Title

Comparative Circulation Dynamics of the Five Main HIV Types in China

Journal

Journal of Virology, 94(23)

ISSN

0022-538X

Authors

Vrancken, Bram

Zhao, Bin

Li, Xingguang

et al.

Publication Date

2020-11-09

DOI

10.1128/jvi.00683-20

Peer reviewed

Title: COMPARATIVE CIRCULATION DYNAMICS OF THE FIVE MAIN HIV TYPES IN CHINA

Authors : Bram Vrancken[1†], Bin Zhao[2†], Xingguang Li[3†], Xiaoxu Han[2#], Haizhou Liu[4], Jin Zhao[5], Ping Zhong[6], Yi Lin[6], Junjie Zai[7], Mingchen Liu[2], Davey M Smith[8], Simon Dellicour[1,9*] and Antoine Chaillon[8*#]

[†]*Contributed equally to this work.*

**Contributed equally to this work.*

[#]*Corresponding authors*

Affiliations:

[1] Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, KU Leuven - University of Leuven, Leuven, Belgium**;**

[2] NHC Key Laboratory of AIDS Immunology (China Medical University), National Clinical Research Center for Laboratory Medicine, The First Affiliated Hospital of China Medical University, Shenyang, 110001, China;

[3] Department of Hospital Office,The First People's Hospital of Fangchenggang, Fangchenggang, 538021, China;

[4] Centre for Emerging Infectious Diseases, The State Key Laboratory of Virology, Wuhan Institute of Virology, University of Chinese Academy of Sciences, Wuhan, 430071, China;

[5] Shenzhen Center for disease control and prevention, Shenzhen，518055 China.

[6] Department of AIDS and STD, Shanghai Municipal Center for Disease Control and Prevention; Shanghai Municipal Institutes for Preventive Medicine, Shanghai, 200336, China

[7] Immunology innovation team, School of Medicine, Ningbo University, Ningbo, Zhejiang 315211, China;

[8] Division of Infectious Diseases and Global Public Health, University of California San Diego, CA;

[9] Spatial Epidemiology Lab. (SpELL), Université Libre de Bruxelles, CP160/12 50, av. FD Roosevelt, 1050 Bruxelles, Belgium.

Corresponding authors

Antoine Chaillon M.D., Ph.D.
Center For AIDS Research (CFAR)
Division of Infectious Diseases,
UCSD & VA San Diego Healthcare System

Stein Clinical Research Building #325
University of California San Diego
9500 Gilman Drive
La Jolla CA, 92093-0679
(+33) 76-779-4224
achaillon@health.ucsd.edu

Xiaoxu Han
NHC Key Laboratory of AIDS Immunology (China Medical University),
National Clinical Research Center for Laboratory Medicine,
The First Affiliated Hospital of China Medical University, Shenyang
hanxiaoxu@cmu.edu.cn

Key words: HIV, phylodynamics, discrete phylogeography; generalized linear model, China

Word Count: 4,919

Figures: 4

ABSTRACT

The HIV epidemic in China accounts for 3% of the global HIV incidence. We compared the patterns and determinants of interprovincial spread of the five most prevalent circulating types. HIV *pol* sequences sampled across China were used to identify relevant transmission networks of the five most relevant HIV-1 types (B, CRF01_AE, CRF07_BC, CRF08_BC and CRF55_01B) in China. From these, the dispersal history across provinces was inferred. A generalized linear model (GLM) was used to test the association between migration rates among provinces and several measures of human mobility. A total of 10,707 sequences between 2004-2017 across 26 provinces were collected, among which 1,962 newly reported here. A mean of 18 (Min-Max:1-54) independent transmission networks involving up to 17 provinces were identified. Discrete phylogeographic analysis largely recapitulate the documented spread of the HIV types which, in turn, to large extent mirror within-China population migration flows. In line with the different spatio-temporal spread dynamics, the identified drivers thereof were also heterogeneous but are consistent with a central role of human mobility. The comparative analysis of the dispersal dynamics of the five main HIV types circulating in China suggests a key role of large populations centers and developed transportation infrastructures as hubs of HIV dispersal. This advocates for coordinated public health efforts in addition to local targeted interventions.

IMPORTANCE

While traditional epidemiological studies are of great interest in describing the dynamics of epidemics, they cannot fully capture the geospatial dynamics and factors driving the dispersal of pathogens such as HIV as they struggle to capture linkages between infections. To overcome this, we used a discrete phylogeographic approach coupled to a generalized linear model extension to characterize the dynamics and drivers of the across-province spread of the five main HIV types circulating in China. Our results indicate that large urbanized areas with dense population and developed transportation infrastructures are facilitators of HIV dispersal throughout China, and highlight the need to consider harmonized country-wide public policies to control local HIV epidemics.

INTRODUCTION

By the end of 2018, the number of people living with HIV (PWH) in China was close to 1.25 million (1, 2). The distribution of HIV-1 subtypes in China is diverse with over 11 circulating genetic variants (3), each with an evolving geographical distribution, prevalence and modes of transmission (3-6). The first nationwide molecular epidemiological survey in 1996-1998 showed that subtype B'/B (47.5%) and subtype C (34.3%) were the most predominant HIV types in China (7). For subtype B', this in part resulted from its high prevalence among plasma donors in China because of unsanitary commercial plasma collection (8). Surveys conducted in 2002-2003 and 2006 indicated that the circulating recombinant forms (CRF) CRF07_BC, CRF01_AE, and CRF08_BC had become the dominant HIV types in China. Founder effects make that CRF07_BC and CRF08_BC mostly circulated among injecting drug users (IDUs) in North-Eastern and South-Eastern China, respectively (9-12), while subtype B' remains dominant among former plasma donors in Central China (13, 14). Meanwhile, CRF01_AE became the dominant type and replaced subtype B as the principal driver of infection among men reporting having sex with men (MSM) (3). The National Sentinel Surveillance System of China revealed that the proportion of MSM transmission increased from 14.7% in 2009 (15) to 27.6% in 2016 (16) with an increased proportion of CRF01_AE and CRF07_BC infections among MSM, while the proportion of HIV-1 subtype B decreased between 2012 and 2016 (6, 17). In addition to these predominant HIV types, CRF55_01B, generated through recombination between CRF01_AE and subtype B variants, has been first identified among MSM in the city of Shenzhen (18, 19). Circulating primarily among MSM, it has now spread throughout most provinces of China with a prevalence ranging from 1.5% to 12.5% (20). Its prevalence has increased in the past five years, especially in South and East China with higher pooled estimated rate in Guangdong (12.22%, 95% CI 10.34–13.17) and Fujian (8.65%, 95% CI 4.98–13.17)(17). It is now circulating mostly in Guangdong and neighboring provinces in China, and across all risk groups (18).

The burden of HIV is also geographically unevenly spread: whereas HIV is present in all provinces, the top six high-prevalence provinces (Yunnan, Guangxi, Henan, Guangdong and Xinjiang) accounted for over 60% of the national number of PWH (21). The recent upsurge of HIV among MSM in large Chinese cities including Beijing, Chongqing, Chengdu, Guangzhou, Shanghai and Shenyang adds to this imbalance (22).

These multiple and diverse epidemics driven by changing risk factor patterns in part result from the inability of treatment, prevention, and control programs to halt the rapid growth of the HIV epidemics, which now account for ~3% of the global HIV prevalence (23). The epidemic growth of ~80,000 new infections per year (1) coincided with intense rural-to-urban migration flows (24-30)

and considerable investments in land and airway transport infrastructures expediting longer-distance human mobility (31). By the end of 2017, the migrant population, seeking better employment opportunities and living conditions in economically more developed areas, reached 244 million (32, 33), and migrant-workers have become the main driver of within-country migration. Importantly, the labor migrant population is also at higher risk for HIV acquisition and transmission because of poor knowledge about self-protection and the transmission routes of HIV (34), and they have been shown to fuel local epidemics (30, 35).

While traditional epidemiological studies are of great interest in describing the dynamics of epidemics, they struggle to fully capture the geospatial dynamics and factors driving the dispersal of pathogens. By merging virus genetic, geospatial and epidemiological data, phylodynamic models allow investigating the migration history of pathogens and its drivers in the absence of detailed contact tracing data and when linkage among infections is not obvious (36-39). Such analyses have been widely adopted both for human (40, 41) and plant viruses (38, 42), and more recently for HIV (43, 44).

The overall goal of the present study is to characterize the dynamics and drivers of the across-province spread of the main HIV types circulating in China. For this purpose, we capitalize on a discrete phylogeographic approach coupled to a generalized linear model extension.

RESULTS

Population characteristics

A total of 6800, 1578 (822/756), 1158, 957 and 211 available sequences were retrieved for CRF01_AE, subtype B (B/B'), CRF07_BC, CRF08_BC and CRF55_01B respectively. The number of provinces included in each final data set varied from 7 (CRF55_01B) to 17 (CRF01_AE and B/B'). See ***Figure 1*** for the distribution of provinces per data set.

Preliminary phylogenetic analysis and subsampling

For CRF01_AE, an initial set of 6,423 HIV-1 CRF01_AE *pol* sequences from 53 countries across the world between 1990 and 2017 retrieved from the Los Alamos National Laboratory HIV Sequence Database (45) was combined with the CRF01_AE data set of 6,800 sequences to delineate clades that capture the epidemic dynamics in transmission networks that pertain to China. We identified 83 of such clades (n=1876 sequences). To obtain data informative of interprovincial migration patterns, these were reduced to the 54 clades (size 3-24 sequences) that included samples from at least 2 Chinese provinces (totaling 454 sequences from 17 provinces). The same rationale was used for the other data sets. Starting from a total of 1578, 1158 and 957

sequences for subtype B/B', CRF07_BC, and CRF08_BC data sets, we obtained 15, 16, and 7 clades from 17, 10 and 8 provinces respectively. For CRF55_01B, which is circulating in China only, we obtained a single clade of 197 sequences collected across 7 provinces.

Discrete phylogeographic inferences

We used Bayesian phylogeographic inference to evaluate the dispersal history of the five main circulating types across Chinese provinces. This allows, for each subtype and CRF, to identify the significant migration events between Chinese provinces, and to estimate their number and directionality (***Figure 1***). The reconstructed patterns of spread based on the identified clades revealed strong evidence (adjusted Bayes Factor [BF_{adj}] ≥20) of migration between provinces for all sampled HIV populations. The relative contribution of each province as source and sink of HIV dispersal throughout China is summarized in ***Table 1***.

CRF01_AE and CRF07_BC clades have become the two predominant HIV CRF in China with an overall prevalence of 46.34% [95% CI: 40.56–52.17%] and 19.16% (95% CI: 15.02–23.66%), respectively (46). Here, the discrete phylogeographic analysis for CRF01_AE supported a complex migration history, with Beijing (Chinese capital with the second highest population density), Guangdong (southern region, capital Guangzhou), Shanghai (the most populous urban area in China) and Anhui (an important part of the Yangtze River Delta and in the top four provinces of China in labor export) being the provinces most involved in the interprovincial spread of migration events, both as major sources (with 24.9% [95%CI: 24.8-25], 16.6% [95%CI: 16.5-16.7], 15.8% [95%CI: 15.7-15.9] and 10.6% [95%CI: 10.5-10.7] of viral diffusion, respectively) and as major sinks (with 17.2% [95%CI: 17.1-17.3], 11.7% [95%CI: 11.6-11.8], 25.7% [95%CI: 25.5-25.8] and 13.3% [95%CI: 13.2-13.4] of introduction, respectively) (***Figures 1 and 2A***). For CRF_07BC, the second most prevalent type, the main sources were Beijing and Shanghai along with Yunnan (northwest-central region, capital Kunming) (***Figures 1 and 2C***), while for CRF08_BC, the main source was the province of Yunnan. Our model also showed robust evidence of viral migration across China for HIV-1 subtypes B/B' with Hubei (Capital Wuhan) being the major source of viral migration accounting for 69.9% [95%CI: 69.7-70.1] of viral dispersal with predominant diffusion toward the province of Henan (capital Zhengzhou) with 55.8% [95%CI: 55.6-56] of all introduction events, acting as a sink for the B/B' epidemic (***Figures 1 and 2B***). Finally, the southern province of Guangdong with the largest population was the only source of migration for CRF55_01B directed toward Anhui and Hunan that is supported by our data (***Figures 1 and 2E***).

From an historical perspective, our analyses showed a higher density of migration events across provinces in the late 1990' and 2000' years for subtype B/B' while migration events in general are more concentrated over the past 15 years for CRF01_AE, CRF07_BC and CRF08_BC. We also found that the historical interprovincial dispersal of CRF55_01B predominantly occurred around 2010 (*data not shown*).

Generalized linear model analyses

We next used the generalized linear model (GLM) extension of the phylogeographic model to evaluate the association between potential predictors and the migration frequencies among provinces (***Figure 3***). For CRF01_AE our model revealed a strong association between migration events and air traffic density as well as connectivity among locations, associations that are robust to randomizing tip-to-location assignments (BF_{adj} >> 100). The conditional effect size for connectivity between major cities over land was negative, meaning that spread between provinces that are easier to travel between over land is less frequent than between provinces that are less well connected over land. For CRF08_BC, a higher HIV prevalence in the province of origin was associated with increased HIV dispersal (BF_{adj}=87.8), which also associates with the number of immigrants at the origin (BF_{adj} =11.3). Whereas a higher HIV prevalence at the province of origin links to more frequent migration from that province, the conditional effect size for the number of immigrants at the origin was negative, implying that for CRF08_BC more immigration towards a province links to less frequent virus migration from that province. The only other predictor that was well-supported is spatial distance for subtype B/B'. For this predictor too, the conditional effect size is negative, indicating that migration is more frequent between closer locations. No other associations are well-supported (i.e. BF_{adj} ≥3) (***Figure 3)***.

DISCUSSION

Starting from >10,000 HIV-1 *pol* sequences from the five main prevalent HIV-1 subtypes and CRFs in China collected between 1996 and 2017, we reconstructed the spatial diffusion of the five most prevalent HIV-1 types across provinces in China. Our reconstructions largely recapitulate their documented spread, which we discuss one by one:

CRF01_AE. CRF01_AE has become the dominant HIV variant in most provinces (3). In line with previous epidemiological studies (3) and molecular analyses (47), our reconstructions capture that most migration events occurred recently between southern and eastern/north-eastern provinces

(see ***Figure 1*** and ***Table 1***). Specifically, we found that Beijing (the political, economic and cultural center of China) was the main source of CRF01_AE dispersal throughout the country and that Guangdong, Shanghai and Anhui are the other major hubs of CRF01_AE dispersal (***Table 1***). This largely matches the geographic scope of within-China population migration flows, which were concentrated within and between the southern and eastern main economic provinces (48, 49). It is of note that CRF01_AE is dominant among MSM (3), and that interprovincial migrants (as compared to intra-provincial migrants) not only are more likely to be male but also tend to be younger and have fewer years of formal education (49), which are factors associated with higher risk behavior (50).

The GLM analyses confirmed a strong association between the intensity of migration events and air traffic density, and an inverse relation of connectivity between major cities over land with the migration intensity. Combined, our results indicate that interprovincial CRF01_AE mobility is driven predominantly by longer-distance migration, possibly MSM-related, a combination that has also been noted in e.g. regions of Canada (51).

HIV Subtype B/B'. After an initial period of dominance, the prevalence of B/B' has declined (17). Consistent with this trend, we found that viral dispersal of HIV-1 subtype B mostly occurred in the 1990s and early 2000s (*data not shown*). Also in line with epidemiological surveys and with previous molecular analyses (52), we found that Hubei and Henan, both with a historically predominant circulation of B/B' among blood donors (53-57), were the major sources of interprovincial dispersal for this HIV-1 type (***Table 1 and Figure 2***).

Zhengzhou (Henan capital) is located at the junction of the major north-south Beijing-Guangzhou and east-west Lanzhou-Lianyungang railways and has evolved into a major national administrative, economic, and transportation hub (58), and Henan and Hubei are among the top five largest 'migration sending areas' (49). This shows that there was ample opportunity for long-distance spread of B/B' in relation to human mobility. In turn, the dominance of shorter-distance spread of B/B' implies that it did not find much fertile ground in highly mobile high risk groups, such as interprovincial migrant workers. This aspect is reflected in the results of our GLM analyses, which showed that migration events occurred more frequently among more nearby provinces (***Figure 3***).

CRF07_BC and CRF08_BC. CRF07_BC was originally reported in the Yunnan province in 1980s and spread quickly among IDUs. In recent years, it has been introduced in MSM populations, which drove its spread to elsewhere in China, particularly to Beijing, Shanghai, Guangdong and Zhejiang (9, 17, 59, 60). CRF08_BC on the other hand was initially reported in Yunnan and Guangxi provinces among IDUs but it has rarely been reported in MSM or other risk populations.

While none of the predictors appear significantly associated with the spread process of CRF07_BC, the origin of CRF08_BC in the south-eastern part of China and its subsequent spread towards economically more developed provinces that are attraction poles for inland migration is reflected in which predictors were found to significantly associate with the migration process, as well as the direction of their effect sizes. Specifically, the migration frequency out of a province increases with increasing HIV prevalence, and the migration intensity is inversely associated with the number of immigrants in the provinces of origin, suggesting that immigration hot spots functioned as a sink for this type (***Figure 3***).

CRF55_01B. This CRF was first identified among MSM in Shenzhen, Guangdong (18, 19). It has now spread throughout most provinces of China (20, 61, 62) although it mostly circulates in Guangdong and neighboring provinces, and across all risk groups (18). In line with this, our results point to Guangdong as main source of the dispersal to the western province of Anhui but also the adjacent province of Hunan (**Table 1**). As Guangdong is an economically well-developed province and attraction pole of migrant workers, it may intuitively seem at odds that it is a source rather than a destination of CRF055_01B spread. This may, however, be explained by return migration, which has become more intense over the years (63).

Prior to the 1980s, rural-urban migration in China was minimal. Since then, China has witnessed an extraordinary internal migration: rural-to-urban migrants increased the urban population by approximately 390 million. Of these rural migrants, approximately 54% were interprovincial migrants, most of which left their home province, but also with many returning after some time, and many visiting their families on a regular basis (e.g. with Chinese New Year). The reconstructed patterns of interprovincial spread for the main HIV types in China support the idea that migrant workers are at least partially involved in their diffusion. Unfortunately, the lack of epidemiological metadata prevented us to more explicitly elucidate the dynamics of spread within and between relevant subpopulations by for example associating epidemiological characteristics with uptake in clusters of closely related viruses (51, 64-66), which can help identify on what aspects to focus screening and prevention efforts. The involvement of migrant workers can be tested more directly within the GLM framework. Regrettably, we could only dispose of the total number of immigrants/emigrants by province instead of more granular pairwise migration flow data. Nonetheless, the identified drivers of HIV dispersal in China are in line with the view that human mobility strongly impacts pathogen epidemic dynamics (67). This combines with the reconstructed patterns of spread that largely reflect within-China population migration flows (that are directed towards and between major population and economic centers), to suggest that the patterns of viral transmission for at least some of the HIV epidemics in China were driven by major

population centers, which can act as gravity attractors before the virus spread to smaller populations (68, 69). This also illustrates that, in the absence of concurrent national prevention efforts with a focus on the most important drivers of ongoing transmission, local epidemics will rapidly be re-seeded, challenging the long-term impact of isolated intervention efforts.

Limitations. One major limitation of our study is that the collection of the HIV-1 *pol* sequences from the five main prevalent HIV-1 types in China has not been performed under a common framework, which may render our analyses prone to sampling bias. To the best of our knowledge, this drawback affects nearly every phylogeographic study of HIV-1 and other viruses. Whereas structured coalescent approaches hold promise for unbiased inferences in the face of biased sampling, inference under high state spaces and large data sets remains challenging for these models. For this reason, we relied on the computationally more efficient discrete trait analysis (70, 71). To counter this model's sensitivity to biased sampling of subpopulations, we (i) adopted a filter based on location state randomizations and (ii) combined the geographical information from different partitions that represent different samples of the same epidemic to minimize the risk of false positive migration linkages and associations with covariates (72, 73). Given that the reconstructed interprovincial spread largely captures the documented spread of the investigated HIV-1 types, we believe that these precautions were effective.

Several factors can explain that only few of the tested predictors associated with the spread patterns. The high-level resolution of our phylogeographic reconstructions makes that potential predictors can only be evaluated against a limited number of migration events between locations, in particular for CRF07_BC, CRF08_BC and CRF55_01B. Also, when only few migration events are observed, the impact of imperfect representations of the location-specific diversity on the ancestral reconstructions will increase and can obfuscate the relevance of potential predictors. Furthermore, our models did not capture potential time-varying dynamics of the selected predictors over the study period. This is particularly important for longstanding epidemics, such as HIV-1 subtype B/B'. Unfortunately, we could not test this hypothesis as we did not dispose of time-variable predictors.

CONCLUSION

The rapid increase of HIV-1 prevalence among migrant populations and the lack of effective intervention strategies is one of the current challenges for China (74, 75). In this study, the combined use of phylogeographic reconstructions and generalized linear model provides insights into the spatial viral dynamics of various HIV epidemics across provinces in China. The role of large urbanized areas with dense population and developed transportation infrastructures as

facilitator of HIV dispersal throughout China illustrates the need to consider harmonized country-wide public policies to control local HIV epidemics.

MATERIALS AND METHODS

Ethics statement

The study was approved by the ethics committee of the First Affiliated Hospital of China Medical University in Shenyang and Wuhan University of Bioengineering.

Data set compilation

We retrieved all publicly available HIV partial *pol* sequences (HXB2 position 2253-3554) of CRF01_AE, CRF07_BC, CRF08_BC, B/B' and CRF55_01B with known sampling date and sampling province of China from the Los Alamos National Laboratory HIV Sequence Database (45).

We additionally collected 1,962 CRF01_AE and HIV-1 subtype B partial *pol* sequences from the NHC Key Laboratory of AIDS Immunology, China Medical University (GenBank accession numbers MT336741:MT336811; MT368039:MT369927). We also retrieved publicly available HIV *pol* sequences from other countries along with sampling time and related geographical information. When multiple sequences were available for one participant, only the closest sequence from the estimated time of infection was kept.

Identification of Chinese clades

The geospatial unit in all phylogeographic analyses was the Chinese Province (first subnational administrative level). For all five subtypes except CRF55_01B, for which only isolates from China are available (19), we applied the step-by-step approach described below.

1. Following the approach of Cuypers et al. (76) and using an as complete as possible background data set (77), we first identified clades that likely correspond to distinct HIV introductions in China. To this end, the sequences for each subtype were first complemented with the publicly available location-annotated HIV *pol* sequences from the same subtype and aligned to a *pol* reference sequence (HXB2, GenBank accession K03455 (78). AliView (79) was used for manually editing the alignments.
2. Next, phylogenetic trees were inferred using FastTree2 (80) under a general GTR+Γ substitution model. These served to identify strongly supported Chinese clades, i.e. clades only including Chinese sequences and associated with a Shimodaira Hasegawa (SH) support of at least 0.9 (81-83).
3. Within these monophyletic clades, well-supported clusters of sequences sampled from the same administrative area (province) were identified. These were downsampled by randomly selecting one sequence from each cluster. This step reduces computational burden while

preserving estimation accuracy for the migration flow quantities between Chinese provinces (40).

Time scale for the evolutionary histories

When sequence data sets lack a clear temporal signal, it is common practice to use empirical evolutionary rate estimates for specifying a suitable prior distribution on the evolutionary rate parameter (e.g.(84, 85).

Subtype B/B'. To obtain plausible priors for the evolutionary rate for HIV-1 subtype B, we considered that various evolutionary rate have been reported for *pol*, varying from ~0.001 to ~0.003 substitutions/site/year (s/s/y) (86-88). For this reason, we specified a normal distribution as prior on the mean clock rate with mean 0.002 s/s/y and standard deviation such that the 95% confidence interval is bound at 0.001 s/s/y and 0.003 s/s/y.

CRF01_AE. We considered data from the literature (mean rate estimate ~0.0015 (87)) as well as the population-level substitution rate estimate of ~0.0027 [95%HPD: 0.0013- 0.0032] s/s/y obtained from clade-based specific analyses of the CRF01_AE data set with a Bayesian hierarchical phylogenetic model (HPM) approach (data not shown, (88)). This led us to specify a normal distribution as the prior on the mean clock rate with mean 0.002 s/s/y and standard deviation of 0.0005.

CRF07_BC, CRF08_BC and CRF55_01B. For these subtypes, a normal distribution was specified as the prior on the mean clock rate of ~0.001 s/s/y and standard deviation of 0.0005 according to clade-based estimates (data not shown).

Many of the clades that represent the HIV epidemics in China are limited in size. As this precludes reliable inference under the parameter-rich uncorrelated relaxed clock model (e.g. (89)), we opted to model the rate of evolutionary change in clades with ≥10 taxa with a relaxed clock model (90) while for the smaller clades a strict clock model was assumed.

Phylogeographic inference

Phylogeographic inference was performed using the discrete diffusion model (70, 91) implemented in the software package BEAST 1.10 (92). To promote estimation accuracy and precision of the transition rates among locations, the substitution model (GTR+Γ) and spread process were shared among clades of the same type (40, 72). A constant size coalescent prior was assumed for all clades of CRF01_AE, subtype B and CRF55_01B, and a non-parametric Bayesian skygrid tree prior was for clades with ≥20 taxa for CRF07_BC and CRF08_BC (93, 94).

To identify the subset of transition rates that was most informative to reconstruct the dispersal history, we used a model averaging procedure (Bayesian stochastic search variable selection – BSSVS) (70). In this procedure, the level of support depends on the *a priori* expected and *a posteriori* noted fraction of time during the Markov Chain Monte Carlo (MCMC) integration that a migration link or predictor helps explain the migration history. In the default setup, however, the *a priori* expectation only depends on the number of locations but does not account for the relative abundance of samples by location. This can bias inference in the presence of uneven sampling. The adjusted Bayes factor (BF_{adj}) (73) improves on this by incorporating information on the relative abundance of samples by location. It also relies on the *a priori* expected and *a posteriori* noted inclusion frequencies under BSSVS but relative to the original test, it requires two analyses: a first one where the trait values remain associated with their respective taxa, and a second one during which the trait values are randomized over the tips of the tree during the MCMC sampling. The latter provides the expectation in the absence of structure in the population, akin to the date randomization test when evaluating the presence of temporal signal(42, 95). As before, support for the significance is calculated as a ratio with the posterior odds as the enumerator, but as denominator we consider the inclusion frequency from the randomized analysis instead of the prior odds. Bayes factor (BF) support for all possible types of location exchanges was calculated with SpreaD3 (96). BF and BF_{adj} between 3-10, 10-20, and above 20 were considered to be substantial, positive and strong supports respectively for the observed transition rates between sampled locations (97). Estimates of the posterior probability of expected number of migration events between all pairs of locations (Markov jumps) were computed through stochastic mapping techniques (98, 99).

MCMC chains were run to ensure adequate mixing. Maximum clade credibility (MCC) trees were obtained with TreeAnnotator 1.10 (92) and convergence and mixing properties were inspected using Tracer 1.7 (100).

Generalized linear model analyses

We used the GLM extension of the discrete phylogeographic model implemented in BEAST 1.10 (39) to investigate the contribution of a series of location-associated variables to the migration rates among Chinese provinces. These variables included socio-demographic indicators (population size, number of emigrants and immigrants), HIV prevalence, sample size, and variables related to connectivity between locations (i.e. air traffic density, travel time by railways, the presence of shared borders, a measure of connectivity based on an accessibility model to major cities, and a proxy for spatial distance). Predictors were considered both at the origin and destination location. The population size, the number of emigrants and immigrants and HIV

444 prevalence were obtained from the National Bureau of Statistics of China (101) and the China
445 National Center for Disease Control and Prevention (102).
446
447 The numbers of sequences sampled at the origin/destination were included in the GLM to account
448 for the potential impact of sampling biases within the analysis (39). The air traffic density was
449 approximated by an air passenger flux matrix that quantifies the number of passengers traveling
450 between each pair of administrative areas (39). We use a data set provided by the OAG (Official
451 Airline Guide; www.oag.com) and containing the annual average number of seats on scheduled on
452 commercial flights between pairs of airports between 2014 and 2016 (103), assuming that the
453 number of seats represents a reasonable proxy for the number of passengers traveling between
454 airports. Travel time by railways was represented by the shortest travel time between the capitals
455 of each province obtained from 12306 China Railway website (104).
456
457 Geospatial connectivity measures included in the GLM were the following: a binary determination
458 if administrative areas share a common border, the average travel time by railways between
459 locations as well as two measures of connectivity among administrative areas computed using an
460 algorithm based on circuit theory and implemented in the program Circuitscape 4.0.5 (105): a
461 measure of connectivity and a proxy of spatial distance, obtained by computing pairwise
462 resistances on an inaccessibility grid and a uniform grid, respectively. For a given pair of locations,
463 Circuitscape computes the pairwise electric resistance based on a geo-referenced grid (or “raster”)
464 covering the study area and defining the local electric resistance values. To compute the proxy of
465 spatial distance, we simply used a homogeneous raster file with cell values uniformly set to “1”,
466 and for the pairwise connectivity measures, we used the inaccessibility raster as in (106) to define
467 the local values of electric resistance. Cell values of this inaccessibility raster indicate the travel
468 time required to reach the nearest urban center, with an urban center defined as a contiguous
469 area with 1,500 or more inhabitants per square kilometer or a population center of at least 50,000
470 inhabitants (106). For computational tractability, the resolution of both the uniform and
471 inaccessibility raster were decreased to ~5 arcmin (original resolution: ~0.5 arcmin). There are
472 several advantages to use pairwise Circuitscape distances computed on a uniform raster alone or
473 in complement to great-circle distances as proxies of spatial distance. First, pairwise Circuitscape
474 distances constitute more realistic measures because the underneath path model does not
475 assume straight-line movements and it also prevents movement through inaccessible areas.
476 Furthermore, given that the uniform raster is the homogeneous version of the inaccessibility raster,
477 pairwise Circuitscape resistances computed on the uniform raster also represent a proper
478 negative control (38, 107) for the inclusion, in the GLM analysis, of pairwise Circuitscape

resistances computed on an heterogeneous raster like the inaccessibility one. Indeed, the inclusion of a GLM predictor that does not have an impact on the dispersal, but for which pairwise distances have been computed using an advanced path model (like the one implemented in Circuitscape), can yield a false positive result in the absence of an appropriate negative control (38). Circuitscape computes pairwise electric resistance between two points or between two sets of points that all have to be associated with precise geographic coordinates. Given that such precise sampling coordinates were not available for sampled sequences, we randomly assign geographic coordinates to each sampled sequence. While this assignment was stochastic, we still used a human population density raster (resolution of ~5 arcmin) to define the sampling probability of all the raster cells within an administrative area. Hence, for each sequence originated from a given administrative area, its probability of being sampled from a particular raster cell was proportional to human population density value assigned to this cell. As this is a stochastic procedure, the sampling coordinate assignment and subsequent Circuitscape analyses were repeated 100 times. Final matrices of pairwise resistances computed on the uniform and inaccessibility rasters were obtained by averaging the 100 matrices computed after each repetition of the above procedure. Note that we used the same procedure to compute the averaged great-circle distances among locations.

To protect against a potential impact of sampling imbalances on the GLM results, support for the need for a predictor to help explain the variation in migration rates across locations was obtained after accounting for the relative abundance of the involved trait states (73).

Data availability. HIV-1 subtype B partial *pol* sequences are available in GenBank under accession numbers MT336741 to MT336811 and MT368039 to MT369927.

FUNDING ACKNOWLEDGMENTS

This work was supported by Mega-Projects of National Science Research for the 13th Five-Year Plan (2018ZX10721102); Central Public-interest Scientific Institution Basal Research Fund (2018PT31042); The Scientific Research Funding Project of Liaoning Province Education Department in 2019 "Raising Seedlings" for Young Science and Technology Talents.
B.V. (grant nr. 12U7118N) is and S.D. was supported by postdoctoral fellowships from the Fonds voor Wetenschappelijk Onderzoek (FWO, Belgium). S.D. is currently supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium). A.C. acknowledges funding from the University of San Diego Center for AIDS Research (CFAR), an NIH funded program. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

514

REFERENCES

1. Lyu P, Chen FF. 2019. [National HIV/AIDS epidemic estimation and interpretation in China]. Zhonghua Liu Xing Bing Xue Za Zhi 40:1191-1196.

2. Ding Y, Ma Z, He J, Xu X, Qiao S, Xu L, Shi R, Xu X, Zhu B, Li J, Wong FY, He N. 2019. Evolving HIV Epidemiology in Mainland China: 2009-2018. Curr HIV/AIDS Rep 16:423-430.

3. He X, Xing H, Ruan Y, Hong K, Cheng C, Hu Y, Xin R, Wei J, Feng Y, Hsi JH, Takebe Y, Shao Y. 2012. A Comprehensive Mapping of HIV-1 Genotypes in Various Risk Groups and Regions across China Based on a Nationwide Molecular Epidemiologic Survey. PLOS ONE 7:e47289.

4. Xiao P, Li J, Fu G, Zhou Y, Huan X, Yang H. 2017. Geographic Distribution and Temporal Trends of HIV-1 Subtypes through Heterosexual Transmission in China: A Systematic Review and Meta-Analysis. Int J Environ Res Public Health 14.

5. Yuan R, Cheng H, Chen LS, Zhang X, Wang B. 2016. Prevalence of different HIV-1 subtypes in sexual transmission in China: a systematic review and meta-analysis. Epidemiol Infect 144:2144-53.

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

7. Shao Y, Su L, Xing H, Shen J, Sun X, Zhang Y, Cheng H, Liu GE. 2000. HIV Molecular epidemic research in China. Bulletin Of Medical Research.

8. Wang Z, Han W, Wang C, LI H, Cui W, Xue X, Su L, Xing H, Gong X, Shao Y. 1999. Subtype and C2-V3region sequence analysis on HIV-1 in Henan. Journal for China AIDS/STD:167-169.

9. Zhang M, Jia D, Li H, Gui T, Jia L, Wang X, Li T, Liu Y, Bao Z, Liu S, Zhuang D, Li J, Li L. 2017. 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 33:1065-1069.

10. Feng Y, Takebe Y, Wei H, He X, Hsi JH, Li Z, Xing H, Ruan Y, Yang Y, Li F, Wei J, Li X, Shao Y. 2016. Geographic origin and evolutionary history of China's two predominant HIV-1 circulating recombinant forms, CRF07_BC and CRF08_BC. Sci Rep 6:19279.

11. Yang R, Kusagawa S, Zhang C, Xia X, Ben K, Takebe Y. 2003. Identification and characterization of a new class of human immunodeficiency virus type 1 recombinants comprised of two circulating recombinant forms, CRF07_BC and CRF08_BC, in China. J Virol 77:685-95.

12. Takebe Y, Liao H, Hase S, Uenishi R, Li Y, Li XJ, Han X, Shang H, Kamarulzaman A, Yamamoto N, Pybus OG, Tee KK. 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 Suppl 2:B39-44.

13. Zhao CY, Li BJ, Chen SL. 2011. Molecular epidemiological investigation of HIV-1 circulating strain infected after blood receiving. Chin J Dis Control Prev 11:36-38.

14. Zhao CY, Zhao HR, Li BJ. 2010. Molecular epidemiology study on HIV infection among paid blood donors. Chin J Health Lab Technol 20:3136-3137.

15. Li D, Ge L, Wang L, Guo W, Ding Z, Li P, Cui Y. 2014. [Trend on HIV prevalence and risk behaviors among men who have sex with men in China from 2010 to 2013]. Zhonghua Liu Xing Bing Xue Za Zhi 35:542-6.

16. NCAIDS, NCSTD, CDC C. 2017. Update on the AIDS/STD epidemic in China in December, 2016. Chin J AIDS STD 23:93.

17. Yin Y, Liu Y, Zhu J, Hong X, Yuan R, Fu G, Zhou Y, Wang B. 2019. The prevalence, temporal trends, and geographical distribution of HIV-1 subtypes among men who have sex with men in China: A systematic review and meta-analysis. Epidemiol Infect 147:e83.

18. Zhao J, Cai W, Zheng C, Yang Z, Xin R, Li G, Wang X, Chen L, Zhong P, Zhang C. 2014. Origin and outbreak of HIV-1 CRF55_01B among MSM in Shenzhen, China. J Acquir Immune Defic Syndr 66:e65-7.

19. Han X, An M, Zhang W, Cai W, Chen X, Takebe Y, Shang H. 2013. Genome Sequences of a Novel HIV-1 Circulating Recombinant Form, CRF55_01B, Identified in China. Genome announcements 1:e00050-12.

20. Wei L, Lu X, Li H, Zheng C, Li G, Yang Z, Chen L, Cheng J, Wang H, Zhao J. 2018. Impact of HIV-1 CRF55_01B infection on CD4 counts and viral load in men who have sex with men naive to antiretroviral treatment. The Lancet 392:S43.

21. Xingyi C. 2017. China's AIDS epidemic in 2017. https://user.guancha.cn/main/content?id=16151. Accessed

22. Qi J, Zhang D, Fu X, Li C, Meng S, Dai M, Liu H, Sun J. 2015. High risks of HIV transmission for men who have sex with men--a comparison of risk factors of HIV infection among MSM associated with recruitment channels in 15 cities of China. PLoS One 10:e0121267.

23. State Council AIDS Working Committee Office UTGoAiC. 2004. A Joint Assessment of HIV/AIDS Prevention, Treatment and Care in China

24. Hong Y, Stanton B, Li X, Yang H, Lin D, Fang X, Wang J, Mao R. 2006. Rural-to-urban migrants and the HIV epidemic in China. AIDS Behav 10:421-30.

25. Hu Z, Liu H, Li X, Stanton B, Chen X. 2006. HIV-related sexual behaviour among migrants and non-migrants in a rural area of China: role of rural-to-urban migration. Public Health 120:339-45.

26. Zhang T, Miao Y, Li L, Bian Y. 2019. Awareness of HIV/AIDS and its routes of transmission as well as access to health knowledge among rural residents in Western China: a cross-sectional study. BMC Public Health 19:1630.

27. Mi G, Ma B, Kleinman N, Li Z, Fuller S, Bulterys M, Hladik W, Wu Z. 2016. Hidden and Mobile: A Web-based Study of Migration Patterns of Men Who Have Sex With Men in China. Clin Infect Dis 62:1443-7.

28. Zong Z, Yang W, Sun X, Mao J, Shu X, Hearst N. 2017. Migration Experiences and Reported Sexual Behavior Among Young, Unmarried Female Migrants in Changzhou, China. Glob Health Sci Pract 5:516-524.

29. Dai W, Gao J, Gong J, Xia X, Yang H, Shen Y, Gu J, Wang T, Liu Y, Zhou J, Shen Z, Zhu S, Pan Z. 2015. Sexual behavior of migrant workers in Shanghai, China. BMC Public Health 15:1067.

30. Su L, Liang S, Hou X, Zhong P, Wei D, Fu Y, Ye L, Xiong L, Zeng Y, Hu Y, Yang H, Wu B, Zhang L, Li X. 2018. Impact of worker emigration on HIV epidemics in labour export areas: a molecular epidemiology investigation in Guangyuan, China. Scientific Reports 8:16046.

31. Hong J, Chu Z, Wang Q. 2011. Transport infrastructure and regional economic growth: evidence from China. Transportation 38:737-752.

32. China Population Publishing House. 2018. Report on China's migrant population development.

33. Lu M, Xia Y. 2016. Migration in the People's Republic of China. Asian Development Bank Institute,

34. Yang B, Wu Z, Schimmele CM, Li S. 2015. HIV knowledge among male labor migrants in China. BMC public health 15:323-323.

35. Zhang L, Chow EP, Jahn HJ, Kraemer A, Wilson DP. 2013. High HIV prevalence and risk of infection among rural-to-urban migrants in various migration stages in China: a systematic review and meta-analysis. Sex Transm Dis 40:136-47.

36. Baele G, Dellicour S, Suchard MA, Lemey P, Vrancken B. 2018. Recent advances in computational phylodynamics. Curr Opin Virol 31:24-32.

37. Müller NF, Dudas G, Stadler T. 2019. Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations. Virus Evolution 5.

607 38. Dellicour S, Vrancken B, Trovao NS, Fargette D, Lemey P. 2018. On the importance of negative
608 controls in viral landscape phylogeography. Virus Evol 4:vey023.

609 39. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, Russell CA, Smith DJ, Pybus OG,
610 Brockmann D, Suchard MA. 2014. Unifying viral genetics and human transportation data to predict
611 the global transmission dynamics of human influenza H3N2. PLoS Pathog 10:e1003932.

612 40. Perez AB, Vrancken B, Chueca N, Aguilera A, Reina G, Garcia-Del Toro M, Vera F, Von Wichman
613 MA, Arenas JI, Tellez F, Pineda JA, Omar M, Bernal E, Rivero-Juarez A, Fernandez-Fuertes E, de
614 la Iglesia A, Pascasio JM, Lemey P, Garcia F, Cuypers L. 2019. Increasing importance of European
615 lineages in seeding the hepatitis C virus subtype 1a epidemic in Spain. Euro Surveill 24.

616 41. Vrancken B, Cuypers L, Perez AB, Chueca N, Anton-Basantas J, de la Iglesia A, Fuentes J, Pineda
617 JA, Tellez F, Bernal E, Rincon P, Von Wichman MA, Fuentes A, Vera F, Rivero-Juarez A, Jimenez
618 M, Vandamme AM, Garcia F. 2019. Cross-country migration linked to people who inject drugs
619 challenges the long-term impact of national HCV elimination programmes. J Hepatol 71:1270-1272.

620 42. Trovão NS, Baele G, Vrancken B, Bielejec F, Suchard MA, Fargette D, Lemey P. 2015. Host
621 ecology determines the dispersal patterns of a plant virus. Virus evolution 1:vev016-vev016.

622 43. Graf T, Vrancken B, Maletich Junqueira D, de Medeiros RM, Suchard MA, Lemey P, Esteves de
623 Matos Almeida S, Pinto AR. 2015. Contribution of Epidemiological Predictors in Unraveling the
624 Phylogeographic History of HIV-1 Subtype C in Brazil. J Virol 89:12341-8.

625 44. Faria NR, Vidal N, Lourenco J, Raghwani J, Sigaloff KCE, Tatem AJ, van de Vijver DAM, Pineda-
626 Pena AC, Rose R, Wallis CL, Ahuka-Mundeke S, Muyembe-Tamfum JJ, Muwonga J, Suchard MA,
627 Rinke de Wit TF, Hamers RL, Ndembi N, Baele G, Peeters M, Pybus OG, Lemey P, Dellicour S.
628 2019. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa.
629 PLoS Pathog 15:e1007976.

630 45. Los Alamos National Laboratory, National Institutes of Health. HIV databases.
631 http://www.hiv.lanl.gov/. Accessed

632 46. Xiao P, Li J, Fu G, Zhou Y, Huan X, Yang H. 2017. Geographic Distribution and Temporal Trends of
633 HIV-1 Subtypes through Heterosexual Transmission in China: A Systematic Review and Meta-
634 Analysis. International journal of environmental research and public health 14:830.

635 47. Wang X, He X, Zhong P, Liu Y, Gui T, Jia D, Li H, Wu J, Yan J, Kang D, Han Y, Li T, Yang R, Han
636 X, Chen L, Zhao J, Xing H, Liang S, He J, Yan Y, Xue Y, Zhang J, Zhuang X, Liang S, Bao Z, Li T,
637 Zhuang D, Liu S, Han J, Jia L, Li J, Li L. 2017. Phylodynamics of major CRF01_AE epidemic
638 clusters circulating in mainland of China. Sci Rep 7:6330.

639 48. Baidu Map Eyes Big Data Team. 2015. Analysis Report of Big Data in Hometown of China.

49. su Y, Tesfazion P, Zhao Z. 2017. Where are migrants from? Inter- vs. intra-provincial rural-urban migration in China. China Economic Review 47.

50. Qiao Y-c, Xu Y, Jiang D-x, Wang X, Wang F, Yang J, Wei Y-s. 2019. Epidemiological analyses of regional and age differences of HIV/AIDS prevalence in China, 2004–2016. International Journal of Infectious Diseases 81:215-220.

51. Vrancken B, Adachi D, Benedet M, Singh A, Read R, Shafran S, Taylor GD, Simmonds K, Sikora C, Lemey P, Charlton CL, Tang JW. 2017. The multi-faceted dynamics of HIV-1 transmission in Northern Alberta: A combined analysis of virus genetic and public health data. Infect Genet Evol 52:100-105.

52. Li Z, He X, Wang Z, Xing H, Li F, Yang Y, Wang Q, Takebe Y, Shao Y. 2012. Tracing the origin and history of HIV-1 subtype B' epidemic by near full-length genome analyses. Aids 26:877-84.

53. Chu XG, Zhang XF, Zhan FX, Tang H, Chen HP, Peng TH, Gong ZJ. 2007. [Study on molecular epidemiology of people infected with human immunodeficiency virus-1 in Hubei province]. Zhonghua Liu Xing Bing Xue Za Zhi 28:992-5.

54. Qian S, Guo W, Xing J, Qin Q, Ding Z, Chen F, Peng Z, Wang L. 2014. Diversity of HIV/AIDS epidemic in China: a result from hierarchical clustering analysis and spatial autocorrelation analysis. Aids 28:1805-13.

55. Shan H, Wang JX, Ren FR, Zhang YZ, Zhao HY, Gao GJ, Ji Y, Ness PM. 2002. Blood banking in China. Lancet 360:1770-5.

56. Zeng P, Wang J, Huang Y, Guo X, Li J, Wen G, Yang T, Yun Z, He M, Liu Y, Yuan Y, Schulmann J, Glynn S, Ness P, Jackson JB, Shan H, Nhlbi Retrovirus Epidemiology Donor Study-Ii IC. 2012. The human immunodeficiency virus-1 genotype diversity and drug resistance mutations profile of volunteer blood donors from Chinese blood centers. Transfusion 52:1041-9.

57. Zhang L, Chen Z, Cao Y, Yu J, Li G, Yu W, Yin N, Mei S, Li L, Balfe P, He T, Ba L, Zhang F, Lin HH, Yuen MF, Lai CL, Ho DD. 2004. Molecular characterization of human immunodeficiency virus type 1 and hepatitis C virus in paid blood donors and injection drug users in china. J Virol 78:13591-9.

58. Government of China. 2016. Chinese long-term railway network plan

59. Luo MY, Pan XH, Fan Q, Zhang JF, Ge R, Jiang J, Chen WJ. 2019. [Epidemiological characteristics of molecular transmission cluster among reported HIV/AIDS cases in Jiaxing city, Zhejiang province, 2017]. Zhonghua Liu Xing Bing Xue Za Zhi 40:202-206.

671 60. Han ZG, Zhang YL, Wu H, Gao K, Zhao YT, Gu YZ, Chen YC. 2018. [Prevalence of drug resistance
672 in treatment-naive HIV infected men who have sex with men in Guangzhou, 2008-2015]. Zhonghua
673 Liu Xing Bing Xue Za Zhi 39:977-982.

674 61. Xiao P, Zhou Y, Lu J, Yan L, Xu X, Hu H, Li J, Ding P, Qiu T, Fu G, Huan X, Yang H. 2019. HIV-1
675 genotype diversity and distribution characteristics among heterosexually transmitted population in
676 Jiangsu province, China. Virology Journal 16:51.

677 62. Yin Y, Liu Y, Zhu J, Hong X, Yuan R, Fu G, Zhou Y, Wang B. 2019. The prevalence, temporal
678 trends, and geographical distribution of HIV-1 subtypes among men who have sex with men in
679 China: A systematic review and meta-analysis. Epidemiology and infection 147:e83-e83.

680 63. Liang Z, Li Z, Ma Z. 2014. Changing Patterns of the Floating Population in China during 2000-2010.
681 Population and development review 40:695-716.

682 64. Poon CM, Wong NS, Kwan TH, Wong HTH, Chan KCW, Lee SS. 2018. Changes of sexual risk
683 behaviors and sexual connections among HIV-positive men who have sex with men along their HIV
684 care continuum. PLoS One 13:e0209008.

685 65. Dennis AM, Volz E, Frost A, Hossain M, Poon AFY, Rebeiro PF, Vermund SH, Sterling TR, Kalish
686 ML. 2018. HIV-1 Transmission Clustering and Phylodynamics Highlight the Important Role of Young
687 Men Who Have Sex with Men. AIDS Res Hum Retroviruses 34:879-888.

688 66. Chaillon A, Delaugerre C, Brenner B, Armero A, Capitant C, Nere ML, Leturque N, Pialoux G, Cua
689 E, Tremblay C, Smith DM, Goujard C, Meyer L, Molina JM, Chaix ML. 2019. In-depth Sampling of
690 High-risk Populations to Characterize HIV Transmission Epidemics Among Young MSM Using PrEP
691 in France and Quebec. Open Forum Infect Dis 6:ofz080.

692 67. Pybus OG, Tatem AJ, Lemey P. 2015. Virus evolution and transmission in an ever more connected
693 world. Proceedings of the Royal Society B: Biological Sciences 282:20142878.

694 68. Holmes EC. 2008. Evolutionary History and Phylogeography of Human Viruses. Annual Review of
695 Microbiology 62:307-328.

696 69. Xia Y, Bjornstad ON, Grenfell BT. 2004. Measles metapopulation dynamics: a gravity model for
697 epidemiological coupling and dynamics. Am Nat 164:267-81.

698 70. Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots.
699 PLoS Computational Biology 5.

700 71. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010. Phylogeography takes a relaxed random walk
701 in continuous space and time. Molecular biology and evolution 27:1877-1885.

72. Faria NR, Hodges-Mameletzis I, Silva JC, Rodés B, Erasmus S, Paolucci S, Ruelle J, Pieniazek D, Taveira N, Treviño A, Gonçalves MF, Jallow S, Xu L, Camacho RJ, Soriano V, Goubau P, de Sousa JD, Vandamme A-M, Suchard MA, Lemey P. 2012. Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. The Journal of general virology 93:889-899.

73. Chaillon A, Gianella S, Dellicour S, Rawlings SA, Schlub TE, Faria De Oliveira M, Ignacio C, Porrachia M, Vrancken B, Smith DM. 2020. HIV persists throughout deep tissues with repopulation from multiple anatomical sources. J Clin Invest doi:10.1172/jci134815.

74. Liu X, Erasmus V, Wu Q, Richardus JH. 2014. Behavioral and Psychosocial Interventions for HIV Prevention in Floating Populations in China over the Past Decade: A Systematic Literature Review and Meta-Analysis. PLOS ONE 9:e101006.

75. Li X, Gao R, Zhu K, Wei F, Fang K, Li W, Song Y, Ge Y, Ji Y, Zhong P, Wei P. 2018. Genetic transmission networks reveal the transmission patterns of HIV-1 CRF01_AE in China. Sex Transm Infect 94:111-116.

76. Cuypers L, Vrancken B, Fabeni L, Marascio N, Cento V, Di Maio VC, Aragri M, Pineda-Pena AC, Schrooten Y, Van Laethem K, Balog D, Foca A, Torti C, Nevens F, Perno CF, Vandamme AM, Ceccherini-Silberstein F. 2017. Implications of hepatitis C virus subtype 1a migration patterns for virus genetic sequencing policies in Italy. BMC Evol Biol 17:70.

77. Vrancken B, Alavian SM, Aminy A, Amini-Bavil-Olyaee S, Pourkarim MR. 2018. Why comprehensive datasets matter when inferring epidemic links or subgenotyping. Infect Genet Evol 65:350-351.

78. Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147:195-7.

79. Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics (Oxford, England) 30:3276-3278.

80. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. PLoS One 5.

81. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-21.

82. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology 52:696-704.

83. Shimodaira H, Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16:1114-1114.

84. Al-Qahtani AA, Baele G, Khalaf N, Suchard MA, Al-Anazi MR, Abdo AA, Sanai FM, Al-Ashgar HI, Khan MQ, Al-Ahdal MN, Lemey P, Vrancken B. 2017. The epidemic dynamics of hepatitis C virus subtypes 4a and 4d in Saudi Arabia. Sci Rep 7:44947.

85. Zhang Y, Vrancken B, Feng Y, Dellicour S, Yang Q, Yang W, Zhang Y, Dong L, Pybus OG, Zhang H, Tian H. 2017. Cross-border spread, lineage displacement and evolutionary rate estimation of rabies virus in Yunnan Province, China. Virol J 14:102.

86. Abecasis AB, Vandamme AM, Lemey P. 2009. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. J Virol 83:12917-24.

87. Patino-Galindo JA, Gonzalez-Candelas F. 2017. The substitution rate of HIV-1 subtypes: a genomic approach. Virus Evol 3:vex029.

88. Vrancken B, Baele G, Vandamme AM, van Laethem K, Suchard MA, Lemey P. 2015. Disentangling the impact of within-host evolution and transmission dynamics on the tempo of HIV-1 evolution. Aids 29:1549-56.

89. de Goede AL, van Deutekom HW, Vrancken B, Schutten M, Allard SD, van Baalen CA, Osterhaus AD, Thielemans K, Aerts JL, Kesmir C, Lemey P, Gruters RA. 2013. HIV-1 evolution in patients undergoing immunotherapy with Tat, Rev, and Nef expressing dendritic cells followed by treatment interruption. Aids 27:2679-89.

90. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4.

91. Edwards CJ, Suchard MA, Lemey P, Welch JJ, Barnes I, Fulton TL, Barnett R, O'Connell TC, Coxon P, Monaghan N, Valdiosera CE, Lorenzen ED, Willerslev E, Baryshnikov GF, Rambaut A, Thomas MG, Bradley DG, Shapiro B. 2011. Ancient hybridization and an Irish origin for the modern polar bear matriline. Current biology: CB 21:1251-1258.

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

93. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307-1320.

764 94. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. 2013. Improving bayesian
765 population dynamics inference: a coalescent-based model for multiple Loci. Molecular biology and
766 evolution 30:713-724.

767 95. Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. 2010. Using time-structured
768 data to estimate evolutionary rates of double-stranded DNA viruses. Mol Biol Evol 27:2038-51.

769 96. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. 2016. SpreaD3: Interactive
770 Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol 33:2167-9.

771 97. Kass RE, Raftery AE. 1995. Bayes Factors. Journal of the American Statistical Association 90:773-
772 795.

773 98. Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov models of
774 evolution. Journal of mathematical biology 56:391-412.

775 99. Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian
776 coalescent-based inference of population dynamics. Molecular Biology and Evolution 25:1459-1471.

777 100. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior Summarization in
778 Bayesian Phylogenetics Using Tracer 1.7. Syst Biol 67:901-904.

779 101. National Bureau of Statistics of China. 2016. Nationwide Population Census
780 http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm. Accessed

781 102. China National Center for Disease Control and Prevention. 2016. The number of HIV
782 infections in China, 2016. http://www.chinacdc.cn/en/. Accessed

783 103. Woolley-Meza O, Thiemann C, Grady D, Lee JJ, Seebens H, Blasius B, Brockmann D.
784 2011. Complexity in human transportation networks: a comparative analysis of worldwide air
785 transportation and global cargo-ship movements. The European Physical Journal B 84:589-600.

786 104. China Railway. 2019. Travel Time by Train in China. https://www.12306.cn/index/.
787 Accessed

788 105. McRae BH, Dickson BG, Keitt TH, Shah VB. 2008. Using circuit theory to model connectivity
789 in ecology, evolution, and conservation. Ecology 89:2712-24.

790 106. Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, Hancher M, Poyart E,
791 Belchior S, Fullman N, Mappin B, Dalrymple U, Rozier J, Lucas TCD, Howes RE, Tusting LS, Kang
792 SY, Cameron E, Bisanzio D, Battle KE, Bhatt S, Gething PW. 2018. A global map of travel time to
793 cities to assess inequalities in accessibility in 2015. Nature 553:333-336.

107. Dellicour S, Rose R, Pybus OG. 2016. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. BMC Bioinformatics 17:82.

FIGURES AND TABLES

Figure 1. Migration events between Province in China. The thickness of the arrows corresponds to the average number of inferred migration events, their curvature indicates the migration direction, and their colors reflect the support for each link (green, orange, purple for respectively $3 \leq BF_{adj} < 10$ (substantial), $10 \leq BF_{adj} < 20$ (positive) and $BF_{adj} \geq 20$(strong)). Provinces are colored according to the number of sequences included in the clusters for each HIV type.

Figure 2. Migration events between provinces in China. Sankey plot showing the proportion of migration events from each source province toward the recipient provinces. Left side of the plots shows the source of migration events. Right side of the plot shows the destination of migration events. Only results with adjusted Bayes Factors (BF_{adj}) ≥3 are shown. Panel A to E for types CRF_01AE, B, CRF_07BC, CRF_08BC and CRF55_01B respectively.

Figure 3. Predictors of transition rates among locations. A. The boxplots report the posterior distribution of each GLM coefficient, i.e. the contribution of each predictor to the model, when included in the model (the conditional effect size). The adjusted BFs after accounting for sampling heterogeneity are reported when ≥3, and the corresponding conditional effect sizes are plotted in darker grey. **B.** Map of the Chinese Provinces. **C.** Variables tested as predictors of dispersal transition rates across locations. (A) Number of HIV Cases per provinces were obtained from the National Bureau of Statistics of China (101) and the China National Center for Disease Control and Prevention(102); (B) the population size (in million), (C) Number of Emigrants and (D) Immigrants were obtained from the National Bureau of Statistics of China (101). The numbers of sequences sampled at the origin/destination (See **Figure 1**) were also included in the GLM to account for the potential impact of sampling biases within the analysis (39).

Table 1. Relative importance of provinces in the interprovincial spread of the main HIV types in China. For each province, the percentage refers to the average proportion that the province is as the source (from) or recipient (to) if the emigration event. For each type, the three most relevant provinces are listed.

Type	From	Mean [95%HPD]	To	Mean [95% HPD]
CRF_01AE	Beijing	24.9% [24.8-25]	Shanghai	25.7% [25.5-25.8]
	Guangdong	16.6% [16.5-16.7]	Beijing	17.2% [17.1-17.3]
	Shanghai	15.8% [15.7-15.9]	Anhui	13.3% [13.2-13.4]
	Anhui	10.6% [10.5-10.7]	Guangdong	11.7% [11.6-11.8]
	Shandong	5.6% [5.5-5.6]	Jiangsu	10.6% [10.5-10.7]
	Zhejiang	5.4% [5.4-5.5]	Henan	7.5% [7.4-7.6]
	Liaoning	5.3% [5.3-5.4]	Guangxi	4.4% [4.4-4.5]
	Guangxi	3.9% [3.8-3.9]	Liaoning	3.1% [3.1-3.2]
	Sichuan	3.3% [3.3-3.4]	Shandong	2.4% [2.4-2.5]
	Jiangsu	3.1% [3.1-3.2]	Zhejiang	2.3% [2.3-2.3]
	Henan	2.3% [2.3-2.3]	Sichuan	1.1% [1-1.1]
	Fujian	1.8% [1.8-1.8]	Chongqing	0.5% [0.5-0.5]
	Hebei	1.2% [1.2-1.2]		
CRF_07BC	Beijing	43.4% [43.1-43.8]	Guangdong	43.4% [43.1-43.8]
	Shanghai	29.9% [29.5-30.3]	Zhejiang	39.2% [38.8-39.6]
	Yunnan	15.7% [15.4-16]	Beijing	10.2% [10-10.5]
	Xinjiang	7.2% [7-7.4]	Xinjiang	6.4% [6.2-6.6]
	Liaoning	3.7% [3.6-3.9]	Ningxia	0.7% [0.7-0.8]
CRF_08BC	Yunnan	85.2% [84.8-85.7]	Guangxi	44.2% [43.6-44.9]
	Guangdong	14.8% [14.3-15.2]	Guangdong	32.4% [31.8-33]
			Hebei	11.3% [10.8-11.7]
			Shanghai	8.6% [8.2-9]
			Sichuan	3.5% [3.3-3.7]
B	Hubei	69.9% [69.7-70.1]	Henan	55.8% [55.6-56]
	Henan	16.7% [16.6-16.9]	Guangdong	10.9% [10.8-11]
	Liaoning	5.4% [5.3-5.5]	Zhejiang	9.8% [9.7-10]
	Zhejiang	3% [2.9-3.1]	Hubei	7.5% [7.4-7.6]
	Hebei	2.1% [2-2.1]	Beijing	7.4% [7.2-7.5]
	Beijing	1.3% [1.3-1.4]	Anhui	5.6% [5.5-5.7]
	Guangdong	0.9% [0.9-0.9]	Shandong	1% [0.9-1]
	Anhui	0.3% [0.3-0.4]	Hebei	0.5% [0.5-0.6]
	Yunnan	0.2% [0.2-0.2]	Guangxi	0.4% [0.3-0.4]
	Shandong	0.1% [0-0.1]	Jiangsu	0.3% [0.3-0.4]
	Fujian	0% [0-0]	Shanghai	0.3% [0.3-0.3]
	Jilin	0% [0-0]	Jilin	0.2% [0.2-0.2]
			Liaoning	0.2% [0.1-0.2]
			Ningxia	0.2% [0.1-0.2]
			Fujian	0% [0-0]
			Yunnan	0% [0-0]
CRF_5501B	Guangdong	100% [99.8-100]	Anhui	73.9% [71.8-75.8]
			Hunan	26.1% [24.2-28.2]