Genetics and Molecular Research

Sequencing of Gag/Env association with HIV genotyping resolution and HIV-related epidemiologic studies of HIV in China

L. Ren[1,2,4]*, H.W. Wang[3]*, Y. Xu[3], Y. Feng[2], H.F. Zhang[2] and K.H. Wang[1,3]

[1]The Affiliated Hospital of Kunming University of Science and Technology, Kunming, Yunnan Province, China
[2]The First People's Hospital of Yunnan Province, Kunming, Yunnan Province, China
[3]Yunnan Institute of Digestive Disease, The First Affiliated Hospital of Kunming Medical University, Kunming, Yunnan Province, China
[4]Medical Faculty of Kunming University of Science and Technology, Kunming, Yunnan Province, China

*These authors contributed equally to this study.
Corresponding author: K.H. Wang
E-mail: kunhuawangProf@163.com

Genet. Mol. Res. 15 (4): gmr15048870
Received June 9, 2016
Accepted July 12, 2016
Published October 24, 2016
DOI http://dx.doi.org/10.4238/gmr15048870

ABSTRACT. HIV genotyping has led to conflicting results between laboratories. Therefore, identifying the most accurate gene combinations to sequence remains a priority. Datasets of Chinese HIV subtypes based on several markers and deposited in PubMed, Metstr, CNKI, and VIP databases between 2000 and 2015 were studied. In total, 9177 cases of amplification-positive samples from 26 provinces of China were collected and used to classify HIV subtypes based on eight individual

genes or a combination thereof. CRF01_AE, CRF07_BC, CRF08_BC and B were the prevalent HIV subtypes in China, accounting for 84.07% of all genotypes. Gag/Env sequencing classified a greater number of HIV subtypes compared to other genes or combination of gene fragments. The geographical distribution of Gag and Gag/Env genotypes was similar to that observed with all genetic markers. Further principal component analysis showed a significantly different geographical distribution pattern of HIV in China for HIV genotypes detected with Gag/Env, which was in line with the distribution of all HIV genotypes in China. Gag/Env sequences had the highest diversity of the eight markers studied, followed by Gag and Gag/Pol/Env; Pol/Env polymorphisms were the least divergent. Gag/Env can serve as a high-resolution marker for HIV genotyping.

Key words: HIV; Molecular Epidemiology; Biomarker; Gag; Env; Pol

INTRODUCTION

Infection with human immunodeficiency virus (HIV) typically leads to development of acquired immunodeficiency syndrome (AIDS). HIV is an enveloped RNA virus (Zhang et al., 2015) that destroys the immune system by attacking and subsequently depleting $CD4^+$ T lymphocytes. Highly-active anti-retroviral therapy is an effective therapeutic strategy for reducing the viral load in AIDS patients; however, it cannot completely remove the HIV reservoir from these individuals (Bao et al., 2014; Dai et al., 2015). The high mutation rate of HIV is one of the major factors that has hindered development of an HIV vaccine; this problem is compounded by the enhanced diversity of the virus resulting from its high recombination rate (Su et al., 2000). Thus, understanding the molecular epidemiology of HIV/AIDS is increasingly important as it will shed light on the origin and global distribution of AIDS, and will facilitate strategies designed to prevent the spread of HIV.

The HIV genome consists of two identical positive-stranded RNA molecules, which can be reverse-transcribed into double-stranded DNA; each of the RNA strands contains approximately 9200-9800 base pairs. HIV contains two long terminal repeats (LTR) at the end of its genome that include cis-regulatory sequences important for the expression of the provirus. In addition, at least nine protein-coding genes are distributed between the LTRs, including three structural proteins (Gag, Pol, and Env), two regulatory proteins (Tat and Rev), and four accessory proteins (Vif, Vpr, Vpu, and Nef) (Sides et al., 2005; Cunha et al., 2012). Whole-genome sequencing of HIV facilitates the identification of HIV subtypes, and this strategy is considered the gold standard for HIV classification (Robertson et al., 2000). However, the high cost of whole-genome sequencing has limited its usage, and thus sequencing of focused genomic regions, such as the Gag, Pol, Env, and Vpr genes or a combination of different gene fragments, is viewed as a relatively cost-effective alternative (Preston et al., 1988; Mansky and Temin, 1995; Lan et al., 2008; Chen et al., 2011; Chen et al., 2012). Three structural genes (Chen et al., 2011) [Gag (Yao et al., 2012; Ye et al., 2012), Env (Pang et al., 2012; Ye et al., 2012; Li et al., 2014a) and Pol (Holguín et al., 2008; Chen et al., 2014; Yan et al., 2015)] have been widely used for genotyping HIV. Previous studies led to the classification of HIV into two main types, HIV-1 and HIV-2 (Zhang et al., 2015),

with HIV-1 being the predominant branch. The HIV-1 class includes M (major), O (outlier), N (non-M, non-O), and P (Vallari et al., 2011; Li et al., 2014b); among these subtypes, group M accounts for most global HIV cases. In total, however, 11 genetic clusters (A-K) and more than 72 circulating recombinant forms (CRFs) have been identified based on variations in the Env gene (http://www.hiv.lanl.gov.html). Significantly, although most HIV strains have been assigned to specific genotypes based on their Gag, Env, and Pol sequences, conflicting genotypes for the same individual have been reported, most likely because of the high mutation rates of the HIV genome (Robertson et al., 1995; Perelson et al., 1996). For example, one patient was reported to be infected with subtype C based on sequencing of the Pol gene and the p17 fragment of Gag gene; however, the same individual was classified as CRF07_BC with the p24 fragment of the Gag gene, as well as subtype B with the Vpu gene (Robertson et al., 2000; Chen et al., 2013). To avoid such discrepancies, simultaneous analysis of different gene fragments, such as Gag/Env (He et al., 2012; Chen et al., 2013), Gag/Pol (Yao et al., 2012), Pol/Env (Chen et al., 2012; Pang et al., 2012), and Gag/Pol/Env (Lan et al., 2008; Ye et al., 2012; Zeng et al., 2012; Li et al., 2014a; Dai et al., 2015), will facilitate the clear identification of potential new inter-subtype recombinants. However, the field still lacks a set of standard markers for identifying HIV strains, which precludes inter-study comparisons and hampers the comprehensive understanding of the panorama of HIV infection. Thus, evaluating the genes or combination of gene fragments that contain sufficient variation to classify individuals into specific haplotypes will be of great benefit for future research in this field, and will more clearly define the landscape of HIV molecular epidemiology.

To solve this problem, we extracted and analyzed datasets of HIV genotypes found in Chinese AIDS patients since 2000 that had been classified based on sequencing of single genes or combinations of multiple gene fragments. The findings were used to describe the general distribution pattern of HIV genotypes in different Chinese populations. Furthermore, we evaluated the spatial frequency distribution of HIV genotypes based on different markers and estimated the genetic diversity associated with each marker. Our study thus sheds light on the molecular epidemiology of HIV and can be used to guide future studies in this clinically relevant field.

MATERIAL AND METHODS

Materials

The datasets for molecular epidemiology of HIV in China between 2000 and 2015 were extracted from the PubMed, Metstr, CNKI, and VIP databases; the last two databases contained data from Chinese journals and were indexed with “HIV”, “HIV-1”, “AIDS”, “molecular epidemiology”, and “China” as keywords. In total, 339 potentially relevant articles were screened, but only 71 articles were subsequently selected based on our strict criteria. Data from the literature include some uncertain subtypes and unique recombinant forms (Tan et al., 2010; Chen et al., 2011; Zhou et al., 2011). We considered 12,613 samples distributed among 43 cities, which covered 26 provinces of China; among them, 9177 cases were classified into certain haplotypes based on Env, Gag, Pol, Vpr, Gag/Env, Gag/Pol, Pol/Env, and Gag/Pol/Env markers. The original data for these 9177 individuals were retrieved and analyzed in this study (**Tables S1** and **S2**).

Data analysis

The datasets were retrieved based on reports in 71 articles. The frequencies of each genotype with different gene markers were calculated and compared using the Microsoft Office software (Version 2007). Counter maps of the spatial frequencies were constructed to elaborate the geographical distribution patterns of haplotypes for each of the markers using the Kriging algorithm of Surfer 8.0 (Golden Software Inc., Golden, CO, USA) (Cavalli-Sforza et al., 1994). To evaluate the genetic diversity represented by each fragment or combination of different fragments (Table 1), we used the Arlequin 3.11 software and considered all of the datasets belonging to the same molecular marker as one group (Excoffier et al., 2007). Further principal component analysis (PCA) was conducted based on the haplotype frequencies as described previously (Yao et al., 2002).

Table 1. Haplotype diversity of HIV in China identified with differential gene or combination of gene fragments.

Gene fragment	Samples	Genotype number	Gene diversity (means ± SD)
Env	825	16	0.7811 ± 0.0087
Gag	1181	14	**0.8043 ± 0.0055**
Pol	1684	21	0.6855 ± 0.0097
Vpr	300	7	0.6804 ± 0.0171
Gag/Env	2235	22	***0.8169 ± 0.0044***
Gag/Pol	706	11	0.6752 ± 0.0154
Pol/Env	723	17	0.4661 ± 0.0225
Gag/Pol/Env	1523	20	**0.8038 ± 0.0040**

The genetic diversity of each gene or combination of two or three gene fragments was calculated based on haplotypes identified with each marker. The combination of Gag/Env fragments is associated with the highest genetic diversity flagged in bold and italicized numbers, followed by Gag, and then combination of Gag/Pol/Env gene fragments flagged in bold numbers.

RESULTS

General profile of HIV molecular epidemiology in China

We classified 9177 samples into discrete haplotypes based on the subtype information of 8 individual genes, or a combination of two or three gene fragments. This classification included Env, Gag, Pol, Vpr, Gag/Env, Gag/Pol, Pol/Env, and Gag/Pol/Env, which have been widely used to assign HIV haplotypes in China. With the exception of some individuals who could not be confidently assigned, most cases were assigned to 40 discrete haplotypes (**Table S3**). Among them, CRF01_AE (37.97%), CRF07_BC (16.02%), CRF08_BC (15.03%), and B (15.05%) were the most prevalent haplotypes of HIV in China, accounting for more than 84.07% of the total haplotypes (Figure 1).

To evaluate the resolution of the aforementioned 8 genes or combination of two or three gene fragments for HIV genotyping, the genotype for each marker was determined, as shown in **Table S3**. Our results indicate that the eight markers can identify 7 to 22 haplotypes of HIV, and that the combination of Gag/Env has the highest resolution (22 HIV genotypes were identified), followed by the combination of Gag/Pol/Env gene fragments (20 genotypes) and the Pol gene alone (19 genotypes). By contrast, analysis of the Vpr gene alone only led to assignment of 7 haplotypes. Several unique HIV lineages were identified following sequencing of different genes or a combination of two or three gene fragments, as shown in

Table S3. Haplotypes D, E, and 01C appeared in the Env group; haplotypes CRF01_BC, U/CRF01_AE, C/CRF57_BC, B/CRF51_01B, B'/C, and U/G were identified using the Pol gene fragment; and CRF03_AB were identified using the Vpr gene. CRF12_BF, G/CRF12_BF, and A'/CRF10_CD were only detected following combined analysis of Gag/Env gene fragments, whereas the recombination of haplotypes CRF02_AG/CRF01_AE and CRF01_AE/B/B' was identified by simultaneous analysis of Pol/Env, and CRF55_01B and B/F were assigned using a combination of Gag/Pol/Env gene fragments.

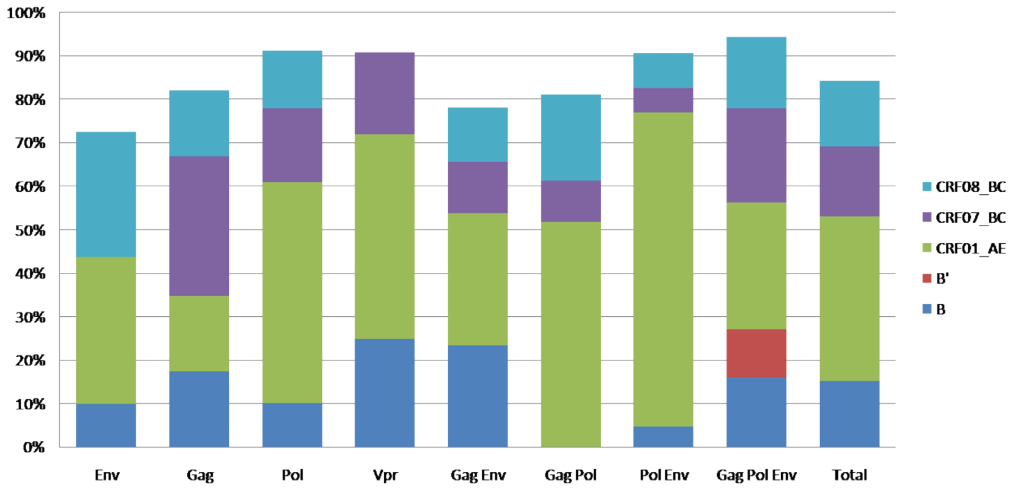

Figure 1. Histogram of the predominant HIV haplotypes in China as identified with different markers.

Population structure based on HIV haplotypes in China with different markers

To determine which marker was optimal in terms of describing the landscape of HIV haplotypes in China, spatial distribution patterns were constructed by considering all of the patients screened with different markers as a group, and by considering the haplotype frequency as the input factor (Figure 2a). The distribution patterns of other markers were also constructed by considering all of the patients detected with the same marker as one group (Figure 2b-d and **Figure S1a-e**), and these patterns were compared with the general distribution pattern of HIV haplotypes in China (Figure 2a in Figure 2). As indicated by Figure 2a, our data showed that certain haplotypes of HIV have higher frequencies in three regions of China, including Yunnan and Guangxi provinces, Shanghai and its surrounding regions, and Beijing and its neighboring regions. A similar pattern was described using Gag/Env (Figure 2b), Gag/Pol/Env (Figure 2c), and Gag (Figure 2d) as markers in comparison to other markers (**Figure S1a-e**).

Spatial distribution pattern of HIV in China

We next examined whether or not the molecular epidemic spectrum of HIV in China as indicated by the Gag/Env marker was consistent with the general spectrum of HIV in China. To this end, samples from 9177 patients, whose viral infections had been classified using

different markers, were considered as a group, and PCA was performed to derive a clustering pattern for HIV-infected groups from different regions of China (Figure 3a). As shown in Figure 3a, a general principal component (PC) map of HIV in China derived from the first two PCs accounted for 68.65% of the total variation. An obvious geographical distribution pattern of HIV haplotypes was observed; the first PC separated groups in eastern China (such as Shanghai, Zhejiang, and Guangdong) from those of other regions, such as populations from northern (Beijing, Shaanxi, Hebei, and Henan) and southern (Yunnan and Guangxi) China. The second PC contributed to the south-to-north cline; groups from Yunnan and Guangxi and those from Beijing, Shaanxi, Henan, Hebei, and Xinjiang were located between the groups from southern and northern China. We also derived a PCA plot by screening groups with other genes or combinations of gene fragments, as shown in Figure 3b, c and d. The PC maps based on the haplotypes identified with Gag/Env (Figure 3b), Gag/Pol/Env (Figure 3c), and Gag (Figure 3d) had distribution patterns similar to those of the general molecular epidemic spectrum of HIV in China, and were different from those identified by analysis of other single genes or combinations, including Env, Pol, Vpr, Gag/Pol, and Pol/Env (**Figure S2a-e**).

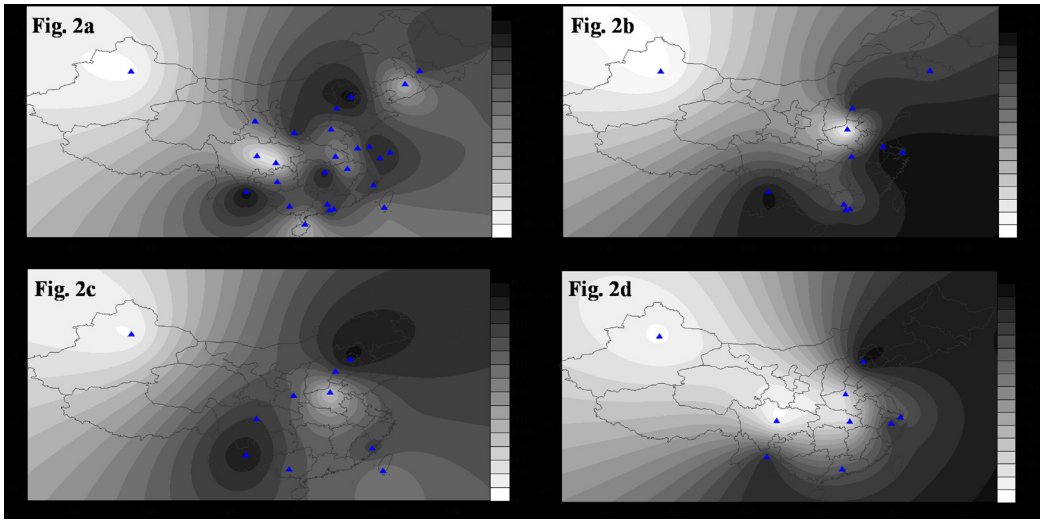

Figure 2. Spatial frequency distributions of HIV haplotypes identified with different gene or combinations thereof. The spatial frequency distributions were created using the Kriging algorithm of the Surfer 8.0 package. The original absolute frequencies are listed in **Table S2**. **a.** Spatial frequency distributions of HIV haplotypes identified based on subtype frequencies of the eight genes described in this study. **b.** Spatial frequency distributions of HIV haplotypes identified with combined analysis of Gag/Env gene fragments. **c.** Spatial frequency distributions of HIV haplotypes identified with combined analysis of Gag/Pol/Env. **d.** Spatial frequency distributions of HIV haplotypes identified following analysis of Gag sequences.

Statistical index of different markers on identifying HIV

As shown in Table 1 and **Table S3**, analysis of the aforementioned eight genes or a combination of two or three gene fragments yielded a diverse number of haplotypes (range 7-22). However, in view of the limited population size, we calculated the genetic diversity of each gene or combination of two or three gene fragments based on the haplotypes identified

with each marker. Our data reveal that the combination of Gag/Env fragments is associated with the highest genetic diversity (0.8169 ± 0.0044), followed by Gag (0.8043 ± 0.0055), and then the combination of Gag/Pol/Env gene fragments (0.8038 ± 0.0040). The combination of Pol/Env gene fragments has the lowest gene diversity (0.4661 ± 0.0225), but screening of the HIV-infected groups based on the Pol/Env combination identified 17 HIV haplotypes, which was higher than the number provided by any other gene permutation (**Table S3**).

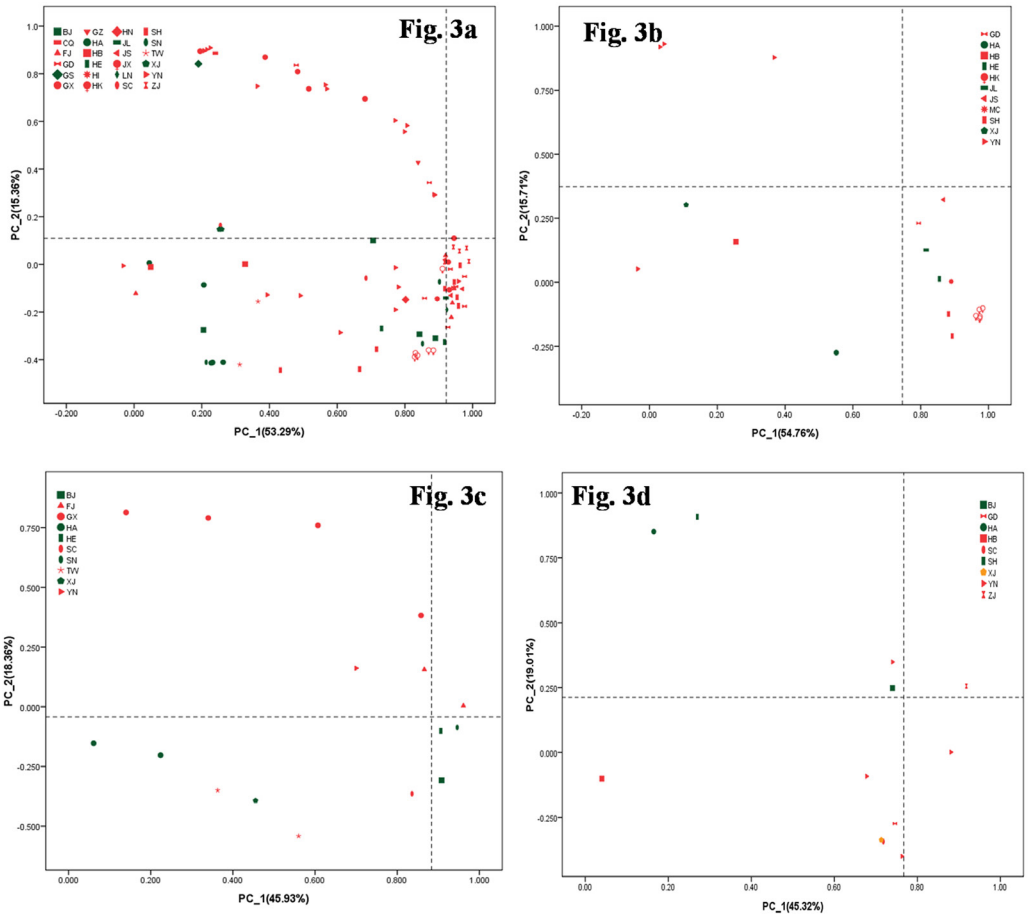

Figure 3. PCA plot of HIV infected populations in China. **a.** PC map of populations based on HIV haplotype frequency of all HIV infected groups in China; **b.** PC map of populations based on HIV haplotype frequency screened with Gag/Env of HIV infected groups in China; **c.** PC map of populations based on HIV haplotype frequency screened with Gag/Pol/Env of HIV infected groups in China; **d.** PC map of populations based on HIV haplotype frequency screened with Gag of HIV infected groups in China.

DISCUSSION

HIV infection is the causative step in the development of AIDS. Thus, genotyping of HIV and further delineating its geographical distribution pattern will help to limit diffusion of

HIV strains, as well as enhance the understanding of the mechanisms underlying resistance to anti-HIV drugs and provide data that can be used to inform vaccine development (Preston et al., 1988; Perelson et al., 1996; Robertson et al., 1995; Robertson et al., 2000). Whole-genome sequencing is the most efficient strategy for genotyping HIV (Robertson et al., 1995; Robertson et al., 2000), but the high cost of this method has limited its widespread adoption. Rather, many groups have opted to sequence single genes or combinations of two or three genes to reduce costs. In this regard, the eight markers Env, Gag, Pol, Vpr, Gag/Env, Gag/Pol, Pol/Env, and Gag/Pol/Env have often been used to subtype HIV in China (Lan et al., 2008; Chen et al., 2012; Pang et al., 2012; Yao et al., 2012; Ye et al., 2012; Zeng et al., 2012; Li et al., 2014a; Dai et al., 2015). However, the high mutation and recombination rates associated with the HIV genome (Robertson et al., 1995; Robertson et al., 2000), have led to ambiguous and sometimes contradictory genotyping results, even in the same individual (Robertson et al., 2000; Qiu et al., 2005; Chen et al., 2013). This has limited the further understanding of the molecular epidemiology of HIV in China, one of the most affected countries in the world.

By extensively dissecting the genotype of 9177 HIV-infected Chinese patients using the eight markers described above, we detected approximately 40 haplotypes of HIV. CRF01_AE (37.97%), CRF07_BC (16.02%), CRF08_BC (15.03%), and B (15.05%) comprised the majority of haplotypes (84.07%). Genotyping based on Gag/Env identified 22 haplotypes, which accounted for 55% of the total haplotypes detected throughout China; this marker was much more robust than any of the other 7 genes or combinations thereof. Furthermore, by considering the dataset detected with differential markers as a whole, we were able to analyze the spatial distribution and regional distribution of HIV strains across China. There was a distinct south-to-north cline pattern, indicating that HIV strains developed independently in China. The intermediate positions in the principle component map for Xinjiang Province imply that admixture of HIV subtypes from northern and southern China occurred at this location. This pattern was supported by the datasets derived from sequencing of Env and Gag/Env, which revealed that Gag/Env accounted for the highest gene diversity among all markers tested. This may be explained by the relative mutation rates of Gag and Env, which encode two of the three core structural proteins of HIV (the core protein and the envelop protein, respectively) (Sides et al., 2005; Cunha et al., 2012). However, these two genes have distinct mutation rates. The former is relatively conserved and has a low mutation rate of 6.0% (Su et al., 2000), which is helpful for identifying basal mutations among the HIV genomes. In contrast, the Env gene is associated with a higher mutation rate than the other genes studied, and indeed has the highest evolutionary rate (30%) (Su et al., 2000). This underlies the diversity of Env sequences, and in part explains why Env is not a strong marker for subtype classification. However, combined analysis of Gag and Env compensates for the drawbacks of each individual gene and significantly improves the subtype marker classification power. Moreover, Pol had a lower mutation rate (3%) than the other genes (Su et al., 2000), suggesting that either Gag or Pol can serve as a key marker for identifying HIV subtypes; data from sequencing of these genes can then be merged with Env data to further increase the robustness of HIV classification. Our PCA of datasets derived from screening of Gag/Env and Gag/Pol/Env revealed a similar geographic distribution pattern of HIV in China relative to the dataset including all the HIV groups. This further confirmed that the combined analysis of Gag/Env provides the highest-resolution marker with respect to HIV genotyping. In addition, the separation of southern, northern, and eastern HIV groups in China, as well as the south-to-north cline of HIV groups, implies that the different strains of HIV were initially

introduced independently of one another and subsequently dispersed throughout China. These data will be valuable in developing strategies to prevent the spread of HIV.

In this study, the polymorphism analysis of Gag/Env of HIV in China, which showed relatively higher genetic diversity, suggested that this fragment may serve as an effective biomarker for genotyping of HIV in this region. However, we only reanalyzed previously published data to gain more detailed information regarding the epidemiology of HIV in China. More sequences of focused genomic regions and of combination fragments are needed to be analyzed and compared with the whole-genome sequencing to develop a database system for genotyping analysis of HIV (Araújo et al., 2006).

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Foundation for Science and Technology Planning Project of Yunnan Provincial Bureau of Health (#2012WS63), the Key Joint Funds of the Natural Science Foundation of Yunnan Province and Kunming Medical University (#2014FB021), the Yunnan Institute of Digestive Disease Institute (#2014NS121), the National Science Foundation of China (#81360069), the Academician Workstation of Yunnan Province, the Foundation for Innovative Group of the Gastrointestinal Surgery of Yunnan Province (#2012HC013), and the Foundation of Medical Leading Talent of Yunnan Province (#L-201205).

REFERENCES

Araújo LV, Soares MA, Oliveira SM, Chequer P, et al. (2006). DBCollHIV: a database system for collaborative HIV analysis in Brazil. *Genet. Mol. Res.* 5: 203-215.

Bao Y, Tian D, Zheng YY, Xi HL, et al. (2014). Characteristics of HIV-1 natural drug resistance-associated mutations in former paid blood donors in Henan Province, China. *PLoS One* 9: e89291. http://dx.doi.org/10.1371/journal.pone.0089291

Cavalli-Sforza LL, Menozzi P and Piazza A (1994). The history and geography of human genes. Princeton, Princeton university press. Princeton.

Chen M, Yang L, Ma Y, Su Y, et al. (2013). Emerging variability in HIV-1 genetics among recently infected individuals in Yunnan, China. *PLoS One* 8: e60101. http://dx.doi.org/10.1371/journal.pone.0060101

Chen S, Cai W, He J, Vidal N, et al. (2012). Molecular epidemiology of human immunodeficiency virus type 1 in Guangdong province of southern China. *PLoS One* 7: e48747. http://dx.doi.org/10.1371/journal.pone.0048747

Chen X, Zheng Y, Li H, Mamadou D, et al. (2011). The Vpr gene polymorphism of human immunodeficiency virus type 1 in China and its clinical significance. *Curr. HIV Res.* 9: 295-299. http://dx.doi.org/10.2174/157016211797635937

Chen Y, Chen S, Kang J, Fang H, et al. (2014). Evolving molecular epidemiological profile of human immunodeficiency virus 1 in the southwest border of China. *PLoS One* 9: e107578. http://dx.doi.org/10.1371/journal.pone.0107578

Cunha LK, Kashima S, Amarante MF, Haddad R, et al. (2012). Distribution of human immunodeficiency virus type 1 subtypes in the State of Amazonas, Brazil, and subtype C identification. *Braz. J. Med. Biol. Res.* 45: 104-112. http://dx.doi.org/10.1590/S0100-879X2012007500003

Dai D, Shang H, Han XX, Zhao B, et al. (2015). The biological characteristics of predominant strains of HIV-1 genotype: modeling of HIV-1 infection among men who have sex with men. *J. Med. Virol.* 87: 557-568. http://dx.doi.org/10.1002/jmv.24116

Excoffier L, Laval G and Schneider S (2007). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1: 47-50.

He X, Xing H, Ruan Y, Hong K, et al.; Group for HIV Molecular Epidemiologic Survey (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. http://dx.doi.org/10.1371/journal.pone.0047289

Holguín A, López M and Soriano V (2008). Reliability of rapid subtyping tools compared to that of phylogenetic analysis for characterization of human immunodeficiency virus type 1 non-B subtypes and recombinant forms. *J. Clin. Microbiol.* 46: 3896-3899. http://dx.doi.org/10.1128/JCM.00515-08

Lan YC, Elbeik T, Dileanis J, Ng V, et al. (2008). Molecular epidemiology of HIV-1 subtypes and drug resistant strains in Taiwan. *J. Med. Virol.* 80: 183-191. http://dx.doi.org/10.1002/jmv.21065

Li L, Sun B, Zeng H, Sun Z, et al. (2014a). Relatively high prevalence of drug resistance among antiretroviral-naive patients from Henan, Central China. *AIDS Res. Hum. Retroviruses* 30: 160-164. http://dx.doi.org/10.1089/aid.2013.0144

Li X, Feng Y, Yang Y, Chen Y, et al. (2014b). Near full-length genome sequence of a novel HIV-1 recombinant form (CRF01_AE/B) detected among men who have sex with men in Jilin Province, China. *AIDS Res. Hum. Retroviruses* 30: 701-705. http://dx.doi.org/10.1089/aid.2014.0008

Mansky LM and Temin HM (1995). Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* 69: 5087-5094.

Pang W, Zhang C, Duo L, Zhou YH, et al. (2012). Extensive and complex HIV-1 recombination between B′, C and CRF01_AE among IDUs in south-east Asia. *AIDS* 26: 1121-1129. http://dx.doi.org/10.1097/QAD.0b013e3283522c97

Perelson AS, Neumann AU, Markowitz M, Leonard JM, et al. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271: 1582-1586. http://dx.doi.org/10.1126/science.271.5255.1582

Preston BD, Poiesz BJ and Loeb LA (1988). Fidelity of HIV-1 reverse transcriptase. *Science* 242: 1168-1171. http://dx.doi.org/10.1126/science.2460924

Qiu Z, Xing H, Wei M, Duan Y, et al. (2005). Characterization of five nearly full-length genomes of early HIV type 1 strains in Ruili city: implications for the genesis of CRF07_BC and CRF08_BC circulating in China. *AIDS Res. Hum. Retroviruses* 21: 1051-1056. http://dx.doi.org/10.1089/aid.2005.21.1051

Robertson DL, Sharp PM, McCutchan FE and Hahn BH (1995). Recombination in HIV-1. *Nature* 374: 124-126. http://dx.doi.org/10.1038/374124b0

Robertson DL, Anderson JP, Bradac JA, Carr JK, et al. (2000). HIV-1 nomenclature proposal. *Science* 288: 55-56. http://dx.doi.org/10.1126/science.288.5463.55d

Sides TL, Akinsete O, Henry K, Wotton JT, et al. (2005). HIV-1 subtype diversity in Minnesota. *J. Infect. Dis.* 192: 37-45. http://dx.doi.org/10.1086/430322

Su L, Graf M, Zhang Y, von Briesen H, 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. *J. Virol.* 74: 11367-11376. http://dx.doi.org/10.1128/JVI.74.23.11367-11376.2000

Tan Y, Chan D, Chan D, Ip PK, et al. (2010). High genetic diversity of HIV-1 viruses in Macao, China. *J. Infect.* 61: 164-172. http://dx.doi.org/10.1016/j.jinf.2010.04.012

Vallari A, Holzmayer V, Harris B, Yamaguchi J, et al. (2011). Confirmation of putative HIV-1 group P in Cameroon. *J. Virol.* 85: 1403-1407. http://dx.doi.org/10.1128/JVI.02005-10

Yan H, Ding Y, Wong FY, Ning Z, et al. (2015). Epidemiological and molecular characteristics of HIV infection among money boys and general men who have sex with men in Shanghai, China. *Infect. Genet. Evol.* 31: 135-141. http://dx.doi.org/10.1016/j.meegid.2015.01.022

Yao X, Wang H, Yan P, Lu Y, et al. (2012). Rising epidemic of HIV-1 infections among general populations in Fujian, China. *J. Acquir. Immune Defic. Syndr.* 60: 328-335. http://dx.doi.org/10.1097/QAI.0b013e31824f19f5

Yao YG, Kong QP, Bandelt HJ, Kivisild T, et al. (2002). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* 70: 635-651. http://dx.doi.org/10.1086/338999

Ye JR, Yu SQ, Lu HY, Wang WS, et al. (2012). Genetic diversity of HIV type 1 isolated from newly diagnosed subjects (2006-2007) in Beijing, China. *AIDS Res. Hum. Retroviruses* 28: 119-123. http://dx.doi.org/10.1089/aid.2011.0012

Zeng H, Sun Z, Liang S, Li L, et al. (2012). Emergence of a new HIV type 1 CRF01_AE variant in Guangxi, Southern China. *AIDS Res. Hum. Retroviruses* 28: 1352-1356. http://dx.doi.org/10.1089/aid.2011.0364

Zhang L, Wang YJ, Wang BX, Yan JW, et al. (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. http://dx.doi.org/10.1177/0956462414543841

Zhou F, Luo AD, Jiang JX, Tao TX, et al. (2011). AIDS molecular epidemiological survey in Laibin. *J. Trop. Med.* 11: 420-423.

Supplementary material

Table S1. Populations analyzed in this work.

Table S2. General information and original HIV haplotype frequency of literature cited in this study.

Table S3. The HIV haplotype frequency for groups assigned with eight gene markers in this study.

Figure S1. The spatial frequency distributions of HIV haplotypes identified with Env, Pol, Vpr, Gag/Pol, Pol/Env, respectively. The spatial frequency distributions were created using the Kriging algorithm of the Surfer 8.0 package. The original absolute frequencies are listed in **Table S2**. **a.** The spatial frequency distributions of HIV haplotypes identified based on subtypes frequency of Env; **b.** The spatial frequency distributions of HIV haplotypes identified with Pol; **c.** The spatial frequency distributions of HIV haplotypes identified with Vpr; **d.** The spatial frequency distributions of HIV haplotypes identified with combined analysis of Gag/Pol; **e.** The spatial frequency distributions of HIV haplotypes identified with Pol/Env.

Figure S2. PCA plot of HIV infected populations in China based on haplotypes identified with Env, Pol, Vpr, Gag/Pol, Pol/Env, respectively. **a.** PC map of populations based on HIV haplotype frequency identified with Env in China; **b.** PC map of populations based on HIV haplotype frequency identified with Pol in China; **c.** PC map of populations based on HIV haplotype frequency screened with Vpr in China; **d.** PC map of populations based on HIV haplotype frequency screened with Gag/Pol in China; **e.** PC map of populations based on HIV haplotype frequency screened with Pol/Env in China.