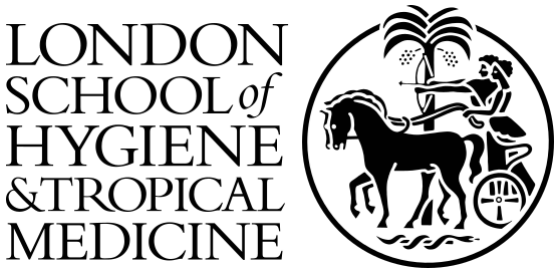


LSHTM Research Online

Yuan, Huangbo; Liu, Zhenqiu; Wu, Xuefu; Wu, Mingshan; Fang, Qiwen; Tully, Damien C; Zhang, Tiejun; (2019) Evolutionary characteristics and genetic transmission patterns of predominant HIV-1 subtypes among MSM in China. International journal of infectious diseases. ISSN 1201-9712 DOI: https://doi.org/10.1016/j.ijid.2019.10.035

Downloaded from: http://researchonline.lshtm.ac.uk/id/eprint/4655262/

DOI: https://doi.org/10.1016/j.ijid.2019.10.035

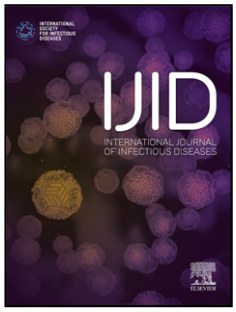
Usage Guidelines:

Please refer to usage guidelines at https://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/

https://researchonline.lshtm.ac.uk

Journal Pre-proof



Evolutionary characteristics and genetic transmission patterns of predominant HIV-1 subtypes among MSM in China

Huangbo Yuan, Zhenqiu Liu, Xuefu Wu, Mingshan Wu, Qiwen Fang, Damien C. Tully, Tiejun Zhang

PII: S1201-9712(19)30429-1

DOI: https://doi.org/10.1016/j.ijid.2019.10.035

Reference: IJID 3811

To appear in: *International Journal of Infectious Diseases*

Received Date: 21 August 2019

Revised Date: 24 October 2019

Accepted Date: 27 October 2019

Please cite this article as: Yuan H, Liu Z, Wu X, Wu M, Fang Q, Tully DC, Zhang T, Evolutionary characteristics and genetic transmission patterns of predominant HIV-1 subtypes among MSM in China, *International Journal of Infectious Diseases* (2019), doi: https://doi.org/10.1016/j.ijid.2019.10.035

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier.

Journal Pre-proof

Original Article

Evolutionary characteristics and genetic transmission patterns of predominant HIV-1 subtypes among MSM in China

Running title：Evolution and transmission of HIV-1 among Chinese MSM

Huangbo Yuan[a], Zhenqiu Liu[a, b], Xuefu Wu[a], Mingshan Wu[a], Qiwen Fang[a], Damien C Tully[c], Tiejun Zhang[a, *]

[a] Department of Epidemiology, School of Public Health, Fudan University, Shanghai, China and Key Laboratory of Public Health Safety (Fudan University), Ministry of Education, China. Postal address: 131 Dong'an Road, Xuhui District, Shanghai City, China.

[b] State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China. Postal address: 2005 Songhu Road, Yangpu District, Shanghai City, China.

[c] Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK. Postal address: Keppel Street, London WC1E 7HT, UK.

* Corresponding Authors: Tiejun Zhang, MD., PhD. Email: tjzhang@shmu.edu.cn; TEL/FAX: +86-21-54237088. Postal address: Room 513 Building FuXing, 131 Dong'an Road, Shanghai, 200032, P.R.China.

Journal Pre-proof

Journal Pre-proof

Highlights

- A greatest expansion of the epidemic among MSM in China occurred between 1999 and 2005.
- CRF01_AE had a higher estimated evolutionary rate and exhibited more sites under positive selection.
- CRF07_BC was more active in interprovince transmission

Abstract

Objectives: Men who have sex with men (MSM) has become one of the major risk groups for HIV-1 infection in China and the predominant subtypes among this population had changed in the last two decades. The objective of this study was to understand the evolutionary characteristics and transmission patterns of HIV-1 dominant strains in the Chinese MSM population.

Methods: We retrieved 4980 published HIV-1 *pol* gene sequences of MSM in China and then conducted comprehensive evolutionary and transmission analyses. Bayesian coalescent-based method and selection pressure analyses were used to reconstruct the time scale and demographic history and estimate other evolutionary parameters. Transmission patterns were characterized by using network analyses.

Results: There were 2546 (51.12%) CRF01_AE, 1263 (25.36%) CRF07_BC, 623 (12.51%) subtype B, accounting for 88.99% of all the sequences. From 2000 to 2016, the prevalence of CRF01_AE stably composed nearly half of all sequences all time (58.33%~45.38%, $P = 0.071$), and CRF07_BC slightly increased from 13.3% to 22.49% ($P < 0.001$), while subtype B dramatically decreased from 41.67% to 9.04% ($P < 0.001$). Demographic reconstruction showed a greatest expansion of the HIV epidemic occurred between 1999 and 2005. CRF01_AE had a higher estimated evolutionary rate (2.97×10^{-3} substitutions/site/year) and exhibited more sites under positive selection (25/351 codons) compared to other subtypes. Network analyses showed CRF07_BC (68.29%, 84/123) had a higher proportion of cross-region networks than that of CRF01_AE (49.1%, 174/354) and subtype B (36.46%, 35/96) ($P < 0.001$).

Journal Pre-proof

Conclusion: The predominant subtypes of HIV-1 in Chinese MSM have different evolutionary characteristics and transmission patterns, which poses a significant challenge to HIV treatment and disease prevention.

Keywords: HIV; MSM; Virus evolution; Transmission

Introduction

Since the first case of a HIV-positive patient reported in 1985 in China, the number of reported HIV/AIDS cases has been increasing annually (Liu et al., 2018). In the meantime, the main drivers of China's HIV epidemic shift considerably, from blood transmission and injecting drug usage (IDU) to sexual transmission, particularly with men who have sex with men (MSM) (Jia et al., 2007, Xiao et al., 2007). Among newly reported HIV infection cases in China, the proportion of MSM increased from 2.5% in 2006 to 25.8% in 2014 (National Health Family Planning Commission of the People's Republic of China, 2015). In some developed areas and cities, MSM account for more than 50% of all HIV cases(Lu et al., 2017, Yang et al., 2017, Zheng et al., 2016). The increasing HIV epidemic among Chinese MSM has now become a key vulnerable population for targeted HIV prevention and intervention strategies.

Human immunodeficiency virus (HIV) is characterized by high genetic variability and extensive heterogeneity, which is caused by its error-prone replication of genome and high rate of recombination. In the last decade, the prevalence of HIV-1 circulating recombinant forms (CRFs) has increased dramatically, and it currently comprises about 20% of all known HIV infections(Robertson et al., 1995, Sharp et al., 1995). CRFs and other recombinants account for approximately 80% of circulating strains in east and southeast Asia, including China (Hemelaar et al., 2019). Furthermore, Chinese MSM generally have more than one sexual partner(Wu et al., 2013) and have a lower rate of condom use(Tang et al., 2018). Due to this high-risk behavior and viral characteristics, the genetic diversity of HIV-1 have become increasingly complex in this population. More CRFs and unique recombinant forms (URFs) have also been detected in this population, which were mainly produced by recombination of the predominant

Journal Pre-proof

subtypes and other strains(Li et al., 2013, Yan et al., 2015). Special attention thus should be given to MSM populations, including the circulating HIV strains among them.

Understanding the evolutionary and transmission history of HIV predominant subtypes among MSM is important for designing effective preventive strategies, and meaningful for future studies in countries where MSM contribute disproportionately to HIV transmission. In the present study, we sought to characterize the evolutionary and transmission characteristics of three major epidemic HIV subtypes (CRF01_AE, CRF07_BC and subtype B) among Chinese MSM applying Bayesian coalescent-based and network methods based on sequences from HIV-1-infected MSM derived from 2000 through 2016. The results illustrate the shifting evolutionary and transmission history of predominant subtypes of HIV among MSM populations in China.

Materials and methods

Data collection

We retrieved all available records of HIV-1 sequences of MSM in China from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov). A total of 8890 sequences were obtained with basic information, such as GenBank accession number, subtype, sampling location and collection date, if provided. For sequences without basic information, we then collected information of each sequence according to the accession number from GenBank database or the original source publication. The *pol* gene region was extracted from these data according to annotation information. And all sequences whose length were less than 1000 bp were removed. In patients with multiple sequences, only the earliest one was retained. Finally, a total of 4980 *pol* gene sequences were kept for later processing.

Sequence alignment and recombination analyses

We aligned all sequences using MAFFT (version 7)(Katoh and Standley, 2013) and manually checked the alignment. The alignment was then edited with MEGA (version 7)(Kumar et al., 2016). All the sequences were manually selected in order to maximize the length and the number of segments for analysis. Segments spanning 1053

Journal Pre-proof

bp of *pol* gene, including the entire protease (PR) and partial reverse transcriptase (RT) regions (nucleotide 2253–3305 by using HXB2 as a calibrator), were finally selected. Intra-subtype recombinant sequences were detected and removed from the dataset utilizing RDP 4 software(Martin et al., 2015).

Selection pressure analyses

To identify codons under positive selection, relative rates of synonymous substitutions per site (*dS*) and nonsynonymous substitutions per site (*dN*) were calculated using the HyPhy program (version 2.2.4)(Pond et al., 2005). The ω ratios (*dN/dS*) of every codon sites were calculated using four different codon-based maximum likelihood approaches, including single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), fast, unconstrained Bayesian approximation (FUBAR), and mixed effects model of evolution (MEME). Codon sites showing evidence of positive selection by at least two of the methods listed above with high statistical significance ($p < 0.1$ or Bayes factor > 50) were considered to be under positive selection. The global ω ratios (*dN/dS*) of entire sequence were also calculated using AnalyzeCodonData package of HyPhy. For all the methods employed for the datasets, the GTR model was used as nucleotide substitution bias model, and trees were inferred by the neighbor-joining method.

Evolutionary analyses and demographic reconstructions

To reduce computing burden and avoid sampling bias due to the heterogeneous number of sequences by location, a down-sampling procedure was performed. Briefly, we stratified the sequences of each province according to the collection year and then randomly sampled at most 3 sequences per year in each province.

We used BEAST (Bayesian evolutionary analysis sampling trees) (version 2.5.1)(Bouckaert et al., 2014) to estimate the evolutionary rate (nucleotide substitutions per site per year, substitutions/site/year). All datasets of the three subtypes were analyzed using HKY nucleotide substitution model for codons partition 1 (position 1+2) and GTR nucleotide substitution model for codons partition 2 (position 3), which were

Journal Pre-proof

selected by jModelTest (version 2.1.10)(Darriba et al., 2012). We used a relaxed uncorrelated lognormal (UCLN) molecular clock model in order to infer the timescale of HIV evolution while accommodating among-lineage rate variation. A Bayesian skyline coalescent tree prior was used to estimate the effective infection population size.

Markov chain Monte Carlo (MCMC) analyses were run for 100 million generations, with parameter values sampled at every 10 000 steps. The resulting log files were imported into the program TRACER v1.7.1 and the first 10% of the output was used as a burn-in. Convergence of the estimates was evaluated with generation vs. log probability plots in TRACER using an effective sample size >150. We report the posterior mean for evolutionary parameters. When reporting substitution rate and the most recent common ancestor (tMRCA) from a relaxed clock model, we give the mean of estimated parameter. The reconstruction of the Bayesian skyline plot was also implemented in TRACER.

Transmission network analyses

The approximately-maximum likelihood phylogenetic tree was estimate in FastTree 2.1.11 using the GTR + G + I nucleotide substitution model. Transmission clusters were extracted from the phylogenetic tree using the software Cluster Picker.(Ragonnet-Cronin et al., 2013) Only those pairs whose intra-cluster maximum pairwise genetic distances less than 3.0% and node support threshold greater than 90% were kept for network analysis. Then the Tamura-Nei 93 pairwise genetic distances of all sequences within the available clusters were calculated. The minimum genetic distances algorithm, as described in a previous study(Li et al., 2016b), was used to define the linkages within a cluster. For visualizing and analyzing network, the network data were processed in the R software utilizing the *ape* and *qgraph* package. To see if drug resistance mutations affect the shape of the network and influence network construction, we then removed the 43 codons associated with drug resistance mutations (CADRM) in *pol* gene (codon 23, 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 76, 82, 83, 84, 85, 88, 90 in PR region, codon 41, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 115, 116, 151, 179, 181, 184, 188, 190, 210, 215, 219, 225, 230 in RT region) and repeated

Journal Pre-proof

this analysis.

Statistical analysis

The linear trend of subtype proportion by year were test using Chi-square test. Differences were tested using Chi-square tests for categorical data and t-tests for continuous variables, respectively. In order to explore the correlation between the number of other provinces that link to the province for each province and the number of migrant people, Pearson's correlation tests were used. The number of migrant people (sum of inflow and outflow population) of each province was collected from the 2010 population census reported by National Bureau of Statistic of China. Statistical significance was defined as $P < 0.05$. All statistical analyses were conducted in R version 3.5.2 software.

Results

Distribution of HIV-1 subtypes among MSM in China

A total of 4980 *pol* gene sequences from HIV infected Chinese MSM were obtained. Among them, the CRF01_AE (2546, 51.12%), CRF07_BC (1263, 25.36%), and subtype B (623, 12.51%) were the predominate strains, accounting for 88.99% of all the sequences. Thus, we further focused on these predominant epidemic HIV-1 subtypes. Meanwhile, intra-subtype recombinant sequences, including 33 CRF01_AE, 47 CRF07_BC and 7 subtype B strains, were detected and removed from the dataset. The final curated *pol* dataset contained 4345 sequences.

The time and geographic distributions of HIV-1 subtypes from this study are described in Table 1 and Figure 1. Briefly, the proportion of HIV subtypes has been changing in the last two decades: the proportion of CRF07_BC slightly increased from 13.3% to 22.49% ($P < 0.001$), CRF01_AE remained constant contributing nearly half of the proportion (58.33%~45.38%, $P = 0.071$), while subtype B decreased from 41.67% to 9.04% ($P < 0.001$). While CRF01_AE was mainly located in the east, north and northeast of China, CRF07_BC was mainly located in southwest and south of China, and subtype B was mainly located in north and central of China.

Journal Pre-proof

Journal Pre-proof

Evolutionary characteristics of the dominant epidemic strains

Selection pressure

The global ω ratios (*dN/dS*) were 0.232, 0.175 and 0.193 for subtype CRF01_AE, CRF07_BC and subtype B, respectively. Nearly 7.1% of codons (25/351, 9 in Pro region and 16 in RT region) in CRF01_AE, 4.6% (16/351, 7 in Pro region and 9 in RT region) in CRF07_BC and 4.0% (14/351, 7 in Pro region and 7 in RT region) in subtype B were under positive selection. (Supplementary table 1)

Evolutionary rates and tMRCA

To reduce computing burden and avoid sampling bias, we sampled 10 datasets of *pol* gene sequences for HIV-1 CRF01_AE, CRF07_BC and subtype B according collection date and sampling location. Each dataset consisted of 215 CRF01_AE subtype sequences, 153 CRF07_BC subtype sequences and 138 subtype B sequences. (Table 2)

Rates of nucleotide substitution were estimated for every datasets using a MCMC method (Supplementary table 2). The mean of estimated median rates of nucleotide substitution of all 10 datasets for each subtype are listed as follows: CRF07_BC was 2.03×10^{-3} substitutions/site/year , subtype B was 2.09×10^{-3} substitutions/site/year , and CRF01_AE was relatively high (2.97×10^{-3} substitutions/site/year) ($P < 0.001$).

With these substitution rates, we estimated the time of the most recent common ancestor for each subtype. The mean of tMRCA of CRF01_AE was estimated at 1987, and that of CRF07_BC was estimated at 1996. However, the probable origin time of subtype B was relatively early (mean of tMRCA: 1972).

Demographic reconstructions

Figure 2 shows the Bayesian skyline plot reconstructing the spread of the three epidemic subtypes of HIV-1 among MSM in China. The CRF01_AE was characterized by a rapid growth period between 1999 and 2005, and stabilized in 2006. The CRF07_BC subtype is characterized by an initial rapid growth period of about 5 years (2000 to 2005), and also reached a stable stage in 2006. The subtype B has a shorter

Journal Pre-proof

growth stage than other subtypes (about 3 years, 2000 to 2003), and it reached a stable stage after 2003. All together, the analyses of all three HIV-1 subtypes shown that the greatest expansion of the epidemic among MSM in China occurred between 1999 and 2005, and the rate of spread reached a plateau after 2006.

Transmission network patterns of the dominant epidemic strains

Of 4345 sequences, 1456 sequences of CRF01_AE were segregated into 354 networks (node size 2-31), 826 sequences of CRF07_BC were segregated into 123 networks (node size 2-68), 335 sequences of subtype B were segregated into 96 networks (node size 2-19). Figure 3 shows the potential transmission clusters and networks of each subtype. The proportion of large networks (above 10 nodes) of CRF07_BC was higher than that of other two subtypes (Table 3). When we removed CADRM, only the large networks of CRF01_AE decreased (Table 3). Among these networks, 49.15% (174/354) CRF01_AE, 68.29% (84/123) CRF07_BC and 36.46% (35/96) subtype B have networks consisting of sequences from different provinces. The proportions of cross-province links of each subtype were 35.1% (387/1102) in CRF01_AE, 44.8% (315/703) in CRF07_BC, and 26.4% (63/239) in subtype B. Both proportions of CRF07_BC above were significantly higher compared with other subtypes.

We also counted the number of other linked provinces in each province , and only provinces who had links to at least 10 other provinces were shown below: in CRF01_AE, Beijing (13), Guangdong (17), Liaoning (10), Shanghai (13); in CRF07_BC, Beijing (10), Chongqing (10), Guangdong (11), Guizhou (11), Henan (10), Shanghai (13), Zhejiang (10); in subtype B, Beijing (10), Shanghai (10). Interestingly, we observed a positive association between linkages to other provinces for each province and the number of migrant people (Pearson correlation coefficient = 0.493, P = 0.044). However, the association was only observed in CRF01_AE after performing stratification (Pearson correlation coefficient = 0.484, P = 0.049). (Figure 4)

Discussion

Journal Pre-proof

Journal Pre-proof

Several nationwide molecular epidemiology survey of HIV in China evidenced that the predominant subtypes of HIV spreading in Chinese MSM population has dramatically changed over the last three decades(Li et al., 2016a, Zhang et al., 2015). The recombinants of HIV-1 (CRF01_AE, CRF07_BC) were becoming the predominant epidemic subtypes. Although several studies have been completed for molecular evolution analyses of HIV among Chinese MSM, most of them were confined to specific areas or specific HIV subtypes(An et al., 2012, Wang et al., 2017, Zhang et al., 2017), so the infection situation could not be depicted as a whole. The present study collected a large number of HIV sequences from 19 Chinese provinces and incorporated the use of multiple evolution methods for investigating the evolution and transmission of HIV-1 dominant strains from MSM. Our research could provide an evolutionary perspective for the molecular characteristics of HIV-1 major subtypes among MSM in China.

Not surprising, we noticed that the subtype proportion of the HIV sequences of Chinese MSM in the last two decades showed the similar trend as previous molecular epidemiology studies(Li et al., 2016a, Zhang et al., 2015). Although we found subtype B was first introduced to China in 1972 and it was known as the predominant strain in the early HIV epidemic in China(Li et al., 2016a), after 2006, it was dramatically decreasing , while CRF01_AE was the most prevalent strain, accounting for half of all sequences in the last decades and CRF07_BC posed a slight increasing trend. A similar phenomenon was also detected in many regions around the world, in which there is a significant displacement of the existing HIV-1 subtype by other new strains, particularly by CRFs (Lau and Wong, 2013). The results of demographic reconstruction were consistent with the trend, which showed a rapid growth period of HIV infection within the population between 1999 and 2005, and CRF01_AE and CRF07_BC had a faster and longer growth period compared with subtype B. The same spreading waves of CRF01_AE and CRF07_BC were also observed by Wang *et al.* and Zhang *et al.*, respectively (Wang et al., 2017, Zhang et al., 2017). These two recombination subtypes have critically contributed to the epidemic of HIV in recent years, and as a result, displaced subtype B infections and have become the predominant strains within this

Journal Pre-proof

specific population. This shift of circulating strains in MSM populations may suggest a replicative and transmission fitness advantage in which a combination of both the host's genetic and virologic factors may play a role in its dissemination. With regards to the contribution of host factors, research conducted in Mekong Delta, Vietnam found a specific HLA pattern associated with weakened immune pressure in local populations which may have facilitated immune evasion of CRF01_AE, and thus escape mutations could accumulate over time to ultimately represent the most prevalent form of the virus (Lazaro et al., 2011). Similar evidence among Chinese populations have not been demonstrated sufficiently and future research in terms of the association between host's genetics and HIV infection are warranted.

HIV is a RNA virus and is characterized by high mutation rate, which is responsible for their enormous adaptive capacity(Elena and Sanjuan, 2005). Furthermore, HIV subtypes normally vary in mutation rate. In the present study, we found the mutation rate of *pol* gene of CRF01_AE was higher than other two subtypes, which was also observed by Zhang *el at.* (Li et al., 2015), indicating CRF01_AE might adapt to changing environmental conditions more readily. The selection pressure analyses showed the *pol* gene of CRF01_AE had more positive selection sites than other two subtypes. The *pol* gene, which functions the generation of viral DNA(Zack et al., 1990), is currently the target of antiviral therapy, and is also a region generating drug resistant mutations(Bennett et al., 2009). Positive selection on these sites as a result of adaptation to a single large environmental change is normally thought to be followed by some degree of fixation of the newly acquired beneficial variants through purifying selection within host or population level(Fu and Akey, 2013, Sabeti et al., 2006). Previous studies reported high prevalence of CXCR4 viruses and fast disease progression in CRF01_AE infections(Chu et al., 2017, Li et al., 2014). Therefore, CRF01_AE might produce more beneficial advantageous mutations on *pol* gene in its evolution process and be advantageous in viral replicative capacity and in the dynamics of resistance acquisition under antiviral therapy selective pressure. These evolutionary characteristics of CRF01_AE will pose a great challenge for the prevention and control efforts in a large number of CRF01_AE-infected MSM in China.

Journal Pre-proof

However, this high adaptive evolution of CRF01_AE may have a negative impact to its onward transmission, which is so called short-sighted evolution, although the high viral load in its acute infection stage is advantageous in transmission to the contrary(Lythgoe et al., 2017). In the network analyses, we found CRF01_AE and CRF07_BC had larger proportion of cross-province networks and links than that of subtype B, indicating these two strains of HIV were more active on cross-regional transmission. Particularly, the proportion in CRF07_BC was extremely high. We also observed a positive association between linkages to other provinces for each province and the number of migrant people in CRF01_AE, indicating the migration had a strong impact on its active cross-regional transmission. Considering the distinct mutation rate and transmission behavior of the two subtypes, we thus hypothesize these two recombinants may have distinct evolution/transmission strategies. CRF07_BC might have higher transmissibility, considering its characteristics of low mutation rate and active cross-regional transmission. This hypothesis needs future researches with more direct evidence to confirm. Given the increasing prevalence of CRF07_BC among MSM nationwide, particularly in Southwest and South China in recent years(Li et al., 2016a), combined with the frequent cross-region transmission, we thus suggest more development prevention and control efforts should be tailored toward to this regional distribution strain. Importantly, the increasing trend of CRF07_BC will pose a challenge in HIV diagnostic laboratories, particularly pertaining to HIV-1 serological tests and RNA assays where the accuracy of results may be influenced by non-B subtype or CRFs sequence variation and the CRF07_BC is rarely included in the standard strains(Luft et al., 2011, Moyo et al., 2015).

Some limitations should be noted here. First, the main limitation of our study is the selection/sampling heterogeneity as all sequences in the study were obtained from a public database. We have conducted a down-sampling procedure to avoid potential sampling bias. Second, sequences in the early stage of HIV outbreak in China were unavailable, which could help to calibrate the molecular clock. Finally, a longer alignment could be better for robust results, but the alignment in our study was relatively short.

Journal Pre-proof

Journal Pre-proof

In conclusion, the study elucidated the molecular evolutionary characteristics of predominant subtypes of HIV circulating among Chinese MSM population. Genetic transmission network analyses further revealed a complexity cross-regional transmission pattern. We thus suggest that this molecular approach, combined with clinical, experimental and public health approaches will help to reveal the phenotype of distinct HIV strains and predict and control the entire HIV epidemic among MSM in China.

Author contribution

Huangbo, Zhenqiu and Tiejun conceptualized the study. Damien, Qiwen, Xuefu and MS contributed to study design. Huangbo, Zhenqiu were involved in data collection. Huangbo analysed the data and wrote the initial draft. Tiejun, Zhenqiu and Damien revised the manuscript before submission, and complemented it with contextual data. All authors read and approved the final manuscript.

Funding

This study was supported by National Natural Science Foundation of China (81772170) and by National Key Research and Development Program of China (No. 2017YFC0211704).

Disclaimer

The funders did not play a role in the design, conduct or analysis of the study, nor in the drafting of this manuscript.

Conflict of interest statement

None declared.

Ethical Approval

Not required.

Journal Pre-proof

Journal Pre-proof

References

2015 China AIDS response progress report. National Health Family Planning Commission of the People's Republic of China; 2015.

An M, Han X, Xu J, Chu Z, Jia M, Wu H, et al. Reconstituting the epidemic history of HIV strain CRF01_AE among men who have sex with men (MSM) in Liaoning, northeastern China: implications for the expanding epidemic among MSM in China. Journal of virology 2012;86(22):12402-6.

Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, et al. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. Plos One 2009;4(3):e4724.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. Plos Comput Biol 2014;10(4):e1003537.

Chu M, Zhang W, Zhang X, Jiang W, Huan X, Meng X, et al. HIV-1 CRF01_AE strain is associated with faster HIV/AIDS progression in Jiangsu Province, China. Sci Rep 2017;7(1):1570.

Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 2012;9(8):772-.

Elena SF, Sanjuan R. Adaptive value of high mutation rates of RNA viruses: Separating causes from consequences. Journal of virology 2005;79(18):11555-8.

Fu W, Akey JM. Selection and adaptation in the human genome. Annual review of genomics and human genetics 2013;14:467-89.

Hemelaar J, Elangovan R, Yun J, Dickson-Tetteh L, Fleminger I, Kirtley S, et al. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. Lancet Infect Dis 2019;19(2):143-55.

Jia Y, Lu F, Sun X, Vermund SH. Sources of data for improved surveillance of HIV/AIDS in China. Se Asian J Trop Med 2007;38(6):1041-52.

Katoh K, Standley DM. MAFFT: multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 2013;30(4):772-80.

Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution 2016;33(7):1870-4.

Lau KA, Wong JJ. Current Trends of HIV Recombination Worldwide. Infectious disease reports 2013;5(Suppl 1):e4.

Lazaro E, Tram LT, Bellecave P, Guidicelli GL, Anies G, Thu HH, et al. Molecular characterization of HIV-1 CRF01_AE in Mekong Delta, Vietnam, and impact of T-cell epitope mutations on HLA recognition (ANRS 12159). PLoS One 2011;6(10):e26244.

Li X, Li W, Zhong P, Fang K, Zhu K, Musa TH, et al. Nationwide trends in molecular epidemiology of HIV-1 in China. AIDS Res Hum Retroviruses 2016a;32(9):851-9.

Li X, Ning C, He X, Yang Y, Xing H, Hong K, et al. Near full-Length genome sequence of a novel HIV Type 1 second-generation recombinant form (CRF01_AE/CRF07_BC) identified among men who have sex with men in Jilin, China. AIDS Res Hum Retroviruses 2013;29(12):1604-8.

Li X, Xue Y, Lin Y, Gai J, Zhang L, Cheng H, et al. Evolutionary dynamics and complicated genetic transmission network patterns of HIV-1 CRF01_AE among MSM in Shanghai, China. Sci Rep 2016b;6:34729.

Li Y, Han Y, Xie J, Gu L, Li W, Wang H, et al. CRF01_AE subtype is associated with X4 tropism and fast HIV progression in Chinese patients infected through sexual transmission. AIDS (London, England) 2014;28(4):521-30.

Journal Pre-proof

Journal Pre-proof

Li Z, Liao L, Feng Y, Zhang J, Yan J, He C, et al. Trends of HIV subtypes and phylogenetic dynamics among young men who have sex with men in China, 2009-2014. Sci Rep 2015;5:16708.

Liu Z, Shi O, Yan Q, Fang Q, Zuo J, Chen Y, et al. Changing epidemiological patterns of HIV and AIDS in China in the post-SARS era identified by the nationwide surveillance system. BMC infectious diseases 2018;18(1):700.

Lu X, Kang X, Liu Y, Cui Z, Guo W, Zhao C, et al. HIV-1 molecular epidemiology among newly diagnosed HIV-1 individuals in Hebei, a low HIV prevalence province in China. PLoS One 2017;12(2):e0171481.

Luft LM, Gill MJ, Church DL. HIV-1 viral diversity and its implications for viral load testing: review of current platforms. Int J Infect Dis 2011;15(10):E661-E70.

Lythgoe KA, Gardner A, Pybus OG, Grove J. Short-sighted virus evolution and a germline hypothesis for chronic viral infections. Trends Microbiol 2017;25(5):336-48.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 2015;1(1):vev003.

Moyo S, Wilkinson E, Novitsky V, Vandormael A, Gaseitsiwe S, Essex M, et al. Identifying Recent HIV Infections: From Serological Assays to Genomics. Viruses 2015;7(10):5508-24.

Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics (Oxford, England) 2005;21(5):676-9.

Ragonnet-Cronin M, Hodcroft E, Hue S, Fearnhill E, Delpech V, Brown AJL, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics 2013;14:317.

Robertson DL, Sharp PM, McCutchan FE, Hahn BH. Recombination in HIV-1. Nature 1995;374(6518):124-6.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science (New York, NY) 2006;312(5780):1614-20.

Sharp PM, Robertson DL, Hahn BH. Cross-species transmission and recombination of 'AIDS' viruses. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 1995;349(1327):41-7.

Tang W, Liu C, Cao B, Pan SW, Zhang Y, Ong J, et al. Receiving HIV Serostatus Disclosure from Partners Before Sex: Results from an Online Survey of Chinese Men Who Have Sex with Men. AIDS and behavior 2018;22(12):3826-35.

Wang X, He X, Zhong P, Liu Y, Gui T, Jia D, et al. Phylodynamics of major CRF01_AE epidemic clusters circulating in mainland of China. Sci Rep 2017;7(1):6330.

Wu Z, Xu J, Liu E, Mao Y, Xiao Y, Sun X, et al. HIV and Syphilis Prevalence Among Men Who Have Sex With Men: A Cross-Sectional Survey of 61 Cities in China. Clinical Infectious Diseases 2013;57(2):298-309.

Xiao Y, Kristensen S, Sun JP, Lu L, Vermund SH. Expansion of HIV/AIDS in China: lessons from Yunnan province. Soc Sci Med 2007;64(3):665-75.

Yan J, Xin R, Li Z, Feng Y, Lu H, Liao L, et al. CRF01_AE/B/C, a novel drug-resistant HIV-1 recombinant in men who have sex with men in Beijing, China. AIDS Res Hum Retroviruses 2015;31(7):745-8.

Yang JZ, Chen WJ, Zhang WJ, He L, Zhang JF, Pan XH. Molecular epidemiology and transmission of HIV-1 infection in Zhejiang province, 2015. Zhonghua liuxingbingxue zazhi 2017;38(11):1551-6.

Zack JA, Arrigo SJ, Weitsman SR, Go AS, Haislip A, Chen ISY. HIV-1 entry into quisecent primary lumphocytes-molecular-analysis reveals a labile, latent viral structure. Cell 1990;61(2):213-22.

Journal Pre-proof

Journal Pre-proof

Zhang L, Wang YJ, Wang BX, Yan JW, Wan YN, Wang J. Prevalence of HIV-1 subtypes among men who have sex with men in China: a systematic review. International journal of STD & AIDS 2015;26(5):291-305.

Zhang M, Jia D, Li H, Gui T, Jia L, Wang X, et al. Phylodynamic analysis revealed that epidemic of CRF07_BC strain in men who have sex with men drove its second spreading wave in China. AIDS Res Hum Retroviruses 2017;33(10):1065-9.

Zheng MN, Ning TL, Gao YJ, Zhao X, Li L, Cheng SH. Molecular epidemiology and transmission of HIV in Tianjin, 2015. Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi 2016;37(8):1142-7.

Journal Pre-proof

Journal Pre-proof

Figure legends

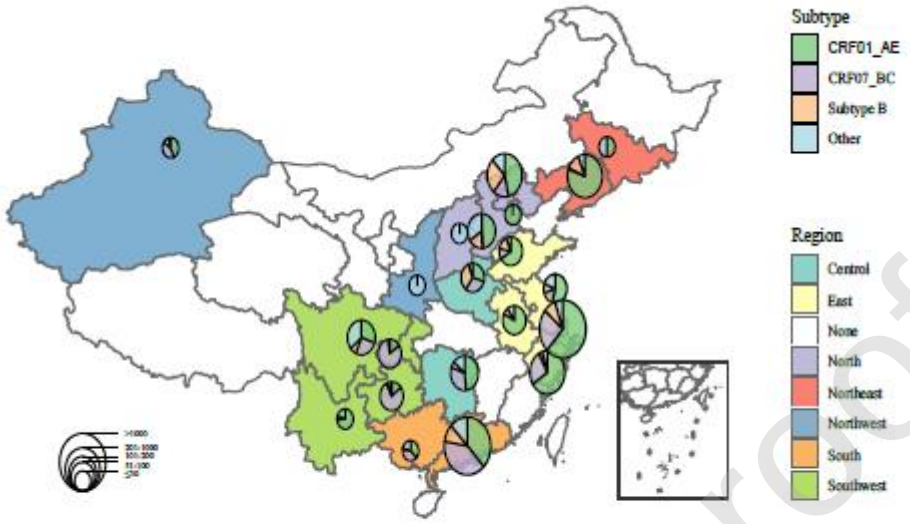


Figure 1. Geographical distribution of HIV-1 subtypes of MSM in China from 2000 to 2016.

Figure 1. Geographical distribution of HIV-1 subtypes of MSM in China from 2000 to 2016.

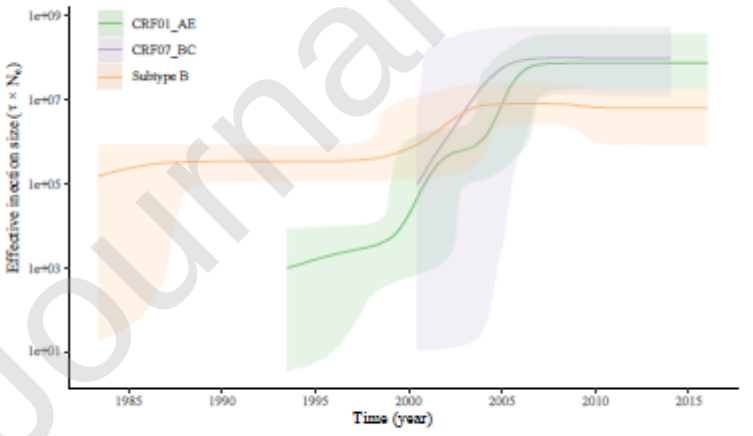


Figure 2. The reconstruction of the Bayesian skyline plot for three strains of MSM in China. Median estimates of the effective number of infections using Bayesian skyline (solid line) are shown together with 95% highest probability density intervals of the Bayesian skyline estimates (translucent area).

Figure 2. The reconstruction of the Bayesian skyline plot for three strains of MSM in China. Median estimates of the effective number of infections using Bayesian skyline (solid line) are shown together with 95% highest probability density intervals of the

Journal Pre-proof

Bayesian skyline estimates (translucent area).

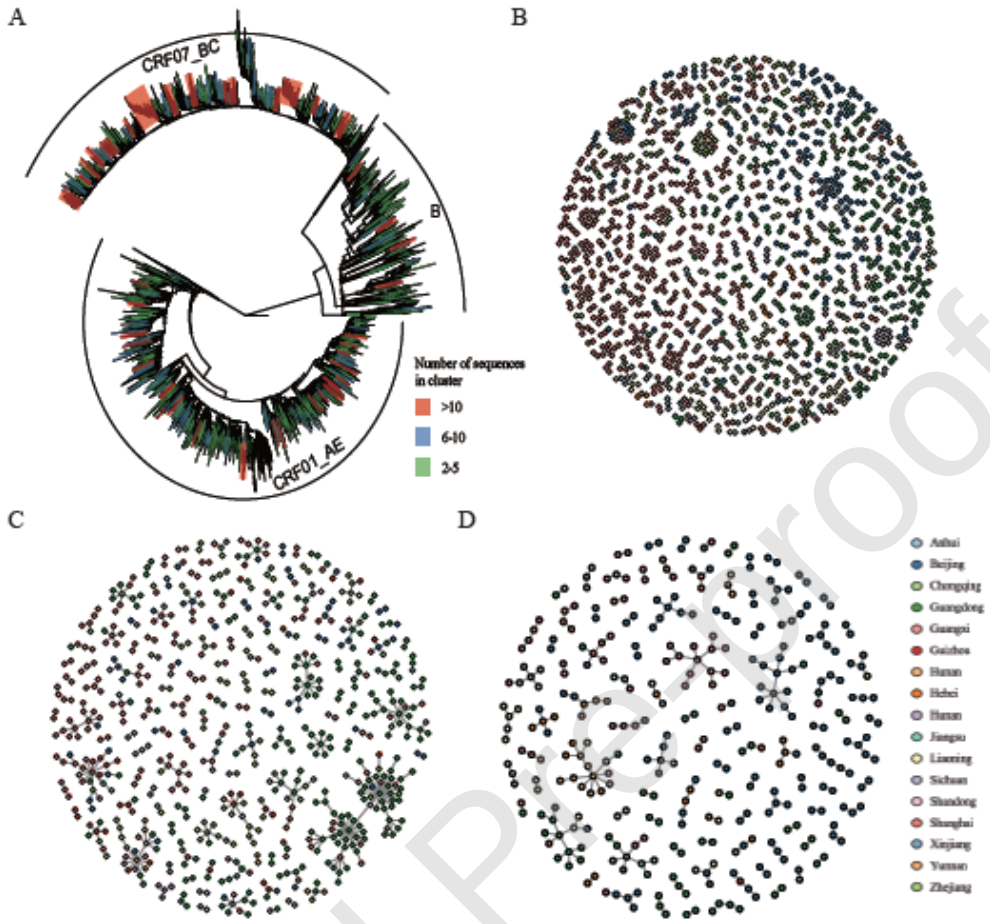


Figure 3. Potential transmission clusters and transmission networks of dominant strains of MSM in China. A) The phylogenetic tree was constructed using approximately maximum likelihood method based on pol region in FastTree 2.1.11. The nucleotide substitution mode was GTR+G+I. B) Clusters of CRF01_AE, C) Clusters of CRF07_BC and D) Clusters of subtype B inferred networks. Various provinces in China are colour coded.

Figure 3. Potential transmission clusters and transmission networks of dominant strains of MSM in China. A) The phylogenetic tree was constructed using approximately maximum likelihood method based on *pol* region in FastTree 2.1.11. The nucleotide substitution mode was GTR+G+I. B) Clusters of CRF01_AE, C) Clusters of CRF07_BC and D) Clusters of subtype B inferred networks. Various provinces in China are colour coded.

Journal Pre-proof

Journal Pre-proof

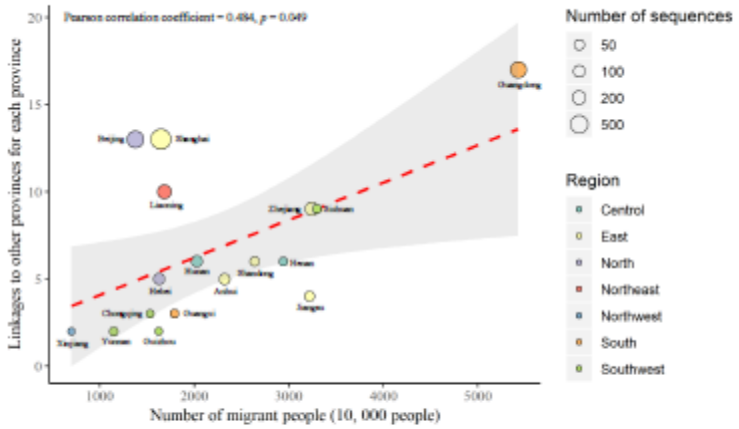


Figure 4. Correlation between the number of other linked provinces for each province and the number of migrant people in CRF01_AE. Regression line (red dash line) together with its 95% confidence interval (grey area) were displayed. Various regions of provinces belonging in China are colour coded.

Figure 4. Correlation between the number of other linked provinces for each province and the number of migrant people in CRF01_AE. Regression line (red dash line) together with its 95% confidence interval (grey area) were displayed. Various regions of provinces belonging in China are colour coded.

Journal Pre-proof

Journal Pre-proof

Table 1. Temporal distribution of predominant HIV-1 subtypes of Chinese MSM from 2000 to 2016

Collection date	CRF01_AE	CRF07_BC	Subtype B	Other	Total
2000-2005	7(58.33)	0(0)	5(41.67)	0(0)	12
2006-2007	60(44.44)	18(13.33)	52(38.52)	5(3.7)	135
2008-2009	449(48.91)	184(20.04)	193(21.02)	92(10.02)	918
2010-2011	909(57.35)	403(25.43)	174(10.98)	99(6.25)	1585
2012-2013	895(48.85)	546(29.8)	154(8.41)	237(12.94)	1832
2014-2016	226(45.38)	112(22.49)	45(9.04)	115(23.09)	498
Trend $\chi 2$	3.26	22.23	135.32	--	--
P value	0.071	<0.001	<0.001	--	--

Number (%), the sequences data is displayed as the number of sequences and its proportion in all sequences in the same period.

Journal Pre-proof

Journal Pre-proof

Table 2. Geographical distribution of sequences in each dataset

Region	City	CRF01_AE (n=215)	CRF07_BC (n=153)	Subtype B (n=138)
Central	Henan	12	11	12
	Hunan	18	16	6
East	Anhui	3	0	3
	Jiangsu	9	9	4
	Shandong	11	5	6
	Shanghai	15	15	15
	Zhejiang	23	11	10
North	Beijing	30	24	30
	Hebei	8	2	8
	Tianjin	1	0	0
Northeast	Jilin	2	0	0
	Liaoning	14	3	12
Northwest	Xinjiang	4	7	3
South	Guangdong	22	19	21
	Guangxi	3	4	1
Southwest	Chongqing	8	9	2
	Guizhou	9	10	2
	Sichuan	9	6	3
	Yunnan	14	2	0

Journal Pre-proof

Journal Pre-proof

Table 3. Network construction of three subtypes in Chinese MSM

Nodes	Complete sequence N (%)				Sequence removed CADRM N (%)		
	CRF01_AE	CRF07_BC	Subtype B		CRF01_AE	CRF07_BC	Subtype B
2-5	285 (80.5)[b]	86 (69.9)[ac]	86 (89.6)[b]		297 (84.4)[b]	81 (71.7)[ac]	81 (86.2)[b]
6-10	46 (13.0)	21 (17.2)[c]	6 (6.3)[b]		45 (12.8)	14 (12.4)	7 (7.4)
>10	**23 (6.5)**[b]	16 (13.0)[ac]	4 (4.2)[b]		**10 (2.8)**[b]	18 (15.9)[a]	6 (6.4)
Total	354	123	96		352	113	94

N (%), quantity of networks and its percentage; CADRM, codons associated with drug resistance mutations; [a, b, c], Compared with CRF01_AE, CRF_BC and subtype B using Chi-square test, respectively (P value less than 0.05);Statistical significance (P value of Chi-square test less than 0.05) were displayed with bold for comparison between counterparts.

Journal Pre-proof