Hepatitis C virus subtype 3a was introduced in the USSR in the early 1980s

Victor A. Manuylov,1,2,4 Elena V. Chub,3 Vera S. Kichatova,4 Natalya V. Soboleva,4 Olga V. Isaeva,4 Andrey A. Zamyatnin Jr1,5 and Sergey V. Netesov3,6

Abstract
A total of 2120 nucleotide sequences of the NS5b region of HCV subtype 3a were analysed, including 310 strains derived from former republics of the USSR (Azerbaijan, Estonia, Lithuania, Russia, Tajikistan and Uzbekistan). Among the viral isolates collected from former regions of the Soviet Union, 294 strains formed 3 sustained phylogenetic clusters, with each having a common origin. Phylogenetic analysis demonstrated that the most recent common ancestors of the current strains inside the three clusters were introduced into the USSR population in 1981±1, 1984±2 and 1985±2, respectively (the confidence intervals were calculated using Student’s t-distribution, P<0.05). The time estimation obtained for HCV subtype 3a correlated well with the historical and epidemiological context of this period, and in particular with the start of widespread injection drug use in the USSR in the first half of the 1980s.

INTRODUCTION
Hepatitis C virus (HCV) is a highly variable RNA virus of the genus Hepacivirus in the family Flaviviridae. The heterogeneity of the HCV genome, which has a length about 9600 nucleotides, varies depending on the genomic region. For example, the long-term mutation rate for HCV genome region NS5b, which codes the viral polymerase, is estimated to be 1.0–1.65×10⁻³ nucleotide substitutions per site per year [1–3], whereas the core region that codes the viral capsid protein is more conservative, with a mutation rate of about 6–10×10⁻⁴ substitutions per site per year [1, 4, 5].

The HCV genome is highly heterogeneous, and the identified viral isolates are classified into seven genotypes and several dozen subtypes (1a, 1b, etc.) [6]. The genetic homology of the isolates of different HCV genotypes is as low as 65–70 %, while the homology of the isolates of different subtypes inside the same genotype is 75–85 % [7].

The HCV subtypes may demonstrate different clinical and epidemiological traits (as reviewed in [8]). In particular, subtypes 1b and 3a, the most common HCV subtypes worldwide, differ in the clinical outcome of the infection (e.g. subtype 1b is more resistant to interferon therapy), as well as in the major means of infection transmission. The transmission route of subtype 1b is generally related to nosocomial risk factors, while subtype 3a is mainly observed among injection drug users (IDUs) [9, 10]. As a result, the viral subpopulations of subtypes 1b and 3a can be studied independently, even if they circulate in the same local (geographical) groups of human populations. The present study aimed to investigate the time of the introduction of HCV subtype 3a into the population of the former USSR and its evolutionary history.

RESULTS
A total of 2120 nucleotide sequences of the NS5b region of HCV subtype 3a were analysed (Table 1, Fig. 1). Among them, 60 sequences were studied in this work for the first time (GenBank accession numbers: KY552746–KY552773, KY548481–KY548487, KY548489–KY548493, KY511562–KY511580 and KY511582, Table 1), and 2060 sequences were retrieved from the HCV database (http://hcv.lanl.gov, [11]). The lengths of the sequences ranged from 234 to 403 nucleotides. Among the entire collection, 310 strains were sampled in the countries of the former Soviet Union (Table 1), while the others were derived from all over the world. Most of the strains from former regions of the USSR
were collected in the Russian Federation in the period 1999–2013. In the analysis, strains from Azerbaijan, Estonia, Lithuania, Tajikistan and Uzbekistan were also used. Table 1 presents brief information concerning the strains, including the country of origin, year of collection and corresponding reference data.

Fig. 1. The overview of the maximum likelihood (ML) tree based on the subtype 3a NS5b sequences (topology only). The locations of the clusters formed by sequences derived from the former USSR are marked by rectangles.
It has been found that according to any basic phylogenetic analysis regardless of the method used [maximum likelihood (ML), unweighted pair group method with arithmetic mean (UPGMA) and neighbour joining (NJ)], most of NS5b sequences collected from former countries of the USSR formed three well-defined clusters inside the clade of subtype 3a. As an example, the overview of the ML tree based on all of the 2120 NS5b sequences of HCV subtype 3a (and one sequence of subtype 1b as an outgroup) with the location of the clusters is depicted in Fig. 1, but both UPGMA and NJ trees (not shown) demonstrate a similar topology, with the same clustering of the former USSR strains into three sub-clades. This fact suggests that the current viral population of HCV subtype 3a circulating in former countries of the Soviet Union is derived from at least three common sources of infection introduced into the population of this territory in the past. Sequence reconstruction of the most recent common ancestors (MRCAs) in the clusters allows one to apply phylodynamic methods and estimate the time of their emergence in the USSR.

At the same time, none of the clusters found received significant statistical support in the bootstrap test [12] with 1000 replicas in any of the generated basic trees (ML, NJ and UPGMA). The Bayesian inference test [13] that was implemented together with the ML analysis also appeared to be spurious, and showed no statistical support for the former USSR clusters, although the clusters themselves could still

Table 1. Data for the studied strains derived from the former USSR and some other countries

<table>
<thead>
<tr>
<th>Country</th>
<th>GenBank accession numbers and references*</th>
<th>Brief information concerning the groups that were sources for the HCV strains in the study</th>
<th>Number of strains within the clusters</th>
<th>Number of strains out of the clusters</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azerbaijan (AZ)</td>
<td>FJ4355xx, Earhart et al., 2008 (unpublished)</td>
<td>Eight cities (Sumgayit, Goygol, Baku, Jaliabad, Lankaran, Massaly, Siazan and Agcabedi), 2000–2001; IDUs only†.</td>
<td>28</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>Estonia (EE)</td>
<td>EF1950xx, Tallinn, 1994–2004: hospitalized patients, blood donors and health-care workers.</td>
<td></td>
<td>38</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>Lithuania (LT)</td>
<td>DQ7902xx, Vilnius: patients of paediatric oncohaematological department. The strains were sampled at the beginning of 2000s, but the patients were infected in 1990s.</td>
<td></td>
<td>20</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>Russia (RU)</td>
<td>AF380xxx, AF506xxx, AJ5072xx, Shustov et al., 2005*</td>
<td>Novosibirsk (western Siberia), 2001–2002: health-care workers, patients from a hospital for drug users, patients from an AIDS centre and outpatients from a local general practice clinic.</td>
<td>186</td>
<td>11</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moscow, 2014: IDUs only.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tajikistan (TJ)</td>
<td>AB330xx, Dushanbe, 2006: clinical patients with chronic hepatitis and/or liver cirrhosis/ hepatocellular carcinoma. 2002–2004: five groups of patients: healthy donors, IDUs, tuberculosis patients, hematological patients and hepatitis patients.</td>
<td></td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>China (CN)</td>
<td>HQ3188xx, Zhenjiang city (Jiangsu province), 2009: IDUs only.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyprus (CY)</td>
<td>EU6846xx, GQ3325xx, [41]</td>
<td>2008: IDUs only.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*For some sequences indexed in GenBank the authors indicated a reference with an ‘unpublished’ status. In such cases, we used another published reference that contained the information about the particular sequences (if available).
†IDU, injection (intravenous) drug user. In the table, IDUs was marked in bold if the particular groups that were studied by the authors of the sequences consisted solely of IDUs.
‡Numbers of strains are only shown only for groups from former USSR countries.
Fig. 2. Scaled ML tree of the 447 HCV 3a strains derived from the former USSR as well as from other countries (see text for commentary). Countries’ designations are given according to the international abbreviations: AZ, Azerbaijan; EE, Estonia; LT, Lithuania; RU, Russia; TJ, Tajikistan; UZ, Uzbekistan; CN, China, etc. (see Table 1). Bold branches lead to root nodes that had high statistical support (>90%) in the aLRT-SH test. Grey arrows point to the nodes of MRCAs of strains that formed clusters I, II and III (see Fig. 1). The strains sampled from known IDUs are marked in grey.

The MRCAs of the strains in the three clusters we identified were reconstructed using a consensus approach from the sequences forming the corresponding clusters. The genetic distances between the MRCAs and the current strains with a known year of sampling were determined. Using an HCV NS5b long-term mutation rate value of $1.3 \times 10^{-3}$ substitutions per site per year that has been reported in the literature [2], it was found that the MRCAs for clusters I, II and III emerged in 1985±2, 1984±2 and 1981±1, respectively (the time confidence intervals were calculated using Student’s t-distribution, $P<0.05$). Thus, it can be concluded that HCV subtype 3a was introduced into the population of the former USSR at least three times in the first half of the 1980s.

Among the other HCV sequences found in the strains from the former USSR, such as the sequences belonging to the subtype 3a core region, or the subtype 1b core and NS5b regions, no clusters were identified (the trees are not shown). One possible reason for this is the small number of sequences in the analysis: only 33 core 3a sequences from the former USSR were available in the international databases, which is insufficient for reliable phylogenetic clustering. On the other hand, for subtype 1b, the number of NS5b sequences from the former USSR was relatively large (229 NS5b sequences indexed in the database), but they did not form clusters and were distributed throughout the tree without any distinct patterns. This probably resulted from a
widespread and long-term distribution of subtype 1b in the global human population (as reviewed in [8]).

**DISCUSSION**

It is well known that the global dissemination of HCV subtype 3a infection is closely related to transmission by means of injection drug use [9, 10], and Russia and other former USSR regions are not an exception [17–19]. In the 1970s and up to the early 1980s, intravenous drug use was not widespread in the Soviet Union. Importantly, the USSR was a state with tough regulation of trans-border migration. As a result, the distribution of various HCV subtypes in the USSR was supposedly quite dissimilar from the rest of the world’s. The situation began to change in the mid-1980s. Indeed, in 1971 the Soviet government officially declared that drug addiction was not a serious social and health problem in the USSR [20]. But in 1986, the CIA reported that the USSR was facing the problem of illicit drugs [21], and in 1990 the severity of this problem was officially recognized by the Soviet government [22]. The availability of opium and then heroin for injection in the Soviet Union increased with the beginning of the Soviet–Afghan War in 1979 [23, 24]. Thus, the estimated time for the emergence and rapid distribution of HCV subtype 3a in the USSR in the mid-1980s is closely related to the historical context, and most probably this subtype was not endemic for the Soviet Union population before the 1980s. Interestingly, the exponential epidemic growth of HCV subtype 3a infection among injection drug users in the Western Europe occurred much earlier, in the 1940s [25]. Therefore, it took about 40 years for this type of virus to penetrate across the ‘iron curtain’ of the Soviet Union’s border.

It is significant that most of HCV 3a strains analysed in our study were sampled from non-IDU patients. Strains derived from exactly known IDUs are shown in Table 1 and in Figs 2, 3 and 4. For example, among 310 strains from former countries of the USSR, only 70 were collected during special studies of drug-user groups (Table 1). As can be seen from Figs 2, 3 and 4, the strains taken from IDUs do not form any statistically supported distinct sub-clusters, and are intermixed with non-IDU strains (although the smallest cluster, cluster I, includes half of all the strains from IDUs). This means that subtype 3a, from the time of its introduction into the USSR in the 1980s up to the time when the strains were sampled in the 2000s, became a widely distributed HCV subtype, which now does not only affect IDUs, as had previously been thought [17–19], but also low-risk groups of the population.

Some of the studied strains from the former USSR (16 out of 310, Table 1) do not join any of the clusters described in Figs 2, 3 and 4. A reasonable explanation for this would be that the ancestors of these strains were introduced into the population of the former Soviet Union after the USSR collapsed in 1991. A continuous transmission of HCV from Western Europe to the east was shown by authors from Lithuania [26], who reported that there have been multiple introductions of HCV into the Lithuanian population (some of these introductions occurred before 1995, while others occurred after 1994). In the tree presented in Figs 2, 20 out of 24 Lithuanian strains (Table 1) joined the former USSR clusters (Fig. 2), and the other 4 formed a small unsupported cluster (close to the root of the whole tree) positioned near the sequences from other countries. Most likely this continuous trans-border transmission of HCV still occurs in most of the former Soviet republics [18, 27]. Now the general viral population of subtype 3a at this territory consists of strains that originated from a viral ancestor, which was first introduced in the 1980s, and the strains that were transmitted later from Western Europe in the 1990s.

The clusters of former USSR strains also include a number of strains from other countries (Fig. 2, Table 1). A possible explanation for this is that the origins of these strains were related to the Soviet population in some way. Indeed, the clusters contain the USSR strains along with strains from China, Cyprus, Egypt, the United Kingdom, Greece, Spain, Thailand and other countries (Table 1) that are traditionally very popular among Russian migrants and tourists. For example, 13 strains from China form a sub-cluster that can be seen at the bottom of cluster III (Fig. 2). These Chinese strains (HQ3188xx 28; Table 1) were sampled in Jiangsu Province in eastern China, but a group of Chinese authors [28] used phylogeographical methods to show that the subtype 3a strains from this region originated from the Xinjiang Uyghur Autonomous Region, which has common borders with Russia, Kazakhstan, Tajikistan, Kyrgyzstan and Afghanistan. There is evidence for active historical and contemporary migrations of populations from these countries into and out of this Chinese region. Another example is the sequences from Cyprus that can be found within the former USSR clusters (Table 1). There is now a large Russian diaspora in Cyprus, and it was shown [29] that HCV variants that are specific for Russia and other countries from the former USSR, e.g. recombinant form 2k/1b [30–33], are found to be circulating in the population of Cyprus.

Finally, another important conclusion should be drawn. To date, multiple models for HCV molecular clocks have been suggested. They mainly differ in the mutation rate across the HCV genome, which depends on the region, genotype and scale of analysis, if within- or between-host evolution of HCV is considered [1]. The long-term mutation speed of variable HCV genome regions (except highly conservative untranslated regions) can differ by two orders of magnitude according to different models, from $10^{-3}$ to $10^{-4}$ substitutions per site per year and even lower. In the present study, we used a rate of $1.3 \times 10^{-3}$, which was previously reported for the NS5b region of subtype 3a [2]. The same value was shown for the NS5b region in HCV subtypes 1a, 1b and genotype 4 [4, 5]. At the same time, somewhat different values of this parameter have been described in the literature. Thus, rates of $1.65 \times 10^{-3}$ and $1.0 \times 10^{-3}$ were reported for the NS5b regions of subtype 3a [3] and genotype 1 [1],
respectively. The fact that the period of the initial introduction of subtype 3a into the USSR in the early 1980s, as calculated in our work, correlates well with the historical and epidemiological context, is an important result for the evolutionary aspects of HCV. In particular, it confirms the correctness of the long-term mutation rate value of $1.3 \times 10^{-3}$ for the NS5b region of subtype 3a. This value can be used for further analysis of the history of HCV transmission in social and geographical groups where injection drug use and consequently HCV subtype 3a circulation have been common. The obtained results are important for understanding the global evolutionary history of HCV in the 20th century.

**METHODS**

**Sequences**

A total of 2120 genomic hepatitis C virus (HCV) cDNA sequences were used in analysis; 2060 sequences were retrieved from the HCV database (http://hcv.lanl.gov, [11]) and 60 sequences were originally obtained in Russia from 2008–2013 and had not previously been submitted to any international database (GenBank accession numbers: KY552746–KY552773, KY548481–KY548493 and KY511 562–KY511582; Table 1). All of these sequences were classified as members of HCV subtype 3a containing a part of NS5b genomic region of HCV. Among the strains, 310 sequences were derived from former republics of the USSR: 29 from Azerbaijan, 38 from Estonia, 24 from Lithuania, 197 from Russia, 3 from Tajikistan and 19 from Uzbekistan (Table 1). For each sequence, the year of sampling was known (Fig. 2). The sequences were aligned using MEGA v. 6 software [34] and trimmed to remove the cross-missed nucleotide sites at the side parts. The final length of the analysed sequences was 251 nucleotides (positions 8316–8566 according to the H77 HCV prototype isolate, AF011753 [35]). For additional analysis, the sequences from the above-mentioned HCV database were used: 4921 sequences of subtype 1b, core HCV region; 4615 sequences of subtype 1b, NS5b region; and 1080 sequences of subtype 3a, core region.

**Phylogenetic analysis**

Phylogenetic analysis of the sequences was performed using the ML method with the general time-reversible (GTR) model and gamma-distributed rates among sites with four categories, as implemented in MEGA v. 6 software [34]. N and UPGMA algorithms were used as implemented in the Phylip v. 3.695 package [36]. For these algorithms, the distance matrixes were calculated using the Kimura 2-parameter method and the transition/transversion ratio 4.0 when sequences of the single HCV subtype were used [37].

The bootstrap analysis [12] was performed using 1000 replicas for each data set. The Bayesian inference test was performed using BEAST v. 1.8 software [13] with the following parameters: SRD06 substitution model [38] and constant population size model with relaxed lognormal molecular clock model. Analysis was completed for over 450 million generations. Tracer v. 1.6 software was used to check the

---

**Fig. 3.** The scaled ML tree of the studied 3a strains (continued from Fig. 2). The designations are the same as in Fig. 2.
convergence, and the effective sample size (ESS) was >200. An approximate likelihood ratio test for branches based on an aLRT-SH procedure was performed using PhyML 3.0 software [15] with the GTR substitution model.

Molecular clocks

The MRCAs of the isolates in the phylogenetic clusters were reconstructed from the sequences forming the corresponding clusters using a consensus approach. The consensus reconstruction algorithm took into account the influence of the sequence sampling date as well as the probability of transition and transversion substitutions. The genetic distances were counted as the ratio of the number of substitutions between the studied sequence and its MRCA to a length of the sequence. For the timing calculation, a long-term mutation rate of $1.3 \times 10^{-3}$ substitutions per site per year was used for the NS5b region of HCV [2]. To estimate the date of MRCA introduction, the time period (in years) calculated from the genetic distances between the consensus MRCA sequence and the studied current sequence was subtracted from the year of the sequence sampling.

Statistics

The confidence intervals of the mean dates for the MRCAs were calculated using Student’s $t$-distribution and a significance level of $P<0.05$. 

---

**Fig. 4.** The scaled ML tree of the studied 3a strains (continued from Figs 2 and 3). The designations are the same as in Fig. 2.
Funding information
The author(s) received no specific grant from any funding agency.

Acknowledgements
The authors are thankful to Helene Norder, Mikhail Mikhailov, Karen Kyuregyan and Anastasia Karlsen for their professional support and assistance during data analysis and discussion of the results.

Conflicts of interest
The author(s) declare that there are no conflicts of interest.

Ethical statement
The authors confirm that this study and the paper complies with COPE standards and ethics.

References
33. Chub EV, Kochneva GV, Granirov VM, Netesov SV. Type 2k/1b recombinants of hepatitis C virus among the population of Altai region. Infect Dis 2007;5:5–11.

---

**Five reasons to publish your next article with a Microbiology Society journal**

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as ‘excellent’ or ‘very good’.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.