**Research Article**                                                                 **Open Access**

# Application of Ambiguous Analysis for Determining Recent Infections of HIV-1 in a China Men Who Had Sex with Men Population

**Xiaoyan Bai[1#], Hongwei Zhang[2#], Wei Shao[3], Lan Li[2], Zheng Zeng[1*] and Hao Wu[2*]**

[1]Laboratory of Virology, Department of Infectious Diseases, Peking University First Hospital, Beijing 100034, P.R. China
[2]Beijing Youan Hospital, Capital Medical University, Beijing 100069, China
[3]HIV Drug Resistance Program, Center for Cancer Research, National Cancer Institute, National Institutes of Health, USA
[#]Both authors contributed equally to this work

## Abstract

**Objective:** To distinguish HIV-1 recent infection (≤ 300 days) from long-term HIV-1 infection (>300 days) in China men who had sex with men (MSM) population. We analyzed the change over time in the proportion of ambiguous nucleotides in patient plasma sequences. We hypothesized that this method could be used to determine recent infections.

**Methods:** HIV-1 sequences and clinical data of MSM were collected from June 2007 to September 2010. All the sequences were obtained by single genome amplification and sequencing. HIV-ambi-count was used to calculate the proportion of ambiguous nucleotides, and the optimal cut-off values to identify recent infection were established using receiver operator characteristic analysis (ROC).

**Results:** A total of 188 sequences from 150 patients were collected (38 patients were sampled twice at different time points), consisting of 118 sequences of subtype CRF01_AE and 70 sequences of subtype B. The optimal cut-off values for determination of early HIV-1 infection in subtypes CRF01_AE and B were 0.31% and 0.347%, respectively. The sensitivity, specificity, positive and negative predictive values were 76.5%, 55%, 89.3% and 32.4% for subtype CRF01-AE (P=0.005), 80.6%, 62.5%, 94.3% and 29.4% for subtype B (P=0.017), respectively. The area under the ROC curve (AUCROC) was 0.652 (95% CI, 0.520-0.783, P=0.03) and 0.75 (95% CI, 0.622-0.878, P=0.022) for subtype CRF01-AE and B, respectively.

**Conclusions:** The proportion of ambiguous nucleotides may be a good marker to distinguish recent infection from long-term HIV-1 infection with high positive predictive value in the China MSM population.

**Keywords:** HIV-1; Recent infection; Long-term infection; Ambiguous nucleotide; MSM

## Introduction

HIV-1 infection is a global health problem, with millions of people newly infected every year. However, it is difficult to determine the difference between a recent infection and a chronic infection. HIV-1 incidence is classically estimated by prospective cohort studies [1], which are expensive and time consuming. What's more, once participants were recruited into cohort studies, interventions would be taken to avoid high-risk behaviors, thus leading to an under-estimated incidence [2]. Based on enzyme immunoassay (EIA), Jassen et al. [3] developed an enzyme immunoassay (EIA) to distinguish recent (4-12 months) from long-term (>12 months) HIV-1 infections [4,5] by modifying sample dilution and incubation time. Subsequently, several laboratory methods have been developed to estimate time of infection, such as BED-capture enzyme immunoassay (BED-CEIA) [6], antibody avidity assay [7], and detection of anti-p24 IgG3 antibodies [8]. Currently, BED-CEIA is widely used to detect recent infection, which is based on the rationales that HIV-1 specific IgG level rises gradually in the early stage of infection, and finally maintains a plateau for many years. Comparing to prospective cohort studies, these methods are more convenient and useful. The sensitivities of 13 serologic assays vary from 42% to 100% with a median of 89%, and the specificities range from 49.5% to 100% with a median of 86.8% [9]. However, there are still many limitations of these methods, including inconsistency between different laboratories results and misclassification problems [10]. Thus, more studies are needed to develop a new method with high sensitivity and reproducibility.

HIV-1 infection is initiated in most cases by a single virus [11], and then viral diversity builds up gradually under the pressure of viral mutation and immune selection [12]. Based on their genetic make-up, human immunodeficiency virus type 1 (HIV-1) is divided into four groups: M, N, O and P. Group M could be further divided into several subtypes, including subtypes A-D, F-H, J and K. The pandemic and high evolutionary rate of HIV-1 as well as the high mutation rate and recombination of viruses, has resulted in emergence of the circulating recombination form CRF01_AE [13,14]. Evolution and recombination of virus could also accelerate disease progression of AIDS, and reduce the effect of drugs [13]. The most prevalent subtype in Europe and North America is subtype B (Los Alamos HIV database, http://www.hiv.1anl.gov). Early studies showed that subtype B was also the dominant subtype in Chinese MSM population, but CRF01_AE

strains were found to surpass subtype B strains lately [15]. In recent years, several methods based on viral sequences have been developed to identify recent HIV-1 infection (≤1 year) [16-18]. Importantly, Kouyos et al. [16] showed that the proportion of ambiguous nucleotides correlated with the time elapsed between HIV infection and sampling for genotyping. The threshold of ≥0.5% ambiguous nucleotides performed well in determining recent infections of HIV subtype B in a Switzerland population with high sensitivity (86.8%), high specificity (70%) and high negative predictive value (98.7%). However, because the subtypes of Chinese patients are mainly CRF01_AE and B, it has yet to be determined if this method of measuring ambiguous nucleotides within a patient between sampling times could be used to determine recent infections.. Here, our study aims to evaluate the performance of this method in Chinese patients.

## Methods

### Patients

All patients in this study were HIV-positive men who had sex with men (MSM). They were recruited from the local HIV/AIDS sentinel surveillance sites in Beijing Youan Hospital, Capital Medical University. Informed consent was given by participating patients during the period of 2007-2010. Eligible participants were adults with one or more sequences and complete clinical data such as age, time of infection, medical history, HIV viral load and $CD_4^+$ cell count. Sequences with uncertain infection days, incomplete clinical data, and from patients with coinfections of multiple HIV subtypes were removed from the analyses. All patients were not treated with anti-HIV medications.

To identify HIV infection, samples of patients were tested by enzyme linked immunosorbent assay (ELISA) (BIO-RAD xmar, HIV1/2 antibody ELISA kit, Shanghai, China) in Beijing Youan hospital. The seropositive samples were sent to Beijing Centers for Disease Control and prevention (CDC) and further confirmation tests were done there [19].

Time of infection was estimated as follows: (1) If the result of ELISA was negative but HIV RNA (VL-EasyQ NucliSens EasyQ Analyzer, NucliSens EasyQ HIV-1 v2.0, Biomerieux) was detected, then the infection time was determined as 14 days before sampling; (2) If the result of ELISA was positive but the result of western blot (WB) was suspicious, then the infection time was determined as 30 days before sampling; (3) If the results of ELISA and WB were all positive, then the infection time was determined as the midpoint of the interval between last negative and first positive HIV tests results [20,21]. Infection within 300 days was defined as recent infection, and if the time interval between HIV infection and sampling was longer than 300 days, the patient was determined to have a long-term infection.

All the sequences were obtained by single genome amplification and sequencing. The sequenced HIV-1 pol region covered 1293 base pairs from the transframe protein gene (p6) to the reverse transcriptase gene (RT) [18].

### Analysis of sequences

JpHMM (jumping profile Hidden Markov Model) and BLAST search from Los Alamos HIV database (http://www.hiv.lanl.gov) were used to determine HIV subtypes.

Sequences were aligned using ClustalW in the MEGA 5.1 software package [22], and they were checked and edited using the BioEdit software package.

The proportion of ambiguous nucleotides within each sequence was calculated by HIV-ambi-count software. This analysis needs no reference sequences.

### Statistical analysis

SPSS 13.0 software was used to analyze the data. Continuous data was expressed as mean ± SD and analyzed using Student's t-test. Categorical data was analyzed by the chi-square test or Fisher's exact test. All P values were 2-sided, and 0.05 was set as the statistical significance value.

Optimal cut-off values of the proportion of ambiguous sites were established to distinguish recent from long-term infections. The classification performance was evaluated with receiver operating characteristic (ROC) analyses. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for each subtype as follows: sensitivity = true positive / (true positive + false negative), specificity = true negative / (true negative + false positive), positive predictive value = true positive / (true positive + false positive), negative predictive value = true negative / (true negative + false negative).

## Results

### Participant characteristics

One hundred and eighty-eight sequences from 150 enrolled patients were collected (38 patients were sampled twice at different time), including 160 sequences of recent infection and 28 sequences of long-term infection (Table 1). The mean age of all participants was 31.24 ± 8.63 years old. The mean age of recent infected and long-term infected were 31.01 ± 8.52 and 32.57 ± 9.28 years old, respectively. No significant statistical difference was observed between ages of these two groups (P=0.38).

### Sequence characteristics

Subtype analysis showed that 118 sequences were CRF01_AE (62.8%), and the remaining 70 sequences were B (37.2%). As shown in Figure 1, the proportion of ambiguous nucleotides spread from 0 to 4.27%, and differences between recent and long-term infection were observed in all subtypes. For subtype CRF01_AE, the median value of ambiguous nucleotides fraction for recent and long-term infections were 0.155% and 0.318%, respectively, P=0.031. For subtype B, the median value of ambiguous nucleotides fraction for recent and long-time infections were 0.080% and 0.425% respectively, P=0.02.

### Optimal cut-off value of ambiguous nucleotides proportion

To establish an optimal cut-off value of the proportion of

| | Recent infection (n=160) * | Long-term infection (n=28) * | P |
|---|---|---|---|
| Age (years) | 31.01 ± 8.52† | 32.57 ± 9.28 | 0.38 |
| Time since infection (days) | 140.28 ± 92.94 | 383.54 ± 112.25 | 0 |
| Viral load (log10 copies/ml) | 4.64 ± 0.92 | 4.34 ± 1.14 | 0.141 |
| Viral load test undone | 18 | 0 | |
| CD4 count (per mm3) | 445.36 ± 193.04 | 488.18 ± 175.82 | 0.27 |
| Subtypes | | | |
| CRF01_AE | 98 | 20 | |
| B | 62 | 8 | |

\*: Number of sequences. †mean ± SD

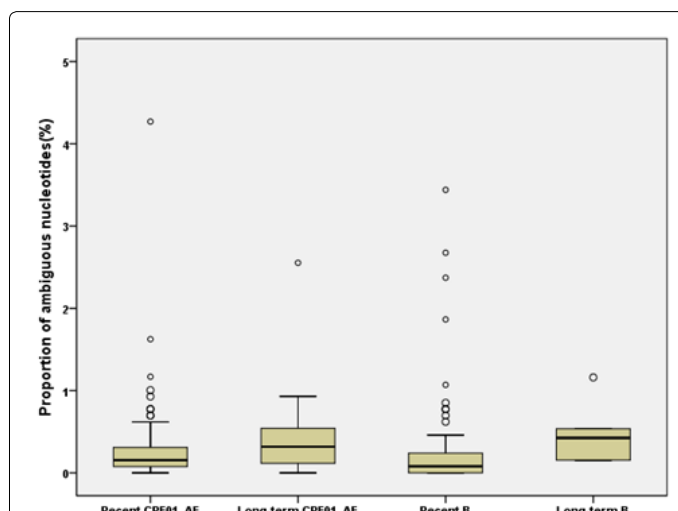**Table 1:** Characteristics of participants.

ambiguous sites determining recent infection, we evaluated series of cut-off values. From Table 2, we could see that a cut-off value of 0.31% performs well for subtype CRF01_AE in distinguishing between recent and long-term infections. That is, if the fraction of ambiguous sites from a sequence was not larger than 0.31%, then the sequence was from a recent-infected patient, otherwise the sequence was from a long-term infected patient. The sensitivity, specificity, PPV and NPV for this cut-off was 76.5%, 55%, 89.3% and 32.4%, respectively (P=0.005). ROC analysis showed area under the ROC curve (AUCROC) was 0.652 (95%CI, 0.520-0.783, P=0.03) (Figure 2 A).

When 0.347% was set as the cut-off value for subtype B, 50 of 62 recent infection sequences were classified correctly (sensitivity, 80.6%). Meanwhile, the specificity was 62.5%. As Table 2 shows, the PPV was 94.3%, and the NPV was 29.4% (P=0.017). The AUCROC was 0.75 (95%CI, 0.622-0.878, P=0.022) (Figure 2 B).
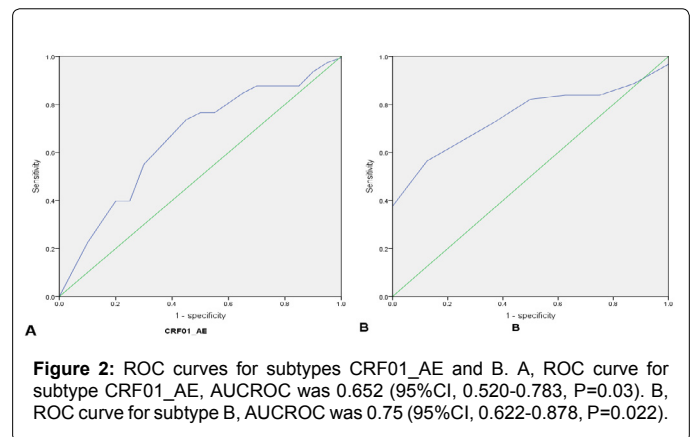
## Discussion

In our study, we found the fraction of ambiguous nucleotides within individual patient sequences was a good indicator to distinguish between recent and long-term infections in a Chinese MSM population. In total, 188 sequences were collected, most of which were subtype CRF01_AE (62.8%) (Table 1), which is consistent with other studies also reporting that CRF01_AE strains have surpassed subtype B strains in Beijing MSM [15]. The proportion of ambiguous nucleotides ranged from 0 to 4.27%, and differences between recent and long-term infections were observed in all subtypes (Figure 1).

We further analyzed the proportion of ambiguous nucleotides of all sequences separately according to their subtypes, and found that the fraction of ambiguous nucleotides varied between recent and long-term infections with the fraction increasing with long term infections. Then we attempted to find the optimal cut-off value of the fraction of ambiguous nucleotides to determine the different between recent infections and long term infections. As Table 2 and Figure



**Figure 2:** ROC curves for subtypes CRF01_AE and B. A, ROC curve for subtype CRF01_AE, AUCROC was 0.652 (95%CI, 0.520-0.783, P=0.03). B, ROC curve for subtype B, AUCROC was 0.75 (95%CI, 0.622-0.878, P=0.022).

| Subtype | Cutoff, % | Sensitivity, % | Specificity, % | PPV, % | NPV, % | P value |
|---|---|---|---|---|---|---|
| CRF01_AE | | | | | | |
| | 0.193 | 55.1 | 70 | 90 | 24.1 | |
| | 0.234 | 74.5 | 55 | 89 | 30.6 | |
| | 0.310 | 76.5 | 55 | 89.3 | 32.4 | 0.005 |
| | 0.395 | 84.7 | 35 | 86.5 | 31.8 | |
| | 0.465 | 87.8 | 30 | 86 | 33.3 | |
| | 0.658 | 90.8 | 10 | 83.2 | 18.2 | |
| B | | | | | | |
| | 0.154 | 56.5 | 87.5 | 97.2 | 20.6 | |
| | 0.235 | 72.6 | 62.5 | 93.8 | 22.7 | |
| | 0.347 | 80.6 | 62.5 | 94.3 | 29.4 | 0.017 |
| | 0.500 | 83.9 | 37.5 | 91.2 | 23.1 | |
| | 0.578 | 83.9 | 12.5 | 88.1 | 9.1 | |
| | 0.735 | 87.1 | 12.5 | 88.5 | 11.1 | |

**Table 2:** Cut-off values for different subtypes.

2 show, the best cut-off value for subtype CRF01_AE was 0.31%. The sensitivity was 76.5% and the specificity was 55%. The positive predictive value was 89.3%. The AUCROC was 0.652. When 0.347% was set as the cut-off value for subtype B, the highest total classification accuracy (78.6%) was achieved. Similar to CRF01_AE, the sensitivity and specificity of subtype B were 80.6% and 62.5%, respectively. The PPV was 97.2%. The AUCROC was 0.75. These results were slightly lower than those described by some of the previous works [16,18]. This might be due to the shorter infection time of long-term infected patients (383.54 ± 112.25) as we only used data sets of precisely timed longitudinal samples. Other mature laboratory methods such as BED-CEIA [6] were reported to be quite reliable with high sensitivity and specificity. However, lots of experimental instruments were needed, and the results may differ in different laboratories [9]. In contrast, sequencing based assays could identify recent infection by analyzing sequences collected from genotyping or drug resistance tests. Extra experimental instruments are not needed here, and the method is less time consuming.

However, there are still limitations of our study. Firstly, since immune selection plays an important role in the process of viral diversity, differences in immune pressure between patients such as HLA types should be taken into consideration. Secondly, all patients in this study were infected from a single founder strain. However, previous study showed that around 36% of MSM were found to be infected by more than one type of virus [23], and recent infections originating from multiple founder strains might be misclassified as long-term infections [24]. Thirdly, antiretroviral therapy also strongly



**Figure 1:** Comparison of proportions of ambiguous nucleotides between recent and long-term infections. The ends of the whiskers mark the 5th and 95th percentile. Outliers are marked by dots. For subtype CRF01_AE, the median value of ambiguous nucleotides fraction for recent and long-term infections were 0.155% and 0.318%, respectively, P=0.031. For subtype B, the median value of ambiguous nucleotides fraction for recent and long-time infections were 0.080% and 0.425% respectively, P=0.02.

influences viral diversity [23], however, the classification value of sequencing based assays on exposure or treatment experienced people is limited.

In conclusion, the proportion of ambiguous nucleotides is a useful marker to distinguish recent from long-term HIV-1 infections, but this new genotypic tool can't be used alone and must be interpreted with clinical and serological data. This method has performed well in the Chinese population. Considering the limited sample size in our study, more investigations are needed to confirm our findings.

### Acknowledgement

### Conflicts of Interest and Source of Funding

### References

1. Vanichseni S, Kitayaporn D, Mastro TD, Mock PA, Raktham S, et al. (2001) Continued high HIV-1 incidence in a vaccine trial preparatory cohort of injection drug users in Bangkok, Thailand. AIDS 15: 397-405.

2. McDougal JS, Parekh BS, Peterson ML, Branson BM, Dobbs T, et al. (2006) Comparison of HIV-1 incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the BED capture enzyme immunoassay. AIDS Res Hum Retroviruses 22: 945-952.

3. Janssen RS, Satten GA, Stramer SL, Rawal BD, O'Brien TR, et al. (1998) New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. JAMA 280: 42-48.

4. UNAIDS/WHO Working Group on Global HIV/AIDS/STI Surveillance (2011) When and how to use assays for recent infection to estimate HIV incidence at a population level.

5. Kim AA, Hallett T, Stover J, Gouws E, Musinguzi J, et al. (2011) Estimating HIV incidence among adults in Kenya and Uganda: a systematic comparison of multiple methods. PLoS One 6: e17535.

6. Parekh BS, Kennedy MS, Dobbs T, Pau CP, Byers R, et al. (2002) Quantitative detection of increasing HIV type 1 antibodies after seroconversion: a simple assay for detecting recent HIV infection and estimating incidence. AIDS Res Hum Retroviruses 18: 295-307.

7. Suligoi B, Galli C, Massi M, Di Sora F, Sciandra M, et al. (2002) Precision and accuracy of a procedure for detecting recent human immunodeficiency virus infections by calculating the antibody avidity index by an automated immunoassay-based method. J Clin Microbiol 40: 4015-4020.

8. Wilson KM, Johnson EI, Croom HA, Richards KM, Doughty L, et al. (2004) Incidence immunoassay for distinguishing recent from established HIV-1 infection in therapy-naive populations. AIDS 18: 2253-2259.

9. Guy R, Gold J, Calleja JM, Kim AA, Parekh B, et al. (2009) Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: a systematic review. Lancet Infect Dis 9: 747-759.

10. Murphy G, Parry JV (2008) Assays for the detection of recent infections with human immunodeficiency virus type 1. Euro Surveill 13.

11. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA 105: 7552-7557.

12. Troyer RM, Collins KR, Abraha A, Fraundorf E, Moore DM, et al. (2005) Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression. J Virol 79: 9006-9018.

13. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, et al. (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58: 19-42.

14. Thomson MM, Nájera R (2005) Molecular epidemiology of HIV-1 variants in the global AIDS pandemic: an update. AIDS Rev 7: 210-224.

15. Wang W, Xu J, Jiang S, Yang K, Meng Z, et al. (2011) The dynamic face of HIV-1 subtypes among men who have sex with men in Beijing, China. Curr HIV Res 9: 136-139.

16. Kouyos RD, von Wyl V, Yerly S, Böni J, Rieder P, et al. (2011) Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis 52: 532-539.

17. Ragonnet-Cronin M, Aris-Brosou S, Joanisse I, Merks H, Vallée D, et al. (2012) Genetic diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. J Infect Dis 206: 756-764.

18. Andersson E, Shao W, Bontell I, Cham F, Cuong do D, et al. (2013) Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. Infect Genet Evol 18: 125-131.

19. National Guideline for Detection of HIV/AIDS (2009) Chinese Center For Disease Control And Prevention.

20. Goujard C, Bonarek M, Meyer L, Bonnet F, Chaix ML, et al. (2006) CD4 cell count and HIV DNA level are independent predictors of disease progression after primary HIV type 1 infection in untreated patients. Clin Infect Dis 42: 709-715.

21. Seng R, Goujard C, Desquilbet L, Sinet M, Rouzioux C, et al. (2008) Rapid CD4+ cell decrease after transient cART initiated during primary HIV infection (ANRS PRIMO and SEROCO cohorts). J Acquir Immune Defic Syndr 49: 251-258.

22. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Bio Evol 28: 2731-2739.

23. Park SY, Love TM, Nelson J, Thurston SW, Perelson AS, et al. (2011) Designing a genome-based HIV incidence assay with high sensitivity and specificity. AIDS 25: F13-19.

24. Joos B, Fischer M, Kuster H, Pillai SK, Wong JK, et al. (2008) HIV rebounds from latently infected cells, rather than from continuing low-level replication. Proc Natl Acad Sci USA 105: 16725-16730.