Complete Genome Sequencing and Evolutionary Analysis of HCV Subtype 6a Including Strains from Guangdong Province, China

RU XU (✉ leng868@163.com)

Guangzhou blood center; The key medical laboratory of Guangzhou <https://orcid.org/0000-0001-5926-9213>

Jieting Huang

Guangzhou Blood Center; The key medical laboratory of Guangzhou

Min Wang

Guangzhou Blood center; The key medical laboratory of Guangzhou

Qiao Liao

Guangzhou Blood center; The key medical laboratory of Guangzhou

Zhengang Shan

Guangzhou Blood Center; The key medical laboratory of Guangzhou

Huishan Zhong

Guangzhou Blood Center; The key medical laboratory of Guangzhou; Southern Medical University

Xia Rong

Guangzhou Blood Center; The key medical laboratory of Guangzhou; Southern Medical university

Yongshui Fu

Guangzhou Blood Center

Research Article

Keywords: HCV, whole genome, evolution, transmission, positive selection sites

Posted Date: July 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-704778/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

In recent years, Hepatitis C virus 6a was not only the predominant subtype in Guangdong but also has the potential of increasing in non-Guangdong in China over time. To understand the virus evolution and transmission mode, the representative HCV 6a specimens in Guangdong province were chosen for deep sequencing and performed evolutionary analysis with all the 6a whole genome reference sequences. Our results showed that less than 5% of the genome was found to be under positive selection. The protein with the highest proportion of sites under positive selection was E1 and E2; A positive association between positive selection sites and the presence of CD8 epitopes was found in non-Guangdong genomes ($\chi^2 = 9.168$, $P < 0.05$). The evolutionary rate of 6a whole genome was 9.59×10^{-4} s/s/y. E2 has the fastest evolutionary rate (9.40×10^{-3} s/s/y), followed by E1, NS3 and NS5A. Spatio-temporal evolutionary analysis showed that Vietnam and other Southeast Asian countries were the origins of global HCV6a. Further, there were two important transmission modes which were all originated from Vietnam and other Southeast Asian countries in Guangdong. One (Group B) indicated that HCV6a was then transmitted to Hong Kong, and then to Guangdong and cross-dissemination between these two places; Another (Group C) revealed that HCV6a was directly disseminate to Guangdong and formed a regional epidemic. Our study firstly revealed the transmission mode of HCV6a by whole genome sequence, indicating that the impact of selective constraints in Guangdong and non-Guangdong, it can be helpful to plan for future prevention and management of HCV6a infection in Guangdong, China.

Introduction

HCV is classified into eight genotypes (1–8) and subdivided into more than 90 subtypes named in alphabetical order from a to z due to a lack of proof reading activity by its viral RNA polymerase (https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/634/table-1). Among them, HCV genotype 6 was the most divergent genotype and predominantly distribute in South China and South-East Asian countries, the latter carrying the greatest burden of HCV gt6 in Asia^{1–3}. Several of HCV gt6 subtypes were identified and classified only recently, suggesting that a considerable proportion of HCV diversity remains undescribed. HCV 6a, the first subtype to be discovered, was endemic in Vietnam, Taiwan, Hongkong and South China^{3–5}, especially in Guangdong Province, which accounts for more than half. Guangdong Province located in the southern part of China, is the center of cultural and commercial exchanges in domestic and international crossroads. It is the most populous province of China (126 million people), 32 million are migrants who are on business or temporarily hired as laborers in the myriad of factories with lack good hygiene conditions, leading to a frequently travel between Guangdong and their rural hometowns (<http://www.stats.gov.cn/tjsj/pcsj/>). Collectively, these factors created an environment to increase HCV prevalence not only in Guangdong but also in non-Guangdong. Our recently study showed that HCV 6a was not only most common in Guangdong Province, but also increased by labor exchange outside Guangdong Province in China⁶, indicating that HCV 6a was transmitted rapidly all over China through some possible propagation networks. In addition, another study also unraveled that the HCV subtype 6a could rapidly spread across China⁷.

It has been reported that transmission was an important force driving HCV viral evolution as the virus infected a novel host immune system, which alters the selection pressure on the viral genome^{8,9}. According to previous study, HCV genome evolves at ~ 0.001 substitutions/site/year (s/s/y), a rate typical of RNA viruses^{10,11}. Studies have revealed that HCV genetic variability of the same gene in different genotype is different¹⁰. Furthermore, HCV genetic variability is not evenly distributed across the viral genome^{7,12}. HCV whole genomes could maximize the evolutionary resolution¹¹, detect positively selected sites along the genomes. Our aim in this study was to reveal HCV evolutionary history using the representative HCV 6a samples in Guangdong Province and archived reference sequences worldwide, acquire detailed view of the action of selection and how it may differentially affect variants of this virus in Guangdong and non-Guangdong in China.

Methods

1. Sample collection

Previously, a BDs—Blood donors, BDs—cohort of 100 HCV 6a were genotyped by E1 region, which revealed that HCV 6a was predominant in Guangdong Province⁵. Among them, 14 representative HCV 6a samples were selected according to the topology structure of maximum likelihood phylogenetic tree (supplementary Figure 1) by mega X. They were all treatment-naïve. This study was approved by the Institutional Review Board at the Guangzhou Blood Center and the guidelines set by this board were strictly followed. Study protocols followed the ethical guidelines set in place by the 1975 Declaration of Helsinki and were approved by the Medical Ethics Committee of Guangzhou Blood Center.

2. Next generation sequencing

Total RNA was extracted using the AgencourtRNAAdvance blood kit (Beckman Coulter) and then reverse transcribed using Superscript III (Invitrogen). Metagenomic libraries were prepared using the KAPA Library Prep kit (KAPA Biosystems) with index tagging by 16 cycles of PCR using KAPAHiFi Hot Start (KAPA Biosystems) and NEBNext Multiplex Oligos (oligonucleotides). Libraries were quantified by Qubit (ThermoFisher) and TapeStation (Agilent) and then pooled at equimolar concentrations for sequencing on the Illumina HiSeq platform.

3. Bioinformatic processing

The raw reads were trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The clean reads were enriched in-silico by removing human reads and rRNA reads (Ribopicker, <http://ribopicker.sourceforge.net/>) and then matched against the 187 whole-genome HCV reference sequences using Tanoti (<http://bioinformatics.cvr.ac.uk/tanoti.php>). Finally, the sam files were generated using Tanoti and a majority consensus sequence was generated for sites with >5 reads. The assemblies were individually inspected in UGene (<http://ugene.net/>).

4. HCV genotype and sequence datasets

The consensus sequences were placed within the context of previously known genotypes(https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification) using MEGA X software package using the maximum likelihood method with 500 bootstrap replicates. RDP4¹³ and GARD¹⁴ were used to test for recombination. All available (n=108) HCV 6a whole genome reference sequences were downloaded from the HCV-GLUE website (<http://hcv.glue.cvr.ac.uk/>) and NCBI (<https://www.ncbi.nlm.nih.gov/gene>). The sequences lacked location and sampling date information were excluded (n=5). Ultimately, 117 sequences (103 6a reference sequences with 14 6a sequences in this study) were used to construct the MCMC tree.

Among reference sequences, 15 were from Vietnam, 10 were Asian_immigrants_Canada (collected from Canada, but they are Asian immigrants), 15 were from Hongkong, 21 were from a community in China(no detailed province information), 28 were from Guangdong Province in China, others countries were less than 10 sequences, including United States, Denmark, Hunan province, Hainan province and Yunnan province in China.

5. Spatio-temporal phylogenetic analysis

The Bayesian Markov Chain Monte Carlo (MCMC) inference method, implemented in BEAST v1.10.4 (<http://beast.community/>) was used to estimate the origin and transmission of HCV 6a all over the world. The SRD06 nucleotide partitioning model, an uncorrelated lognormal relaxed molecular clock and a Bayesian Skyline coalescent model with 10 groups were employed. A lognormal prior distribution (mean, 0.001; standard deviation, 0.002) for the prior substitution rate to achieve MCMC convergence. The estimated effective sampling sizes (ESSs) were inspected and traces with 10% burn-in using Tracer v1.7.1 (<http://beast.community/tracer>). When all ESSs were ≥ 200 , a sufficient sampling was considered to have been achieved. Maximum clade credibility trees were calculated and annotated using TreeAnnotator 1.7.5 (<https://www.beast2.org/treeannotator/>). FigTree v1.4.1 (<http://tree.bio.ed.ac.uk/software/figtree/>) was used to colour lineages according to their sampling location.

6. Positive selection sites for HCV 6a genomes between Guangdong and Non-Guangdong

The total of HCV-6a whole genome sequences were divided into two groups (Guangdong and non-Guangdong) according to their geographic location. we ensured that all HCV sequences derived from DAA-naïve patients, excluding the reference sequences from DAA-treated patients described in literature. Mixed Effects Model of Evolution (MEME), implemented in datamonkey package (<http://www.datamonkey.org/>), has more superior performance over older models such as the two-rate Fixed Effects likelihood (FEL and Single-Likelihood Ancestor Counting (SLAC) under a broad range of scenarios. Therefore, MEME was used to detect adaptive evolution in the codon alignment between Guangdong genomes and non-Guangdong genomes. CD4 and CD8 T-cell epitope positions were retrieved 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>).

7. Resistance-associated substitutions (RAS) prevalence in Guangdong sequences

The prevalence of HCV 6a RASs was determined for NS3, NS5A and NS5B in patients at baseline. RASs were detected with 95%, 5% and 1% threshold in a patient using phdrSamReporter module within HCV-GLUE project¹⁵. The majority and minority variants for each subtype were reported relative to the amino acid position of GT1a H77 reference.

8. Accession number

The nucleotide sequences reported in this study were deposited in Genbank with the following accession numbers: MZ161145-MZ161158.

Results

1. Consensus sequences were obtained

In order to get reliable consensus sequences for each sample, we defined clear criteria for consensus calling based on quality score and depth. The criteria were: 1) Phred quality score should be above 32; 2) the minimum depth at a position of the HCV genome should be greater or equal to 5. After trimming the sequences using above criteria, all the 14 samples had complete open-reading-frames (Table 1). The average depth of each position was greater than 1000 (mean:4399, mix-max:1143-7399). After removing human reads and rRNA reads, the percentage of mapped reads to total reads was significantly increased from $3.05\% \pm 1.58\%$ to $43.11\% \pm 14.82\%$. The reads unmapped to HCV contain bacterial and fungi, no

other virus was found. Neither RDP4 nor GARD showed evidence of recombination. The genotyping results of the whole genome sequences were the same as those of E1 gene in previous study⁶.

Table 1

The consensus calling results of HCV 6a samples

Samples ID	Depth Average (min-max)	Phredquality score	coverage	Mapped reads	Total reads	Total clean reads	Mapped rates/total reads	Mapped reads/total clean reads
1	3517(1266-10080)	34	100.00%	226,860	4,706,368	521,530	4.82%	43.5%
2	5369(2333-14850)	34	100.00%	343,165	9,045,252	731,988	3.79%	46.9%
3	5402(2100-15736)	36	100.00%	344,665	11,791,174	740,465	2.92%	46.6%
4	1143(44-4443)	35	100.00%	74,126	7,792,022	292,310	0.95%	25.4%
5	5695(2280-12416)	34	100.00%	362,732	8,969,770	809,218	4.04%	44.8%
6	3361(1235-8593)	34	100.00%	214,872	10,828,376	511,484	1.98%	42.0%
7	7399(2386-20927)	34	100.00%	475,274	7,295,824	810,932	6.51%	58.6%
8	2938(800-10299)	34	100.00%	188,510	9,298,778	279,718	2.03%	67.4%
9	5289(480-12689)	34	100.00%	339,379	10,633,734	585,044	3.19%	58.0%
10	5330(1716-11770)	35	99.7%	341,742	10,428,600	604,714	3.28%	56.5%
11	3990(631-8608)	34	100.00%	256,739	11,735,954	903,942	2.19%	31.9%
12	4407(257-9725)	34	100.00%	284,240	12,624,228	839,265	2.25%	33.9%
13	6963(2141-23422)	34	100.00%	448,583	10,596,556	1,207,650	4.23%	37.2%
14	794(258-2760)	34	100.00%	50,758	10,380,432	467,878	0.49%	10.9%

2. The origin and evolutionary history of HCV 6a for whole genome genome

The evolutionary rate of HCV 6a whole genome was 9.59×10^{-4} s/s/y. The COV (coefficient of variation) was 0.593, the posterior was -1.373×10^{-5} . All the 6a worldwide formed two clades (A and B), which were originated from 1905 (95%CI:1031-1943). Clade A were concentrated by reference from Vietnam, Asian_immigrants_Canada and a community in China. On the root of clade B, the location of reference sequences was consistent with clade B, showed that the HCV 6a

worldwide may from Vietnam and other Asian countries. The Guangdong HCV 6a were concentrated into 3 groups, indicating that there were three transmission events in Guangdong Province. Group A was originated from the year of 1964 (95%CI:1871-1988), while group B was originated from the year of 1966 (95%CI:1941-1975). The topology of beast tree showed that Guangzhou HCV 6a in Group B were originated from Hongkong, then formed a local endemic. Group C was originated from the year of 1968 (95%CI:1884-1981), and majority of sequences in this group were from Guangdong Province.

3. The evolutionary rate for HCV 6a partitioned gene

For HCV 6a partitioned genes, E2 has the fastest evolutionary rate (9.40×10^{-3} s/s/y) among all the partition genes, followed by E1 (1.00×10^{-3} s/s/y), NS3 (1.00×10^{-3} s/s/y), NS5A (1.00×10^{-3} s/s/y), NS4A (1.00×10^{-3} s/s/y), and P7 (9.99×10^{-4} s/s/y), NS4B (9.96×10^{-4} s/s/y), NS2 (9.94×10^{-4} s/s/y), NS5B (9.84×10^{-4} s/s/y), NS4A (9.58×10^{-4} s/s/y), Core (3.95×10^{-4} s/s/y) and 5'UTR (2.80×10^{-4} s/s/y). The Coefficient of Variance for all regions was below 0.5 (0.360 ± 0.122) except for 5'UTR which the substitution rate heterogeneity among lineages was much greater (5.158). The covariance for all regions was close to zero (range, -0.0141 in core to 0.0324 in the 5'UTR).

4. Analysis of positively selected positions for HCV 6a genomes between Guangdong and Non-Guangdong

Totally, MEME analyses detected 102 and 105 positively selected sites among Guangdong and non-Guangdong sequences, respectively. There was a significant difference of positive selection sites between Guangdong and non-Guangdong region ($\chi^2=9.168$, $P<0.05$). The protein with the highest proportion of sites under positive selection was E1 and E2 (Fig 2); The proteins with the lowest proportion of sites under positive selection were NS3. Non-Guangdong HCV 6a tended to present more positively in E2 (28 sites in non-Guangdong vs 25 sites in Guangdong) and NS2 (7 sites in non-Guangdong vs 4 sites in Guangdong), Guangdong HCV 6a presented more positively sites in E1 (18 sites in Guangdong vs 12 sites in non-Guangdong), P7 (5 sites in Guangdong vs 1 sites in non-Guangdong) and Core (6 sites in Guangdong vs 2 sites in non-Guangdong). Fifteen homologous sites were found to be positively selected in both regions, which located in Core (1 site), E1 (3 sites), E2 (8 sites), NS5A (1 site) and NS5B (2 sites). A map of Guangdong and non-Guangdong genomes representing the different layers of data analyzed (positive selection, CD8 and CD4 T cells) was shown in Fig 3. A positive association between positive selection sites and the presence of CD8 epitopes was found in non-Guangdong genomes ($\chi^2=9.168$, $P<0.05$). In Guangdong genomes, there was no difference between positive selection sites, CD8 T cell epitopes and CD4 T cell epitopes.

5. Polymorphisms associated with resistance to NS3 Protease, NS5A and NS5B Polymerase inhibitors in whole genome sequences of HCV 6a

We characterized the prevalence of baseline NS3, NS5A and NS5B RASs with different sensitivities using 95%, 5% and 1% cut-offs for the frequency of a RAS in intra-host viral population sequences. Baseline RAS frequencies are summarized in Table 2. Mutations Q80K and 158V/I in NS3 region that confer resistance to siméprévir and boceprevir were found in 100.0% (14/14) and 7.14% (1/14) of the isolates, respectively. Even mutation V36G conferring low-level resistance to telaprevir and asunaprevir were found in 28.57% (4/14) of the patients, the cutoff for each patient was less than 5%. In NS5A region, M28F, Q30R and Y93T/S that confer resistance to daclatasvir, ombitasvir, ledipasvir and ledipasvir were found in 42.9% (6/14), 100.0% (14/14) and 64.3% (9/14) with 95% as a cutoff value, respectively. Mutation K24R was

found in 14.3% (2/14) of the patients, the cutoff for each patient was less than 5%. In NS5B, only one mutation C445F that confer resistance to dasabuvir was found in 71.4% (10/14) with 95% as a cutoff value.

Table 2

Polymorphisms associated with resistance to NS3 Protease, NS5A and NS5B Polymerase inhibitors in whole genome sequences of HCV 6a in Guangdong region

Protein	Amino Acid Position	Reference Amino acid	Drug	Substitution	The number of pctPresent* >95%	The number of 1%<pctPresent* <5%
NS3	36	V	Telaprevir, Asunaprevir	G	0	4
NS3	80	Q	Siméprévir	K	14	0
NS3	158	V	Boceprevir	I	1	0
NS5A	24	K	ledipasvir, velpatasvir	R	0	2
NS5A	28	M	Daclatasvir, ombitasvir, ledipasvir	F	6	2
NS5A	30	Q	Daclatasvir	R	14	0
NS5A	93	Y	Daclatasvir, Ledipasvir, Ombitasvir	T/S	9	1
BS5B	445	C	Dasabuvir	F	10	0
*: The prevalence in the viral population in one sample						

Discussion

With the development of next-generate sequencing (NGS) technology and bioinformation analysis methods, whole-genome sequencing of HCV has until now been a convenient, easy and accurate procedure that has now been adopted in scientific research and clinical applications^{16,17}. The NGS methodology we used in this study was able to generate HCV whole genomes from blood donor's samples, furthermore, it could accurately and reliably define the HCV subtype as the high average depths and strict access standards.

The molecular evolution of HCV 6a has previously only been investigated using single genome region or two concatenated partial genome regions¹⁸. Utilizing Bayesian MCMC phylogenetic methods, we analyzed 107 whole genome sequences and found that HCV 6a worldwide evolved from a most recent common ancestor (tMRCA) in 1905 (95%CI:1031-1943), which was younger than any other subtypes of HCV gt6¹⁹. The topology of the MCMC tree showed that Vietnam and other Asian countries could be the origin of 6a all over the world, which was consistent with hypothesis of previous studies^{3,12}. Even though there is no spatio-temporal phylogenetic analysis, they provided considerable evidence to support their hypothesis²⁰. In addition, we found that HCV 6a was introduced into Guangdong around 1960-1970 by two important transmission events. Originated from Vietnam, HCV 6a was transmitted to Hongkong. There were some evidences for the association between Hong Kong and Vietnamese Immigration²¹, our study also showed that Hongkong 6a was originated from Vietnam. One of the important transmission events was that HCV 6a in Guangdong were originated from Hongkong in 1966 (1941-1975), and then formed a local endemic in Guangdong province. We suspected that was because of the

exchange between Guangdong Province and Hongkong in the early founding of the country²². Another transmission event showed that HCV6a was directly disseminate to Guangdong and formed a regional epidemic. From the topology of tree, we found that majority of sequences in group C was from Guangdong province not only in this study but also in reference sequences. Therefore, we conclude that this clade is clustered together, because the basic characteristics or lifestyles of the population in this area are similar, which resulting in a certain concentration of infected virus sequences.

Our study showed that E2 has the fastest evolutionary rate, roughly ten times than the majority partitioned genes of HCV 6a. It is well known that E2 contains three hypervariable regions (HVR), known as HVR1-3²³, which are under continuous selection pressure exerted by the host immune system. Therefore, it is reasonable that E2 evolve fastest among all the partitioned genes. The evolutionary rate of E1, NS3, NS5A and NS4A was the same as a rate typical of RNA viruses 1.00×10^{-3} s/s/y, almost equal to whole genome evolutionary rate. E1, an HCV gene, usually represent whole-genome for HCV genotyping, as the variability of E1 can approximately replace HCV whole-genome²². Our result also supported this insight. NS3 is the HCV protease and NTPase/helicase, NS4A acts as a cofactor for the NS3 protease²³, which can explain why the NS3 and NS4A have the same evolutionary rate. The evolutionary rate of NS2 was 9.94×10^{-4} s/s/y, NS5B and NS4A was slightly slower than NS2 gene. Core and 5'UTR had the slowest evolutionary rate among all the HCV 6a partitioned genes. Study also showed that HCV core protein was the lowest variability and have most pan-genotypic consensus positions²⁴. Combination with these two results, we speculated that was because HCV core protein forms the viral capsid and interacts with multiple cellular proteins, which may require a high level of conserved residues. 5'UTR was so a conserved gene that do not distinguish some of the HCV-6 subtypes prevalent in Southeast Asia, and that these subtypes should instead be classified as genotype 1 or 1b²⁵ when based on primers from the 5'UTR. Using core sequencing data as well as 5' UTR data, it shown a remarkable improvement as it in genotyping accuracy and differentiation between HCV-1 and HCV-6 variants²⁶.

Most published studies detecting positively selected sites in this virus have analyzed for different major subtypes such as HCV 1a and HCV 1b, no considering the region of the sequence source^{27,28}. In this study, we found that the proportion of positively selected sites along HCV genomes was difference between Guangdong and non-Guangdong sequences, which indicating that geographical location is related viral adaptation between HCV and the cellular immune response. Analyses at the genomic level also revealed that Guangdong and non-Guangdong HCV 6a evolve under difference forces and constraints (CD8 T cell epitopes favoring selection in non-Guangdong HCV 6a). We speculate that the virus transmission pattern together with host genetic background in different regions leading to the different positive sites. The highest proportion of positive selection sites was found in E1 and E2 in both geographical locations, which is consistent with previous published studies^{24,29,30}. These regions were reported to have the high genetic variability along the HCV genomes³¹, which reflect the generation of escape mutants whereby the interaction between positive selection and host immune response. As we know, E1 and E2 glycoproteins determines binding and fusion properties of HCV during cell entry³². Consequently, the positive site was the consequence of the human immune response and virus mutation when infected a new host. On the contrary, the proteins with the lowest proportion of sites under positive selection were NS3 which is a potential target for direct-acting antivirals that function to inhibit virus replication³³. In addition, NS3 was also one of the most frequently HCV protein targets of the CD4 and CD8 T cell response^{34,35}. The lowest positive selection sites in this gene would make DAA treatment more effective and facilitate viral escape from the immune system whereby CD4 and CD8 T mediated cell immune response.

The frequency of RASs varied by both HCV subtype and geographic regions³³, therefore, we analyzed the frequency and distribution of naturally occurring RASs of HCV 6a in Guangdong regions using NGS sequence data. The RASs of HCV 6a has previously only been investigated using sanger sequencing, which can only detect more than 15%-25% of HCV variations, while deep sequencing can detect as low as 1% of HCV variation^{36,37}. In this study, mutations V36G in NS3 region and K24R in NS5A region were not found when the cutoff >95%, but after adjusting the cutoff to 5%, the prevalence

was as high as 28.57% (4/14) and 14.3% (2/14) of the patients. Other RASs reported in this study were also found in other region's 6a sequences³⁸. Even HCV 6a RASs could compromise treatment in robust infectious cell culture models³⁹, but 100% SVR rates have been reported in a few reports on the outcomes of DAA therapy in gt6-infected patients^{40, 41}. More functional and clinical trials are needed to validate the RASs found in this study.

In conclusion, this study firstly analyzed the propagation mode and evolution rate of HCV 6a by whole genome sequence, revealed the transmission mode of HCV 6a in Guangdong. We have shown that, Guangdong and non-Guangdong HCV 6a share different selective pressures along their genomes, which give information about the effect of some of the interactions between HCV and its host on HCV variability. It may be useful for antiviral research against this virus. Overall, the results from this study provided data support for the formulation of HCV 6a prevention strategy in Guangdong, China.

Declarations

Conflicts of Interest

The authors declare no conflict of interest.

Availability of data and material

Genome sequences have been submitted to the GenBank database.

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. All authors read and approved the manuscript.

Acknowledgements

This work was supported by grants from National Science Foundation of China (32000666) and Guangdong Basic and Applied Research Foundation (2020A1515010118).

References

1. Dunford L, Carr MJ, Dean J, Waters A, Nguyen LT, Ta Thi TH 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, Guo X, Wu S, Hou J (2017) Hepatitis C virus genotypes and subtypes circulating in Mainland China. *Emerging microbes infections* 6:e95
3. Thong VD, Akkarathamrongsin S, Poovorawan K, Tangkijvanich P, Poovorawan Y (2014) Hepatitis C virus genotype 6: virology, epidemiology, genetic variation and clinical implication. *World journal of gastroenterology* 20:2927–2940
4. Yan J, Fu XB, Zhou PP, He X, Liu J, Huang XH et al. Complicated HCV subtype expansion among drug users in Guangdong province, China. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*. 2019;73:139–45
5. Wang M, Liao Q, Xu R, Song D, Huang J, You Q et al (2019) Hepatitis C virus 3b strains in injection drug users in Guangdong Province, China, may have originated in Yunnan Province. *Archives of virology* 164:1761–1770

6. Rong X, Xu R, Xiong H, Wang M, Huang K, Chen Q et al (2014) Increased prevalence of hepatitis C virus subtype 6a in China: a comparison between 2004–2007 and 2008–2011. *Archives of virology* 159:3231–3237
7. Li Q, Yao Y, Shen Y, Cao D, Li Y, Zhang S et al (2017) Assessment of HCV genotypes in Yunnan Province of Southwest China. *Virus Genes* 53:190–196
8. Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S et al (2018) QUENTIN: reconstruction of disease transmissions from viral quasiespecies genomic data. *Bioinformatics* 34:163–170
9. Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A et al (2011) Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS pathogens* 7:e1002243
10. 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 evolutionary biology* 11:131
11. Yuan M, 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
12. Iles JC, Raghwanji J, Harrison GLA, Pepin J, Djoko CF, Tamoufe U et al. Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa. *Virology*. 2014;464–465:233 – 43
13. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution* 1:vev003
14. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22:3096–3098
15. Singer JB, Thomson EC, McLauchlan J, Hughes J, Gifford RJ (2018) GLUE: a flexible software system for virus sequence data. *BMC Bioinform* 19:532
16. Janiak M, Caraballo Cortes K, Perlejewski K, Kubicka-Russel D, Grabarczyk P, Demkow U et al. Next-Generation Sequencing of Hepatitis C Virus (HCV) Mixed-Genotype Infections in Anti-HCV-Negative Blood Donors. *Advances in experimental medicine and biology*. 2018;1096:65–71
17. Bellocchi MC, Aragri M, Carioti L, Fabeni L, Pipitone RM, Brancaccio G et al. NS5A Gene Analysis by Next Generation Sequencing in HCV Nosocomial Transmission Clusters of HCV Genotype 1b Infected Patients. *Cells*. 2019;8
18. Huang K, Chen J, Xu R, Jiang X, Ma X, Jia M et al (2018) Molecular evolution of hepatitis C virus in China: A nationwide study. *Virology* 516:210–218
19. Pybus OG, Barnes E, Taggart R, Lemey P, Markov PV, Rasachak B et al (2009) Genetic history of hepatitis C virus in East Asia. *Journal of virology* 83:1071–1082
20. Pham VH, Nguyen HD, Ho PT, Binh DV, Pham HL, Pham PH 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
21. 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
22. McCormick AL, Macartney MJ, Abdi-Abshir I, Labbett W, Smith C, Irish D 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. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology* 66:56–59
23. Troesch M, Meunier I, Lapierre P, Lapointe N, Alvarez F, Boucher M et al (2006) Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein. *Virology* 352:357–367
24. Cuypers L, Li G, Libin P, Piampongsant S, Vandamme AM, Theys K (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

25. Ross-Thriepfand D, Harris M (2015) Hepatitis C virus NS5A: enigmatic but still promiscuous 10 years on! *J Gen Virol* 96:727–738
26. Bouchardeau F, Cantaloube JF, Chevaliez S, Portal C, Razer A, Lefrere JJ et al (2007) Improvement of hepatitis C virus (HCV) genotype determination with the new version of the INNO-LiPA HCV assay. *J Clin Microbiol* 45:1140–1145
27. Patino-Galindo JA, Gonzalez-Candelas F (2017) Comparative analysis of variation and selection in the HCV genome. *Infection genetics evolution: journal of molecular epidemiology evolutionary genetics in infectious diseases* 49:104–110
28. Cuevas JM, Gonzalez M, Torres-Puente M, Jimenez-Hernandez N, Bracho MA, Garcia-Robles I et al (2009) The role of positive selection in hepatitis C virus. *Infection genetics evolution: journal of molecular epidemiology evolutionary genetics in infectious diseases* 9:860–866
29. Humphreys I, Fleming V, Fabris P, Parker J, Schulenberg B, Brown A et al (2009) Full-length characterization of hepatitis C virus subtype 3a reveals novel hypervariable regions under positive selection during acute infection. *Journal of virology* 83:11456–11466
30. Torres-Puente M, Cuevas JM, Jimenez-Hernandez N, Bracho MA, Garcia-Robles I, Wrobel B et al (2008) Using evolutionary tools to refine the new hypervariable region 3 within the envelope 2 protein of hepatitis C virus. *Infection genetics evolution: journal of molecular epidemiology evolutionary genetics in infectious diseases* 8:74–82
31. Forni D, Cagliani R, Pontremoli C, Pozzoli U, Vertemara J, De Gioia L et al (2018) Evolutionary Analysis Provides Insight Into the Origin and Adaptation of HCV. *Frontiers in microbiology* 9:854
32. Douam F, Dao Thi VL, Maurin G, Fresquet J, Mompelat D, Zeisel MB et al (2014) Critical interaction between E1 and E2 glycoproteins determines binding and fusion properties of hepatitis C virus during cell entry. *Hepatology* 59:776–788
33. Wu R, Geng D, Chi X, Wang X, Gao X, Xu H 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. *Infection drug resistance* 12:2987–3015
34. Schulze Zur Wiesch J, Ciuffreda D, Lewis-Ximenez L, Kaspruwicz V, Nolan BE, Streeck H 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. *The Journal of experimental medicine* 209:61–75
35. Fitzmaurice K, Petrovic D, Ramamurthy N, Simmons R, Merani S, Gaudieri S et al (2011) Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus infection. *Gut* 60:1563–1571
36. Hernandez D, Zhou N, Ueland J, Monikowski A, McPhee F (2013) Natural prevalence of NS5A polymorphisms in subjects infected with hepatitis C virus genotype 3 and their effects on the antiviral activity of NS5A inhibitors. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology* 57:13–18
37. Sarrazin C (2016) The importance of resistance to direct antiviral drugs in HCV infection in clinical practice. *Journal of hepatology* 64:486–504
38. Li Z, Liu Y, Zhang Y, Shao X, Luo Q, Guo X et al (2017) Naturally Occurring Resistance-Associated Variants to Hepatitis C Virus Direct-Acting Antiviral Agents in Treatment-Naive HCV Genotype 6a-Infected Patients. *BioMed research international* 2017:9849823
39. Pham LV, Ramirez S, Gottwein JM, Fahnoe U, Li YP, Pedersen J et al. HCV Genotype 6a Escape From and Resistance to Velpatasvir, Pibrentasvir, and Sofosbuvir in Robust Infectious Cell Culture Models. *Gastroenterology*. 2018;154:2194 – 208 e12.
40. Puoti M, Foster GR, Wang S, Mutimer D, Gane E, Moreno C et al (2018) High SVR12 with 8-week and 12-week glecaprevir/pibrentasvir therapy: An integrated analysis of HCV genotype 1–6 patients without cirrhosis. *Journal of hepatology* 69:293–300
41. Feld JJ, Jacobson IM, Hezode C, Asselah T, Ruane PJ, Gruener N et al (2015) Sofosbuvir and Velpatasvir for HCV Genotype 1, 2, 4, 5, and 6 Infection. *N Engl J Med* 373:2599–2607

Figures

Figure 1

A MCMC tree estimated from the whole genome sequences of HCV 6a. Coloured branches represent the geographic distribution of the sequences (reference sequences and 14 sequences in this study). HCV 6a isolates identified in this study are labeled in red solid circle. The posterior value and tMRCA were labeled in the main clusters.

Figure 2

The distribution of positively selected sites along the different regions of the H77 reference polyprotein for the Guangdong and non-Guangdong subsets. PSC represented positively selected sites.

Figure 3

Map of the HCV-6a Guangdong (Green), non-Guangdong (Yellow) genomes, indicating the location of totally positively selected sites, CD8 T cell epitopes (Blue) and CD4 T cell epitopes (Purple).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryfigure1.tif](#)