SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

April 2012 Vol.55 No.4: 328–335
doi: 10.1007/s11427-012-4312-0

# A new pattern-based method for identifying recent HIV-1 infections from the viral *env* sequence

YANG Jing[1], XIA XiaYu[1], HE Xiang[2], YANG SenLin[2], RUAN YuHua[2], ZHAO QuanBi[2], WANG ZhiXin[1], SHAO YiMing[2*] & PAN XianMing[1*]

[1]*Key Laboratory of Bioinformatics, Ministry of Education, School of Life Sciences, Tsinghua University, Beijing 100084, China;*
[2]*State Key Laboratory for Infectious Disease Prevention and Control, National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China*

Received January 13, 2012; accepted February 16, 2012

The long asymptomatic stage of HIV infection poses a great challenge in identifying recent HIV infections. This is a bottleneck for monitoring HIV epidemic trends and evaluating the effectiveness of national AIDS control programs. Several serological methods were used to address this issue with some success. Because of high false-positive rates in patients with advanced infection or in ART treatment, UNAIDS still hesitates to recommend their use in routine surveillance. We developed a new pattern-based method for measuring intra-patient viral genetic diversity for determination of recent infections and estimation of population incidence. This method is verified by using several datasets (424 subtype B and 77 CRF07_BC samples) with clearly identified HIV-1 infection times. Pattern-based diversities of recent infections are significantly lower than that of chronic ones. With larger window periods varying from 200 to 350 days, a higher accuracy (90%–95%) not affected by advanced disease nor ART treatment could be obtained. The pattern-based genetic method is supplementary to the existing serology-based assays, both of which could be suitable for use in low and high epidemic regions, respectively.

**determination of HIV recent infections, estimation of population incidence, viral genetic diversity, pattern-based distance**

**Citation:** Yang J, Xia X Y, He X, *et al*. A new pattern-based method for identifying recent HIV-1 infections from the viral *env* sequence. Sci China Life Sci, 2012, 55: 328−335, doi: 10.1007/s11427-012-4312-0

Since the acquired immune deficiency syndrome (AIDS) caused by the human immunodeficiency virus type 1 (HIV-1) was identified 30 years ago, the pandemic has resulted in extremely heavy human costs of over 60 million infections, including nearly 30 million deaths worldwide. Understanding the current state and future trends of HIV-1 infection is essential to halt the AIDS epidemic [1]. However, the very long asymptomatic stage of HIV infection poses a great challenge in identifying recent HIV infections, and is a bottleneck for both monitoring HIV epidemic trends and evaluating the effectiveness of national AIDS control programs. Therefore, HIV-1 incidence has never been correctly estimated in a timely fashion, although various strategies have been applied to address this problem. For instance, the Joint United Nations Program on HIV/AIDS (UNAIDS) decreased their previous estimates of the worldwide annual HIV-1 incidence by 42%, whereas the CDC of the United States increased theirs by 40% [2]. Without such critical information, the public health agencies are unable to determine the rising HIV/AIDS case reports at the end of the year due to improvement of the reporting system, or more new infections.

An ideal way to monitor epidemic trends is to determine

*Corresponding author (email: pan-xm@mail.tsinghua.edu.cn; yshao08@gmail.com)

infection duration of individual carriers/patients from their blood samples, which can be collected with ease. In particular, a huge pool of existing samples can be used to trace historical epidemic trends. In the past decade, many scientists worked on various serological methods to identify recent HIV-1 infections, named Serological Testing Algorithms for Recent HIV Seroconversion (STARHS) [3–6]. The most widely used STARHS method is the one developed by the US CDC: the BED HIV-1 capture enzyme immunoassay. The BED is based on measuring the ratio of anti-HIV IgG to the total IgG in the blood, with a measurable window of 153 days (the increasing linear curve during acute HIV infection). The BED assay had been used in the field in many developed and developing countries, especially after release of the US national HIV incidence estimate [1]. However, its performance was not satisfactory, with unacceptably high false-positive rates due to a drop in specificity in both the late stage of HIV infection (AIDS) and antiviral treatment. The UNAIDS HIV incidence study task force recommended the BED should not be used in routine surveillance, or in absolute incidence estimates [7–9]. Thus, there is an urgent need to improve the current assays and provide complimentary methodology to overcome this obstacle.

It has been known for some time that viral genetic diversity in newly infected persons is much lower than in chronic infections [10–15], but no systematic effort had been conducted to distinguish recent from chronic HIV infection by measuring HIV genetic diversity. First of all, the HIV viral gene actively undergoes variations in many ways, including point mutations (annual mutation rate of about 1% for the *env* gene), insertions and deletions, which result in a small mean value difference between recent and chronic infection, and large deviation within each category. Second, about one quarter to one third of the subjects are infected by more than one founder virus [16], which can cause confusion about the molecular clock of HIV infection duration. All these aspects make it difficult to use the HIV genetic diversity as a reliable estimator. Here, we developed a novel methodology by carefully combining suitable tools of HIV-1 genetic diversity measurement and statistical method to overcome the previously mentioned challenges. The new algorithm was validated using a panel of datasets with clearly-defined HIV infection time available in the international database (subtype B dataset) and by our own cohort study (CRF07_BC dataset). The new methodology is shown to be able to detect recent HIV infection with high accuracy (above 90%) in patients with both early and advanced stages of infections. Therefore, this new methodology has the potential to be used as a complementary method to serology-based assays, such as BED, and provide the combined tools to public health agencies for identifying recent HIV infections and monitoring AIDS epidemic trends.

# 1 Materials and methods

## 1.1 Data collection

### *1.1.1 Subtype B dataset*

A total of 424 samples of 257 subtype B subjects were selected from the CHAVI, MACS, Seattle primary infection cohort, MGH cohort, Women's HIV Interdisciplinary Network, and other projects. Selection criteria were: (i) Patients were from a longitudinal cohort or diagnosed with HIV infections more than one year ago. (ii) Viral loads were more than 50 copies $mL^{-1}$ (range: 300–$10^6$). (iii) More than ten sequences in the *env* C2–V5 region (length about 180 amino acids) are available in the sequence database and quality control measures were used in the experimental procedure for deriving the sequence from plasma samples to ensure that the sequences of a sample were representative of the HIV-1 population *in vivo*. All the C2–V5 *env* sequences of these samples were collected from the Los Alamos HIV database. The window period to classify recent versus chronic infection was defined as 200 days after seroconversion. The database editors suggested that the labeled number of days were mostly estimates. Therefore, we went back to the original papers to confirm the study-specific timing definitions. When sequences were labeled as Fiebig Stages I–V (infected less than 100 days), samples were called "recent infections". Fiebig Stage VI was excluded. Of all samples, 160 were regarded as recent infections and the remaining 264 as chronic infections, details are provided in Appendix Tables 1–3 in the electronic version.

### *1.1.2 Subtype CRF07_BC dataset*

A total of 77 plasma samples were used that were previously collected from 27 drug naïve HIV-1 infected individuals in the HPTN033 program [18]. The window period to classify recent versus chronic infection was defined as 365 days after seroconversion. Of all samples, 21 were regarded as recent infections, and the remaining 56 as chronic infections. Sequences for the calculation of the intra-patient viral genetic diversity were derived from about 280 collected plasma samples by the SGA method [16,19]. Details are provided in Appendix Table 4. The nucleotide sequences from this study are available in GenBank under accession number HQ668526-HQ668930, HQ668941-HQ668974, HQ668984-HQ669461, and HQ669468-HQ670224.

RNA templates were extracted from 280 μL plasma using the QIAamp Viral RNA Mini kit (Qiagen, Hilden, Germany), and eluted with 60 μL elution buffer. The cDNAs were then synthesized by an oligo (dT) primer using the SuperScript III first-strand synthesis system kit (Invitrogen; Carlsbad, California, USA). In brief, 25 μL reaction mixture with 20 μL RNA template, 2.5 μmol $L^{-1}$ primer oligo (dT) and 0.5 mmol $L^{-1}$ dNTPs were denatured at 65°C for 5 min. Then the 25 μL cDNA synthesis mixture containing 1× re-

verse transcription buffer, 5 mmol $L^{-1}$ $MgCl_2$, 10 mmol $L^{-1}$ DTT, 2 units $\mu L^{-1}$ RNaseOUT and 10 units $\mu L^{-1}$ SuperScript III reverse transcriptase, were added into the reaction mixture and incubated at 50°C for 50 min. The reaction was terminated at 85°C for 5 min and RNase H was added to remove the RNA template at 37°C for 20 min. The resulting cDNA was used immediately or stored at −20°C for further use.

The single genome amplification (SGA) of the *env* V1−V5 fragment was performed as previously described [16,19]. Briefly, the serial diluted cDNA templates (1:5 to 1:45) distributed among wells of replicate 96-well plates were amplified to identify the optimal dilution with no more than 30% of positive wells.

The first round PCR amplification was carried out in 25 μL mixture containing 1.5 μL cDNA dilutions, 0.2 μmol $L^{-1}$ primer F1X (5′-GATGGATGAGGATGTAATCAGTTTATGGGA-3′, HXB2: 6533-6562) and R1X (5′-ATTGACGCYGCGCCCATAGTGCT-3′, HXB2: 7828-7806), and 12.5 μL Premix Ex Taq (Takara Bio Inc., Tokyo, Japan). The PCR thermo-cycling profiles are 94°C for 3 min followed by 35 cycles of 94°C for 30 s, 62°C with a decrement of 0.3°C per cycle for 30 s, 72°C for 1 min 30 s; with a final extension of 72°C for 10 min. The second round PCR was performed in 25 μL mixture containing 1.5 μL first round products, 0.2 μmol $L^{-1}$ primer F2X (5′-AGTTTATGGGATCAAAGCCTAAAGCCATGT-3′, HXB2: 6552-6581) and R2X (5′-GCTCTTTTTTCTCTCTCCACCACTCTCCT-3′, HXB2: 7759-7731). The PCR conditions were as follows: 94°C for 3 min followed by 30 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 1 min 30 s; with a final extension of 72°C for 10 min. The PCR products were inspected using the 1% Pre-cast E-Gel 96 (Invitrogen).

To avoid inter-specimen contamination for the massive operation in the amplification procedures, and to reduce the labor intensity in serial dilution and single genome amplification, the JANUS Automated Workstation (PerkinElmer, Waltham, Massachusetts, USA) was used to setup the PCR reaction, to perform the serial dilution, and to distribute the reagent and first round templates for determining both optimal dilution for SGA and following SGA experiments. The second round PCR templates were added into the reaction mixture using multichannel pipettes manually.

The PCR products were purified using the QIAquick PCR Purification kit (Qiagen, Hilden, Germany) and directly sequenced in both directions using an ABI 3100 Genetic Analyzer (BigDye V3.1; Applied Biosystems, Foster City, California, USA). The sequences for each amplicon were assembled and cleaned with a Sequencher V4.9 (Gene Codes, USA). Any sequence with evidence of mixed bases was excluded from further analysis. Finally, the 10 to 39 sequences were obtained from each sample.

The BED assay was performed following the manufacturer's instructions (Calypte, Portland, Oregon, USA). The specimens with a normalized optical density less than 0.8 were classified as recent infections.

### 1.2 Definition of diversity

Here we define a new pair-wise distance to measure the genetic diversity. For a given sequence of length $L$, a segment of length $k$ is designated as a $k$-mer (a test run shows that $k$=14 is the suitable word length for the dataset used in this work). The number of such $k$-mers (denoted as $N$) can be determined by sequentially searching the sequence step by step, such that

$$N=L-k+1. \tag{1}$$

Given two sequences $A$ and $B$, $k$-mers derived from sequence $A$ are referred to as "$A$ $k$-mers". The genetic distance from sequence $A$ to sequence $B$ is determined by the number of "$A$ $k$-mers" appearing in sequence $B$. Thus, the normalized pair-wise distance from sequence $A$ to sequence $B$ (termed as $D_{A\to B}$) is defined as

$$D_{A\to B}=1-\frac{N_B^{(A)}}{(L_B-k+1)}, \tag{2}$$

where $N_B^{(A)}$ is the number of "$A$ $k$-mers" appearing in sequence $B$.

Since it is likely that some $k$-mers will be repeated within a given sequence, the value of $D_{A\to B}$ is possibly different from that of $D_{A\to B}$, so that the diversity is defined as the average value across all pair-wise distances:

$$D=\frac{\sum_{x=1}^{m}\sum_{\substack{y=1\\(y\neq x)}}^{m}D_{x\to y}}{m(m-1)}, \tag{3}$$

where $m$ is the number of viral sequences per sample at a given time.

### 1.3 Sequence length requirements

While using full-length gp120 sequences will clearly give the most reliable and accurate estimates of infection duration, it would also increase the cost of the method. To reduce the cost of the method as much as possible, we analyzed the effect of sequence length on accuracy.

Since the diversity has normalized values ranging from 0 to 1, independent of sequence length, changes in sequence length affect the variance. Longer gp120 sequences result in more consistent numbers of mutations occurring in each sequence, resulting in a smaller variance. According to the binomial distribution, the probability ($p$) of at least one mutation occurring per year in a sequence is as follows:

$$p=1.0-(1.0-a)^n, \tag{4}$$

where $a$ ($10^{-2}$ mutations/site/year) is the replication error frequency and $n$ is the sequence length. According to eq. (4), about 180 amino acids (C2–V5) is a suitable length for high accuracy.

### 1.4 Sequences per sample requirement

According to eq. (3) the change of intra-patient diversity is a function of the number of sequences with occurring mutation. If more than one half of the sampled sequences (denoted as $m$) were with at least one mutation occurring after the window period, this sample could be identified as chronic infection. The probability (Pr) of more than one half of the sampled sequences occurring with at least one mutation could be calculated by the binomial distribution:

$$\Pr(i \geqslant m/2) = \sum_{i \geqslant m/2}^{m} C_m^i p^i (1-p)^{m-i}, \quad (5)$$

where $i \geqslant m/2,\ldots, m$ represents the number of sequences occurring with at least one mutation, $C_m^i$ denotes the number of combinations of $m$ things taken $i$, $p$=0.85 is the probability of at least one mutation occurring (calculated by eq. (4)). According to eq. (5), more than 10 sequences per sample are necessary to accurately identify recent infection.

### 1.5 Accuracy measurements

We defined "recent infections" as true positives and "chronic infections" as true negatives. The accuracy ($Ac$) is defined as follows:

$$Ac = \frac{TP + TN}{TP + FP + TN + FN}, \quad (6)$$

where $TP$ is the number of positive cases that were correctly identified, $TN$ is the number of negative cases that were correctly rejected, $FP$ is the number of over-identified cases, and $FN$ is the number of under-identified cases.

In addition, other general parameters that are used in most other studies were applied. These included sensitivity ($Sn$), specificity ($Sp$), predicted positive value ($PPV$), and predicted negative value ($NPV$), which were defined as follows:

$$Sn = \frac{TP}{TP + FN},\ Sp = \frac{TN}{TN + FP},$$

$$PPV = \frac{TP}{TP + FP},\ NPV = \frac{TN}{FN + TN}. \quad (7)$$

### 1.6 Estimating the number of the recent infections in a dataset

Assume that $N_0$ and $N_1$ samples are identified as recent infections and chronic infections in a dataset, respectively. Since the number of false negatives and false positives could often not cancel each other out, the number of recent infections ( $N_0'$ ) in the dataset should be adjusted as

$$\begin{aligned} N_0' &= N_0 \times PPV \ + N_1 \times (1 - NPV\ ) \pm \Delta, \\ \Delta &= \left( t_{(N_0-1),0.95} \times \sqrt{N_0 \times PPV \times \left(1 - PPV\ \right)} + \right. \\ &\quad \left. t_{(N_1-1),0.95} \times \sqrt{N_1 \times NPV\ \times \left(1 - NPV\ \right)} \right). \end{aligned} \quad (8)$$

The uncertain interval (95% CI) in eq. (8), $\Delta$, is the function of parameters, $PPV$, $NPV$, the rate of recent infection in a dataset and the sample size of the dataset ($N_0$+$N_1$). The sample size required for high accuracy with accepted relative uncertain interval could be estimated by eq. (8).

## 2 Results

### 2.1 Identifying recent infection in the subtype B dataset

Samples in the subtype B dataset were derived from a different cohort study. The detailed information of subjects from each cohort are summarized in Table 1. The pattern-based diversity value ($D$) of each sample in the subtype B dataset was calculated from the available C2–V5 sequences of the *env* gene and was plotted in Figure 1A and B and summarized in Appendix Tables 1–3 in the electronic version. With a window period of 200 d, the mean values (standard deviation) were $D_0$=0.08 ($S_0$=0.10) for the recent infections, $D_1$=0.48 ($S_1$=0.16) for chronic infections, showing a statistically significant difference ($P$<0.001). Moreover, the average diversities of chronic infections from 55 samples of antiretroviral-treated patients and 89 samples of untreated patients were 0.48 and 0.52, respectively, showing a statistically insignificant difference ($P$=0.13). The influence of CD4 counts (range: 2–7195) on diversity were also statistically insignificant. The average diversities of chronic infections from 25 samples of <200 CD4 counts, 33 samples of 200–500 CD4 counts and 40 samples of >500 CD4 counts were 0.47, 0.47 and 0.58, respectively ($P$=0.43).

It has been reported that about one quarter to one third of the subjects are infected by more than one founder [12]. In the subtype B dataset, 32 recent infections are documented to be infected by multiple founders [16], diversity values of which are larger than that infected by one founder. Simply using the diversity values as an estimator, about 7 of 32 were identified as chronic infections. To differentiate the recent infections of multiple founders from chronic ones, we analyzed the $CV$ values of diversity (standard deviation of pair-wise distance/diversity), showing that the $CV$ values of the chronic infections approximately follow a normal distribution, while that of the recent infections do not. The $CV$ value could not be used to identify recent versus chronic infections because of the high false negative rate, but could

**Table 1** Information of samples in different cohorts[a)]

| Subtype | Cohort | Risk factor | | | | Average sequence/ sample | Infection duration | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | | IDU | MSM | Hetero | Other | | <200 d | 201–365 d | >1 years | |
| B | MACS | 0 | 78 | 0 | 1 | 16.7 | 5 | 4 | 70 | 79 |
| | CHAVI | 1 | 36 | 34 | 77 | 26.2 | 118 | 0 | 30 | 148 |
| | WHIN | 2 | 0 | 21 | 1 | 15.2 | 0 | 0 | 24 | 24 |
| | SPIC | 0 | 15 | 0 | 0 | 21.2 | 2 | 0 | 13 | 15 |
| | MGH | 0 | 10 | 0 | 0 | 23.5 | 0 | 0 | 10 | 10 |
| | Others | 35 | 54 | 7 | 52 | 19.2 | 35 | 5 | 108 | 148 |
| B'C | HPTN033 | 73 | 4 | 0 | 0 | 21.7 | 12 | 9 | 56 | 77 |
| | Total | 111 | 197 | 62 | 131 | 21.2 | 172 | 18 | 311 | 501 |

a) MACS, Multicenter AIDS cohort study; CHAVI, the Center for HIV/AIDS Vaccine Immunology cohort; MAHS, Malawi Award for Human Settlements cohort; WHIN, Women's HIV Interdisciplinary Network cohort; SPIC, Seattle primary infection cohort; MGH, Massachusetts General Hospital cohort.

**Table 2** The results of identifying recent infections in the subtype B dataset[a)]

| Cohorts | Total samples | Cohort defined | | Alignment-based method | | | Pattern-based method | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Recent (≤200 d) | Established (>200 d) | *Sn* (%) | *Sp* (%) | *Ac* (%) | *Sn* (%) | *Sp* (%) | *Ac* (%) |
| MACS | 79 | 5 | 74 | 100.0 | 81.1 | 87.8 | 60.0 | 98.6 | 96.2 |
| CHAVI | 148 | 118 | 30 | 94.0 | 80.0 | 91.2 | 98.3 | 93.3 | 97.3 |
| WHIN, SPIC, MGH | 49 | 2 | 47 | 100.0 | 85.1 | 85.7 | 100.0 | 91.5 | 91.8 |
| Others | 148 | 35 | 113 | 92.8 | 74.7 | 79.0 | 91.4 | 92.0 | 91.9 |
| Total | 424 | 160 | 264 | 94.4 | 78.8 | 84.7 | 95.6 | 93.9 | 94.6 |

a) Window period is 200 days.

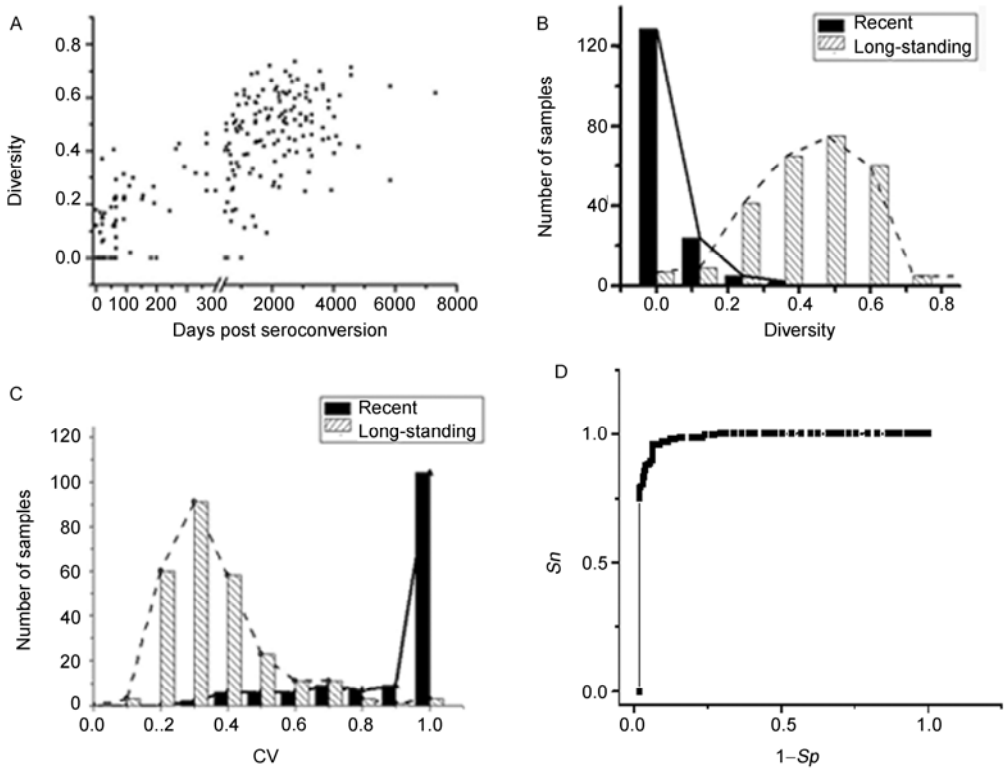


**Figure 1** Diversity versus infection duration for 424 subtype B samples. A, The diversity versus infection duration for 316 samples labeled with infection days and Fiebig Stages I–V in the dataset of 424 samples. The sample points included in circle are false positive. B, The distributions of diversity values of total 424 samples included those marked with "late" (more than two years after seroconversion). C, The distribution of *CV* values of 424 samples. D, ROC curve of distinguishing recent and chronic infections.

be used to correctly exclude chronic infections (Figure 1C). Thus, *CV*=0.95 (mean value plus 3 folds of the standard deviation of the chronic infection) was used as cutoff to exclude chronic infections as the first step before using the average diversity for further identification. Any sample with a *CV* value larger than 0.95, without regard to its diversity value, was identified as a recent infection.

To show the success of our method in identifying recent from chronic infections, we carried out the receiver operating characteristics (ROC) analysis over the whole subtype B database; the area under the ROC curve was 0.97 (Figure 1D). The diversity value of 0.24 was selected as the best cutoff to distinguish recent from chronic infections. We correctly identified 153 in 160 (95.6%) recent infection, including 30 of 32 samples from multiple infection founders, and 248 in 264 (93.9%) chronic infections, overall, the accuracy was 94.6% for the individual carriers/patients (Table 2).

The change of window period barely affected the performance of our method. We have used different window periods to classify recent versus chronic infections. When the window periods were varied from 150 to 350 d, their accuracy rates ranged from 90% to 95% (Figure 2).

For comparison, the alignment-based diversity of each sample was calculated using the CLUSTALW [20] and PHYLIP [21]. The mean values (standard deviation) were $D_0$=0.01 ($S_0$=0.02) for the recent infections, $D_1$=0.08 ($S_1$=0.04) for chronic infections. When the value of 0.045 was used as cutoff, the accuracy was about 84.6% (Table 2), showing that our method has a big advantage over the alignment-based method.

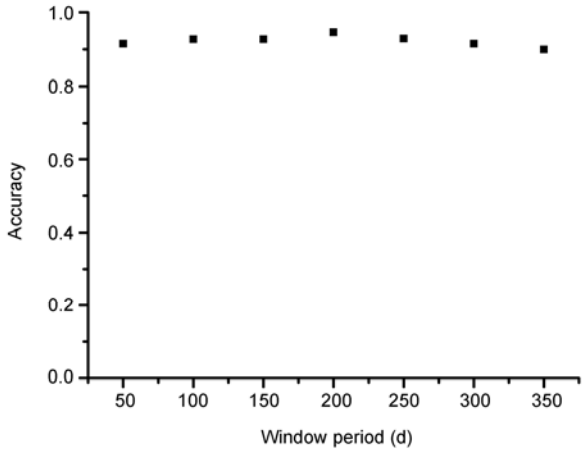


**Figure 2** Window period versus accuracy of 424 sample dataset. As the window period ranges from 150 to 350 days, accuracy of our method varies from 90% to 95%.

The identification performance over the sub-dataset from different cohorts shows that the accuracy (97.3%) of the Center for HIV/AIDS Vaccine Immunology (CHAVI) project was clearly higher than that of the others (91.9%). The higher accuracy derived from the CHAVI Project could be due to its higher stringency for the definition of seroconversion with shorter sampling intervals. In all, the accuracies of our method in different cohorts were above 91%. Therefore, our method is less influenced by different populations.

## 2.2 Identifying recent infection in the CRF07_BC dataset

Applying our method to the CRF07_BC database, with a window period of one year, we identified 17 in 21 (80.9%) as recent infections and 52 in 56 (92.8%) as chronic infections. Overall, the accuracy was 89.6% (Table 3), almost equal to that in the subtype B dataset (with one year as the window period).

For comparison, we have also applied the BED assay to run over the CRF07_BC dataset. The BED method identified 15 in 21 (71.4%) as recent infections and 45 in 56 (80.3%) as chronic infections. Overall, the accuracy was 77.9% (Table 3).

## 2.3 Determination of recent infections and the minimal size of the dataset for estimating HIV incidence

HIV incidence is estimated by using statistical approaches with adjustment from the recent infection rate of a sample collection extrapolating to the population level [1]. Using eq. (8) requires all the accuracy parameters of the assay method measurable, and the higher the accuracy of the assay method, the smaller the uncertain interval of the adjusted recent infection. The number of recent infections in the subtype B dataset was estimated by eq. (4) to be 160±13 (95% CI), suggesting that our method can be used to estimate the number of recent infections from a sample collection with high accuracy (100%) and accepted uncertain interval (8%). By incorporating incidence, estimated in the US population

**Table 3** The results of identifying recent infections in the subtype CRF07_BC dataset[a)]

| Subtype | Source | Interval (months) | Cohort defined | | *Sn* (%) | *Sp* (%) | *Ac* (%) | Method |
|---|---|---|---|---|---|---|---|---|
| | | | Recent | Chronic | | | | |
| B | – | – | 169 | 255 | 98.8 | 84. 5 | 89.9 | This work |
| CRF07_BC | China CDC (HPTN033) | 6 | 21 | 56 | 81.0 | 92.9 | 89.6 | This work |
| | | | 21 | 56 | 71.4 | 80.3 | 77.9 | BED* |

a) Window period is 365 days. *, The parameters measuring accuracy of BED method are calculated with widow period of 365 d. If calculated with a window period of 170 d, the parameters measuring accuracy of BED method are sensitivity (*Sn*)=89%, specificity (*Sp*)=73%, and accuracy (*Ac*)=75%.

in 2006 (about 0.3) [1] and *PPV*=0.905, *NPV*=0.973, respectively, into eq. (8), the sample size required for estimating the number of recent infections with 10% relative uncertain interval (95% CI) was about 300, and with 5% about 600.

## 3 Discussion

Prior to the XVIII International AIDS Conference in 2010, *Science* magazine published an editorial entitled "AIDS Response at a Crossroad". This article pointed out that given the situation that there is no cure for AIDS and no vaccine for HIV prevention, lack of accurate ways to distinguish recent from chronic HIV infection at the individual or population levels is one of the three major obstacles in understanding the trends of the HIV/AIDS epidemic and mobilizing a successful global HIV/AIDS control campaign [22]. Timely identification of such strategic information is key for global and national AIDS prevention programs to allocate resources and coordinate successful control measures to stop the epidemic [23].

Several serological methods were used to address the issues with some success. In serological assays, infected people have different windows in producing HIV-specific antibodies, ranging from two to twelve months [4]. In other words, some infected people exhibit high titer of antibody as early as within two months, and hence can be misjudged as chronic infections. Conversely, some chronic carriers can have lowered or even "disappeared" antibody levels in their blood, and can be mistakenly judged as recent infections. The broader problem for the serological assays is lack of not only reliable knowledge about the relationship between assay specificity and infection duration [24], but also information on the window periods of immunological response for different HIV subtypes and population groups [4].

It was found in the recent systematic reviews in the field by the WHO Working Group on HIV Incidence-Assays and academia [1,2] that even though the serology-based assays have reasonable sensitivity, they are vulnerable to misclassifying established HIV infection as recent ones. Because of high false-positive rates in patients with advanced infection or in ART treatment, UNAIDS still hesitates to recommend their use in routine surveillance.

Compared with the serological assays, our new method has an advantage with high-accuracy; distinguishing the recent infections from the chronic ones and a low false-positive rate. 98.5% of 136 samples from patients infected more than 8 years (average incubation period) were correctly identified, whereas only 89.1% of 128 samples from patients infected 200 d–8 years were correctly identified. It is shown that all false positives occur within four year after seroconversion, see Figure 1A. Furthermore, our method can provide a highly accurate identification of recent and long-term infections with a window period of up to one year instead of the ~160–200 d window cutoff used in the BED method, which has the advantage of choosing a suitable window period to ensure that new infections are not missed. In addition, it is shown that our method is less affected by viral loads, host response, and treatment.

A widely used method, referred to as alignment-based viral sequences analysis, has been used for estimating the origin and history of the HIV-1 epidemic [20–22]. However, this method is particularly problematic for analyzing genes such as the *env* gene, which actively undergoes variations, including insertions and deletions. Pair-wise distance defined by the alignment-based method depends on the reliability of sequence alignments. However, since the gap penalty is chosen by experience, alignment results are variable. By contrast, our pattern-based distance gives nearly identical values to different evolutionary events, such as point mutations, insertions, and deletions [23]. In preparation for our manuscript, Park *et al*. published an HIV incidence assay using the alignment-based *env* gene diversity [25]. The accuracy of their method was as high as that of our method, but they used a small and biased dataset containing 182 recent and 43 chronic cases. Using a larger dataset (424 cases) with the recent to chronic ratio close to the real situation (about 0.3 [1] ), our results show that the accuracy of the assay using alignment-based diversity is much lower than that of the pattern-based one described in this study. The coefficients of variation (standard deviation/mean) of the alignment-based and pattern-based approaches are 2.44 and 1.37 for recent infections, and 0.54 and 0.41 for long-standing infection, respectively, showing that the measurement constructed in the pattern-based approach is more sensitive and reliable than the alignment-based approach for identification of recent infection.

This method uses the viral *env* sequence to identify recent infection. In actual application, it is important to ensure that the sequences of a sample are representative of HIV-1 populations *in vivo*. This experimental procedure is labor intensive and difficult in application in large scale. Fortunately, owing to the advent of automation technology, generating sequences from blood samples has become much faster, easier and cheaper than ever before. In the course of adopting the automation technology, we have modified it into an automated liquid handling workstation, and were able to overcome the rate-limiting step in attaining accurate viral sequences. Furthermore, samples collected in resource-limited areas could be analyzed in a well-established laboratory through international collaboration.

Recently, systematic reviews of the fields of serology-based HIV measuring assays were conducted by the WHO Working Group on HIV Incidence-Assays and academia. Both reviews concluded that there is an urgent need to improve the current serology-based HIV incidence assays, and the solution should come from combined approaches of complimentary methodology using vinous biomarkers [1,2]. There is much diversity in the level of global HIV/AIDS

epidemics; the adult HIV prevalence ranging from a high of up to 30% in some African countries to as low as 0.01% in Europe and middle-eastern countries. The annually reported HIV/AIDS cases are also varied widely in national surveillance of different countries, from just a few hundreds to hundreds of thousands. Therefore, various countries in the high and low HIV epidemic regions can selectively use serology-based or pattern-based genetic diversity methods according to the purpose of their surveillance needs. In fact, the China CDC has begun to test the combined use of both BED and the pattern-based genetic diversity method in the field to accurately evaluate the recent HIV infection and monitor AIDS epidemic trends as well, in both high and low epidemic regions in China.

*The authors are grateful to Prof. Li Peng from the School of Life Science, Tsinghua University for his helpful suggestions in the manuscript. We also thank Ma ZheQin, Ma PengFei, and Zhang Bin from the National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention for partly participating in the SGA amplification, viral load detection and BED assay. This work was supported in part by the National Natural Science Foundation of China (Grant No. 30870475), Ministry of Science and Technology of China (Grant No. 2009CB918801), Ministry of Health of China (Grant No. 2008ZX10001-003), and the International Development Research Center, Ottawa, Canada (Grant No. 104519-010).*

1 Hall H I, Song R, Rhodes P, *et al*. Estimation of HIV incidence in the United States. JAMA, 2008, 300: 520–529
2 Brookmeyer R. Measuring the HIV/AIDS epidemic: approaches and challenges. Epidemiol Rev, 2010, 32: 26–37
3 Dobbs T, Kennedy S, Pau CP, *et al*. Performance characteristics of the immunoglobulin G-capture BED-enzyme immunoassay, an assay to detect recent human immunodeficiency virus type 1 seroconversion. J Clin Microbiol, 2004, 42: 2623–2628
4 Guy R, Gold J, Calleja J M, *et al*. Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: a systematic review. Lancet Infect Dis, 2009, 9: 747–759
5 Hargrove J W, Humphrey J H, Mutasa K, *et al*. Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. AIDS, 2008, 22: 511–518
6 Loschen S, Batzing-Feigenbaum J, Poggensee G, *et al*. Comparison of the human immunodeficiency virus (HIV) type 1-specific immunoglobulin G capture enzyme-linked immunosorbent assay and the avidity index method for identification of recent HIV infections. J Clin Microbiol, 2008, 46: 341–345
7 Karita E, Price M, Hunter E, *et al*. Investigating the utility of the HIV-1 BED capture enzyme immunoassay using cross-sectional and longitudinal seroconverter specimens from Africa. AIDS, 2007, 21: 403–408
8 Hayashida T, Gatanaga H, Tanuma J, *et al*. Effects of low HIV type 1 load and antiretroviral treatment on IgG-capture BED-enzyme immunoassay. AIDS Res Hum Retroviruses, 2008, 24: 495–498
9 UNAIDS Reference Group. UNAIDS Reference Group on estimates, modelling and projections--statement on the use of the BED assay for the estimation of HIV-1 incidence for surveillance or epidemic monitoring. Wkly Epidemiol Rec, 2006, 81: 40
10 Herbeck J T, Nickle D C, Learn G H, *et al*. Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host. J Virol, 2006, 80: 1637–1644
11 Shankarappa R, Margolick J B, Gange S J, *et al*. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol, 1999, 73: 10489–10502
12 Lee H Y, Giorgi E E, Keele B F, *et al*: Modeling sequence evolution in acute HIV-1 infection. J Theor Biol, 2009, 261: 341–360
13 Korber B, Muldoon M, Theiler J, *et al*. Timing the ancestor of the HIV-1 pandemic strains. Science, 2000, 288: 1789–1796
14 Worobey M, Gemmel M, Teuwen D E, *et al*. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature, 2008, 455: 661–664
15 Yusim K, Peeters M, Pybus O G, *et al*. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philos Trans R Soc Lond B Biol Sci, 2001, 356: 855–866
16 Keele B F, Giorgi E E, Salazar-Gonzalez J F, *et al*. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA, 2008, 105: 7552–7557
17 Haubold B, Pfaffelhuber P, Domazet-Loso M, *et al*. Estimating mutation distances from unaligned genomes. J Comput Biol, 2009, 16: 1487–1500
18 Zhang Y, Shan H, Trizzino J, *et al*. HIV incidence, retention rate, and baseline predictors of HIV incidence and retention in a prospective cohort study of injection drug users in Xinjiang, China. Int J Infect Dis, 2007, 11: 318–323
19 Palmer S, Kearney M, Maldarelli F, *et al*. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J Clin Microbiol, 2005, 43: 406–413
20 Larkin M A, Blackshields G, Brown N P, *et al*. Clustal W and Clustal X version 2.0. Bioinformatics, 2007, 23: 2947–2948
21 Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 1985, 39: 783–791
22 Justman J, El-Sadr W M. AIDS response at a crossroads. Science, 2010, 329: 120
23 WHO, UNAIDS, UNICEF. Towards universal access: scaling up priority HIV/AIDS interventions in the health sector. 2009.
24 Hallett T B, Ghys P, Barnighausen T, *et al*. Errors in 'BED'-derived estimates of HIV incidence will vary by place, time and age. PLoS ONE, 2009, 4: e5720
25 Park S Y, Love T M, Nelson J, *et al*. Designing a genome-based HIV incidence assay with high sensitivity and specificity. AIDS, 2011, 25: 1–7