VIRUS EVOLUTION

Virus Evolution, 2022, **8(1)**, 1–12

DOI: https://doi.org/10.1093/ve/veac014

Advance access publication date: 17 February 2022

Research Article

The emergence and transmission dynamics of HIV-1 CRF07_BC in Mainland China

Xingguang Li,[1,2,†,‡,*] Yanping Li,[3,†] Haizhou Liu,[4,†,§] Nídia S. Trovão,[5,**] and Brian T. Foley[6,*]

[1]Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo 315010, China, [2]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo 315000, China, [3]College of Chemistry Biology and Environment, Yuxi Normal University, Yuxi 653100, China, [4]National Virus Resource Center, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China, [5]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA and [6]HIV Databases, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[†]These authors contributed equally to this work.

[‡]https://orcid.org/0000-0002-3470-2196

[§]https://orcid.org/0000-0002-4727-088X

[**]https://orcid.org/0000-0002-2106-1166

*Corresponding authors: E-mail: xingguanglee@hotmail.com; btf@lanl.gov

Abstract

A total of 1155 partial *pol* gene sequences of human immunodeficiency virus (HIV)-1 CRF07_BC were sampled between 1997 and 2015, spanning 13 provinces in Mainland China and risk groups [heterosexual, injecting drug users (IDU), and men who have sex with men (MSM)] to investigate the evolution, adaptation, spatiotemporal and risk group dynamics, migration patterns, and protein structure of HIV-1 CRF07_BC. Due to the unequal distribution of sequences across time, location, and risk group in the complete dataset ('full1155'), subsampling methods were used. Maximum-likelihood and Bayesian phylogenetic analysis as well as discrete trait analysis of geographical location and risk group were carried out. To study mutations of a cluster of HIV-1 CRF07_BC (CRF07-1), we performed a comparative analysis of this cluster to the other CRF07_BC sequences ('backbone_295') and mapped the mutations observed in the respective protein structure. Our findings showed that HIV-1 CRF07_BC most likely originated among IDU in Yunnan Province between October 1992 to July 1993 [95 per cent hightest posterior density (HPD): May 1989–August 1995] and that IDU in Yunnan Province and MSM in Guangdong Province likely served as the viral sources during the early and more recent spread in Mainland China. We also revealed that HIV-1 CRF07-1 has been spreading for roughly 20 years and continues to cause local transmission in Mainland China and worldwide. Overall, our study sheds light on the dynamics of HIV-1 CRF07_BC distribution patterns in Mainland China. Our research may also be useful in formulating public health policies aimed at controlling acquired immune deficiency syndrome in Mainland China and globally.

Key words: CRF07_BC; spatiotemporal and risk group dynamics; Mainland China; CRF07-1; phylogeography.

1. Introduction

The World Health Organization (WHO) reported that the estimated number of people living with human immunodeficiency virus (HIV) (PLWH) in the world in 2020 is around 37.6 million (30.2–45.0 million), of which 1.5 million (1.1–2.1 million) people were newly infected; 0.69 million (0.48–1.0 million) people died of HIV-related causes; and 73 per cent of adults living with HIV received lifelong antiretroviral therapy (ART) (https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics). The WHO also reported that the number of PLWH who know their status in China was 861,000 by the end of 2018, with 718,000 PLWH receiving ART and 677,000 PLWH who have achieved viral load suppression (https://cfs.hivci.org/country-factsheet.html). For key populations, the WHO also reported that HIV prevalence among men who have sex with men (MSM) was 6.9 per cent in China in 2018 and HIV prevalence among injecting drug users (IDU) was 5.9 per cent and also reported 94 per cent condom use with the most recent client among sex workers (https://cfs.hivci.org/country-factsheet.html), mostly through heterosexual behavior (Hetero).

The major HIV-1 subtypes and circulating recombinant forms (CRFs) in China are CRF07_BC, CRF01_AE, CRF08_BC, CRF55_01B, and B' (Thai B) (Vrancken et al. 2020). The first CRF to be discovered in China was CRF07_BC (Su et al. 2000), which was generated by the recombination of HIV-1 subtypes B' and C among IDU (Takebe et al. 2010). It was first reported in 1997 in Xinjiang Province, although its origin can be traced to Yunnan Province (Takebe et al. 2010). HIV-1 CRF07_BC initially spread along the northwestern drug-trafficking route from Yunnan to Sichuan, Ningxia, and Xinjiang (Su et al. 2000) and has since become the most prevalent and widespread subtype/CRF in China. HIV-1 CRF07_BC currently

accounts for more than 40 per cent of all infections in China, mostly among MSM. In the recent decade and last few years, the prevalence of HIV-1 among MSM in China has been increasing, especially in some big cities (Ma et al. 2007; Zhong et al. 2011; Han et al. 2015; Li et al. 2015a; Zhao et al. 2016; Song et al. 2018; Zai et al. 2020; Chen et al. 2021; Fan et al. 2021; Zhang et al. 2021). Previous studies showed that HIV-1 CRF07_BC accounted for more than 41 per cent of all newly diagnosed cases in Shenzhen sampled between January 2011 and December 2018 (Zhang et al. 2021), for more than 50 per cent in Guangxi between January 2013 and December 2018 (from 44 per cent in 2012–3) (Li et al. 2018; Pang et al. 2021), 43 per cent in Shenzhen in 2018 (from 34 per cent in 2011) (Zhang et al. 2021), and 41 per cent in Fujian in 2013 (from 19 per cent in 2012) (Chen et al. 2018) among MSM population. CRF07_BC presents as one of the most concerning subtypes/CRFs of HIV-1, since it is present in all risk groups (Fan et al. 2021; Zhang et al. 2021) and has spread to other countries (Australia, Germany, Indonesia, Kazakhstan, Malaysia, Myanmar, South Korea, the UK, and the USA) (Pang et al. 2012; Lapovok et al. 2014; Chin et al. 2015; Chow et al. 2016; Chen et al. 2017; Dennis et al. 2018; Yebra et al. 2018; Machnowska et al. 2019; Ueda et al. 2019; Di Giallonardo et al. 2020). Despite this current knowledge, we still have an incomplete understanding of why HIV-1 CRF07_BC spread widely and quickly among the MSM population in China.

The present work employs cutting-edge systematic and comparative approaches to investigate the mutations, adaptation, spatiotemporal and risk group dynamics, migration patterns, and protein structure of HIV-1 CRF07_BC based on 1155 partial *pol* gene sequences of HIV-1 CRF07_BC with known sampling year (1997–2015), location (13 provinces), and risk group (Hetero, IDU, and MSM) in Mainland China. For the first time, based on these data and analyses, we show that CRF07-1 has been diversifying in Mainland China for roughly 20 years [e.g. E248V and K249Q in reverse transcriptase (RT) protein region] (Li et al. 2014). The findings discovered in this study may help design effective HIV surveillance and public health prevention strategies focused on clusters of concern as well as provide valuable information to support the development of diagnosis, antiretroviral drugs, and broadly neutralizing antibodies and vaccines to control the spread of HIV-1 CRF07_BC and other subtypes/CRFs in Mainland China and globally.

2. Materials and methods

2.1 Sequence dataset compilation

All publicly available partial *pol* sequences of HIV-1 CRF07_BC (HXB2 genome position 2253–3401, with minimal fragment length of 1,000, known year of sampling, and province of sampling in Mainland China, from Hetero, IDU, and MSM risk groups were retrieved from the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://www.hiv.lanl.gov). Problematic sequences, as defined by LANL (https://www.hiv.lanl.gov/components/sequence/HIV/search/help.html#bad_seq), were removed, and only one sequence per patient was selected before download. We performed HIV-1 Sequence Quality Analysis (https://www.hiv.lanl.gov/content/sequence/QC/index.html) from the LANL site to analyze the quality of all sequences. The genotype assignment of all downloaded sequences was performed by the Recombinant Identification Program (RIP) Analysis (https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html) using RIP v.3.0 (Siepel et al. 1995) from the LANL site. We used Hypermut v2.0 (Rose and Korber 2000) to analyze and detect the apolipoprotein B mRNA editing complex (APOBEC)-induced hypermutation of all sequences (https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html) from the LANL site. We first performed a multiple sequence alignment analysis using MAFFT v7.427 (Katoh and Standley 2013) under an automatic algorithm and subsequently adjusted using BioEdit v7.2.5 (Hall 1999). The final dataset ('full1155') includes 1,155 publicly available partial *pol* sequences of HIV-1 CRF07_BC with known sampling time (1997–2015), locations (13 provinces), and risk groups (Hetero, IDU, and MSM) in Mainland China.

2.2 Nucleotide substitution model selection and maximum-likelihood phylogenetic tree construction

We identified the best-fitting nucleotide substitution model for 'full1155' according to the Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc), Bayesian information criterion (BIC), and Decision Theory Selection (DT) methods, with three substitution schemes (24 candidate models) in jModelTest v2.1.10 (Darriba et al. 2012). A phylogenetic tree was estimated using maximum-likelihood (ML) method for the dataset ('full1155') using RAxML v8.2.12 (Stamatakis 2014) under $GTR+\Gamma+I$ nucleotide substitution model. Node support was simultaneously inferred using 1,000 bootstrap replicates (Felsenstein 1985). The ML phylogenetic tree was visualized and annotated with geographic location, phylogenetic cluster, and risk group using the web-based tool Evolview v2 (He et al. 2016). Genetic distances between and within risk groups were calculated in MEGA v11.0.10 (Tamura, Stecher, and Kumar 2021) using the maximum composite likelihood model (Tamura, Nei, and Kumar 2004). Rate variation among sites was modeled with a gamma distribution. We then performed a regression analysis of root-to-tip genetic divergence obtained from the ML phylogeny against sampling year for 'full1155' using TempEst (Temporal Exploration of Sequences and Trees) v1.5.3 (Rambaut et al. 2016), assuming a strict molecular clock model, to investigate the temporal signal.

2.3 Nucleotide bases or protein amino acids changes for CRF07-1

A subclade of CRF07_BC, previously described as the CRF07-1 (Li et al. 2014) cluster, was identified in the 'full1155' dataset and singled out as the 'CRF07-1_860' subset. This dataset included 860 publicly available partial *pol* sequences of HIV-1 CRF07_BC from three risk groups: Hetero ($n=53$), IDU ($n=4$), and MSM ($n=803$). To spot which changes were responsible for the branch between the previously described CRF07-1 (Li et al. 2014) cluster (mostly among MSM population; named 'CRF07-1_860') and the other HIV-1 CRF07_BC sequences (mostly circulating in the IDU population; named 'backbone_295'), we performed a simple consensus maker for 'CRF07-1_860' and 'backbone_295', respectively, using common consensus conventions (https://www.hiv.lanl.gov/content/sequence/CONSENSUS/SimpCon.html) from the LANL site and subsequently visualized the consensus nucleotide and amino acid sequence changes for 'CRF07-1_860' and 'backbone_295' using Jalview v2.11.1.4 (Waterhouse et al. 2009). To view three-dimensional structures of amino acid sequence changes for 'CRF07-1_860' and 'backbone_295', homology modeling for the three-dimension structure of RT protein was carried out using the SWISS-MODEL (Waterhouse et al. 2018). The visualization, annotation, and rendering of the protein structure were performed in PyMOL v2.4.0 (Schrödinger) using the '1vrt.1.B' protein structure as a model template. Next, we searched the RT protein CTL/CD8+ epitopes for 'CRF07-1_860' and 'backbone_295' using POL CTL/CD8+ Epitope Map

(https://www.hiv.lanl.gov/content/immunology/maps/ctl/Pol.html) from the LANL site.

2.4 Subsampling sequence dataset

To mitigate potential sampling bias in Bayesian inference estimates caused by over-sampling in the dataset ('full1155'), we randomly subsampled sequences based on year of sampling, province, and/or risk group. We subsampled five sequences per geographic location per year in the 'full1155', giving rise to 'locdate163'. We also subsampled five sequences per risk group per year in the 'full1155' to generate 'riskdate133'. Finally, we subsampled five sequences per geographic location per risk group per year in the 'full1155' to generate 'locrisk229'. Therefore, in the subsequently analyses, we can compare Bayesian inference estimates among the three datasets ('riskdate133', 'locdate163', and 'locrisk229'), for instance, the evolutionary rate, the time to the most recent common ancestor (TMRCA), and the past population dynamic. We can also compare discrete trait analyses, for instance, discrete trait analysis of geographic location between 'locdate163' and 'locrisk229' and discrete trait analysis of risk group between 'riskdate133' and 'locrisk229'. We also investigated the temporal signal of the three datasets ('riskdate133', 'locdate163', and 'locrisk229') using TempEst v1.5.3 (Rambaut et al. 2016), which allows comparison of the TMRCA and evolutionary rate estimates across the three datasets ('riskdate133', 'locdate163', and 'locrisk229') and among different methods (linear regression inference and Bayesian inference).

2.5 Bayesian coalescent phylogeny construction

We employed a Bayesian phylogenetic approach to estimate TMRCA and the evolutionary rate for the three datasets ('riskdate133', 'locdate163', and 'locrisk229') under a GTR + Γ + I nucleotide substitution model with an uncorrelated lognormal relaxed molecular clock (UCLN) model (Drummond et al. 2006), which allows each branch of the tree to have its unique evolutionary rate, independent of the evolutionary rate of neighboring branches, and a Bayesian skyline coalescent tree prior (Drummond et al. 2005), which estimates the past population dynamics of fast-evolving pathogens over time, implemented in BEAST (Bayesian Evolutionary Analysis by Sampling Trees) v1.10.4 (Suchard et al. 2018), a cross-platform software for Bayesian analysis of genetic sequence data using Markov chain Monte Carlo (MCMC) (Yang and Rannala 1997) framework to allow the estimation of time-sampled phylogenetic trees. A non-informative continuous-time Markov chain (CTMC) reference prior (Ferreira and Suchard 2008) was used for the molecular clock rate. Bayesian analyses were run using BEAGLE v3.1.2 (Ayres et al. 2012) for accelerated, parallel likelihood evaluation. The molecular clock was calibrated using the tip-dating method. MCMC chains were run for 500, 800, and 800 million states with sampling every 50,000, 80,000, and 80,000 states, for 'riskdate133', 'locdate163', and 'locrisk229' datasets, respectively, to ensure adequate mixing of all model parameters (including the trees). The convergence of the MCMC chains was inspected using Tracer v1.7.1 (Rambaut et al. 2018). All parameters of interest achieved convergence as determined by effective sample size $\geq$200. A maximum clade credibility (MCC) summary tree was generated using TreeAnnotator v1.10.4 (Drummond et al. 2012) from BEAST v1.10.4 (Suchard et al. 2018) software package after discarding the first 10 per cent as burn-in. This MCC summary tree was visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Empirical trees (consisting of 9,000 time-calibrated trees) evenly sampled from the posterior distributions were generated using LogCombiner v1.10.4 (Drummond et al. 2012) from BEAST v1.10.4 (Suchard et al. 2018) software package after discarding the first 10 per cent as burn-in for subsequent discrete trait analysis for the three datasets ('riskdate133', 'locdate163', and 'locrisk229'), respectively. We also performed Bayesian model selection through marginal likelihood estimation (MLE) using path-sampling (PS) and stepping-stone sampling (SS) (Gelman and Meng, 1998; Baele et al. 2012, 2013) to confirm the best-fitting molecular clock model for each of the three datasets ('riskdate133', 'locdate163', and 'locrisk229'). We explored two clock models (a strict clock and an UCLN (Drummond et al. 2006)) under a GTR + Γ + I nucleotide substitution model and a Bayesian skyline coalescent tree prior (Drummond et al. 2005), implemented in BEAST v1.10.4 (Suchard et al. 2018). The molecular clock rate was set with a CTMC reference prior (Ferreira and Suchard 2008). The Bayesian inferences for each of the three datasets ('riskdate133', 'locdate163', and 'locrisk229') were run for 100 million states and sampled every 10,000th states. We employed MLE through PS and SS (Gelman and Meng, 1998; Baele et al. 2012, 2013) by running 100 path steps, each comprising 1 million states, sampling every 1,000th states, with power posteriors determined from evenly spaced quantiles of a beta (0.3, 1.0) distribution (Xie et al. 2011).

2.6 Discrete trait analysis

Using the samples of 9,000 empirical trees generated for the three datasets ('riskdate133', 'locdate163', and 'locrisk229') from the Bayesian phylogenetic analyses described above, we modeled the phylogenetic history of geographic location and risk group. We treated geographic location and risk group as discrete evolutionary traits (Lemey et al. 2009). We coded geographic locations as Beijing, Guangdong, Guizhou, Hebei, Henan, Jiangsu, Liaoning, Ningxia, Shanghai, Sichuan, Xinjiang, Yunnan, and Zhejiang. The risk group trait was coded as Hetero, IDU, and MSM. To do this, we modeled two types of discrete traits (e.g. geographic location and risk group) as a diffusion process among discrete states for the dataset ('locrisk229') and one of the two types (e.g. geographic location or risk group) for the two datasets ('locdate163' and 'riskdate133'), respectively, in BEAST v1.10.4 (Suchard et al. 2018) with BEAGLE v3.1.2 (Ayres et al. 2012) for MCMC chains of 100, 150, and 200 million iterations, sampling every 10,000th, 15,000th, and 20,000th states of all parameters for 'riskdate133', 'locdate163', and 'locrisk229' datasets, respectively, and sampled 10,000 trees for each of the three datasets. The expected number of discrete trait changes among the branches of the posterior tree distribution was jointly estimated using a 'robust counting' approach (Minin and Suchard 2008a,b; O'Brien et al. 2009) implemented in BEAST v1.10.4 (Suchard et al. 2018). Diffusion among discrete traits was modeled using a non-reversible (or asymmetric) CTMC (Lemey et al. 2009) with an approximated CTMC conditional reference prior to the overall rate scaler and a uniform prior distribution (Drummond et al. 2012). Specifically, we inferred on a branch-by-branch basis the history of discrete trait changes between each pair of geographic location and/or each pair of risk groups. We used a Bayesian stochastic search variable selection procedure with a hierarchical prior on geographic location and risk group indicators (0–1) that allows the CTMC rates to reduce to zero with some probability to identify a subset of well-supported transition events and infer the geographic location and/or risk group states of the ancestral nodes of the trees. Well-supported transition events were identified by Bayes factor (BF) with a cut-off of 3 using SpreaD3 v0.9.7 (Bielejec et al. 2016). The well-supported transition events over

Table 1. Geographic location, year of sampling, and risk factor for HIV-1 CRF07_BC strains used in the present study.

Geographic location	Sampling year	*N*	Risk factor[a]		
			Hetero	IDU	MSM
Beijing	2007–2010	103	13	36	54
Guangdong	2006–2012	554	73	90	391
Guizhou	2007	1			1
Hebei	2007–2015	11	3		8
Henan	2010	4			4
Jiangsu	2007	1	1		
Liaoning	2000–2009	8	2	3	3
Ningxia	2007	1		1	
Shanghai	2009–2013	328			328
Sichuan	2006	6		6	
Xinjiang	1997–2015	33	26	6	1
Yunnan	1997–2007	14	3	11	
Zhejiang	2004–2013	91	31	11	49
Total		1,155	152	164	839

[a]Risk factor: Hetero, heterosexual; IDU, injecting drug user; MSM, men having sex with men.

time were summarized using a Perl script (designated as 'collect_times'), which can be downloaded from the BEAST website (http://beast.community/tutorials/markov_jumps_rewards/files/collect_times). We used R v4.1.0 (Team 2021), which is a free software environment for statistical computing and graphics, to summarize the posterior probability distribution for well-supported transition events and used the ggplot2 package (Wickham 2016) to plot the estimated number of changes over time to and from a particular geographic location and risk group.

3. Results

3.1 Social-demographic characteristics of this dataset

We investigated the distinct geographic location and risk group patterns of spread of HIV-1 CRF07_BC in Mainland China using *protease* (*prot*) and *RT* coding regions sequenced from 1,155 strains sampled between 1997 and 2015, from three risk groups: Hetero ($n = 152$), IDU ($n = 164$), and MSM ($n = 839$). The 'full1155' sequences were obtained from 13 provinces in Mainland China: Beijing ($n = 103$), Guangdong ($n = 554$), Guizhou ($n = 1$), Hebei ($n = 11$), Henan ($n = 4$), Jiangsu ($n = 1$), Liaoning ($n = 8$), Ningxia ($n = 1$), Shanghai ($n = 328$), Sichuan ($n = 6$), Xinjiang ($n = 33$), Yunnan ($n = 14$), and Zhejiang ($n = 91$) (Table 1). The main risk group of the dataset ('full1155') was MSM (839/1,155, 72.6 per cent). 'Full1155' was primarily sampled from Guangdong (554/1,155, 48.0 per cent) and Shanghai (328/1,155, 28.4 per cent). The geographic location and risk group distributions of HIV-1 CRF07_BC over time for 'full1155' are shown in Fig. S1. Of note, the sequence distributions of the 'full1155' through time, geographic location, and risk group were uneven; the sequences of the dataset ('full1155') were mainly distributed between 2007 and 2013 (1,079/1,155, 93.4 per cent). We also showed the risk group distribution among 13 provinces in Mainland China in Fig. 1.

3.2 ML phylogenetic analyses

For 'full1155', the best-fit model was a GTR + Γ + I nucleotide substitution model based on the three substitution schemes (i.e. 24 candidate models) according to the AIC, AICc, BIC, and DT methods using jModelTest v2.1.10 (Darriba et al. 2012) and was thus used in subsequent Bayesian phylogenetic analyses. ML phylogenetic analysis of the 'full1155' dataset identified a large cluster (previously designated as CRF07-1) (Li et al. 2014) with bootstrap support of 94 per cent (Fig. 2). The sequences in the 'CRF07-1 _860' subset were collected in nine provinces in Mainland China: Beijing ($n = 48$), Guangdong ($n = 429$), Guizhou ($n = 1$), Hebei ($n = 8$), Henan ($n = 4$), Liaoning ($n = 3$), Shanghai ($n = 311$), Xinjiang ($n = 1$), and Zhejiang ($n = 55$) (Table 2). The main risk groups of the 'CRF07-1_860' were MSM (803/860, 93.4 per cent), in accordance with the risk group distribution results from the whole dataset ('full1155'). The geographic location and risk group distributions of HIV-1 CRF07-1 over time for 'CRF07-1_860' are shown in Fig. S2. The distributions of 'CRF07-1_860' over time, geographic location, and risk group were uneven with this subset being predominantly found in Guangdong and Shanghai, and it emerged from the main CRF07_BC clade later in time, spreading between 2009 and 2013, mostly among MSM. We estimated the evolutionary divergence within and between each of the three risk groups (IDU, MSM, and Hetero). MSM and Hetero had the smallest (1.70 per cent) and largest (3.29 per cent) genetic distances, respectively. The smallest genetic distance separated MSM and Hetero (3.03 per cent), whereas the largest was between MSM and IDU (3.44 per cent).

3.3 Mutation analysis of CRF07-1

We observed nucleotide base mutations between the major HIV-1 CRF07_BC clade ('backbone_295') and CRF07-1 cluster ('CRF07-1_860'), as shown in html. S1 and html. S3, namely 5 in *Prot* (C30T, C33T, C201T, C207T, and C285T) and 13 in *RT* (C12T, T34C, G82A, A94G, G125A, T168C, G198A, G297A, A324G, C480T, T537C, A743T, and A745C). However, as shown in html. S1–html. S4, only four nucleotide base mutations in *RT* (G82A, A94G, A743T, and A745C) were non-synonymous and could change their amino acids (E28K, K32E, E248V, and K249Q, respectively). The amino acid mutations E28K and K32E are far from the catalytic core of the RT protein in 3D structure (Fig. S3) and not in a site that is often targeted by CTL/CD8+ epitopes (https://www.hiv.lanl.gov/content/immunology/maps/ctl/Pol.html). However, the amino acid mutations E248V and K249Q are at the catalytic core of the RT protein in 3D structure (Fig. S3) and are targeted by at least 12 CTL/CD8+ epitopes (B7, B57, B58, B*5801, B*5703, B*5702, B*5701, B*5301, B*3508, B*3503, B*3502, and B*3501) (https://www.hiv.lanl.gov/content/immunology/maps/ctl/Pol.html).

3.4 Temporal signal analysis

Root-to-tip linear regression analysis of genetic divergence against year of sampling showed that the phylogeny of HIV-1 CRF07_BC from 'full1155' had a moderate positive temporal signal ($R^2 = 0.35$; correlation coefficient = 0.59) using the best-fitting root method, thus suggesting a clock-like pattern of molecular evolution (Fig. 3). Therefore, 'full1155' is suitable for phylogenetic molecular clock analysis in BEAST v1.10.4 (Suchard et al. 2018). The evolutionary rate estimate of HIV-1 CRF07_BC for 'full1155' amounts to 2.48×10^{-3} substitutions per site per year, and the origin of HIV-1 CRF07_BC for 'full1155' is estimated to be approximately in 1 May 1993, which in line with previous estimates for HIV-1 CRF07_BC (Feng et al. 2016). The TMRCA for the other datasets ('riskdate133', 'locdate163', and 'locrisk229') were generally estimated to be older, but still between mid- to late-1980s. The respective evolutionary rates presented higher discrepancies and were estimated to be 1.7–2.7 times lower than that of the 'full1155' (Table S1).

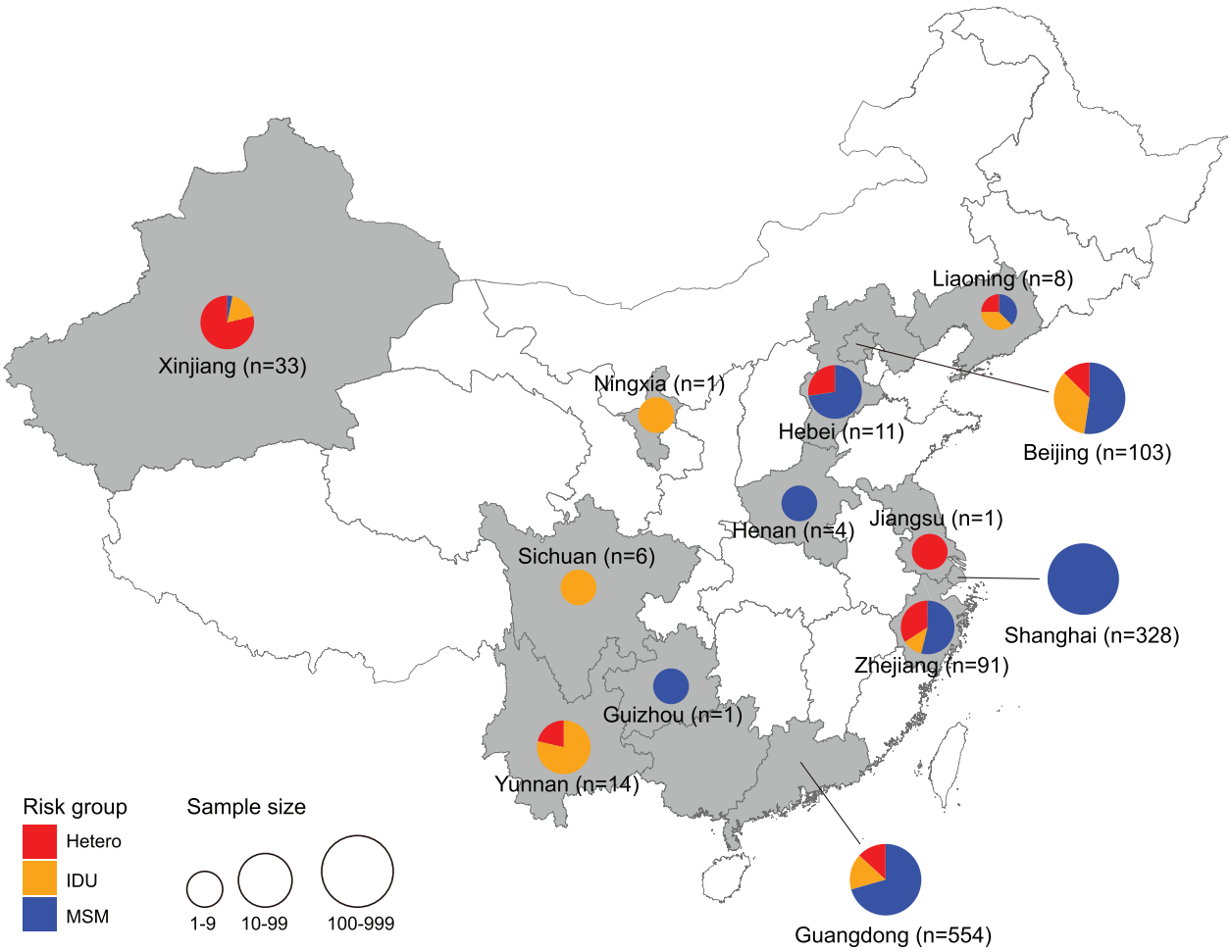

Figure 1. Geographic location distribution of HIV-1 CRF07_BC in Mainland China using 'full1155'. The three risk groups are color-coded, as shown on the left.

3.5 Bayesian phylogenetic analyses

We estimated that an UCLN molecular clock model is a better fit than a strict molecular clock model for all three datasets ('riskdate133', 'locdate163', and 'locrisk229') based on the MLEs of PS and SS (Gelman and Meng, 1998; Baele et al. 2012, 2013) (Table S2). Based on Bayesian time-scaled phylogenetic inference, we estimate the date of the common ancestor of HIV-1 CRF07_BC to be around 16 November 1992 [95 per cent Bayesian highest posterior density (95 per cent HPD): 23 December 1989 to 17 February 1995] for 'riskdate133', 9 July 1993 (95 per cent HPD: 30 October 1990 to 29 July 1995) for 'locdate163', and 6 October 1992 (95 per cent HPD: 30 May 1989 to 27 August 1995) for 'locrisk229' (Table S3). Even though these estimates were not consistent with the dataset's respective root-to-tip regression results using TempEst v1.5.3 (Rambaut et al. 2016), they were closer to the root-to-tip regression results estimated for 'full1155' (Table S1). We also estimated the evolutionary rate of HIV-1 CRF07_BC to be around 1.42×10^{-3} substitutions per site per year (95 per cent HPD: 1.17×10^{-3}–1.69×10^{-3}) for 'riskdate133', 1.46×10^{-3} substitutions per site per year (95 per cent HPD: 1.15×10^{-3}–1.75×10^{-3}) for 'locdate163', and 1.54×10^{-3} substitutions per site per year (95 per cent HPD: 1.20×10^{-3}–1.89×10^{-3}) for 'locrisk229' (Table S3). The estimates and respective 95 per cent HPD intervals for 'riskdate133' and for 'locrisk229' were consistent with the root-to-tip regression results using TempEst v1.5.3 (Rambaut et al. 2016) (Table S1). However, the 95 per cent HPD interval for 'locdate163' did not cover the estimate obtained with the root-to-tip regression results using TempEst v1.5.3 (Rambaut et al. 2016) (Table S1). In addition, we also estimated the date of the most recent common ancestor of HIV-1 CRF07-1 to be around 21 April 2000 (95 per cent HPD: 5 May 1997 to 21 January 2003) for 'riskdate133', 2 March 2001 (95 per cent HPD: 14 March 1998 to 23 September 2003) for 'locdate163', and 18 January 2000 (95 per cent HPD: 1 March 1997 to 17 December 2002) for 'locrisk229'.

We further estimated past population dynamics of HIV-1 CRF07_BC for 'riskdate133', 'locdate163', and 'locrisk229' datasets using a Bayesian skyline plot (BSP) analysis, which reflects the changes in effective population size (*Ne*) over time. The dynamic of the *Ne* for 'riskdate133', 'locdate163', and 'locrisk229' showed at least two distinct phases between 2004 and 2010, a declining phase followed by an increasing phase (Fig. S4).

3.6 Discrete trait dispersal of HIV-1 CRF07_BC in Mainland China

We sought to investigate the spatiotemporal spread and risk group changes of HIV-1 CRF07_BC in Mainland China using an asymmetric discrete trait diffusion model. First, we investigated the geographic location and risk group origins of HIV-1 CRF07_BC. Our discrete trait analysis revealed that the most

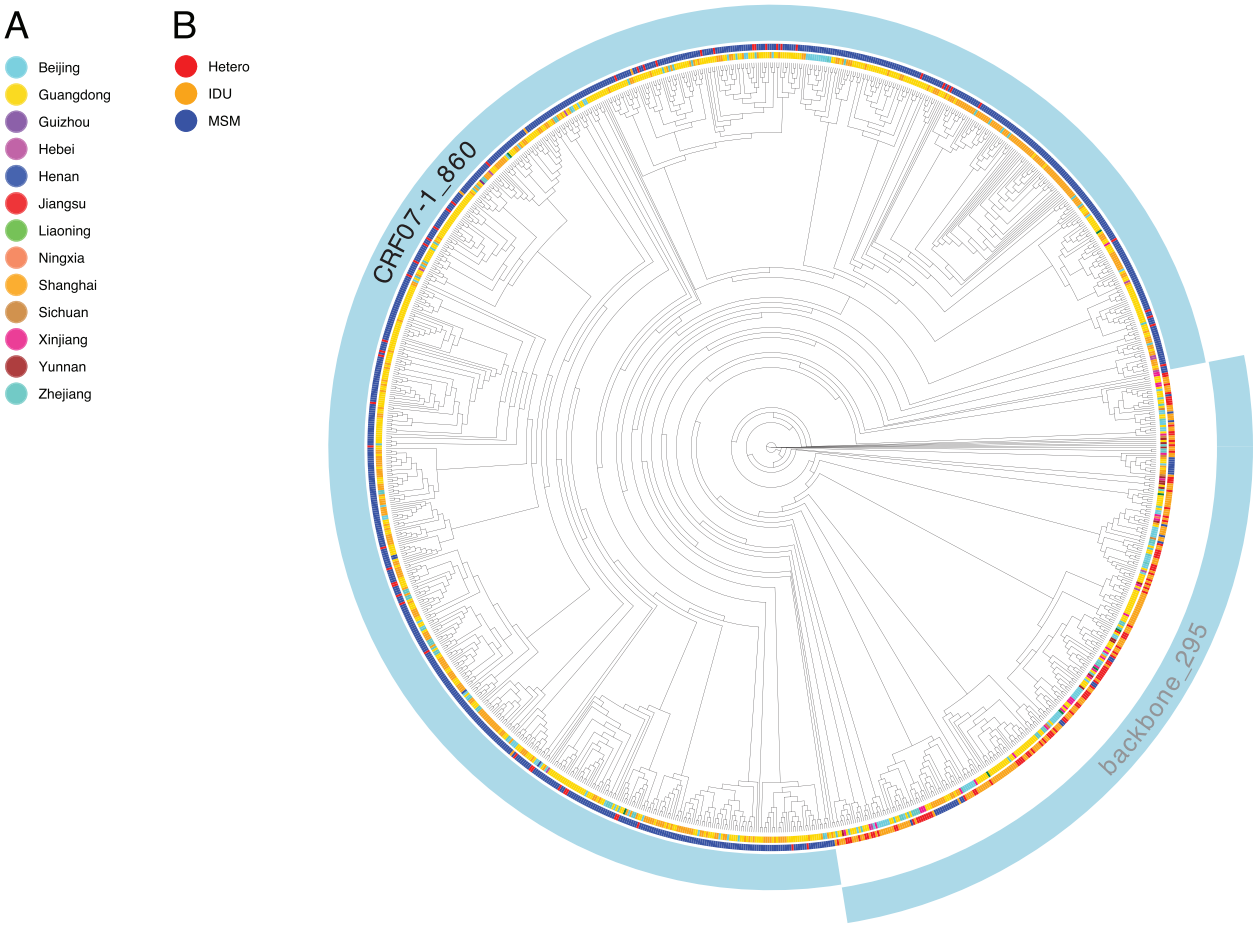

Figure 2. Estimated ML phylogenetic tree of HIV-1 CRF07_BC from Mainland China using 'full1155'. ML phylogenetic tree of HIV-1 CRF07_BC from Mainland China for 'full1155' is shown. The two circles of colored cells show geographic location (inner circle, A) and risk group (outer circle, B).

Table 2. Geographic location, year of sampling, and risk factor for a large cluster of concern of HIV-1 CRF07-1 strains used in the present study.

			Risk factor[a]		
Geographic location	Sampling year	*n*	Hetero	IDU	MSM
Beijing	2007–2010	48	2		46
Guangdong	2006–2012	429	42	4	383
Guizhou	2007	1			1
Hebei	2008–2015	8			8
Henan	2010	4			4
Liaoning	2008–2009	3			3
Shanghai	2009–2013	311			311
Xinjiang	2015	1	1		
Zhejiang	2007–2013	55	8		47
Total		860	53	4	803

[a]Risk factor: Hetero, heterosexual; IDU, injecting drug users; MSM, men having sex with men.

probable root geographic location and risk group of the HIV-1 CRF07_BC ancestor were in Yunnan Province among the IDU population for 'riskdate133','locdate163', and 'locrisk229' datasets (Figs S5 and S6), which was consistent with the previous study (Su et al. 2000).

Second, we investigated the total number of migration events that originated in each geographic location and risk group (exportation events) and the total number of migration events received by a given geographic location and risk group (importation events). For the geographic location, our discrete trait analysis revealed that the most frequent exportation events of HIV-1 CRF07_BC were all from Guangdong Province and Yunnan Province for 'locdate163' and 'locrisk229', respectively (Fig. S7). Our discrete trait analysis also revealed that the most frequent importation events of HIV-1 CRF07_BC were all to Guangdong Province for 'locdate163' and 'locrisk229' datasets (Fig. S7). For the risk group, our discrete trait analysis revealed that the most frequent exportation events of HIV-1 CRF07_BC were all from IDU and MSM for 'riskdate133' and 'locrisk229', respectively (Fig. S8). Our discrete trait analysis also revealed that the most frequent importation events of HIV-1 CRF07_BC were all to Hetero for 'riskdate133' and 'locrisk229' datasets (Fig. S8).

Third, we explored estimates of the total number of migration events that originated in a given geographic location and risk group to another given geographic location and risk group with BF >3. For the geographic location, a number of most inferred migration events were all from Yunnan Province to Guangdong Province, and secondary, from Guangdong Province to Beijing, Shanghai, and Zhejiang Provinces, and from Yunnan Province to Xinjiang Uyghur Autonomous Region and Zhejiang Province, for 'locdate163' and 'locrisk229' datasets (Fig. S9). For the risk group, a number of most inferred migration events were all from IDU population to Hetero population, and secondary, from MSM population to Hetero population, for 'riskdate133' and 'locrisk229' datasets (Fig. S10).

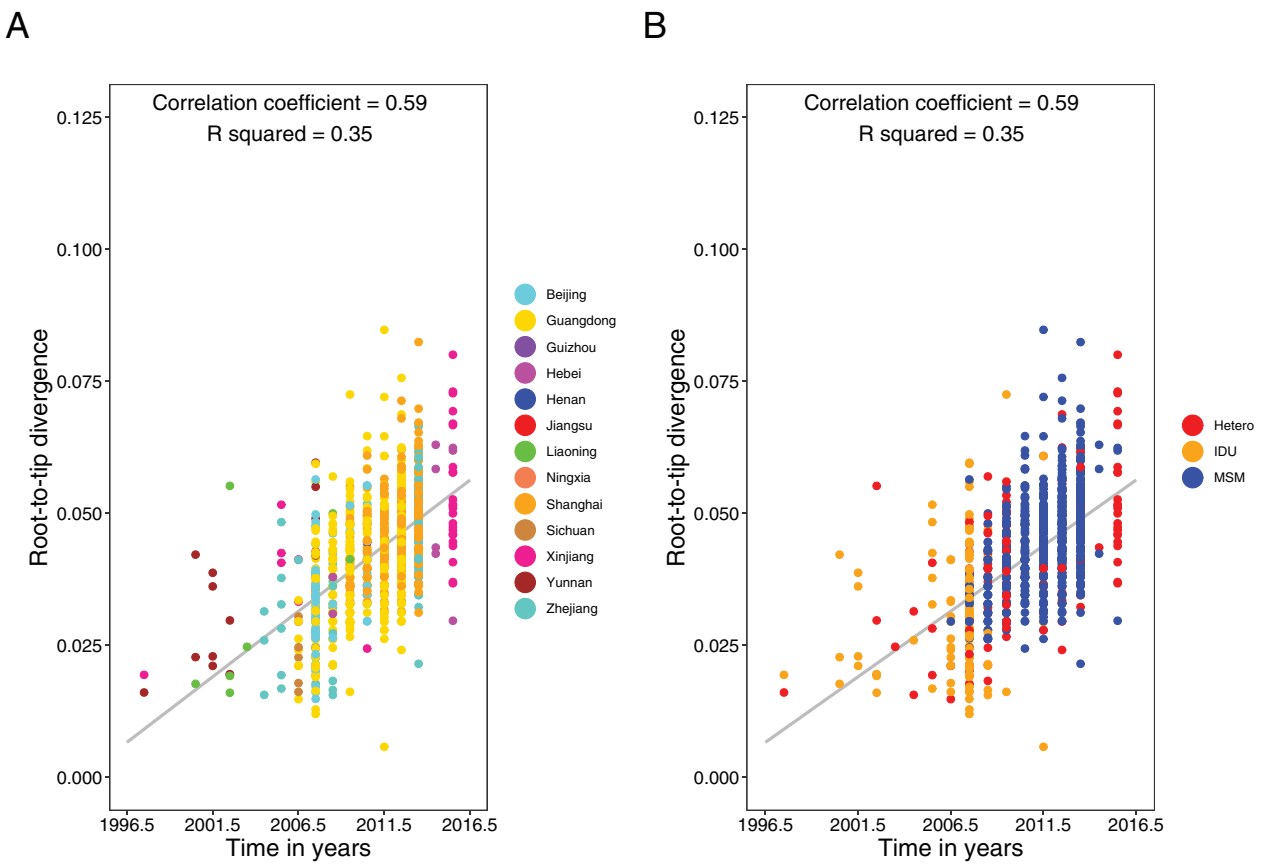

Figure 3. Root-to-tip genetic divergence plot of HIV-1 CRF07_BC in Mainland China using 'full1155'. Root-to-tip genetic divergence for 'full1155' in Mainland China in the ML tree (as shown in Fig. 2) plotted against year of sampling. (A) Geographic location is color-coded, as shown on the right. (B) Risk group is color-coded, as shown on the right. Gray color indicates linear regression line.

Fourth, we explored estimates of the total number of migration events that originated in a given geographic location and risk group to another given geographic location and risk group over time with BF >3. For the geographic location, the inferred migration events from Yunnan Province to other provinces mostly occurred during the early stages of the HIV-1 CRF07_BC epidemic, for 'locdate163' and 'locrisk229', respectively (Fig. 4). However, the inferred migration events from Guangdong Province to other provinces mostly occurred during later stages of the HIV-1 CRF07_BC epidemic, for 'locdate163' and 'locrisk229', respectively (Fig. 4). For the risk group, the inferred migration events from the IDU population to other risk groups mostly occurred during the early stages of HIV-1 CRF07_BC epidemic for both 'riskdate133' and 'locrisk229' datasets (Fig. 5). However, the inferred migration events from the MSM population to other risk groups mostly occurred during later stages of the HIV-1 CRF07_BC epidemic for both 'riskdate133' and 'locrisk229' datasets (Fig. 5).

Fifth, we visualized the geographic location and risk group dispersals of HIV-1 CRF07_BC in Mainland China using SpreaD3 v0.9.7 (Bielejec et al. 2016). The results of geographic location and risk group dynamics of HIV-1 CRF07_BC in Mainland China were similar for 'locdate163' and 'locrisk229' (Fig. 6) and for 'riskdate133' and 'locrisk229' (Fig. S11), respectively. We also made videos for the visualization of geographic location and risk group spread of HIV-1 CRF07_BC in Mainland China over time for 'riskdate133' and 'locrisk229' (Video. S1 and Video. S1) and for 'riskdate133' and 'locrisk229' (Video. S3 and Video. S4), respectively.

4. Discussion

To explore the different spatial and risk group dynamics of HIV-1 CRF07_BC transmission in Mainland China, we used all publicly available partial *pol* sequences (HXB2 genome position 2,253–3,401) obtained from 1,155 strains collected from 13 provinces (Beijing, Guangdong, Guizhou, Hebei, Henan, Jiangsu, Liaoning, Ningxia, Shanghai, Sichuan, Xinjiang, Yunnan, and Zhejiang) of Mainland China and three risk groups (Hetero, IDU, and MSM) between 1997 and 2015.

The distributions of 'full1155' and 'CRF07-1_860' over time, geographic location, and risk group were heterogeneous (Tables 1 and 2, Fig. 1, and Figs S1–S2). The sequences of 'full1155' were mainly distributed in Guangdong and Shanghai among the MSM population between 2007 and 2013. The sequences of 'CRF07-1_860' were also mainly distributed in Guangdong and Shanghai among the MSM population, but during a narrower period (2009–2013). Notably, Hetero had the largest (3.29 per cent) genetic distance, and the genetic distances between IDU and Hetero (3.08 per cent) and between MSM and Hetero (3.03 per cent) were very similar. Therefore, it is likely that the spread of HIV-1 CRF07_BC from the IDU and MSM populations into the general populations could have been facilitated by the Hetero population in China. CRF07-1 spread rapidly, as observed by the large monophyletic cluster, to all provinces of Mainland China, where in many cases it has completely replaced previous circulating variants, fueling a second wave throughout China, which was associated with a high number of cases and deaths.

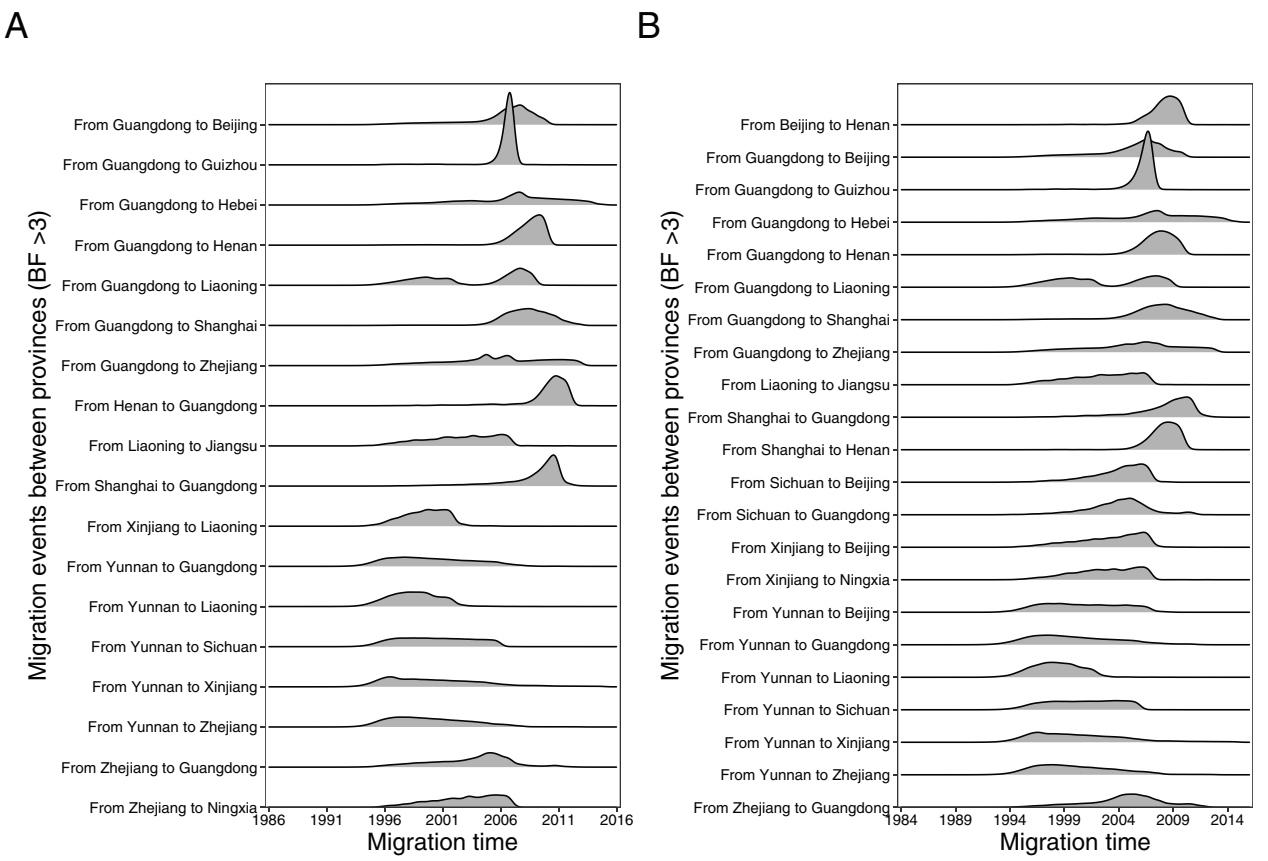

Figure 4. Total number of transitions between provinces in Mainland China over time. Total number of transitions between provinces in Mainland China over time for 'locdate163' (A), and 'locrisk229' (B) with BF >3 is shown.

The mutation of two nucleotide bases (html. S1 and html. S3) in the *RT* gene region, located at the catalytic core of the RT protein in 3D structure (Fig. S3) and targeted by at least 12 CTL/CD8+ (https://www.hiv.lanl.gov/content/immunology/maps/ctl/Pol.html), indicated the viral genetic differences between 'backbone_295' and 'CRF07-1_860' and may be important for the increased transmissibility and/or partial immune evasion properties of HIV-1 CRF07_BC. Notably, non-virological factors, such as host, environmental, and sociological factors, may also be important and could not be ruled out in the current investigation (Hill, Rosenbloom, and Nowak 2012; Faria et al. 2016; Li et al. 2016).

Furthermore, a Bayesian time-scaled phylogenetic analysis along with the BSP tree prior model plot (Drummond et al. 2005) was obtained. The median TMRCA and evolutionary rates estimated for HIV-1 CRF07_BC for 'riskdate133', 'locdate163', and 'locrisk229' ranged between 6 October 1992 and 9 July 1993 and from 1.42×10^{-3} to 1.54×10^{-3} substitutions per site per year, respectively (Table S3). The estimated TMRCA of HIV-1 CRF07_BC for 'riskdate133', 'locdate163', and 'locrisk229' using Bayesian time-scaled phylogenetic analysis did not agree with the dataset's respective root-to-tip regression results using TempEst v1.5.3 (Rambaut et al. 2016) (Table S1); meanwhile, the estimated 95 per cent HPD interval of evolutionary rates for 'locdate163' based on Bayesian time-scaled phylogenetic analysis did not cover the estimate obtained with the root-to-tip regression findings obtained with TempEst v1.5.3 (Rambaut et al. 2016) (Table S1), most likely because the latter assumes a strict molecular clock, whereas the Bayesian analysis was performed using a relaxed molecular clock which allows rate variation through time and thus better characterized changes in the substitution rate throughout HIV-1 CRF07_BC evolutionary history. The estimated TMRCA of HIV-1 CRF07-1 using 'riskdate133', 'locdate163', and 'locrisk229' ranged from 18 January 2000 to 2 May 2001, pointing that it originated in Mainland China approximately 20 years ago (Li et al. 2015b).

Our BSP analysis indicated that between 2004 and 2010, the genetic diversity of HIV-1 CRF07_BC experienced at least one phase of population decrease followed by a phase of population expansion (Fig. S4). The declining period most likely correlates with the reduction of the IDU population in Mainland China, since nowadays most drug users mainly consume drugs by non-injection methods (Figs S1 and S2), making high-risk sexual behavior the main route for HIV infection and transmission. The increasing period was most likely due to the increased circulation and diversification of CRF07-1 in the MSM population in Mainland China (Figs S1 and S2), in accordance with our mutation analysis results (Fig. S3). However, we cannot exclude the possibility that the genetic diversity of HIV-1 CRF07_BC remained constant during the 2000s. Notably, the main risk group distribution of HIV-1 CRF07_BC changed from IDU population to MSM population over time. In addition, the Hetero population was usually intermingled with the IDU or MSM population; however, the IDU population was not usually intermingled with the MSM population (Fig. 2).

Our discrete trait analyses indicate that IDU in Yunnan Province is the most likely risk group and geographic location

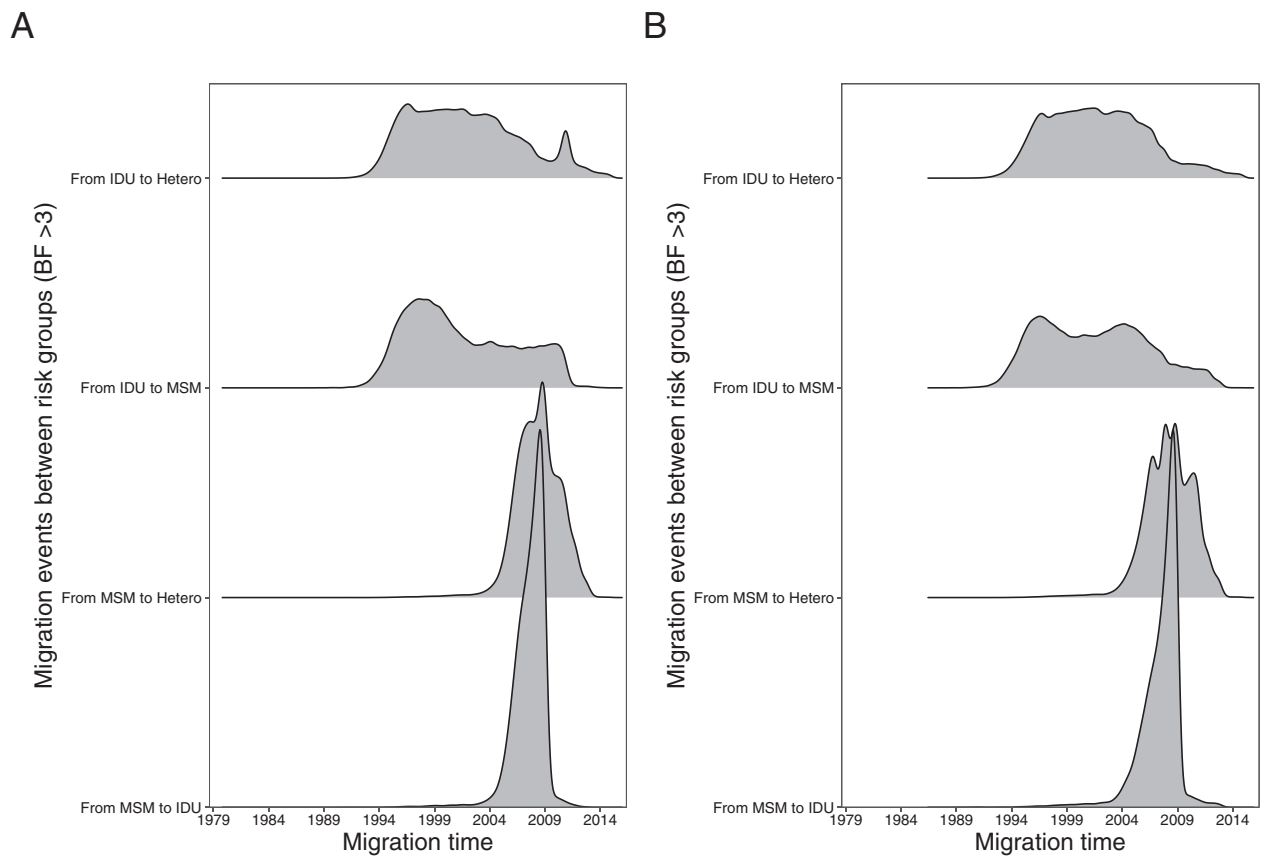

Figure 5. Total number of transitions between risk groups in Mainland China over time. Total number of transitions between risk groups in Mainland China over time for 'riskdate133' (A), and 'locrisk229' (B) with BF >3 is shown.

of origin of HIV-1 CRF07_BC in Mainland China (Fig. S5 and Fig. S6). We found that Yunnan Province and Guangdong Province were likely the most important sources of dissemination of HIV-1 CRF07_BC in Mainland China and were the origin of the epidemics in Zhejiang and other provinces (Figs 4 and 6, Figs S7 and S9, Video S1, and Video S2). We also discovered that IDU and MSM populations were likely the most important sources of HIV-1 CRF07_BC dispersion in Mainland China, and we demonstrated again that HIV-1 CRF07_BC may have spread from IDU and MSM populations into the general population through Hetero population in China (Fig. 5, Figs S8, S10, and S11, Video. S3, and Video. S4).

This implies that the large-scale surveillance and intervention measures conducted in Mainland China are required for effective prevention of local transmission, and eventually management of the epidemic, and reduction of the potential for spread to other nations or regions. It would be valuable to focus surveillance and tracking of new clusters of concern like CRF07-1(Li et al. 2014). CRF07-1 has been detected in nine provinces (Beijing, Guangdong, Guizhou, Hebei, Henan, Liaoning, Shanghai, Xinjiang, and Zhejiang) in Mainland China in the present study, which indicates that CRF07-1 has emerged as a cluster of concern for roughly 20 years and is still causing ongoing local transmission in Mainland China. CRF07-1 rapidly and completely replaced the previous circulating variants throughout China, which could likely be in part due to mutations such as E248V and K249Q (html. S2 and html. S4) in the RT protein region and/or immune evasion properties that altered the dynamics of HIV-1 in China. As a result, we must prioritize these clusters of concern of HIV-1 in Mainland China in our real-time epidemic surveillance and intervention activities, including risk reduction, testing, and treatment. As a result, genetic epidemiology has become increasingly valuable in informing real-time epidemic surveillance and intervention efforts. It is worth noting that our study was based on low and variable sampling of HIV-1 CRF07_BC strains among different provinces and risk groups in Mainland China, and HIV-1 CRF07_BC strains from other provinces and risk groups have not been sampled or sequenced, nor are available at the LANL HIV Sequence Database website (https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html) and thus not included in the present study based on our dataset collection strategy.

Nationally, the publicly accessible partial *pol* sequences of HIV-1 CRF07_BC strains reflect just a small percentage of the total number of infections in Mainland China, with an uneven sampling regime across provinces and risk groups, which could potentially have impacted our estimates based on this relatively limited dataset. As a result, continuous and structured sampling of HIV-1 CRF07_BC with epidemiological data is required for a better understanding of the spatiotemporal dynamics of HIV-1 CRF07_BC epidemic in Mainland China and worldwide, as well as the creation of public policies aimed at eventually controlling acquired immune deficiency syndrome.

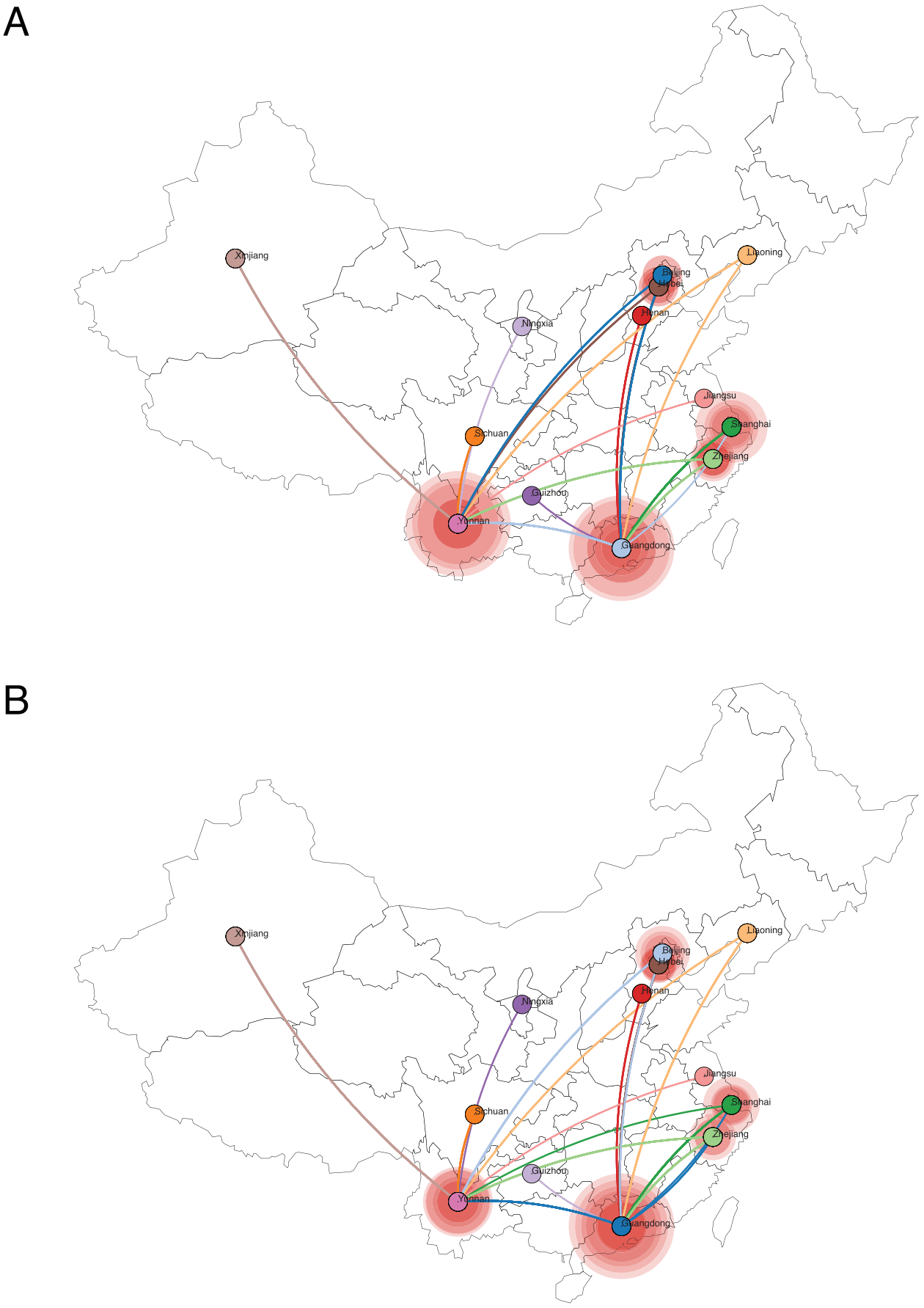

Figure 6. Map of geographic location transitions for HIV-1 CRF07_BC in Mainland China. Points are color-coded by the geographic location of origin. Lines are color-coded by the geographic location of destination. (A) Visualization for 'locdate163'. (B) Visualization for 'locrisk229'.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

The authors would like to thank the Core Facility and Technical Support, Wuhan Institute of Virology, Prof. Philippe Lemey, Prof. Guy Baele, and Prof. Simon Y. W. HO for their technical support. The content is solely the responsibility of the authors and does not represent official views of the National Institutes of Health.

Funding

This work was supported by two grants from the National Key Research and Development Program of China (2018YFC1603803 and 2021YFC2301304) to Dr. Haizhou Liu, and a grant from the National Natural Science Foundation of China (No.81904040) to Dr. Yanping Li.

Conflict of interest: The authors declare no competing interests.

Author contributions

X.L. conceived and designed the study and drafted the manuscript. X.L., N.T., H.L., and B.F. analyzed the data. X.L., N.T., H.L., Y.L., and B.F. interpreted the data and provided critical comments. All authors reviewed and approved the final manuscript.

References

Ayres, D. L. et al. (2012) 'BEAGLE: An Application Programming Interface and High-performance Computing Library for Statistical Phylogenetics', *Systematic Biology*, 61: 170–3.

Baele, G. et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison while Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29: 2157–67.

——— et al. (2013) 'Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics', *Molecular Biology and Evolution*, 30: 239–43.

Bielejec, F. et al. (2016) 'SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes', *Molecular Biology and Evolution*, 33: 2167–9.

Chen, X. et al. (2017) 'First Description of Two New HIV-1 Recombinant Forms CRF82_cpx and CRF83_cpx among Drug Users in Northern Myanmar', *Virulence*, 8: 497–503.

Chen, Y. et al. (2021) 'HIV-1 Subtype Diversity and Transmission Strain Source among Men Who Have Sex with Men in Guangxi, China', *Scientific Reports*, 11: 8319.

Chen, Z. W. et al. (2018) 'Surging HIV-1 CRF07_BC Epidemic among Recently Infected Men Who Have Sex with Men in Fujian, China', *Journal of Medical Virology*, 90: 1210–21.

Chin, B. S. et al. (2015) 'Short Communication: Increase of HIV-1 K103N Transmitted Drug Resistance and Its Association with Efavirenz Use in South Korea', *AIDS Research and Human Retroviruses*, 31: 603–7.

Chow, W. Z. et al. (2016) 'Extensive Genetic Diversity of HIV-1 in Incident and Prevalent Infections among Malaysian Blood Donors: Multiple Introductions of HIV-1 Genotypes from Highly Prevalent Countries', *PLoS One*, 11: e0161853.

Darriba, D. et al. (2012) 'jModelTest 2: More Models, New Heuristics and Parallel Computing', *Nature Methods*, 9: 772.

Dennis, A. M. et al. (2018) 'HIV-1 Transmission Clustering and Phylodynamics Highlight the Important Role of Young Men Who Have Sex with Men', *AIDS Research and Human Retroviruses*, 34: 879–88.

Di Giallonardo, F. et al. (2020) 'Increased HIV Subtype Diversity Reflecting Demographic Changes in the HIV Epidemic in New South Wales, Australia', *Viruses*, 12: 1402.

Drummond, A. J. et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4: e88.

——— et al. (2005) 'Bayesian Coalescent Inference of past Population Dynamics from Molecular Sequences', *Molecular Biology and Evolution*, 22: 1185–92.

——— et al. (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29: 1969–73.

Fan, Q. et al. (2021) 'Analysis of the Driving Factors of Active and Rapid Growth Clusters among CRF07_BC-Infected Patients in a Developed Area in Eastern China', *Open Forum Infectious Disease*, 8: ofab051.

Faria, N. R. et al. (2016) 'Zika Virus in the Americas: Early Epidemiological and Genetic Findings', *Science*, 352: 345–9.

Felsenstein, J. (1985) 'Confidence Limits on Phylogenies: An Approach Using the Bootstrap', *Evolution*, 39: 783–91.

Feng, Y. et al. (2016) 'Geographic Origin and Evolutionary History of China's Two Predominant HIV-1 Circulating Recombinant Forms, CRF07_BC and CRF08_BC', *Scientific Reports*, 6: 19279.

Ferreira, M. A. R., and Suchard, M. A. (2008) 'Bayesian Analysis of Elapsed Times in Continuous-time Markov Chains', *Canadian Journal of Statistics*, 36: 355–68.

Gelman, A., and Meng, X.-L. (1998) 'Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling', *Statistical Science*, 13: 163–85.

Hall, T. A. (1999) 'BioEdit: A User-friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT', *Nucleic Acids Symposium Series*, 41: 95–8.

Han, X. et al. (2015) 'A Large-scale Survey of CRF55_01B from Men-Who-Have-Sex-with-Men in China: Implying the Evolutionary History and Public Health Impact', *Scientific Reports*, 5: 18147.

He, Z. et al. (2016) 'Evolview V2: An Online Visualization and Management Tool for Customized and Annotated Phylogenetic Trees', *Nucleic Acids Research*, 44: W236–41.

Hill, A. L., Rosenbloom, D. I., and Nowak, M. A. (2012) 'Evolutionary Dynamics of HIV at Multiple Spatial and Temporal Scales', *Journal of Molecular Medicine (Berlin, Germany)*, 90: 543–61.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Lapovok, I. et al. (2014) 'Short Communication: Molecular Epidemiology of HIV Type 1 Infection in Kazakhstan: CRF02_AG Prevalence Is Increasing in the Southeastern Provinces', *AIDS Research and Human Retroviruses*, 30: 769–74.

Lemey, P. et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520.

Li, J. et al. (2018) 'HIV-1 Transmissions among Recently Infected Individuals in Southwest China are Predominantly Derived from Circulating Local Strains', *Scientific Reports*, 8: 12831.

Li, X. et al. (2015a) 'HIV-1 Genetic Diversity and Its Impact on Baseline CD4+T Cells and Viral Loads among Recently Infected Men Who Have Sex with Men in Shanghai, China', *PLoS One*, 10: e0129559.

——— et al. (2016) 'The 2014 Ebola Virus Outbreak in West Africa Highlights No Evidence of Rapid Evolution or Adaptation to Humans', *Scientific Reports*, 6: 35822.

——— et al. (2014) 'Molecular Epidemiology of HIV-1 in Jilin Province, Northeastern China: Emergence of a New CRF07_BC Transmission Cluster and Intersubtype Recombinants', *PLoS One*, 9: e110738.

Li, Z. et al. (2015b) 'Trends of HIV Subtypes and Phylogenetic Dynamics among Young Men Who Have Sex with Men in China, 2009-2014', *Scientific Reports*, 5: 16708.

Ma, X. et al. (2007) 'Trends in Prevalence of HIV, Syphilis, Hepatitis C, Hepatitis B, and Sexual Risk Behavior among Men Who Have Sex with Men. Results of 3 Consecutive Respondent-driven Sampling Surveys in Beijing, 2004 through 2006', *Journal of Acquired Immune Deficiency Syndromes*, 45: 581–7.

Machnowska, P. et al., German HIVSSG. (2019) 'Prevalence and Persistence of Transmitted Drug Resistance Mutations in the German HIV-1 Seroconverter Study Cohort', *PLoS One*, 14: e0209605.

Minin, V. N., and Suchard, M. A. (2008a) 'Counting Labeled Transitions in Continuous-time Markov Models of Evolution', *Journal of Mathematical Biology*, 56: 391–412.

——— (2008b) 'Fast, Accurate and Simulation-free Stochastic Mapping', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363: 3985–95.

O'Brien, J. D., Minin, V. N., and Suchard, M. A. (2009) 'Learning to Count: Robust Estimates for Labeled Distances between Molecular Sequences', *Molecular Biology and Evolution*, 26: 801–14.

Pang, W. et al. (2012) 'Extensive and Complex HIV-1 Recombination between B', C and CRF01_AE among IDUs in South-east Asia', *AIDS*, 26: 1121–9.

Pang, X. et al. (2021) 'Patterns and Risk of HIV-1 Transmission Network among Men Who Have Sex with Men in Guangxi, China', *Scientific Reports*, 11: 513.

Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901–4.

——— et al. (2016) 'Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen)', *Virus Evolution*, 2: vew007.

R Core Team. (2021). 'R: *A language and environment for statistical computing*', *R Foundation for Statistical Computing*. Vienna, Austria.

Rose, P. P., and Korber, B. T. (2000) 'Detecting Hypermutations in Viral Sequences with an Emphasis on G—> A Hypermutation', *Bioinformatics*, 16: 400–1.

Schrödinger, LLC. 'The PyMOL Molecular Graphics System, Version 1.8'

Siepel, A. C. et al. (1995) 'A Computer Program Designed to Screen Rapidly for HIV Type 1 Intersubtype Recombinant Sequences', *AIDS Research and Human Retroviruses*, 11: 1413–6.

Song, Y. X. et al. (2018) 'Prevalence of Transmitted Drug Resistance among HIV-1 Treatment-naive Patients in Beijing', *Epidemiology and Infection*, 146: 339–44.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Su, L. et al. (2000) 'Characterization of a Virtually Full-length Human Immunodeficiency Virus Type 1 Genome of a Prevalent Intersubtype (C/B') Recombinant Strain in China', *Journal of Virology*, 74: 11367–76.

Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10', *Virus Evolution*, 4: vey016.

Takebe, Y. et al. (2010) 'Reconstructing the Epidemic History of HIV-1 Circulating Recombinant Forms CRF07_BC and CRF08_BC in East Asia: The Relevance of Genetic Diversity and Phylodynamics for Vaccine Strategies', *Vaccine*, 28: B39–44.

Tamura, K., Nei, M., and Kumar, S. (2004) 'Prospects for Inferring Very Large Phylogenies by Using the Neighbor-joining Method', *Proceedings of the National Academy of Sciences of the United States of America*, 101: 11030–5.

Tamura, K., Stecher, G., and Kumar, S. (2021) 'MEGA11: Molecular Evolutionary Genetics Analysis Version 11', *Molecular Biology and Evolution*, 38: 3022–7.

Ueda, S. et al. (2019) 'Genetic Diversity and Drug Resistance of HIV-1 Circulating in North Sulawesi, Indonesia', *AIDS Research and Human Retroviruses*, 35: 407–13.

Vrancken, B. et al. (2020) 'Comparative Circulation Dynamics of the Five Main HIV Types in China', *Journal of Virology*, 94: e00683–20.

Waterhouse, A. et al. (2018) 'SWISS-MODEL: Homology Modelling of Protein Structures and Complexes', *Nucleic Acids Research*, 46: W296–303.

Waterhouse, A. M. et al. (2009) 'Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench', *Bioinformatics*, 25: 1189–91.

Wickham, H. (2016) *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Xie, W. et al. (2011) 'Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection', *Systematic Biology*, 60: 150–60.

Yang, Z., and Rannala, B. (1997) 'Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method', *Molecular Biology and Evolution*, 14: 717–24.

Yebra, G. et al. (2018) 'A High HIV-1 Strain Variability in London, UK, Revealed by Full-genome Analysis: Results from the ICONIC Project', *PLoS One*, 13: e0192081.

Zai, J. et al. (2020) 'Tracing the Transmission Dynamics of HIV-1 CRF55_01B', *Scientific Reports*, 10: 5098.

Zhang, D. et al. (2021) 'Molecular Surveillance of HIV-1 Newly Diagnosed Infections in Shenzhen, China from 2011 to 2018', *The Journal of Infection*, 83: 76–83.

Zhao, J. et al. (2016) 'The Dynamics of the HIV Epidemic among Men Who Have Sex with Men (MSM) from 2005 to 2012 in Shenzhen, China', *Scientific Reports*, 6: 28703.

Zhong, F. et al. (2011) 'Possible Increase in HIV and Syphilis Prevalence among Men Who Have Sex with Men in Guangzhou, China: Results from a Respondent-driven Sampling Survey', *AIDS and Behavior*, 15: 1058–66.