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

Yutaka Takebea,∗, Huanan Liaoa, Saiki Hasea, Rie Uenishia, Yue Li, Xiao-Jie Li, Xiaoxu Han, Hong Shang, Adeeba Kamarulzaman,e, Naoki Yamamotob, Oliver G. Pybusf, Kok Keng Teea,e

aLaboratory of Molecular Virology and Epidemiology, AIDS Research Center, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan
bAIDS Research Center, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan
cDepartment of Dermatology, The First People’s Hospital, Shanghai Jiao Tong University, 85 Wujin Road, Shanghai 200080, The People’s Republic of China
dKey Laboratory of AIDS Immunology, Ministry of Health, The First Affiliated Hospital, China Medical University, Shenyang, Liaoning 110001, The People’s Republic of China
eCenter of Excellence for Research in AIDS (CERiA), University of Malaya, 50603 Kuala Lumpur, Malaysia
fDepartment of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom

Article history:
Received 18 June 2009
Received in revised form 21 July 2009
Accepted 24 July 2009

Keywords:
HIV-1
Circulating recombinant form (CRF)
Phylodynamics
BEAST
Time of the most recent common ancestor (tMRCA)
East Asia
Vaccine strategies

1. Introduction

The high genetic diversity of HIV-1 adds considerable complexity to the development of efficacious vaccines [1,2]. The complex genetic structure of HIV-1 epidemics can be seen in regions such as Asia, where various HIV-1 strains are co-circulating within and among different risk populations. Widely circulating HIV-1 strains in Asia [3] include subtypes B, B′ (a Thailand variant of subtype B, also referred to as Thai-B) [4,5] and subtype C, as well as the circulating recombinant forms (CRFs) CRF01_AE [6], CRF07_BC, and CRF08_BC [8]. New CRFs have been reported in several regions in Asia, including CRF15_01B and CRF34_01B in Thailand [9,10] and CRF33_01B in Malaysia [11]. Moreover, many unique recombinant forms (URFs) have been detected in Asia, particularly in Myanmar and the western Yunnan province of China [3,12,13].

CRF07_BC and CRF08_BC are two closely related recombinants derived from Thai subtype B′ and Indian subtype C lineages that have had generated epidemics among injecting drug users (IDUs) in China since their discoveries in 1997 in Xinjiang and Guangxi [7,8], respectively. Their origins have been traced back to the Yunnan province [12,14] of southwestern China that borders the ‘Golden Triangle’ region of Southeast Asia and which was the world’s largest heroin producing region. Yunnan plays an important role as an entry point for heroin smuggling into China [15] and is considered to be an epicenter of HIV/AIDS in China: an HIV-1 outbreak was first detected among injecting drug users (IDUs) in Yunnan in 1989 [12,13,16,17].

The early phase of the HIV-1 epidemic in Yunnan was initiated by subtype B strains of both North American and Southeast Asian (B′) origin [4,5], with subtype B′ subsequently becoming the dominant strain among IDUs in the region [18,19]. The subtype distribution changed in the early 1990s after a subtype C lineage of Indian origin was introduced, which then became the predominant circulating strain [20]. Co-circulation of subtypes B and C in the region led to the generation of various phylogenetically distinct B/C recombinants, most notably CRF07_BC and CRF08_BC,
which subsequently spread outside Yunnan province. Drug trafficking activities have been implicated in the spread of CRF07_BC and CRF08_BC across China [15,17].

Fig. 1 illustrates the chronology of the identification of HIV-1 outbreaks in IDUs in Asia. The epidemic began among IDUs in Thailand during late 1987 and early 1988 and subsequent epidemics were detected almost concurrently among IDUs in Myanmar and Yunnan in 1989. By 1994–1995, Yunnan province accounted for more than 80% of HIV-1 case reports in China. Since then, a series of outbreaks have been reported in various parts of China outside Yunnan: in 1996 in Xinjiang; in 1997 in Guangxi and Sichuan; and in 1998 in Guangdong and other regions. In the early 2000s, Taiwan experienced a dramatic increase in HIV-1 infections among IDUs, initially in 2002 in Tainan, then later spreading northwards to Nantou and Taipei in 2004 [21,22]. HIV-1 subtype B and CRF01_AE are responsible for HIV-1 epidemics among IDUs in Thailand [4,5], Myanmar [13] and Vietnam [23,24], while CRF07_BC and CRF08_BC are responsible for epidemics among IDUs in East Asia (i.e. mainland China [7,8] and Taiwan [21]) (Fig. 1).

In recent years, the development of new statistical analysis techniques and the growing size of HIV sequence databases have enabled researchers to trace the origins of HIV [25], estimate the time of the most recent common ancestor (tMRCA) of epidemic strains, and estimate past trends in rates of infection [26,27]. These developments have coincided with new experimental methods in immunology [28], which have expanded our understanding of host defenses against the virus and the consequences of immune pressure (e.g. immune escape). As a result, it is now understood that HIV vaccinology must take into account global heterogeneity in virus diversity and host immune pressure (e.g. HLA haplotype). Evolutionary analyses of HIV genomes have the potential to provide in-depth knowledge of the evolutionary and epidemiological dynamics of HIV viruses, such as those currently circulating in Asia. Such information enhances and informs our efforts in vaccine design, development and potential deployment.

Here we report an investigation of the temporal and spatial dynamics of HIV-1 CRF07_BC and CRF08_BC across East Asia. This was achieved by reconstructing the evolutionary and epidemic histories of these strains using recently developed Bayesian phylogenetic methods. In addition, we also assessed the recent evolutionary history of subtype C in order to gain a more comprehensive picture of the movement of subtype C and its descendant recombinants in Asia.

2. Materials and methods

2.1. HIV-1 CRF07_BC and CRF08_BC sequence information

A total of 138 nucleotide sequences belonging to CRF07_BC (mainland China: n = 22; Taiwan: n = 35) and CRF08_BC (mainland China: n = 81) with known sampling dates were retrieved from the Los Alamos HIV Sequence Database (www.hiv.lanl.gov). All CRF08_BC sequences were from various parts in mainland China: Yunnan (n = 31); Guangxi (n = 44); Liaoning (n = 5); Gansu (n = 1) [8,12,29–32]. CRF07_BC sequences were from Yunnan (n = 10), Xingjiang (n = 6) and Liaoning (n = 6) provinces [7,8,12,29,31–33], and from three cities in Taiwan, Tainan (southern: n = 19), Nantou (central: n = 8) and Taipei (north: n = 8). Most of the sequences from Yunnan and Liaoning have been previously reported by our laboratory [12,34].

For phylogenetic and evolutionary analyses of CRF07_BC and CRF08_BC, we selected the non-recombinant gag region of subtype C origin, encompassing the p17 and partial p24 proteins (HXB2 790–1218), and the non-recombinant gag-pol region, spanning from p2 in gag to the reverse transcriptase in pol (HXB2 1918–2852). For the phylogenetic reconstruction of CRF07_BC strains in Taiwan (con-
ducted to provide a comparison with the mainland China analysis), we used the hypervariable region-striped env gene (HXB2 7077-7665 nt) [22], because sufficient numbers of env sequences were only available for CRF07_BC from Taiwan. Additionally, subtype C or subtype C-related sequences from India and China (n = 124) were also analyzed in order to determine the date of origin of these lineages.

### 2.2. Bayesian MCMC evolutionary analyses

Rates of evolution in different parts of the HIV-1 genome were estimated from a reference set of subtype C sequences using BEAST v1.4 [27], a program that employs a Markov chain Monte Carlo (MCMC) algorithm to estimate evolutionary parameters. To perform these analyses, subtype C regions of gag and gag-pol that correspond to the non-recombinant subtype C segments of CRF07_BC and CRF08_BC, respectively, were retrieved from the HIV Sequence Database. The resulting data set contains 41 subtype C sequences sampled at known times between 1989 and 2005. Maximum likelihood phylogenies were estimated for each sub-genomic region of CRF07 BC and CRF08 BC and also for the heterogenous subtype C reference data set using PAUP* v4.0 beta [35] and BEAST v1.4 [27]. Strict and relaxed molecular clock analyses were performed under the general time-reversible (GTR) [37] and Hasegawa-Kishino-Yano (HKY) [38] nucleotide substitution models, with a gamma-distributed model of among-site rate variation, approximated using four rate categories (Γ4) [39], thereby providing estimates of the most of the most common ancestor (tMRCA) of various nodes [34,40]. In order to estimate the tMRCA of Taiwan BC, the hypervariable region-striped env gene (HXB2 7077-7665 nt) was subjected to analysis [22].

### 2.3. Evolutionary analysis of subtype C

To estimate the times of divergence of subtype C in India and China, non-recombinant subtype C genome regions from 12 near full-length subtype C sequences of Indian origin (including a near full-length sequence from Myanmar [13]) were retrieved from the HIV Sequence Database, together with 112 subtype C env sequences from China (HXB2 6984-7328). The methods described above were used to estimate phylogenies, genome region-specific evolutionary rates, and tMRCA for the African, Indian, and Chinese strains. Furthermore, subtype C sequences from other locations (31 full-length nucleotide sequences from Africa and Brazil) were also studied to assess the time of origin of the global HIV-1 subtype C [34].

### 3. Results

#### 3.1. Estimated timeline of overland expansion of CRF08_BC in China

Phylogeneties estimated from the gag-pol alignment showed that all CRF08_BC sequences grouped in a single clade containing strains from four Chinese regions (Yunnan, Gansu, Liaoning and Guangxi) and were descended from a parental subtype C lineage of Indian origin. Whilst sequences from Yunnan [12] were intermingled with those from other regions (including Gansu, Liaoning and Baise in Guangxi province), the sequences from Binyang and Pingxiang formed a distinct monophyletic group that diverged from the Yunnan/Baise strains, and which could be split into three clusters (designated clusters 1–3). Cluster 1 contained sequences from Binyang (n = 30) and Pingxiang (n = 2), while clusters 2 and 3 contained seven and two Binyang sequences, respectively. CRF08_BC sequences from Liaoning also formed a single cluster (except one subject, who acquired infection in Yunnan). The robustness of each cluster was supported by high posterior clade probabilities (p > 0.95) [34]. The rate of evolution (μ) of the CRF08_BC gag-pol genome region was calculated using Bayesian analysis under various evolutionary models, and was estimated to be μ = (1.8–1.9) × 10⁻³ substitutions/site/year. As summarized in Fig. 2, when the HKY+Γ4 substitution and constant size model were used, the Yunnan (and Gansu) clusters were dated to 1990.3 (95% credible region, CR: 1988.6–1991.9). CRF08_BC was introduced later into Baise at 1995.5 (95% CR: 1994.3–1996.5) and Binyang at 1997.1 (95% CR: 1996.3–1997.9). 1998.3 (95% CR: 1997.0–1999.4) and 1998.5 (95% CR: 1997.3–1999.5) for clusters 1, 2 and 3, respectively (Fig. 2). Hence, the subsequent spread of CRF08_BC from Baise to Binyang and Pingxiang in Guangxi involved at least three lineages (clusters 1–3) that descended from the Yunnan/Baise strains between 1997 and 1999 (Fig. 2). Finally, the tMRCA of CRF08_BC in Pingxiang was dated to 1999.3 (95% CR: 1998.5–1999.9). Interestingly, the tMRCA of the Liaoning cluster is estimated to be 1995.6 (95% CR: 1993.4–1997.5), comparable to that of Guangxi [34]. This suggests that the epidemics in Liaoning, an area with low HIV/AIDS prevalence, and Guangxi could have started at almost the same time.

#### 3.2. Estimated timeline of the expansion of CRF07 BC in East Asia

Phylogenetic reconstruction revealed that all CRF07_BC sequences from mainland China (Yunnan, Xinjiang and Liaoning) as well as those from Taiwan formed a single cluster. Indian subtype C lineages are paraphyletic with respect to CRF07_BC, confirming an Indian subtype C origin for CRF07_BC. These CRF07_BC sequences were grouped into three distinct phylogenetic clusters: mainland China (Yunnan, Xinjiang and Liaoning); southern Taiwan (Tainan); central-northern Taiwan (Nantou-Taipei) clusters. Bayesian coalescent analyses dated the tMRCA of CRF07_BC in mainland China to 1993.3 (95% CR: 1991.2–1995.2), and the tMRCA in southern and central-northern Taiwan to 1999.7 (95% CR: 1998.4–2001.1) and 2002.1 (95% CR: 2001.3–2002.9) [22,34] (Fig. 2). The dates were estimated using a constant size population model and the following evolutionary rates: μ = 4.4 × 10⁻³ substitutions/site/year for the gag region [34] and μ = (4.7–5.0) × 10⁻³ substitutions/site/year for the hypervariable region-striped env gene [22].

#### 3.3. Phylodynamics of HIV-1 subtype C in Asia

In order to elucidate the evolutionary relationships between HIV-1 subtype C and closely related recombinants in Asia, we performed phylogenetic and coalescent analyses using env sequences, since the env regions of CRF07_BC and CRF08_BC belong to subtype C. Maximum clade credibility trees (which summarize the phylogenetic output of the Bayesian coalescent analysis) revealed that subtype C env sequences from India, Myanmar and China grouped in a single cluster, together with CRF07_BC, CRF08_BC and other related recombinants from China. The most basal outgroup strains were subtype C sequences of Indian origin [34]. Taken together, the phylogenetic and coalescent results map the eastward movement of subtype C from India to China, probably through Myanmar [15], between mid-1970s (1976.9; 95% CR: 1972.1–1981.5) and the early 1980s (1981.2; 95% CR: 1976.7–1985.9), respectively. In addition, we estimated the date of origin of HIV-1 subtype C worldwide to be 1967.6 (95% CR: 1962.5–1972.0), highly consistent with previous estimates [41].
Fig. 2. Estimated timescale of the spread of HIV-1 subtype C, CRF07_BC and CRF08_BC in Asia. The times of the most recent common ancestors of HIV-1 strains circulating in various regions in Asia are shown. These were estimated using phylogenetic and Bayesian coalescent analyses (see Methods). In this particular illustration, we adopted the data set based on HKY+4 constant size model in our previous study [34].

Fig. 3. Plausible origin and migration routes of CRF07_BC, CRF08_BC and subtype C in Asia. The temporal and spatial dynamics of the spread of subtype C and related recombinants (CRF07_BC and CRF08_BC) is illustrated. HIV-1 subtype C of Indian origin entered Yunnan province in southwestern China in the early 1980s, possibly via Myanmar. Co-circulation of subtype C and the endemic subtype B' [18,19] led to the generation of B'/C inter-subtype recombinants, CRF07_BC [7] and CRF08_BC [8] in Yunnan. Further spread of CRF07_BC and CRF08_BC into other regions during the 1990s-mainly through injecting drug use [15] is thought to be the major force spurring the HIV/AIDS epidemic in China. The times of the most recent common ancestors (tMRCA) of CRF07_BC, CRF08_BC and parental subtype C strains representing different geographical locations are indicated (see Fig. 2).
4. Discussion

In the present study, we conducted an extensive phylogeny-based study of the dates of origin and geographic migration patterns of CRF07_BC, CRF08_BC and their parental subtype C lineages across Asia (Fig. 3). Prior to the emergence of CRF07_BC and CRF08_BC [7,8], subtype C of Indian origin predominated in southwestern China [20], which appears to have been introduced in the early 1980s (~1981.2). Later, subtype C recombinated with subtype B’ [18,19,42] to form these recombinants. Both recombinants are thought to have originated from Yunnan province [12,14] and later dispersed to other areas, primarily through drug trafficking activities [15,17] (Fig. 3).

The Bayesian coalescent analyses show that CRF08_BC emerged in Yunnan in the early 1990s (~1990.3). The virus then spread eastward to neighboring Guangxi (in Baise city near the Yunnan-Guangxi border) in the mid-1990s (~1995.5) (Figs. 2 and 3), consistent with the historical accounts of CRF08_BC infection among IDUs in this region [43]. Subsequently, highly homogeneous CRF08_BC strains from Baise [8,44] were multiply introduced into the IDU population in Binyang county during 1996–1999, forming multiple clusters (1 through 3) (Figs.2 and 3) [29]. CRF08_BC was then introduced into Pingxiang (near the China–Vietnam border) in the late 1990s, although CRF01_AE strains closely related to northern Vietnam strains account for the majority of IDU infections in this region [23,24,29]. It is interesting to note that the CRF08_BC epidemic in Liaoing province, an area of low HIV prevalence in northeastern China, started at almost the same time as that in Guangxi (around 1995–1996). Our results imply a rapid and simultaneous expansion of CRF08_BC from a common origin in Yunnan, to Guangxi and as far as Liaoing [34] (Fig. 3).

Our results show that CRF07_BC strains from different regions in mainland China (including Xinjiang, Liaoing, and probably Guangdong and Sichuan) likely share a common ancestor that existed around 1993 in Yunnan province. This suggests that CRF07_BC spread almost simultaneously to various regions in China soon after its generation, which probably occurred in Yunnan. Furthermore, CRF07_BC has disseminated beyond mainland China: CRF07_BC was first introduced into southern Taiwan (Tainan) around 2000 (~1999.7) (Fig. 2) probably via southeastern provinces in China (i.e. Fujian and Guangdong, which have strong social and demographic ties with Taiwan) (Figs. 2 and 3) [22,34]. CRF07_BC later spread to the central-northern part of Taiwan (Nantou and Taipei) in the early 2000s (~2002.1), resulting in the largest ever HIV epidemic among IDUs in Taiwan [21,22].

As described in Section 2.2, we performed analyses under various evolutionary and substitution models and across different regions of HIV-1 genome. In most cases, the choice of different models and the regions used in the analysis did not have a significant impact on tMRCA estimation [34]. We perform analyses on multiple regions and under different models in order to demonstrate that our conclusions are robust to the models we chose.

Our estimated dates of common ancestry should properly be considered as conservative, most recent, possible dates, particularly for clusters containing small number of sequences. The start of an outbreak may predate our estimates either because extant virus diversity has been partially or non-randomly sampled, or because the founding/parental lineages of the outbreak have since gone extinct.

Because of the similarity of their recombinant structures, McCutchan et al. proposed that CRF07_BC and CRF08_BC were generated from a putative common ancestral B/C recombinant [14]. Our studies of all available CRF07_BC and CRF08_BC sequences suggest that CRF08_BC has an earlier evolutionary history than CRF07_BC, at least for the env subtype C region. It is thus tempting to speculate that CRF08_BC (estimated tMRCA: ~1990) was probably generated 10 years after subtype C of Indian origin was first introduced to Yunnan province in ~1981, and that CRF07_BC was subsequently formed from CRF08_BC in the same province a few years later, around 1993 (Figs.2 and 3), by additional recombination events with one or more other subtype B’ strains. More samples are necessary to estimate the exact origin of CRF07_BC and CRF08_BC, particularly isolates that were sampled early during the epidemic. The study of Qui et al. [45] has shown that the origins of CRF07_BC and CRF08_BC appear to be complex. Therefore, more evidence is necessary to resolve our understanding of the various probabilities of recombination among HIV-1 subtypes B’ and C in southwestern China and to reconstruct the history of HIV-1 in the region with greater confidence.

Although our study indicates that CRF08_BC may have emerged earlier than CRF07_BC (Fig. 2), CRF08_BC appears not to have spread beyond Yunnan province before ~1995. This suggests that CRF08_BC was circulating within Yunnan, yet confined there, until it spread in 1995 to nearby Guangxi province and to Liaoing. In contrast, CRF07_BC, whose origin in Yunnan is dated to around 1993, appears to have spread very soon after to various geographically disparate regions of China. This may reflect a difference in the epidemiologic characteristics of the IDU transmission networks that carry these two CRFs.

We also note that the early Indian subtype C strains play no major role in today's epidemic in China and appear to have been replaced with related recombinant strains, i.e. CRF07_BC and CRF08_BC. This might suggest the presence of an as yet unidentified selective advantage of subtype C-related recombinants (CRF07_BC and CRF08_BC) over the parental (non-recombinant) subtype C strains. Alternatively, this may simply reflect epidemiologic factors or founder events that have given CRF07_BC and CRF08_BC the opportunity to spread more rapidly than other subtype C lineages.

Finally, the genetic diversity of HIV is one of the key issues in HIV vaccinology. Our results may directly relate to this issue, by explaining the origin and maintenance of this diversity in Asian populations. In order to foretell the future of HIV in Asia, we also need to understand its past. Phylogenies and molecular epidemiology reveal whether different strains are circulating in one location, and how fast strains are moving among locations. These factors need to be known if strain-specific HIV vaccines are to be successfully deployed.

In conclusion, our study highlights the relevance of using evolutionary and molecular epidemiological studies in understanding the HIV/AIDS epidemic in Asia. When employed alongside traditional surveillance surveys, phylogenetic analysis can provide additional information pertaining to regional and global public health issues, including the design and development of effective vaccine strategies.

Acknowledgments

The authors gratefully acknowledge Tim Mastro and anonymous reviewers for critical reading of manuscript. This study was supported in part by grants from the Ministry of Health, Labour & Welfare (H18-AIDS-General-016), Ministry of Education, Culture, Sports, Science and Technology (Overseas Research Fund) to Y.T., and the Japanese Foundation for AIDS Prevention (JFAP), and the Ministry of Science, Technology & Innovation (eScienceFund 02-01-03-SF0379, Malaysia) to A. K.; and the Royal Society International Project Fund to O.G.P. and Y.T. K.K.T., H.L., and R.U. are recipients of the JFAP research resident fellowship. S.H. is a recipient of the research resident fellowship from the Human Sciences Foundation.

Conflict of interest statement

The authors state that they have no conflict of interest.
References


