occur in HCV, it is an uncommon event, at least in comparison to HIV-1. Indeed, the 2k/1b strain discovered in St. Petersburg is the only known circulating recombinant form (CRF) of HCV and has therefore been designated CRF01_1b2k, following a naming scheme similar to that developed for HIV-1 recombinants (23). A CRF is a recombinant form that is found repeatedly in different patients. Since its discovery, CRF01_1b2k has been isolated from patients in many countries, including Ireland, France, Cyprus, Azerbaijan, Uzbekistan, and Russia (20, 25, 26, 29, 37–39). CRF01_1b2k is the only recombinant strain of HCV to have transmitted widely; therefore, it is important to investigate its genesis and dissemination in order to understand why it might be unique and to evaluate the likelihood that other HCV recombinant forms could increase in prevalence in the future.

Evolutionary analysis of viral genomes using methods based on molecular clocks and coalescent theory has previously proved useful in reconstructing the epidemic history of various HCV strains, including subtypes 1a and 1b worldwide (33, 46) and subtype 1b in Japan (66). Similar analyses of HCV genotype 4 in Egypt have estimated the timescale of the large HCV epidemic in that country and have confirmed its iatrogenic cause (48, 65). Very little is known about the evolutionary history of HCV subtype 2k, most likely because of the lack of sequence data for the strain.

In order to address the lack of information about the epidemiological and transmission history of HCV CRF01_1b2k, we have conducted a comprehensive evolutionary analysis of all available viral genome sequences using a well-established Bayesian framework (11). Since previously published sequence data on CRF01_1b2k are limited, we sought to increase the sample size by isolating and sequencing a panel of 21 new recombinant isolates from patients of Russian origin resident in Amsterdam, The Netherlands.

By combining information from different genomic regions in a single analysis, we provide the first estimate of the date of the recombination event that generated CRF01_1b2k. The date that we obtained considerably predates the discovery of the strain and requires a reevaluation of the circumstances surrounding its origin. This is the first time that a recombination event has been dated for any virus other than HIV-1 (32, 51, 67) or influenza A virus (27). Further, we estimate the CRF’s past rate of transmission and its pattern of global geographic spread. To obtain more precise parameter estimates when dating recombination events, we employed a joint phylogenetic approach that improves on methods previously applied to HIV-1 CRFs (51, 67). The methods introduced here should serve as a model for future phylogenetic investigations of genetic-exchange events in RNA virus populations.

MATERIALS AND METHODS
Identification and sequencing of new HCV 2k/1b isolates from Amsterdam. In the course of a study of HCV-infected patients resident in Amsterdam (unpublished data), it was found that HCV genotyping results from the 5’ untranslated and NS5B regions were discordant for 6 (out of 200) patients. Of these, five were male and one was female, and their mean age was 34 years (Table 1). For this study, the 5’-end sequences were not used.

HCV RNA was isolated from 200 μl plasma using the purification method described by Boom et al. (4). cDNA was generated using random hexamer primers as described before (4). The amplification was performed using a conventional PCR with the following cycling conditions: 2 min at 50°C and 10 min at 95°C, followed by 15 cycles, each consisting of 30 s at 95°C, 30 s at 55°C, and 1 min at 72°C. Amplicons were purified from a 1% agarose gel as described by Boom et al. (4). Amplification of a 724-nucleotide (nt) fragment (unpublished data) of the NS5B region was performed as previously described (40). Amplification of core/E1 was performed in a 25-μl volume using HCV1b/2 F (GGGTGAGRCTCCTG GAG) as the forward primer and HCV1b/2 R (TGCCARCARTANGGY TCAT) (positions 292 to 312; all primer locations are indicated relative to the H77 reference genome) as the reverse primer, using the same amplification conditions mentioned above. Amplicons were also purified from a 1% agarose gel as described by Boom et al. (4). Sequencing was performed using HCV1b/2F seq (CCTCTTACAGCTCTYTTGTCTT; positions 128 to 112) as the forward sequencing primer and HCV1b/2R seq (TGCCCAATCTGTTCGTTGTTTGC) as the reverse sequencing primer.

To confirm that the viral variants were recombinants, a 234-base fragment harboring the known breakpoint in NS5B was also amplified and sequenced. Amplification and sequencing of the NS2 breakpoint were performed using HCV2_F (GGACGCGCATACTTGCTCAG) as the forward primer and HCV1B_R (CAGGGAATGATCTTTGCTCAG) as the reverse primer, also using the same cycling conditions mentioned above.

In addition to new CRF01_1b2k sequences from Amsterdam, 6 recombinant isolates from an unpublished study were also included. These sequences were sampled from IDUs in Azerbaijan (Table 1). Most of the remaining CRF01_1b2k isolates came from a study of IDUs in Uzbekistan (25) and from a cohort study of HCV-positive patients from seven countries (26); detailed demographic information on isolates from these two studies is provided here (Table 1). In the latter study, all individuals infected with CRF01_1b2k came from either Russia or Uzbekistan, and 6 of these patients were linked to high-risk groups, namely, blood transfusion recipients or intravenous drug users.

Collation of HCV sequence alignments. To investigate the evolutionary origin and spread of HCV CRF01_1b2k, a data set comprising all subtype 2k/1b (n = 27), 2k (n = 15), and 1b (n = 71) isolates for which both core/E1 and NS5B sequences were available was collated from GenBank and the HCV Sequence Database (24) (see Table S1 in the supplemental material). This collection included the 6 newly sequenced isolates obtained from HCV patients in Amsterdam (see above). Alignments for both genome regions were constructed manually, and each alignment contained exactly the same set of taxa. The date and sampling location of each sequence were obtained from the literature or via personal communication (Table 1).

Estimation of genome region-specific rates of evolution. The evolutionary rates of the core/E1 and NS5B regions used in this study could not be estimated directly from our data, because the sample size and range of sample dates were not large or wide enough. In line with previous studies of HCV epidemic history (e.g., see reference 48), we employed an independent data set with significant temporal information to provide the
substitution rates of the subgenomic regions of interest. Specifically, we used the data and analysis strategy of a recent study that reported rates of evolution for all HCV genome regions for subtypes 1a and 1b (14). That study utilized an alignment-partition approach, which was implemented in the BEAST program, to estimate region-specific rates (11). We applied a codon-structured nucleotide substitution model (55), an uncorrelated relaxed lognormal molecular clock (10), and a Bayesian skyline coalescent model (12) to both subtype 1a and 1b whole-genome alignments, from which we obtained rate estimates for the precise subgenomic regions of interest. Specifically, we used the data and analysis strategy of a recent study that reported rates of substitution models were estimated for the core/E1 and NS5B regions. The prior distributions for the evolutionary rates for the core/E1 and NS5B regions that contained the CRF isolates, plus all available closely related parental subtype reference sequences (belonging to subtype 2k for the core/E1 region and to subtype 1b for the NS5B region). A hierarchical phylogenetic model (62) was used to combine both data sets and thereby provide a joint estimate of the time to the most recent common ancestor (TMRCA) of the CRF clade, while accounting for uncertainty in both genome regions.

As the CRF clade in the different genome regions is known to represent a common evolutionary history, jointly estimating the age of this clade will maximize the explanatory power of the data (62) and is thus more powerful than analyzing the regions independently, as has been done previously (e.g., see reference 67). However, due to the recombination of the different subtypes, we cannot simply assume that a single tree represents the entire genome. Instead, we allow independent trees for each genomic region but maintain the TMRCA of the CRF clade in each region to lie within a small time of each other, while estimating the mean of these as the parameter of interest.

Specifically, separate phylogenies, molecular clock models, and substitution models were estimated for the core/E1 and NS5B regions. The genome region-specific rates (estimated as described above) were used as prior distributions for the evolutionary rates for the core/E1 and NS5B regions. For the NS5B region, the rates estimated from subtype 1b were available closely related parental subtype reference sequences (belonging to subtype 2k for the core/E1 region and to subtype 1b for the NS5B region). A hierarchical phylogenetic model (62) was used to combine both data sets and thereby provide a joint estimate of the time to the most recent common ancestor (TMRCA) of the CRF clade, while accounting for uncertainty in both genome regions.

As the CRF clade in the different genome regions is known to represent a common evolutionary history, jointly estimating the age of this clade will maximize the explanatory power of the data (62) and is thus more powerful than analyzing the regions independently, as has been done previously (e.g., see reference 67). However, due to the recombination of the different subtypes, we cannot simply assume that a single tree represents the entire genome. Instead, we allow independent trees for each genomic region but maintain the TMRCA of the CRF clade in each region to lie within a small time of each other, while estimating the mean of these as the parameter of interest.

Specifically, separate phylogenies, molecular clock models, and substitution models were estimated for the core/E1 and NS5B regions. The genome region-specific rates (estimated as described above) were used as prior distributions for the evolutionary rates for the core/E1 and NS5B regions. For the NS5B region, the rates estimated from subtype 1b were available closely related parental subtype reference sequences (belonging to subtype 2k for the core/E1 region and to subtype 1b for the NS5B region). A hierarchical phylogenetic model (62) was used to combine both data sets and thereby provide a joint estimate of the time to the most recent common ancestor (TMRCA) of the CRF clade, while accounting for uncertainty in both genome regions.

As the CRF clade in the different genome regions is known to represent a common evolutionary history, jointly estimating the age of this clade will maximize the explanatory power of the data (62) and is thus more powerful than analyzing the regions independently, as has been done previously (e.g., see reference 67). However, due to the recombination of the different subtypes, we cannot simply assume that a single tree represents the entire genome. Instead, we allow independent trees for each genomic region but maintain the TMRCA of the CRF clade in each region to lie within a small time of each other, while estimating the mean of these as the parameter of interest.

Specifically, separate phylogenies, molecular clock models, and substitution models were estimated for the core/E1 and NS5B regions. The genome region-specific rates (estimated as described above) were used as prior distributions for the evolutionary rates for the core/E1 and NS5B regions. For the NS5B region, the rates estimated from subtype 1b were used (because a subtype 2k-specific rate was not available). For each pair of sampled phylogenies (core/E1 and NS5B) in the posterior distribution of the MCMC, three node dates were obtained (as labeled in Fig. 2): A, the joint TMRCA of the CRF clade; B, the date of the parental node of the CRF clade in the core/E1 subtype 2k phylogeny; and C, the date of the parental node of the CRF clade in the NS5B subtype 1b phylogeny. The former date, together with the more recent of the last two
dates, therefore defines a time range during which the recombination event must have occurred (see Fig. 2). The posterior distribution of this time range was then compiled by repeating the above-described procedure for each pair of phylogenies in the MCMC output. The BEAST analysis model settings were the same as those outlined in the section on genome region-specific rates above.

**CRF01_1b2k transmission history.** Further Bayesian MCMC phylogenetic analyses were performed solely on the CRF01_1b2k isolates in order to estimate the epidemic history and basic reproductive number, \( R_0 \), of the strain since its emergence. BEAST model settings were the same as those outlined above, except that different coalescent models were employed to reconstruct the transmission history of the CRF. Both the GMRF skyride (36) and exponential-growth coalescent models were used.

**RESULTS**

**Estimation of genome region-specific rates of evolution.** The rates of evolution of the core/E1 and NS5B genomic regions used in this study were estimated from subtype 1a and 1b whole-genome data sets (see Materials and Methods) and are given in Table 2. The rates for the NS5B region corresponded between subtypes, with similar 95% highest posterior density (HPD) intervals, while the rate estimates for the core/E1 region did show some variation among subtypes (Table 2).

**Phylogenetic analysis.** To establish the evolutionary origins of the HCV 2k/1b strain, we analyzed the core/E1 (644 nt) and NS5B (741 nt) regions of 27 CRF 1b/2k, 15 subtype 2k, and 71 subtype 1b isolates. The 2k/1b isolates were sampled between 1999 and 2007 and were from the following locations: Ireland, Uzbekistan, Azerbaijan, Cyprus, Amsterdam, France, and Russia (Table 1). Since the BF test supported the hypothesis that the 2k/1b isolates were monophyletic, a single origin of the CRF was inferred (BFs for the comparisons of monophyly and nonmonophyly models were 0.31 for the core/E1 data set and −0.68 for the NS5B data set). Furthermore, the monophyletic origin of the CRF clade was supported in both genome regions when no phylogenetic constraints were imposed. However, monophyly of CRF 1b/2k isolates was not supported by a high posterior probability (0.67) in the NS5B maximum clade credibility (MCC) tree, most likely reflecting the uncertainty associated with the star-like phylogeny and relatively short sequence length. However, the recombinant nature of many of the isolates in this study was confirmed by direct observation of the breakpoint in the NS2 gene region (confirmed isolates are represented by filled circles in Fig. 1).

**Molecular clock analysis.** Figure 1 and Table 3 provide the estimated TMRCAs of the CRF clade obtained using the joint and independent molecular clock analyses. The estimates obtained separately from the core/E1 and NS5B data sets were in close agreement and exhibit overlapping 95% HPD intervals. These estimates also agree with the joint estimate of TMRCA of the CRF clade and its most closely related parental isolate (node A, which was 1946 (1932 to 1959; Table 3). We also estimated the date of the most recent common ancestor of the CRF clade and its most closely related parental isolate (Fig. 2). These estimates were again similar for both genome regions, at about 1933 (Table 3). Lastly, by comparing the HPDs of the node A date with the dates of the more recent of the two parental nodes (either B or C), we were able to generate bounds for the time of the

<table>
<thead>
<tr>
<th>Genomic region</th>
<th>No. of substitutions/site/yr (10^{-3})a</th>
<th>HCV subtype 1a</th>
<th>HCV subtype 1b</th>
<th>Average rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core/E1</td>
<td>1.75 (1.41, 2.10)</td>
<td>1.36 (1.01, 1.76)</td>
<td>1.56 (1.21, 1.93)</td>
<td></td>
</tr>
<tr>
<td>NS5B</td>
<td>0.89 (0.71, 1.07)</td>
<td>0.91 (0.68, 1.14)</td>
<td>0.90 (0.70, 1.11)</td>
<td></td>
</tr>
</tbody>
</table>

a The estimates were obtained from HCV subtype 1a and 1b whole genomes using a genomic partition model (see Materials and Methods). The numbers inside parentheses represents the 95% highest posterior density credibility interval.

**FIG 1** Molecular clock phylogenies of CRF01_1b2k and its parental subtypes, estimated from the core/E1 alignment and the NS5B alignment. The horizontal bars in each phylogeny contain the dating estimates for two nodes: the common ancestor of the CRF clade (1931 or 1932 to 1958) and the common ancestor of the CRF and the most closely related parental strain (1910 to 1912 to 1952 or 1953). Filled circles, 2k/1b isolates that were confirmed as being recombinant by sequencing of the breakpoint in the NS2 region; open circles, those isolates for which NS2 sequences were not available. The trees shown are the maximum clade credibility trees from the Bayesian MCMC analysis.
recombination event that generated the CRF, which was between 1923 and 1956.

**CRF01_1b2k transmission history.** Figure 3 shows the estimated epidemic history of CRF01_1b2k, as estimated using the Bayesian skyride plot method, which depicts the effective population size of the CRF epidemic over time. The plot indicates approximately constant exponential growth since the emergence of the CRF lineage (Fig. 3) until the mid-1990s, after which the effective population size declines or stabilizes (either is plausible, given the size of the credible region of the estimate). This decrease/stabilization coincides with the advent of screening for HCV in blood donors, which greatly reduced the risk of HCV infection via blood transfusion (31, 52, 70).

To ascertain the CRF’s exponential growth rate ($r$), the CRF data set was also analyzed using an exponential-growth coalescent model. The estimated growth rate was 0.116 year$^{-1}$ (95% HPD interval, 0.079 to 0.159). This estimate was subsequently used to calculate $R_0$ values for the CRF01_1b2k strain under a plausible range of average durations of infectiousness ($D$), using the equation $R_0 = rD + 1$ (46). Both the estimated growth rate ($r = 0.1$ year$^{-1}$) and the estimated $R_0$ values ($R_0 = 2$ to 4) are compatible with a number of equivalent estimates for other HCV subtypes, including those from IDU risk groups (46–48, 66).

The MCC tree of the CRF01_1b2k clade (Fig. 4), when combined with all available epidemiological information (Table 1), indicates a clear pattern of phylogenetic clustering according to the geographic location and risk factor of each patient. For example, where these details were available, 14 out of 15 isolates were associated with IDU, while only 1 was isolated from a patient with a history of blood transfusion (Table 1). These observations further support the view that CRF01_1b2k transmission is strongly linked with the IDU transmission route. Although it is not possible to reconstruct the location of the common ancestor of the CRF lineage with any certainty (because our sample size is small and the basal branches of the phylogeny are poorly supported), all of the isolates included in this study are from or have an epidemiological link to the former Soviet Union, and the oldest strain was sampled in Russia in 1999. A well-supported cluster of strains from Azerbaijan originated in about 1970, and therefore, the CRF has likely circulated there since that time. Hence, it appears that CRF01_1b2k disseminated throughout the Soviet Union before its dissolution in the late 1980s and for some time prior to the discovery of the CRF in St. Petersburg in 1999.

**DISCUSSION**

As the only known HCV recombinant in widespread circulation, the existence and emergence of CRF01_1b2k present an interesting question in HCV epidemiology and evolution. Investigating its evolutionary origins and transmission history helps to understand the circumstances that led to its unique properties. In contrast to HIV, which has 49 known CRFs and a much greater number of unique recombinant forms (30), recombination typically contributes little to the generation and maintenance of HCV genetic diversity. Given that HCV has a higher global prevalence than HIV and, thus, all else being equal, there is a high likelihood of dual infections with divergent HCV strains, it is unlikely that epidemiological factors are

### TABLE 3 Estimates of TMRCA of the CRF clade obtained from separate genome regions and from joint (hierarchical) phylogenetic analysis

<table>
<thead>
<tr>
<th>Clade (node$^a$)</th>
<th>TMRCA (yr)$^b$</th>
<th>NS5B</th>
<th>Joint (hierarchical) estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Core/E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF (A)</td>
<td>1945.2 (1931.6, 1958.3)</td>
<td></td>
<td>1946.0 (1932.5, 1959.0)</td>
</tr>
<tr>
<td>CRF + 2k (B)</td>
<td>1932.0 (1909.7, 1952.8)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CRF + 1b (C)</td>
<td>NA</td>
<td>1933.1 (1912.0, 1952.1)</td>
<td>NA</td>
</tr>
</tbody>
</table>

$^a$ The node labels highlighted in Fig. 2.

$^b$ The numbers inside parentheses represent the 95% highest posterior density credibility interval. NA, not available.
restricting the opportunities for HCV to generate CRFs. Mixed infections with divergent HCV strains have been reported for many different populations and are noted to be prevalent among high-risk groups, particularly IDUs and some hemophiliacs (3, 5, 16, 50, 53).

Since the opportunities for HCV recombination are not limited, it is more likely that fundamental molecular and evolutionary differences between HIV and HCV explain why HIV has many CRFs and HCV has few. These could include differences in the rate of template switching or differences in genomic or immunological constraints, such that HCV recombinants have, on average, lower fitness than HIV recombinants and therefore are rarely transmitted (72). Although both viruses are associated with chronic infections, unlike HIV, HCV can be spontaneously cleared by the host. This may explain in part the differences in the number of recombinants between HIV and HCV, where partial protective immunity against the latter reduces that chance of in vivo recombination of HCV strains (1, 43, 69). However, the high rate of mixed infections observed suggests that this is likely at best to play a minor role in HCV recombination. The low frequency of HCV recombinants is more likely to reflect mechanistic constraints on viral replication. There is evidence that template switching in HCV is
especially rare and that the replication complex is typically encoded on the same genomic strand that it will replicate and transcribe (2). It is also interesting that when replication complexes are exchanged between different genotypes, the replication efficiency is substantially reduced (15). The pseudodiploidy of the HIV genome certainly increases the likelihood of recombination occurring due to the ability of the virus to package two RNA templates (17), while the secondary RNA structure in the HCV genome may limit the production of viable hybrid HCVs (59, 68).

Our study of previously reported and newly obtained HCV isolates provides the first estimates of the date of the recombination event that generated CRF01_1b2k. We estimated the time of origin of CRF01_1b2k to be between 1923 and 1956, which is not much later than the origin and global spread of the parental subtype 1b (33). This date is significantly earlier than we expected: we expected that the CRF’s creation might be linked to the dramatic increase in IDU behavior following the breakup of the former Soviet Union. This result is robust to the manner of isolate sampling: if, because of nonrandom sampling, our isolates are more closely related to each other than under random sampling, then the TMRCA of CRF01_1b2k would be biased toward more recent dates. Furthermore, despite its small size, our data set provides a relatively short time window during which the recombinant must have arisen (Fig. 2). The involvement of subtype 1b in the recombinant is not surprising, as it is one of the most prevalent subtypes worldwide. However, to fully appreciate the origin of CRF01_1b2k, we need to consider the evolutionary history of both parental subtypes and of the recombinant lineage itself. Genotype 2 harbors considerable genetic diversity, especially in West Africa, which is where the genotype is thought to have originated (34). Although the small number of subtype 2k isolates sampled to date likely underestimates the true extent of subtype 2k distribution, such viruses have been isolated from Martinique and Madagascar, implicating a role for the historical trans-Atlantic slave trade in the dissemination of the virus from West Africa (34).

The current distribution of subtype 2k is associated with francophone regions and former Soviet Union countries. In contrast, CRF01_1b2k is more spatially limited, with all isolates being directly or indirectly linked to the former Soviet Union. As the nonrecombinant subtype 2k isolates that are most closely related to CRF01_1b2k are from Moldova and Azerbaijan (see Fig. S1 in the supplemental material), it seems most likely that CRF01_1b2k was generated in the former Soviet Union. An equivalent analysis of subtype 1b viruses provides no reliable phylogeographic linkage, due to the low phylogenetic resolution of the NS5B data set.

Our estimated date of CRF origin coincides with an interesting period of history in the former Soviet Union, which was an early leader in transfusion technology. Under the leadership of Alexander Bogdanov in the 1920s, a nationwide network of blood transfusion centers and research institutes, as well as the Central Institute of Hematology in Moscow, Russia, in 1926, was established throughout the Soviet republics (61). This expanded into a network of ~1,500 blood donating centers across the republics (18). The Soviets also adopted blood storage and preservation techniques at an early stage. They established more than 60 primary and 500 subsidiary blood storage centers by the mid-1930s, which shipped blood across the entire Soviet Union (61). During the Second World War, these networks were swiftly readapted to support the front line; in Moscow alone, about 2,000 blood donations were given per day (18, 61). The impressive scale of the blood service in the former Soviet Union is likely to have favored HCV transmission by increasing the efficiency and geographic range of the virus’s dissemination. Whether specific medical practices at this time increased the probability of mixed viral infections remains unknown. It is interesting to note that Bogdanov himself was fascinated by the ideological interpretation of blood sharing and frequently practiced what he called “physiological collectivism”: the exchange of blood with others through mutual transfusions (61).

Although unscreened blood transfusions can provide a credible hypothesis for the origin of CRF01_1b2k in the Soviet Union some time from 1923 to 1956, we must also attempt to explain how subtype 2k or the CRF itself arrived in the Soviet Union from West Africa or the Caribbean. Migration from Africa to the former Soviet Union did occur during the late 1950s and 1970s as a result of alliances forged by the Soviet government with newly independent African states such as Ghana and Angola (35). However, these connections are too late to have contributed to the emergence of CRF01_1b2k, according to our dating estimates. Although we cannot reject the hypothesis that the CRF was formed in West Africa and subsequently moved to the Soviet Union, our results are more consistent with the recombination event occurring in the latter. This uncertainty is likely to be reduced with further samples, especially subtype 2k viruses, from African and former Soviet Union locations.

The epidemic history CRF01_1b2k (Fig. 3) since its emergence is similar to that estimated for other epidemic subtypes of HCV (e.g., see reference 33). The growth in the CRF01_1b2k effective population sizes coincides with a substantial increase in blood transfusion, including during the Second World War, and with the subsequent rise in intravenous drug usage. CRF01_1b2k transmission seems to have slowed or stabilized since the early 1990s, coinciding with the onset of the anti-HCV screening of donors. In the absence of any data to the contrary, the transmission of this recombinant and its spread from the former Soviet Union reflect the peculiar epidemiological properties of the risk groups that it has been associated with rather than any intrinsic properties of the virus.

We demonstrate the practicality and benefits of using a hierarchical phylogenetic model to jointly estimate parameters of interest when analyzing multipartite sequence data that result from genetic exchange. This method yields more accurate parameter estimates than previous methods (e.g., see references 27 and 67) by incorporating the phylogenetic information and uncertainty in different genomic regions. We recommend that this improved statistical framework be used in future investigations of recombination in fast-evolving RNA viruses.

This study has made significant steps in understanding the epidemic history and spread of the unique circulating HCV recombinant 2k/1b. Most significantly, we show that this strain originated many decades before the post-Soviet rise in injection behavior with which it is currently associated. On the basis of the date of its origin and its molecular epidemiology, there are reasonable grounds to suppose that the Soviet Union’s revolutionary blood service was instrumental in the CRF’s early generation and continental-scale spread. This infrastructure may have facilitated the pan-Eurasian spread of other parenterally transmitted blood-borne infections, and this is an interesting question for future research.
REFERENCES