Genetic Characterization and Recombinant History of a Novel Emerging HIV-1 Second-Generation Circulating Recombinant Form (CRF120_0107) Identified Among Sexually Transmitted Infections in Shenzhen, China

Bo Zhu  
Anhui Medical University

Jin Zhao  
Shenzhen Center for Disease Control and Prevention

Xiaorui Wang  
Beijing Institute of Microbiology and Epidemiology

Hanping Li  
Beijing Institute of Microbiology and Epidemiology

Yongjian Liu  
Beijing Institute of Microbiology and Epidemiology

Chenli Zheng  
Shenzhen Center for Disease Control and Prevention

Lei Jia  
Beijing Institute of Microbiology and Epidemiology

Tianyi Li  
Beijing Institute of Microbiology and Epidemiology

Xiaolin Wang  
Beijing Institute of Microbiology and Epidemiology

Lin Chen  
Shenzhen Center for Disease Control and Prevention

Jingyun Li  
Beijing Institute of Microbiology and Epidemiology

Bohan Zhang  
Beijing Institute of Microbiology and Epidemiology

Jingwan Han  
Beijing Institute of Microbiology and Epidemiology

Lin Li  
(Bearewood@sina.com)  
Beijing Institute of Microbiology and Epidemiology
Abstract

Under the background of the main epidemic HIV strains (CRF01_AE and CRF07_BC) co-circulation in China, more HIV second-generation recombinant (SGR) strains with CRF01_AE and CRF07_BC as the backbone are emerging. In this study, we analyzed the characteristics and evolutionary history of a newly emerging HIV-1 CRF120_0107 composed of CRF01_AE and CRF07_BC based on the near full-length genome (NFLG) in Shenzhen, Guangdong Province, China. NFLG phylogenetic analysis revealed that these sequences formed a distinct monophyletic branch with a high bootstrap value (>90%), distantly related to all known HIV-1 genotypes. Recombination analysis showed that CRF120_0107 was composed of the predominant HIV-1 strains in China: CRF01_AE and CRF07_BC. Further subregional phylogenetic analysis was performed that possible parental lineages of CRF07_BC segments (8, 9, and 10) belonged to the CRF07_BC men who have sex with men cluster (MSM cluster), other CRF01_AE segments also mainly belonged to MSM Cluster (such as CRF01_AE Cluster 5). Bayesian analysis results inferred that CRF120_0107 placed its emergence in Shenzhen approximately between 2009-2011. The appearance of CRF120_0107 further highlights that more and more HIV-1 SGR strains containing CRF01_AE and CRF07_BC will be more generated frequently and will most likely be more conducive to accelerating the spread of HIV in China. This highlighted it is necessary to monitor MSM high-risk individuals with HIV-1 CRF01_AE and CRF07_BC dual infection to prevent the generation of CRF01_AE/CRF07_BC recombinant strains, thus reducing the possibility of HIV-1 genotype resistance and the complexity of treatment in China.

Background

The co-circulation of multi-strains of the human immunodeficiency virus (HIV) facilitates the generation of various inter-genotype unique recombinant forms (URFs) and novel circulating recombination forms (CRFs)[1] which will have an important impact on HIV genetics and evolution. The main characteristics of HIV are extremely high genetic variation and high-speed replication ability. Up to now, 118 CRFs and abundant URFs have been reported worldwide constantly and all of them have been included in the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://www.hiv.lanl.gov/). Similarly, in China, the distribution of HIV-1 genotypes has changed dramatically over time, and especially two CRFs, CRF01_AE and CRF07_BC, have led to a national or regional HIV-1 pandemic in recent years[2].

Shenzhen, as one of the four central cities in the Guangdong-Hong Kong-Macao Greater Bay Area (GBA), is located along the coast of Guangdong Province in South China, where the epidemic trend of HIV-1 has become increasingly complicated. It has been found that since 2010, the most prevalent strain among men who have sex with men (MSM) infected with HIV-1, CRF01_AE has been replaced by CRF07_BC in Shenzhen[3]. In addition, novel HIV-1 second-generation recombinant (SGR) strains with CRF01_AE or CRF07_BC as the backbone have also been discovered in Shenzhen[4], suggesting that CRF01_AE and CRF07_BC are undergoing higher frequency recombination during the HIV-1 epidemic, and adaptive mutation and evolution occurred constantly. In this study, we analyzed the characteristics and
evolutionary history of a newly emerging HIV-1 CRF120_0107 composed of CRF01_AE and CRF07_BC in Shenzhen, Guangdong Province, China.

**Methods**

Three patients (Patient ID: LS14873, LS15083, and LS16846) infected with the novel HIV-1 recombinant strains were included in this study. All participants signed written informed consent. They were infected through sexual transmission in Shenzhen, China. All of them were males with ages between 31 and 38 years old. There was no direct epidemiological linkage among these individuals. The demographic and epidemiological data of these patients are shown in Table 1.

| Patient ID | Sampling year | Region | Gender | Age | Marital status | Risk factor | Accession no. |
|------------|---------------|--------|--------|-----|----------------|-------------|---------------|
| LS15083    | 2015          | SZ, GD | Male   | 31  | unmarried      | MSM         | OK662598      |
| LS16846    | 2015          | SZ, GD | Male   | 32  | married        | SH          | OK662596      |
| LS14873    | 2015          | SZ, GD | Male   | 38  | unmarried      | MSM         | OK662597      |
| 13SJ011    | 2013          | SJ, HE | Male   | UK  | UK             | SH          | KX198574      |

*SZ: Shenzhen city; GD: Guangdong province; SJ: Shijiazhuang city; HE: Hebei province; UK: Unknown; MSM: men who have sex with men; SH: Heterosexual.

To obtain the near full-length genome (NFLG) sequences of the three patients in Shenzhen, viral RNAs were extracted from plasma samples and then reverse transcribed to cDNA. Two halves genomic segments with 1000bp overlapping regions were amplified by nested polymerase chain reaction (PCR) with Platinum™ Taq DNA Polymerase High Fidelity (Invitrogen) as described before[5] and sequenced. All sequence fragments of each sample were assembled by using the ContigExpress software. Sequence quality analysis was carried out by the LANL HIV Sequence Database tool: quality control. Finally, two NFLG sequences about 8800 bp (LS14873: 8877bp, HXB2: 762-9613nt; LS16846: 8894bp, HXB2: 762-9613nt) and a 5′-half genome sequence (LS15083: 4271bp, HXB2: 790-5096nt) were obtained. These sequences were submitted to GenBank and the accession numbers are OK662596-OK662598.

In order to search for the sequences with the highest similarity (> 95%), HIV Basic Local Alignment Search Tool (BLAST) was performed using all sequences obtained in this study as the query set. Another patient (Patient ID: 13SJ011) was a heterosexual (SH) male in Shijiazhuang of Hebei province, China, the sequence named 13SJ011 (2630bp, HXB2:790-3455nt, access number: KX198574) was obtained and exhibited a high degree of genetic similarity with three strains in this study. Then, these 4 sequences were further aligned with HIV-1 reference sequences (major pure subtypes, CRFs, and HIV-1 O Group) by Multiple Alignment using Fast Fourier Transform (MAFFT) online service[6]. NFLGs maximum-likelihood (ML) phylogenetic tree was produced by PhyML 3.0 (maximum-likelihood phylogeny) under the General
Time Reversible (GTR) and Gamma distributed with Invariant sites (G+I) model with approximate Shimodaira-Hasegawa-like likelihood ratio test (aLRT SH-like) values[7].

To demonstrate possible unique inter-subtype recombinant structures, the jumping profile hidden Markov model (jpHMM) (http://jphmm.gobics.de/submission_hiv)[8] and Similarity Plotting (SimPlot v3.5.1)[9] were employed to perform recombination analysis. Recombinant pattern of the novel recombinant genome was generated by Recombinant HIV-1 Drawing Tool. Subregion phylogenetic tree was constructed by the neighbor-joining method based on the Kimura 2-parameter model with 1000 bootstrap replicates in Molecular Evolutionary Genetics Analysis (MEGA6) to verify the origins of these subregions[10]. The program TempEst v1.5.1 (TEMPoral Exploration of Sequences and Trees) was performed to evaluate the sampling time signal for data firstly (R-square > 0.7)[11]. Next, the time of the most recent common ancestor (tMRCA) of recombinant segments was conducted by Bayesian Evolutionary Analysis Sampling Trees (BEAST v1.10)[12] with a GTR+G+I nucleotide substitution model, an uncorrelated lognormal relaxed clock model, a Bayesian Skyline tree prior, 2×10^8 length of chain sampling frequency of 1000. All parameters’ effective sample size (ESS) was greater than 200. All phylogenetic trees were visualized by Interactive Tree Of Life (iTOL) online [13].

Results And Discussion

Phylogenetic analysis of these 4 sequences showed that they did not cluster with any known HIV-1 subtypes/CRFs and formed an independent cluster with high bootstrap value (>90%), indicating that they may belong to a novel genotype (Fig. 1A). And then both the BLAST and recombinant analysis revealed that these strains comprised of HIV-1 CRF01_AE and CRF07_BC segments may be a new CRF with 5 identical breakpoints: 3 CRF01_AE and 3 CRF07_BC segments (Fig. 1B).

Recombination breakpoints were displayed as follows: CRF07_BC (790-1817nt), CRF01_AE (1818-2022nt), CRF07_BC (2023-5759nt), CRF01_AE (5760-6166nt), CRF07_BC (6167-8367nt), CRF01_AE (8368-9411nt) using HXB2 as a reference (Fig. 1C). As shown in Fig. 2, further subregional phylogenetic analysis results indicated that possible parental lineages of CRF07_BC segments (, , and ) belonged to the CRF07_BC MSM cluster. CRF01_AE segments ( and ) also belonged to MSM Cluster (CRF01_AE Cluster 5)[14]. But segment may be too short or insufficient gene similarity, only 13SJ011 sample was located in CRF01_AE Cluster4 (MSM Cluster), other sequences form a cluster by themselves among CRF01_AE Cluster. Subregional phylogenetic analysis revealed that all recombinant segments are closely related to the clusters associated with the sexually transmitted population, especially the MSM population.

Bayesian analysis was performed to estimate the time to the most recent common ancestor (tMRCA). Only the CRF01_AE segments (, , and ) of the two NFLGs (LS14873 and LS16846) were selected for the next time-evolution analysis, because the only CRF01_AE segment ( ) the other two non-NFLG sequences (13SJ011 and LS15083) contain is too short to be detected sufficient time signal by the TempEst software (R-square<0.5). Similarly, we selected (HXB2:790-1817nt) and (HXB2: 2023-3455nt) segments of CRF07_BC recombination segments to do the Bayesian analysis, for the reasons that all sequences
include the 2 segments and the sequence set of these 2 segments reached enough time signals (R-square >0.7). Maximum clade credibility (MCC) trees were reconstructed based on the recombinant regions of CRF01_AE segments (1, 9, and 12) and CRF07_BC segments (2:790-1817nt and 3: 2023-3455nt) respectively. CRF01_AE and CRF07_BC sequences with sampling years were both downloaded from the LANL HIV Sequence Database and obtained from the HIV-infected blood samples collected by our laboratory. All these sequences were subjected to the phylodynamics analysis together with the target sequences in this study. As shown in Fig. 3, the tMRCA of novel recombinant strains for CRF01_AE regions (1, 9, and 12) and CRF07_BC regions (2 and 3: 2023-3455nt) were estimated around 2009.3 (95% highest probability density (HPD): 2007.0-2011.2) and 2011.7 (95%HPD: 2010.8-2012.5), respectively. However, a sequence (Patient ID: 13SJ011) in the CRF07_BC region was estimated to be around 2008.2 (95%HPD: 2007.4-2009.0) and not clustered with other three sequences obtained in Shenzhen, which may be the earlier ancestor of the three Shenzhen sequences, or it may be of distinct origin but sharing the same recombination breakpoints with others. The sequence is too short for us to know for sure. Since all patients have no direct epidemiological relationship and the strains that they infected share the same mosaic structure, according to the HIV-1 CRF nomenclature proposal[15], three recombinant strains in Shenzhen are recognized and included by the HIV database and designated as the new CRF (CRF120_0107). Together with the fact that CRF120_0107 has not previously been described in other countries/regions, the results inferred that CRF120_0107 placed its emergence in Shenzhen approximately between 2009-2011 (Fig. 3).

With the rapid economic development and large population flow base, Shenzhen has great potential for the spread of HIV locally and even across the country. The origin and outbreak of CRF55_01B among MSM in Shenzhen has changed HIV-1 molecular epidemiology patterns in the region, and perhaps across the nation[16, 17]. In addition, since the discovery of CRF01_AE/CRF07_BC recombinants, the MSM population has contributed to the emergence of most URFs _0107 and CRFs_0107 Second-generation CRFs like CRF120_0107 have been identified especially among the MSM population, which has become the special high-risk group of the form and prevalence of the recombinants. At present, it is obvious that the advantages of these strains in the generation and spread among MSM in the region may be the accumulation of power for its high transmission in the future. At the same time, the generation of HIV-1 second-generation recombinant forms (CRF01_AE/CRF07_BC) also acts as an evolutionary force to increase HIV diversity and complexity in this region.

**Conclusion**

In this study, we firstly identified and analyzed the characteristics of the newly emerging HIV-1 CRF120_0107 composed of CRF01_AE and CRF07_BC in Shenzhen, China, and analyzed its evolutionary history. The emergence of CRF120_0107 indicated that more and more strains with CRF01_AE and CRF07_BC as parents form SGR strains, suggesting that CRF01_AE/CRF07_BC SGR strains may have become the dominant strains in China. In view of the prevalence of most URFs _0107 and CRFs_0107 in MSM population, we emphasize that MSM high-risk individuals with CRF01_AE and CRF07_BC dual infection should be closely monitored and treated in advance to prevent the generation of
CRF01_AE/CRF07_BC recombinant strain from causing more widespread transmission and epidemic, so as to avoid increasing the possibility of genotype resistance and the complexity of treatment work.

**Abbreviations**

aLRT SH-like: approximate Shimodaira-Hasegawa-like likelihood; BEAST: Bayesian Evolutionary Analysis Sampling Trees; BLAST: Basic Local Alignment Search Tool; CRFs: circulating recombination forms; ESS: effective sample size; GBA: Guangdong-Hong Kong-Macao Greater Bay Area; GTR: General Time Reversible; G+I: Gamma distributed with Invariant sites; HIV: human immunodeficiency virus; iTOL: Interactive Tree Of Life; jpHMM: jumping profile hidden Markov model; LANL: Los Alamos National Laboratory; MAFFT: Multiple Alignment using Fast Fourier Transform; MCC: Maximum clade credibility; MSM: men who have sex with men cluster; NFLG: near full-length genome; PCR: polymerase chain reaction; PhyML: maximum-likelihood phylogeny; SimPlot: Similarity Plotting; MEGA: Molecular Evolutionary Genetics Analysis; ML: maximum-likelihood; MSM: men who have sex with men; SGR: second-generation recombinant; SH: heterosexual; TempEst: TEMPoral Exploration of Sequences and Trees; tMRCA: the most recent common ancestor; URFs: unique recombinant forms;

**Declarations**

**Acknowledgments**

The authors thank all of the participants and peer workers.

**Author Contributions**

BZ, JH and LL designed the study; BZ and JZ performed the experiments; JZ, CZ and LC collected samples; XR W and YL and participated in sequences assembly. HL and LJ participated in phylogenetic analyses; BZ and XR W participated in recombinant analysis; TL, XL W, BH Z collected the demographic data and HIV reference sequences; BZ, XR W, JL, JH and LL participated in the writing process and manuscript correction. All authors read and approved the final manuscript.

**Funding**

This study was supported by the National Key Research and Development Program of China (2020YFA0907000), the NSFC (81773493, 31800149, 31900157), the State Key Laboratory of Pathogen and Biosecurity (AMMS), the National 13th Five-Year Grand Program on Key Infectious Disease Control (2018ZX10721102, 2018ZX10732101-001-003), Guangdong Basic and Applied Basic Research Foundation (2019B1515120003), San-Ming Project of Medicine in Shenzhen (SZSM201811071).

**Availability of data and materials**

The gene sequences of LS16846, LS14873, and LS15083 were deposited in the GenBank with the accession number: OK662596-OK662598. The sequence 13SJ011 was obtained from Los Alamos
National Laboratory HIV Sequence Database with the accession number: KX198574.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no competing interests.

**References**

1. Gao Q, Feng Y, Gao L, Zhang M, Liu Y, Xiao M, Dong X, Xia X: Genetic characteristics of a New HIV-1 subtype B/C intersubtype circulating recombinant form (CRF118_BC) identified in Yunnan, China. *J Infect* 2021, 83:237-279.

2. Li X, Li W, Zhong P, Fang K, Zhu K, Musa TH, Song Y, Du G, Gao R, Guo Y, et al: Nationwide Trends in Molecular Epidemiology of HIV-1 in China. *AIDS Res Hum Retroviruses* 2016, 32:851-859.

3. Zhao J, Chen L, Chaillon A, Zheng C, Cai W, Yang Z, Li G, Gan Y, Wang X, Hu Y, et al: The dynamics of the HIV epidemic among men who have sex with men (MSM) from 2005 to 2012 in Shenzhen, China. *Sci Rep* 2016, 6:28703.

4. Wang X, Zhao J, Li X, Li H, Zhang Y, Liu Y, Chen L, Zheng C, Jia L, Han J, et al: Identification of a novel HIV-1 second-generation Circulating Recombinant Forms CRF109_0107 in China. *The Journal of infection* 2020.

5. Li L, Liang S, Chen L, Liu W, Li H, Liu Y, Bao Z, Wang Z, Zhuang D, Liu S, Li J: Genetic characterization of 13 subtype CRF01_AE near full-length genomes in Guangxi, China. *AIDS Res Hum Retroviruses* 2010, 26:699-704.

6. Katoh K, Rozewicki J, Yamada KD: MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* 2019, 20:1160-1166.

7. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010, 59:307-321.

8. Schultz AK, Zhang M, Bulla I, Leitner T, Korber B, Morgenstern B, Stanke M: jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res* 2009, 37:W647-651.

9. Zhang C, Ding N, Wei JF: Different sliding window sizes and inappropriate subtype references result in discordant mosaic maps and breakpoint locations of HIV-1 CRFs. *Infect Genet Evol* 2008, 8:693-697.
10. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 2013, 30:2725-2729.

11. Rambaut A, Lam TT, Max Carvalho L, Pybus OG: Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution* 2016, 2:vew007.

12. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A: Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 2018, 4:vey016.

13. Letunic I, Bork P: Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021, 49:W293-w296.

14. Feng Y, He X, Hsi JH, Li F, Li X, Wang Q, Ruan Y, Xing H, Lam TT, Pybus OG, et al: The rapidly expanding CRF01_AE epidemic in China is driven by multiple lineages of HIV-1 viruses introduced in the 1990s. *Aids* 2013, 27:1793-1802.

15. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, et al: HIV-1 nomenclature proposal. *Science (New York, NY)* 2000, 288:55-56.

16. Zai J, Liu H, Lu Z, Chaillon A, Smith D, Li Y, Li X: Tracing the transmission dynamics of HIV-1 CRF55_01B. *Sci Rep* 2020, 10:5098.

17. Zhang D, Zheng C, Li H, Li H, Liu Y, Wang X, Jia L, Chen L, Yang Z, Gan Y, et al: Molecular surveillance of HIV-1 newly diagnosed infections in Shenzhen, China from 2011 to 2018. *J Infect* 2021, 83:76-83.

**Figures**

**Figure 1**

See image above for figure legend

**Figure 2**

Subregion phylogenetic analysis based on CRF01_AE or CRF07_BC segments of HIV-1 CRF120_0107. Subregion neighbor-joining tree was constructed by the Kimura 2-parameter model of nucleotide substitution with 1000 bootstrap replicates in MEGA v6 and visualized by iTOL. Region , , and represent CRF07_BC segments, region , , and represent fragment CRF01_AE segments. Subtype A-D, F-H, J-L, CRF01_AE, CRF07_BC, and HIV-1 O group were used as references. The CRF01_AE or CRF07_BC segments sequences of CRF120_0107 and suspected CRF120_0107 sequence (13SJ011) were marked with “■” and the corresponding sequence labels were also marked with a cyan and pink background respectively. Only bootstrap values≥90% were presented at the nodes of the tree marked with a light purple solid circle.
Figure 3

Evolution analysis based on CRF01_AE or CRF07_BC segments of HIV-1 CRF120_0107. MCC trees show recombinant fragments of CRF01_AE regions (,) and CRF07_BC regions (790-1817nt and 2023-3455nt). Timescale is shown below the tree. The mean tMRCA and 95% highest probability density (HPD) for the CRF120_0107 and suspected CRF120_0107 sequence (13SJ011) nodes are displayed. CRF120_0107 strains and suspected CRF120_0107 strain (13SJ011) are highlighted cyan for the CRF01_AE region and pink for the CRF07_BC region.