Archives of Virology (2022) 167:591–596
https://doi.org/10.1007/s00705-021-05358-7

BRIEF REPORT

Complete genome sequencing and evolutionary analysis of hepatitis C virus subtype 6a, including strains from Guangdong Province, China

Ru Xu[1,2] · Hao Wang[1,2] · Jieting Huang[1,2] · Min Wang[1,2] · Qiao Liao[1,2] · Zhengang Shan[1,2] · Huishan Zhong[1,2] · Xia Rong[1,2,3] · Yongshui Fu[1,2,3,4]

Received: 9 July 2021 / Accepted: 26 November 2021 / Published online: 20 January 2022

Abstract

We performed an evolutionary analysis using whole genome sequence isolates of hepatitis C virus (HCV) 6a from Guangdong Province and reference sequences from various countries. Less than 5% of the HCV genome was found to be under positive selection. The E1 and E2 proteins had the highest proportion of positively selected sites both within and outside of CD8 T cell epitopes in all of the strains. Regions corresponding to CD8 T cell epitopes were under negative selection except in the isolates from Guangdong. Furthermore, we found evidence of three introductions of the virus into Guangdong from Vietnam and other Southeast Asian countries. Thus, this study provides information about the transmission of HCV 6a by comparison of full-length sequences, indicating the impact of selective constraints in Guangdong and across China.

Introduction

Hepatitis C virus (HCV) is divided into eight genotypes (1–8) and subdivided into more than 90 subtypes, which are named in alphabetical order (https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/634/table-1). Of them, genotype 6 (gt6) is the most divergent genotype and is predominantly distributed in South China and other Southeast Asian countries [1–3]. HCV 6a, the first subtype of gt6 discovered, is endemic in Vietnam, Hong Kong, and South China [3, 4], especially in Guangdong Province. In the southern part of China, Guangdong is the most populous province, with 126 million people. Approximately 32 million of these are migrants who work in business or are temporarily hired as laborers (http://www.stats.gov.cn/tjsj/pcsj/). Guangdong Province also has a large population of drug users, which account for the highest proportion in China. Unsafe injection practices have posed an increasing risk for HCV infection in Guangdong in recent years. Collectively, these factors have created an environment that has allowed an increase in HCV transmission in Guangdong and elsewhere in China. Our recent study showed that HCV 6a is the most prevalent subtype in Guangdong Province and has become more prevalent in other provinces [5], indicating that it is being transmitted rapidly all over China. It has been shown that subtype 6a has spread rapidly across China [6].

It has been reported that transmission is an instrumental force driving HCV evolution [7]. When the virus infects a novel host, selection pressure on the viral genome drives the virus to mutate to escape the host immune system [8]. Therefore, positive selection analysis can be used to identify specific sites involved in immune escape. Comparison of whole genome sequences can maximize the evolutionary resolution [9] and help to identify positively selected sites in the genome. The aim in this study was to investigate the evolutionary history of HCV 6a, obtain a detailed view of the mechanism of selection, and understand how this may differentially affect variants of this virus in Guangdong and elsewhere in China.

Handling Editor: Ioly Kotta-Loizou.

Ru Xu and Hao Wang should be considered joint first authors.

✉ Xia Rong
joyjoy@126.com

✉ Yongshui Fu
fuyongshui@gzbc.org

[1] Institute of Clinical Blood Transfusion, Guangzhou Blood Center, 31 Lu yuan Rd, Guangzhou, Guangdong, China

[2] The Key Medical Laboratory of Guangzhou, Guangzhou, Guangdong, China

[3] School of Laboratory Medicine and Biotechnology, Southern Medical University, Guangzhou, Guangdong, China

[4] Zhujiang Hospital of Southern Medical University, Guangzhou, Guangdong, China

Materials and methods

Previously, the E1 genotypes of 100 HCV 6a isolates from a cohort of blood donors (BDs) were determined, revealing that 6a is predominant in Guangdong Province [4]. Fourteen representative 6a sequences were selected according to their position in a maximum-likelihood phylogenetic tree (Supplementary Fig. S1). All of these isolates were from treatment-naïve individuals. The Institutional Review Board at the Guangzhou Blood Center approved this study, and the guidelines set by this board were strictly followed. In addition, the study protocol followed the ethical guidelines set in place by the 1975 Declaration of Helsinki and was approved by the Medical Ethics Committee of Guangzhou Blood Center.

Total RNA was extracted using an Agencourt RNAdvance Blood Kit (Beckman Coulter) and then reverse transcribed using Superscript III (Invitrogen). Metagenomic library construction, Illumina HiSeq sequencing, and bioinformatic analysis were performed as described previously [10]. RDP4 [11] and GARD [12] were used to identify evidence of recombination.

All 108 available whole-genome reference sequences of HCV 6a were downloaded from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/gene). Sequences that lacked location and sampling date information were excluded (n = 5). Ultimately, 117 sequences (103 reference sequences and 14 sequences from this study) were used to construct a Bayesian Markov chain Monte Carlo (MCMC) tree using the SRD06 nucleotide partitioning model, an uncorrelated lognormal relaxed molecular clock, and a Bayesian Skyline coalescent model.

We classified the whole genome sequences of HCV 6a into two groups (Guangdong and non-Guangdong) for positive selection analysis according to their geographic location. We ensured that all sequences were derived from DAA-naïve patients, excluding the reference sequences from DAA-treated patients described in the literature. The mixed-effects model of evolution (MEME), implemented in the Datamonkey package (http://www.datamonkey.org/), has superior performance over older models. Therefore, MEME was used to detect adaptive evolution between Guangdong genomes and non-Guangdong genomes. Information about the positions of CD4 and CD8 T-cell epitopes was obtained from the Los Alamos National Laboratory website (http://hcv.lanl.gov/content/immuno/immuno-main.html) and the Immune Epitope Database and Analysis Resource (http://www.iedb.org). The nucleotide sequences reported in this study were deposited in the GenBank database with the accession numbers MZ161145-MZ161158.

Results

To obtain reliable consensus sequences, we defined clear criteria for consensus base calling based on quality score and depth. The criteria were as follows: 1) Phred quality score should be above 32, and 2) the minimum depth at a position of the HCV genome should be greater than or equal to 5. After trimming the sequences, all 14 sequences had a complete open reading frame. Neither RDP4 nor GARD showed evidence of recombination. The genotyping results based on whole genome sequences were identical to those obtained based on the E1 gene sequences in our previous study [4].

The evolutionary rate of the HCV 6a whole genome was found to be 9.59×10^{-4} substitution/sites/year (s/s/y). For partitioned genes, E2 had the fastest evolutionary rate (9.40×10^{-3} s/s/y), followed by E1, NS3, NS5A, NS4A, P7, NS4B, NS2, NS5B, NS4A, Core, and the 5'UTR (Supplementary Table S1). The HCV 6a sequences formed two clades (I and II) in an MCMC tree (Fig. 1). The reference sequences located at the root of clade I and all of those in clade II were from Vietnam, Asian immigrants (Canada), or a community in China, indicating that the HCV 6a strains circulating worldwide may have originated from Vietnam and other Asian countries. Except for HCV072, the Guangdong HCV 6a isolates were concentrated in three groups, indicating three transmission events in Guangdong Province. Group A originated around 1964 (95% CI: 1871–1988), while group B originated around 1966 (95% CI: 1941–1975). HCV 6a was transported from Vietnam to Hong Kong, then to Guangdong, where it became endemic. Group C originated around 1968 (95% CI: 1884–1981), and most of the sequences in this group were from Guangdong Province.

MEME analysis detected a total of 102 and 105 positively selected sites (PSS) among Guangdong and non-Guangdong sequences, respectively. The proteins with the highest proportion of PSSs were E1 and E2. The one with the lowest proportion was NS3, both in the Guangdong and non-Guangdong sequences (Fig. 2). There was a significant difference in the proportion of PSSs between Guangdong and the non-Guangdong region ($\chi^2 = 39.257$, $P < 0.05$). Non-Guangdong HCV 6a isolates tended to have more PSSs in E2 (28 sites vs. 25 sites) and NS2 (seven sites vs. four sites), whereas Guangdong HCV 6a had more PSSs in E1 (18 sites vs. 12 sites), P7 (five sites vs. one site), and Core (six sites vs. two sites). Fifteen homologous sites were found to be positively selected in both regions, located in Core (one site), E1 (three sites), E2 (eight sites), NS5A (one site), and NS5B (two sites). A map of Guangdong and non-Guangdong genomes representing the different layers of data analyzed (PSSs, CD8

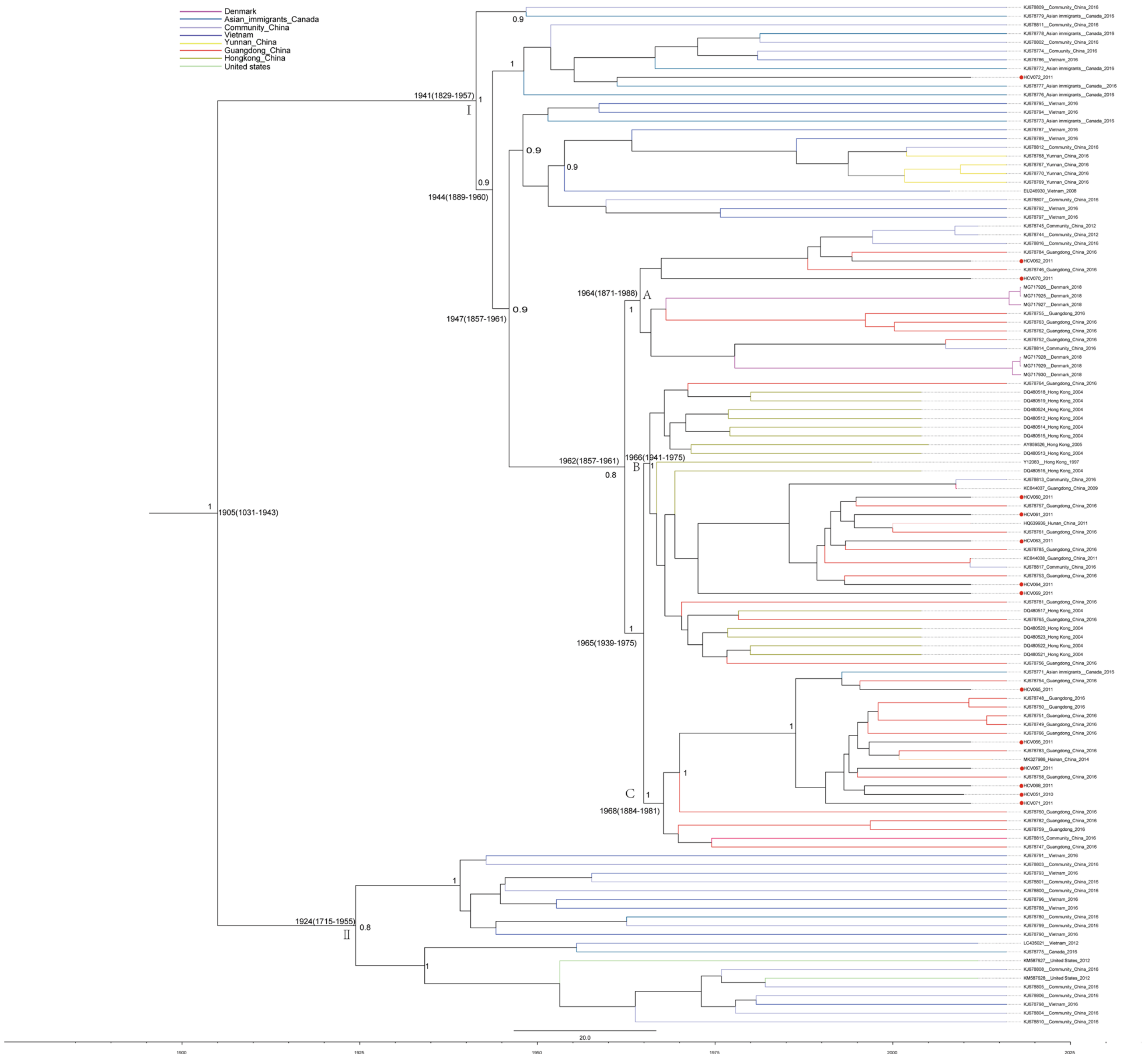


Fig. 1 An MCMC tree based on whole-genome sequences of HCV 6a isolates. Colored branches represent the geographic distribution of the sequences. HCV 6a isolates identified in this study are indicated by a solid red circle. The posterior value and the time to the most recent common ancestor (tMRCA) are shown for the main clusters.

cell epitopes, and CD4 T cell epitopes) is shown in Figure 3. An association between PSSs and the presence of CD8 epitopes was found ($\chi^2 = 9.675$, $P < 0.05$). Regions corresponding to CD8 T cell epitopes were under negative selection except in the isolates from Guangdong. There was no association between PSS, CD8 T cell epitopes, and CD4 T cell epitopes in the Guangdong genomes. The proteins with the highest proportion of PSSs within CD8 T cell epitopes were E1 and E2 in both locations (Supplementary Table S2). Seven amino acid sites located within CD8 T cell epitopes were found to be positively selected both in the Guangdong and non-Guangdong subsets. Among these, site 158 is located in the core region, which is targeted by cytotoxic T lymphocyte (CTL)-restricted HLA type A*02:01. Amino acid positions 227, 235, and 241 are located in the E1 region, targeted by HLA type A*02:01(227) and B35 (235 and 241). Sites 402 and 405 in E2 are within a known epitope targeted by HLA type A2. Site 2274 in NS5A is located in a region targeted by HLA type B60. However, heterogeneity between genotypes and mutations in the above epitopes where these seven PSSs are located has not been considered. Although

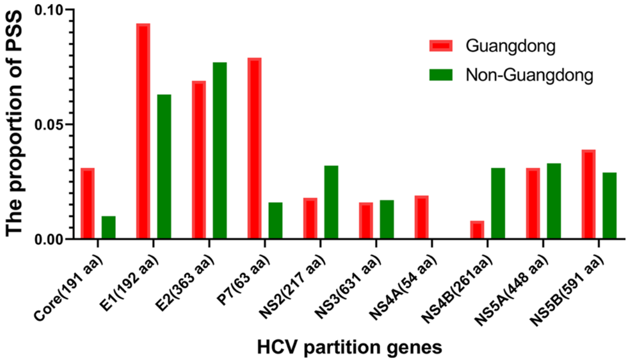


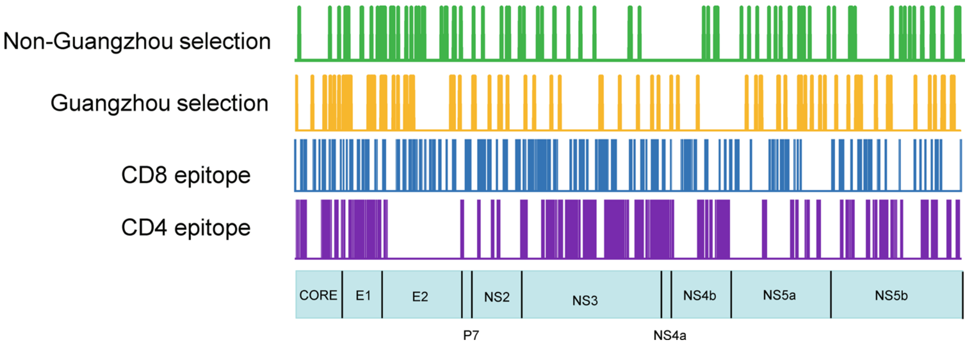
Fig. 2 Distribution of PSSs in the polyprotein of H77 reference sequence for the Guangdong and non-Guangdong subsets. “PSS” indicates a positively selected site

two positively selected sites, 532 in E2 and 1078 in NS3, which were targeted by HLA type B60 and A2 within a CD8 T cell epitope, were only found in Guangdong. T cells specific for these two epitopes (GENDTDVFVL, aa 530-539, and AINGVMWTV, aa 1071-1079) displayed HCV genotype 6 reactivity according to records in the IEDB database (http://www.iedb.org/).

Discussion

We performed a comparative analysis of evolutionary forces that shape the genomes of HCV 6a. Our results reveal a history of separate introductions of HCV 6a into Guangdong Province. In addition, we found that there are several differences as well as similarities between Guangdong and non-Guangdong genomes. These results help us understand the underlying factors driving HCV 6a evolution at the genome level. The molecular evolution of HCV 6a has previously only been investigated using a single genome region or two concatenated partial genome regions [13]. In this study, we analyzed 117 whole-genome sequences and found that HCV 6a worldwide evolved from a common ancestor in 1905, making it younger than any other subtypes of HCV gt6 [14]. The topology of the MCMC tree showed that Vietnam and other Asian countries could be the origin of 6a worldwide, which is consistent with the previous studies [3, 15]. Although there has been no spatiotemporal phylogenetic analysis, the data provide considerable evidence to support this hypothesis [15]. In addition, we found that HCV 6a was introduced into Guangdong around 1960–1970 via three transmission events. First, HCV 6a was introduced from Vietnam into Denmark and Guangdong in 1964, and HCV 6a strains were probably transmitted separately to Guangdong and Denmark, as there is no evidence of population migration between Guangdong and Denmark. In the second introduction, HCV 6a was apparently transported from Vietnam to Hong Kong and then to Guangdong, with cross-dissemination between these two places. It has been speculated that HCV 6a came from Vietnam to Hong Kong and then to Guangdong due to the fact that Hong Kong acted as the "first collecting port" for refugees from Vietnam [16]. However, because there is a lack of time-stamped sequence data from Hong Kong from that time period, this hypothesis could not be tested. In this study, we found evidence in the phylogeographic tree that the transmission route was from Vietnam to Hong Kong and then to Guangdong. Two historical events also corroborate this scenario: one is the association between Hong Kong and Vietnamese immigration, and the other is the population exchange between Guangdong Province and Hong Kong in the initial stage after the foundation of the People’s Republic of China [17]. In the third introduction, HCV 6a appears to have been introduced directly from Vietnam into Guangdong, causing a regional epidemic, as suggested in a previous study [16]. Guangdong Province, which is very close to Vietnam, accepted many refugees from Vietnam during the Vietnam War, and a fraction of those people might have been infected. We also found that most isolates in group C were from Guangdong Province. We speculate that the members of this clade clustered together because of similarities in the basic characteristics or lifestyle of the population in this area, resulting in a concentration of specific virus sequences. It is worth noting that HCV072 does not cluster into clade A, B, or C like other Guangdong

Fig. 3 Map of the HCV-6a Guangdong (yellow) and non-Guangdong (green) genomes, indicating the location of PSSs, CD8 T cell epitopes (blue), and CD4 T cell epitopes (purple)



isolates. The other isolates clustering with HCV072 were mainly from one community in China and Asian immigrants residing in Canada. Although detailed information on this specific community is lacking, we speculate that HCV072 was imported from Asian immigrants when they moved to Canada. Genetic drift may also account for why HCV072 did not cluster with the other Guangdong isolates. Most published studies detecting PSSs in this virus have analyzed different subtypes, such as HCV 1a and HCV 1b, without considering the region from which the sequences were obtained [8, 18]. In this study, we found that the proportion of PSSs differed between Guangdong and non-Guangdong genomes. Non-Guangdong HCV 6a strains have more PSSs in the E2 and NS2 regions, while Guangdong HCV 6a strains have more PSSs in the E1, P7, and Core regions. We speculate that geographical location is related to viral adaptation between HCV and the host's cellular immune response. The main reason for this speculation is that the virus transmission pattern and host genetic background in different regions led to the different positive sites. Seven amino acid sites within CD8 T cell epitopes were found to be positively selected both in the Guangdong and non-Guangdong subsets. However, when considering the heterogeneity between genotypes and mutations in the epitopes, only two sites are located in epitopes that are targeted by HLA type B60 and A2 in Guangdong. HLA-A2 accounts for 30.28%, and HLA type B60, which is broadly specific for the antigen epitope B*40:01 motif, account for 13-14.4% of the population of Guangdong Province (http://allelefrequencies.net/). The difference in selectivity between Guangdong and non-Guangdong subsets revealed that HCV 6a evolved under different forces and constraints in Guangdong and other regions. This information could be beneficial for finding constrained codon sites (non-PSSs) in the CD8 T cell epitope for viral vaccination to limit amino acid substitutions and reduce the likelihood of viral escape in different regions [19]. The negative selection of CD8 T cell epitopes outside Guangdong suggested that HCV 6a was better able to escape cellular immune defenses in other regions than it was in Guangdong, which was consistent with the fact that HCV 6a is the prevalent subtype in Guangdong Province [5]. The highest proportion of PSSs both within and outside of CD8 T cell epitopes was found in E1 and E2 in both locations, consistent with a previous study [20]. Furthermore, these regions were reported previously to have especially high genetic variability [21], suggesting the generation of escape mutants due to interactions between positive selection and the host immune response. The E1 and E2 glycoproteins are involved in the binding and fusion of HCV during cell entry. Therefore, the positively selected sites within CD8 T cell epitopes are likely to be a consequence of the human immune response and virus mutation after infection of a new host. On the other hand, the protein with the lowest proportion of sites under positive selection was NS3

(Fig. 1), a potential target for direct-acting antivirals that inhibit virus replication [22]. When considering epitopes, the proportion of PSSs in NS3 within the CD8 T cell epitope was higher in Guangdong (1.11%) than elsewhere (0.32%) (Supplementary Table S2). Since NS3 is one of the most frequent HCV protein targets of the CD8 T cell response [23, 24], the larger number of PSSs within the CD8 T cell epitope in Guangdong Province would make the CD8 T-cell-mediated immune response less effective, leading to chronic infection.

In conclusion, we have analyzed the propagation route and evolution rate of HCV 6a in Guangdong by comparing whole-genome sequences. Guangdong and non-Guangdong HCV 6a sequences exhibit different selective pressures in different parts of their genomes. This potentially provides information about the interactions between HCV and its host. The locations of PSSs, especially within epitopes, will provide useful information for choosing CD8 T cell epitopes for vaccine development and for the formulation of an HCV 6a prevention strategy in Guangdong, China.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00705-021-05358-7.

Acknowledgments This work was supported by grants from Guangdong Basic and Applied Research Foundation (2020A1515010118), the National Science Foundation of China (32000666 and 81772208), and The Key Medical Laboratory of Guangzhou (2021–2023). We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Authors' contributions RX, JTH, MW, QL, ZGS, and HSZ participated in the study design, wrote the draft, and collected the documentation materials. RX, XR, and YSF participated in the study design and helped revise the draft. HW collected the samples and revised the manuscript. All authors read and approved the manuscript.

Availability of data and material Genome sequences have been submitted to the GenBank database.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Dunford L, Carr MJ, Dean J et al (2012) Hepatitis C virus in Vietnam: high prevalence of infection in dialysis and multi-transfused patients involving diverse and novel virus variants. PLoS ONE 7:e41266

2. Chen Y, Yu C, Yin X et al (2017) Hepatitis C virus genotypes and subtypes circulating in Mainland China. Emerg Microbes Infect. 6:e95
3. Thong VD, Akkarathamrongsin S, Poovorawan K et al (2014) Hepatitis C virus genotype 6: virology, epidemiology, genetic variation and clinical implication. World J Gastroenterol 20:2927–2940
4. Wang M, Liao Q, Xu R et al (2019) Hepatitis C virus 3b strains in injection drug users in Guangdong Province, China, may have originated in Yunnan Province. Arch Virol 164:1761–1770
5. Rong X, Xu R, Xiong H et al (2014) Increased prevalence of hepatitis C virus subtype 6a in China: a comparison between 2004–2007 and 2008–2011. Arch Virol 159:3231–3237
6. Li Q, Yao Y, Shen Y et al (2017) Assessment of HCV genotypes in Yunnan Province of Southwest China. Virus Genes 53:190–196
7. Bull RA, Luciani F, McElroy K et al (2011) Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog 7:e1002243
8. Cuevas JM, Gonzalez M, Torres-Puente M et al (2009) The role of positive selection in hepatitis C virus. Infect Genet Evol 9:860–866
9. Yuan M, Lu T, Li C et al (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
10. Xu R, Huang JT, Du RS et al (2020) High-throughput next generation sequencing technology for the comprehensive assessment of full-length hepatitis C viral genomes in IDU population of Guangdong province. Chin J Blood Transfus 33:565–568 (**article in Chinese**)
11. Martin DP, Murrell B, Golden M et al (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1:vev003
12. Kosakovsky Pond SL, Posada D, Gravenor MB et al (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098
13. Huang K, Chen J, Xu R et al (2018) Molecular evolution of hepatitis C virus in China: a nationwide study. Virology 516:210–218
14. Pybus OG, Barnes E, Taggart R et al (2009) Genetic history of hepatitis C virus in East Asia. J Virol 83:1071–1082
15. Pham VH, Nguyen HD, Ho PT et al (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
16. Fu Y, Qin W, Cao H et al (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
17. McCormick AL, Macartney MJ, Abdi-Abshir I et al (2015) Evaluation of sequencing of HCV core/E1, NS5A and NS5B as a genotype predictive tool in comparison with commercial assays targeting 5’UTR. J Clin Virol 66:56–59
18. Patino-Galindo JA, Gonzalez-Candelas F (2017) Comparative analysis of variation and selection in the HCV genome. Infect Genet Evol 49:104–110
19. Hanada K, Gojobori T, Li WH (2006) Radical amino acid change versus positive selection in the evolution of viral envelope proteins. Gene 385:83–88
20. Cuypers L, Li G, Libin P et al (2015) Genetic Diversity and Selective Pressure in Hepatitis C Virus Genotypes 1–6: Significance for Direct-Acting Antiviral Treatment and Drug Resistance. Viruses 7:5018–5039
21. Forni D, Cagliani R, Pontremoli C et al (2018) Evolutionary analysis provides insight into the origin and adaptation of HCV. Front Microbiol 9:854
22. Wu R, Geng D, Chi X et al (2019) Computational analysis of naturally occurring resistance-associated substitutions in genes NS3, NS5A, and NS5B among 86 subtypes of hepatitis C virus worldwide. Infect Drug Resist. 12:2987–3015
23. Schulze Zur Wiesch J, Ciuffreda D, Lewis-Ximenez L et al (2012) Broadly directed virus-specific CD4+ T cell responses are primed during acute hepatitis C infection, but rapidly disappear from human blood with viral persistence. J Exp Med 209:61–75
24. Fitzmaurice K, Petrovic D, Ramamurthy N et al (2011) Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus infection. Gut 60:1563–1571

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.