The Key Laboratory of Bioinformatics, Ministry of Education, School of Life Sciences, Tsinghua University, Beijing, China, Xia-Yu Xia1

Using a Sequence Clustering Based Diversity Measure

High-Accuracy Identification of Incident HIV-1 Infections

It is known that the intra-patient HIV-1 viral diversity gradually builds up through time: diversity increases in an approximately linear mode for the first several years after infection, then at decreasing rates until a plateau is reached, and often declines in the late stage of infection [13–15]. Recently, two studies have found success in designing accurate incidence assays using such viral genetic diversity measures. Park et al. used the 10% quantile (Q10) of the pairwise Hamming distance (HD) and achieved greater than 96.0% accuracy when tested on a dataset containing 225 samples [16]. As well, Yang et al. developed a pattern-based method and achieved 94.6% accuracy using a dataset of 424 samples [17].

The main difficulties of using viral diversity to determine HIV-1 incidence are in correctly identifying incident infections with multiple transmitted/founder (T/F) strains and chronic infections with declined viral genetic diversity [16–18], as well as in reducing the disproportional impact of insertion/deletions (indels) and recombination on calculated diversity value. At the early phase of HIV-1 infection, distinct T/F strains would evolve into highly homologous sequence clusters with relatively low intra-cluster and high inter-cluster diversity [19,20]. Since the infection duration is only correlated with intra-cluster diversity, overall comparisons of the mean diversity are likely to misclassify early infections with
multiple T/F strains. Indeed, while the HD Q_{20} and pattern-based methods were able to conquer this obstacle well for incident infections with two T/F strains [16,17], their performances were both much reduced for cases involving three or more T/F viruses. Moreover, the env gene region of HIV-1 experiences high rates of indels and recombination, and diversity measures based on simple sequence alignment and pattern matching may be disproportionately impacted [21]. Efforts at improving sequence-based incidence assays must therefore seek to address these shortfalls.

In the present work, we developed a novel intra-patient diversity-based method to determine HIV-1 incident infections. Through utilizing the fact that viral sequences derived from distinct T/F strains tend to form separate homologous clusters, and that only the intra-cluster diversity is correlated with time since HIV-1 infection at the early phase, we created a sequence clustering based diversity (SCBD) metric that clearly and accurately identifies incident infections from either single or multiple T/F strains. We employed a modified dot-matrix pairwise comparison method to eliminate the disproportional impact of indels and recombination, and used the proportion of clustered sequences ($P_c$) as an index to identify late chronic infections with declined viral genetic diversity.

**Materials and Methods**

**Dataset Construction**

We searched the Los Alamos HIV database [http://www.hiv.lanl.gov/; last modified on February 8, 2012] for all SGA sequencing based samples with ≥5 sequences (median, 26 sequences; range, 5 to 166) containing the env gene gp120 C2-V5 region (HXB2: 7050–7590). All samples were obtained from blood serum, plasma, or peripheral blood mononuclear cells (PBMC) of HIV-1 infected individuals. The resulting dataset of 561 samples, named D561, contained all 225 samples used by Park et al. (referred to as D225) and 336 others. In total, D561 comprised 12,778 env SGA sequences obtained from 462 individuals.

All cases were identified as incident or chronic infections using combined information from the original sources, including Fiebig stage [22], clinical records of time since diagnosis (incident infection defined as under one year), and/or symptoms of acute infection. Thus, 398 of the 561 samples were determined by their sources as incident infections (including 103 cases of multiple infections), with 228 subtype B infections from 196 individuals, 136 subtype C from 99 individuals, and 34 others; the remaining 163 samples are chronic infections, comprising 96 subtype B infections from 96 individuals, 57 subtype C from 49 individuals, and 10 others. The routes of exposure included 257 samples (163 incident and 94 chronic) from 231 heterosexual transmissions; 127 samples (81 incident and 46 chronic) from 109 men who have sex with men (MSM); 41 samples (31 incident and 10 chronic) from 34 intravenous drug users (IDU); and 136 unknown ones. Most of the samples were from treatment naïve individuals, except for 9 that were collected from individuals on or after antiretroviral therapy (ART). Detailed information of D561 is listed in the Table S1.

**Sequence Alignment and Diversity-based Clustering**

During multiple HIV-1 infection, viruses arisen from the same T/F strain tend to display lower diversity compared with those delineated from other T/F strains, particularly in early infection [19,20]. We therefore performed clustering of viral sequences for each sample, and defined the SCBD value as the metric for distinguishing incident and chronic infections, as represented in the flowchart (Figure 1).

First, pairwise sequence alignments were performed for all samples in the dataset. Simple measures of pairwise evolutionary distance, such as Hamming distance, relied on alignment methods that did not take into account the nature of the evolutionary events experienced, and can sometimes inflate the calculated diversity [23]. We therefore used iterative dot-matrix alignment, followed by a modified dynamic programming method to find the optimal pairwise alignment. Briefly, two sequences were written along the top row and leftmost column of a two-dimensional matrix, placing a dot where only if continuous 10 nucleotide matches occurred. The longest diagonal separating the sequence pair into unaligned segments was collected, while a segment pair was defined as an interval if the length of either segment was within the window size of 50 nucleotides. The 10 nucleotide match criterion captured most of the non-random nucleotide matches during modeling at mutation rates up to 10% (the upper limit for the HIV-1 env gene), while the 50-nucleotide window size was set to capture at least one indel or recombination event in the unaligned segment pair. The unaligned segments were then iteratively analyzed using above steps, until no diagonal could be found. An interval whose two segments were of equal length was analyzed base by base to detect point mutations; an interval segment pair of unequal length was considered to be the result of an indel or recombination event.

Previous studies have determined that the mean pairwise sequence diversity of the HIV-1 env gene region increases at a roughly constant rate of ~1.0% per year during early infection, with a standard deviation of ~0.2% [13]. Thus, we defined an incident infection as having occurred within one year to date and a chronic infection as having occurred more than one year ago, and calculated the pairwise sequence diversity and the overall mean value for each sample. If this mean diversity was less than 0.6% (the lower bound of the 95% confidence interval for an incident infection’s diversity value), it was directly recorded as the sample's SCBD. Otherwise, sequences were gradually grouped into clusters such that pairwise diversity for every sequence pair within the cluster was less than 1.0% (equivalent to the diversity at one year of infection). Only clusters that contain more than 20% of the sample’s total sequences and three or more sequences each were retained for further consideration, as most multiple infections of HIV-1 originate from approximately 2 to 5 T/F strains [19,20].

Additionally, it has been suggested that pairwise diversities of early infections originating from single T/F strain follow a Poisson distribution [20], with equal mean diversity and standard deviation values and hence a CV value (standard deviation/mean diversity) of 1.0. In contrast, the CV value of early infections caused by multiple, diverse T/F strains is expected to be higher than 1.0 and descends quickly as time goes on [17]. To correctly recognize chronic infections, we set a cutoff CV value of 0.65; this approximates the limit of CV in a chronic sample after 2 years of infection with two equally sized clusters. For any case with a lower CV value, the overall mean diversity ($D_{overall}$) was recorded as its SCBD value.

Furthermore, as some late-stage infections exhibit a decline in the overall viral diversity [13,14], care must be taken to distinguish such cases from incident infections [10]. Since the viral sequences in these late-stage infections have experienced longer evolutionary times and have undergone reproductive selection, their sampled sequences are mostly unable to form clusters, while those from incident multiple infections are mostly able to. We thus supplemented the calculation of SCBD values using the following metric. Using $N_i$ to denote the number of sequences in the $i$th cluster, and $N$ as the total number of sequences in the sample, we calculated the clusterable proportion, $P_c$, of each sample as:
Figure 1. Flowchart for identifying incident from chronic HIV-1 infections using the sequence clustering-based diversity (SCBD) assay.
doi:10.1371/journal.pone.0100081.g001
### Table 1. Homologous clusters identification for cases in datasets D225 and D561.

| Dataset | Clusters* | $D_{online} \leq 0.6\%$ | Chronic | $D_{online} > 0.6\%$ | Chronic |
|---------|-----------|--------------------------|---------|------------------------|---------|
|         |           | Incident** | Multiple |          | Incident** | Multiple |
|         |           | Single | Multiple |          | Single | Multiple |
| D225    | S         | 136 | 16 | 0 | 0 | 9 | 0 |
|         | M         | 0 | 0 | 0 | 0 | 21 | 0 |
|         | D         | 0 | 0 | 0 | 0 | 0 | 0 |
|         | -         | 0 | 0 | 0 | 0 | 43 | 0 |
|         | Total     | 136 | 16 | 0 | 0 | 30 | 43 |
| D561    | S         | 282 | 45 | 1 | 10 | 15 | 0 |
|         | M         | 0 | 0 | 0 | 3 | 43 | 3 |
|         | D         | 0 | 0 | 0 | 0 | 0 | 4 |
|         | -         | 0 | 0 | 0 | 0 | 155 | 0 |
|         | Total     | 282 | 45 | 1 | 13 | 58 | 162 |

*Homologous clusters identified by the SCBD assay.
'S': single; 'M': multiple; 'D': declined; '-' null.

**Incident infections were identified either as 'single' or 'multiple' infected according to resource information.
doi:10.1371/journal.pone.0100081.t001
Clusters were considered valid only when $P_c \geq 0.8$ for a sample with mean diversity greater than 1.0%, and the averaged value of the intra-cluster mean diversities were recorded as the SCBD values. Clusters were considered not valid when $P_c \leq 0.5$ for a sample with mean diversity lower than 1.0%, and identical sequences were removed from each sample, followed by calculation of the SCBD value using the remaining sequences. Otherwise, the SCBD value was calculated using all sequences.

Binary Classification of Incident and Chronic Infections

We defined the SCBD cutoff value as 1.0%, equivalent to the result of one year of HIV-1 viral evolution [13], and classified infections as either incident or chronic based on the calculated SCBDs. The sensitivity, specificity, and the overall accuracy for the dataset were calculated respectively. The Mathew Correlation Coefficient (MCC) value [24], which is related to the chi-square statistic for a binary classification, was also calculated as a quality measure of the classification scheme. To compare the performance of our method with those of the existing sequence-based approaches, we also evaluated the performance metrics of the HD $Q_{10}$ method with a cutoff value of 0 as suggested by Park et al [16], as well as the pattern-based method with 0.32 as suggested by Yang et al [17].

Results

Homologous Clusters of Cases in Datasets D225 and D561

We performed pairwise sequence comparison, $D_{small}$ calculation, and highly homologous sequence clustering for all cases in
D561, including the full set of 225 cases from Park et al [16]. As shown in Table 1, 152 incident cases of D225 and 327 of D561 (≈80%) had \(D_{\text{overall}}\) lower than 0.6%, which could be easily distinguished from the chronic ones. Among them, most were infections with a single T/F virus (136 of D225 and 282 of D561), while the others were originated from two or more closely homologous T/F strains (16 of D225 and 43 of D561), according to the original resources. The other 30 incident cases of D225 and 71 ones of D561 with higher than 0.6% diversity were mainly infections with multiple T/F strains (30 of D225 and 58 of D561), of which our method further identified ~75% of such cases (21 of D225 and 43 of D561) as incident infections containing distant homologous clusters. The remaining few incident cases were determined to consist of one major homologous cluster. On the other hand, the great majority of the chronic cases could not be clustered (43 of D225 and 155 of D561), indicating that sequences had greatly diverged over time.

Sequence Clustering Based Diversity (SCBD) Distribution
We further calculated the SCBD value for all cases in D561, including the full set of D225. Figure 2A plots the intra-patient HIV-1 SCBD dynamics, and Figure 2B shows the distribution of SCBDs of incident and chronic cases separately. There was a clear distinction between the SCBDs of the 398 incident samples and those of the 163 chronic samples, which allowed us to define a binary classification test using a simple cutoff SCBD value. In contrast, the HD Q10 method yielded a high degree of overlap between the diversity distributions of the incident samples and the chronic ones (Figure 2C). This suggested a strong advantage of SCBD over HD Q10 as a metric for distinguishing incident and chronic HIV-1 infections.

Performance on the Datasets D225 and D561
To assess the validity of our SCBD-based classification test, we first examined its performance on the 225-sample dataset employed by Park et al [16]. Meanwhile, the results were compared with those of the HD Q10, and our previous pattern-based methods. As shown in Table 2, our method correctly classified all 225 samples, while the HD Q10 method misclassified 7 of the 182 incident infections as chronic ones, and the pattern-based method misclassified 3 incident and 4 chronic infections. Cross-checking with source information revealed that 6 of the 7 samples misclassified by the HD Q10 method, and all 3 incident samples misclassified by the pattern-based method were from multiple-infected individuals, confirming the particular weakness of both the HD Q10 (\(p<0.01, \chi^2\) test) and pattern-based (\(p=0.02, \chi^2\) test) methods in correctly identifying the source of sequence diversity in multiple incident infections (Table 2). The SCBD method, in contrast, surmounted this challenge. The difference in classification accuracies between the three methods was statistically significant (\(p=0.02, \chi^2\) test).

We further tested the performance of our method and those of HD Q10 and pattern-based diversity on an expanded SGA-based dataset, D561. As shown in Table 2, our SCBD assay correctly classified all but 2 of the 398 incident infections and 2 of the 163 chronic infections, maintaining a high performance with 99.5% sensitivity, 98.3% specificity, 99.3% overall accuracy and an MCC value of 0.98. In contrast, while the HD Q10 method also yielded a high specificity value of 96.9%, it misclassified 40 of the 398 incident infections and yielded a sensitivity value of only 89.9%, bringing the overall accuracy and the MCC value down to 92.0% and 0.83. The pattern-based diversity yielded a high sensitivity value of 96.7%, but misclassified 25 of the 163 chronic infections and yielded a low specificity value of 84.7%, bringing the overall...
accuracy and the MCC value down to 93.2% and 0.83, which outperformed the HD Q10 method but was still much lower than our SCBD method. The difference in performance on this dataset was statistically significant \((p<0.01, \chi^2 \text{ test})\). Moreover, we continued to observe an advantage in the SCBD method in correctly classifying multiple-strain incident infections: sensitivity values for multiple and single incident infections were 99.0% and 99.7% \((p=0.98, \chi^2 \text{ test})\), respectively, compared to 84.5% and 91.9% from HD Q10 \((p=0.02, \chi^2 \text{ test})\), and 87.4% and 100.0% from the pattern-based method \((p<0.01, \chi^2 \text{ test})\).

Finally, we assessed the influences of viral subtype and transmission routes on the SCBD assay's performance. As shown in Table 3.

### Table 3. Influence of viral subtype and transmission route over the SCBD assay's performance on the dataset D561.

| Factor | Category | Accuracies | Sensitivity | Specificity | Accuracy | mcc |
|--------|----------|------------|-------------|-------------|----------|-----|
| Subtype | B        | 227/228 (99.6%) | 95/96 (99.0%) | 322/324 (99.4%) | 0.99     |
|        | C        | 135/136 (99.3%) | 56/57 (98.2%) | 191/193 (99.0%) | 0.98     |
| Others |          | 34/34 (100.0%) | 10/10 (100.0%) | 44/44 (100.0%) | 1.0      |
| Exposure | SH       | 162/163 (99.4%) | 94/94 (100.0%) | 256/257 (99.6%) | 0.99     |
|        | MSM      | 81/81 (100.0%) | 46/46 (100.0%) | 127/127 (100.0%) | 1.0      |
|        | IDU      | 31/31 (100.0%) | 10/10 (100.0%) | 41/41 (100.0%) | 1.0      |
|        | Unknown  | 122/123 (99.2%) | 11/13 (84.6%) | 133/136 (97.8%) | 0.87     |

The influence of viral subtype and transmission route over the SCBD assay's performance on the dataset D561.
**Distribution of Misclassified Cases**

The infection time distribution of the cases misclassified by our SCBD and the HD $Q_{10}$ method was shown in Figure 2D. The majority of misclassified cases in HD $Q_{10}$ were incident infections, and the distribution peak was located at 3–6 months. In contrast, most of the misclassified cases (3/4) under the SCBD method had time since infection between 6–18 months, with a nearly equal number of cases on either side of the one-year binary classification threshold. The minority of misclassified cases in HD $Q_{10}$ were incident infections, and the distribution peak was located at 3–6 months. In contrast, most of the misclassified cases (3/4) under the SCBD method had time since infection between 6–18 months, with a nearly equal number of cases on either side of the one-year binary classification threshold. This, in fact, agrees with our results. In contrast, the cases misclassified by HD $Q_{10}$ were mainly incident infections, peaking at the 3–6 month time frame (Figure 2D). The SCBD method can therefore be more reliably employed than methods such as HD $Q_{10}$ in the estimation of incidence on a population level, as the normally distributed false negative and false positive cases may largely compensate each other during rate calculations.

One foreseeable issue of our SCBD assay is the labor intensiveness and high cost of SGA sequencing techniques compared with serology methods, which may limit its large scale application. However, our method offers substantial improvements in assay specificity and sensitivity, which may effectively offset its deficits when taken into account the concerns of serological methods in reliability and reproducibility. In addition, technical challenges inherent to SGA have often curtailed the rates of successful amplifications and hence may bring to question the generalizability of diversity-based methods on a population level. However, failed SGA amplifications are typically due to poor preservation of sample nucleic acids, and not tendentious bias towards particular types of samples. At the same time, deep sequencing methods which produce more than 10,000 reads from a single sample lends a highly promising alternative for addressing the problems of cost, bias, and efficiency, and future research should therefore seek to apply the SCBD method on deep-sequence data. As well, the current availability of large sets of Sanger-sequenced data in the public domain on the HIV-1 pol gene, with ambiguous base calls preserved for genotypic drug resistance analysis, also holds promise as a basis for diversity-based incidence assays [15,25].

However, the high performance of our assay demonstrates the potential of sequencing based techniques as a strong alternative for identifying incident HIV infections. Its substantial improvement in assay specificity and sensitivity may effectively counter its material and labor costs as compared with serology methods, particularly in...
research settings and for areas under intensive incidence surveillance. Consequently, the SCBD method may be most advantageous to settings with low to moderate incidence relative to available resources, where extensive coverage of the targeted or sample population can be achieved through only a small number of sequencing assays [26,27]. In contrast, SCBD may be more valuable to regions of higher HIV incidence as a validation assay on a subset of all STARHS-screened samples. Strategic employment of the SCBD method in combination with serology based methods will help to efficiently clarify HIV epidemic trends in differently affected areas.

References

1. Hall HI, Song R, Rhodes P, Prejean J, An Q, et al. (2008) Estimation of HIV incidence in the United States. JAMA 300: 320–329.
2. Guy R, Gold J, Calleja JM, Kim AA, Parekh R, 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.
3. Justman J, El-Sadr WM (2010) AIDS response at a crossroads. Science, 329, 120.
4. Brookmeyer R (2009) Should biomarker estimates of HIV incidence be adjusted? AIDS 23: 485–491.
5. Janssen RS, Satten GA, Stramer SL, Rawal BD, O’Brien TR, et al. (1998) New testing strategy to detect HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. JAMA 280: 42–48.
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: 285–307.
7. Chawla A, Murphy G, Donnelly C, Booth CL, Johnson M, et al. (2007) Human immunodeficiency virus (HIV) antibody avidity testing to identify recent infection in newly diagnosed HIV type 1 (HIV-1)-seropositive persons infected with diverse HIV-1 subtypes. J Clin Microbiol 45: 415–420.
8. Murphy G, Parry JV (2008) Assays for the detection of recent infections with human immunodeficiency virus type 1. Euro Surveill 13.
9. Karita E, Price M, Hunter E, Chomba E, Allen S, et al. (2007) Investigating the utility of the HIV-1 BED capture enzyme immunoassay using cross-sectional and longitudinal seroconverter specimens from Africa. AIDS 21: 403–408.
10. Hayashida T, Gatanaga H, Tanuma J, Oka S (2008) Effects of low HIV type 1 load and antiretroviral treatment on IgG-capture BED-enzyme immunoassay. AIDS Res Hum Retroviruses 24: 495–498.
11. UNAIDS (2006) 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 81: 40.
12. Incidence Assay Critical Path Working Group (2011) More and better information to tackle HIV epidemics: towards improved HIV incidence assays. PLoS Med 8: e1001045.
13. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73: 10489–10502.
14. Lee HY, Perelson AS, Park SC, Leitner T (2008) Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput Biol 4: e1000240.
15. Koyou RD, von Wyl V, Yerly S, Boni 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.
16. 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–F19.
17. Yang J, Xiu X, He X, Yang S, Ruan Y, et al. (2012) A new pattern-based method for identifying recent HIV-1 infections from the viral env sequence. Sci China Life Sci 55: 328–335.
18. Allam O, Samaranay, S, Ahmad A (2011) Hammering out HIV-1 incidence with Hamming distance. AIDS, 25, 2047–2048.
19. Li H, Bar KJ, Wang S, Decker JM, Chen Y, et al. (2010) High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men. PLoS Pathog 6: e1000909.
20. Kelle BF, Giorgi EE, Sazawal-Gonzalez JF, Decker J, 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: 7522–7527.
21. Mansky LM, 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.
22. Fiebig EW, Wright DJ, Rawal BD, Garret PE, Schumacher RT, et al. (2003) Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. AIDS 17: 1871–1879.
23. Loyteroja A, Goldman N (2008) Phylogenetics-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320: 1632–1635.
24. Balb P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16: 412–424.
25. Ragan-Metronin M, Aris-Brosou S, Joamie I, Merck H, Vallee D, et al. (2012) General diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. J Infec Dis 206: 756–764.
26. Tey J, Ang LW, Tay J, Cutter JL, James L, et al. Determinants of Late-Stage HIV Disease at Diagnosis in Singapore, 1996 to 2009. Ann Acad Med Singapore 41: 194–199.
27. Chana MOH, UNAIDS WHO (2011) Estimates for the HIV/AIDS epidemic in China. Available: http://www.chinaids.org.cn/n1971/n2151/n777994.files/n777993.pdf.

Supporting Information

Table S1 Detailed information of the dataset D561.

Author Contributions

Conceived and designed the experiments: XP YS. Performed the experiments: XX MG JH. Analyzed the data: XX MG JH. Contributed reagents/materials/analysis tools: XX MG JH. Wrote the paper: XX MG JH ZW XH YR.