Published in final edited form as:

AIDS. 2011 October 23; 25(16): F13–F19. doi:10.1097/QAD.0b013e328349f089.

Designing a Genome-Based HIV Incidence Assay with High Sensitivity and Specificity

Sung Yong Park^a, Tanzy M. T. Love^a, Jeremy Nelson^a, Sally W. Thurston^a, Alan S. Perelson^b, and Ha Youn Lee^a

^aDepartment of Biostatistics and Computational Biology, University of Rochester, Rochester, NY

^bTheoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Abstract

Objective—Considerable inaccuracy in estimates of human immunodeficiency virus (HIV) incidence has been a serious obstacle to the development of efficient HIV/AIDS prevention and interventions. Accurately distinguishing recent or incident infections from chronic infections enables one to monitor epidemics and evaluate the impact of HIV prevention/intervention trials. However, serological testing has not been able to realize these promises due to a number of critical limitations. Our study is to design a novel scheme of identifying incident infections in a highly accurate manner, based on the characteristics of HIV gene diversification within an infected individual.

Methods—We perform a comprehensive meta-analysis on 5596 full envelope HIV genes generated by single genome amplification-direct sequencing from 182 incident and 43 chronic cases. We devise a binary classification test based on the tail characteristics of the Hamming distance distribution of sequences.

Results—We identify a clear signature of incident infections, the presence of closely related strains in the sampled HIV envelope gene sequences in each HIV infected patient, in both single-variant and multi-variant transmissions. The sequence similarity used as a biomarker is found to have high specificity and sensitivity, greater than 95%, and is robust to viral and host specific factors such as the clade of the viral strain, viral load, and the length and location of sequences in the HIV envelope gene.

Conclusions—Because of rapid and continuing improvements in sequencing technology and cost, sequence based incidence assays hold great promise as a means of quantifying HIV incidence from a single blood test.

Keywords

HIV Incidence; Viral Evolution; Envelope Gene; Hamming Distance Distribution

Correspondence to Dr. Ha Youn Lee, PhD, 601 Elmwood Ave. Box 630 Rochester, NY 14642-0001 Department of Biostatistics and Computational Biology, University of Rochester Hayoun_Lee@urmc.rochester.edu Tel: 585-275-0041 Fax: 585-273-1031.

AUTHOR CONTRIBUTIONS

S.Y.P and H.Y.L. designed the project and formulated the assay. T.M.T.L. and S. W. T. provided statistical input and J.N. performed sequence analysis. A.S.P contributed to the initiation and formulation of the project. All authors participated in the writing of the manuscript.

INTRODUCTION

To assess how many people have been recently infected in a given area is an important task in HIV/AIDS prevention [1]. Accurately distinguishing recent or incident infections from chronic infections enables one to monitor epidemics, evaluate the impact of antiretroviral treatment, and assess the efficacy of HIV prevention trials including vaccination [2], microbicides [3], and other types of interventions [4]. Precise estimates of HIV incidence are essential to allocate HIV-related health care resources properly. In particular, incident estimates based on single blood draws from cross-sectional sampling of populations are in great need in resource-limited settings.

The approximate window period of HIV incident infections is the first one year post transmission, which covers the eclipse phase and the stages of the Fiebig classification based on orderly appearance of viral RNA, viral antigens such as p24 and p31, and HIV-specific antibodies [5]. This period is characterized by rapid expansion and decline of viral RNA and gradual increase of HIV-specific antibody titers (Figure 1A). Current HIV incidence assays are based on the idea that antibody level or avidity rise in a predictable pattern during the first 4 to 6 months post transmission, eventually reaching a plateau that stays roughly constant for many years (Figure 1A). This approach includes the Serologic Testing Algorithm for Recent HIV Seroconversion (STARHS) [6, 7], the BED capture enzyme immunoassay (BED) [8], and the guanidine-based antibody avidity assay [9, 10]. However, the serologic assays are found to have a number of critical limitations including difficulty in the standardization, complication in the reproducibility, and strong dependence on the infecting virus clade [9, 11]. The limitations result in notable inaccuracy. The sensitivity, the proportion of incident infections correctly identified as incident, varies in the range of 42% and 100% with median of 89%, across 13 serologic assays [12]. The specificity, the proportion of chronic infections correctly identified as chronic, ranges from 49.5% to 100% with median of 86.8%. The tendency of the misclassification of long-standing infections as recent is pronounced among patients on anti-retroviral treatment [12]. The substantial false recent rate is reported to be one of the limitations of the serologic assays as it results in the overestimation of incidence.

Here, we turn to utilizing recent advances in understanding early HIV infections [13, 14] and demonstrate that information derived from a set of HIV envelope genes obtained from a single blood sample can accurately distinguish incident infections from chronic ones. Single genome amplification and sequencing of HIV envelope genes have shown that the majority of HIV infections originates from a single strain for both subtype B and subtype C infections [13-15]. Not all infections originate from a single founder strain. Two risk groups, men who have sex with men (MSM) and injection drug users (IDUs), show a high chance of being productively infected by more than one type of virus [16, 17]; around 36% of MSM and around 60% of IDUs were found to be multivariant transmission. Figure 1B illustrates the difference between a single-variant and multi-variant transmissions. As the infection propagates within an individual, as a consequence of the dynamic interplay between viral mutation and immune selection [18], the HIV population diversifies (Figure 1B). This study suggests a novel assay with high levels of sensitivity and specificity based on the characteristics of HIV gene diversification originating from either a single-variant or multiple-variant founder strains.

METHODS

Sequence Data Sources

The HIV env sequences of 182 incident and 43 chronic patients were collected from the published data set in Refs. [13], [14] and [17]. Geographic locations of the cohorts were US,

Trinidad, South Africa, Malawi, and Canada. All of the 5596 strains we analyzed were obtained by single genome amplification and sequencing. The incident subjects were sub-staged according to the Fiebig classification [5]: 1 subject was in stage I, 74 in stage II, 24 in stage III, 23 in stage IV, 44 in stage V, and 16 in stage VI. The routes of exposure include 92 heterosexual transmissions, 16 MSM subjects, 12 IDU subjects, and others of unknown route.

ROC Analysis

As in Ref. [19], the proportion of incident infections being correctly identified as incident, sensitivity, is plotted against the proportion of misclassification of chronic infections as incident, 1-specificity, as we incrementally change the putative Q_{10} cutoff value. The optimal cut-off value is determined by the isocost line, maximizing the sum of sensitivity and specificity with equal consideration.

RESULTS

We performed a meta-analysis by collecting 5596 previously published sequences generated by single genome amplification-direct sequencing [20, 21] from 182 incident and 43 chronic cases [13, 14, 17]. The incident subjects were classified as recent HIV infections either by symptoms of acute infection or serologic evidence and the chronic subjects were reported to have an infection period of longer than 1 year (see Methods for more). Incident infections were categorized into either single-variant or multi-variant transmission [13, 14, 17]. The diversification can be quantified using the number of base differences between a pair of sequences, i.e., their Hamming distance (HD): HIV *env* diversity is the average number of base differences among all possible pairs of sequences sampled from a patient, divided by the sequence length. The *env* variance is the variance of the number of base differences among the sequences divided by the sequence length.

The high level of viral sequence diversity associated with multi-variant transmissions suggests that a simple measure of the diversity or variance might misclassify early stages of individuals whose infection started with multiple founder viruses as being chronically infected. Indeed, as shown in Figure 2A and B, the level of *env* diversity of around one third of the incident multi-variant cases overlaps with those of chronic subjects. Furthermore, the third quartile of the *env* variance of the incident subjects with multiple founder strains is greater than the median *env* variance of the chronic subjects. Neither envelope gene diversity nor variance provides clear discrimination between incident infections originating from multiple founder strains and chronic infections.

We sought an alternative signature in the HD distribution that discriminates chronic and incident infections. At an early phase, there should exist a fair number of identical or nearly identical sequences in each lineage of transmitted strain. Indeed, Figure 3A shows that the first peak of the HD distribution of incident cases including both single founder (Figure 3A top left) and multiple founder infections (Figure 3A top right), is located in the region of very low Hamming distances, implying the presence of closely related sequences. As infection progresses and the HIV population diversifies, the proportion of similar sequences should decline (Figure 1B); in fact, the proportion of identical sequences has been found to decrease exponentially as a function of time post infection [13, 15]. Figure 3A confirms that chronic subjects have a negligible frequency of sequence pairs in the region of low HD values, suggesting the absence of closely related sequences. We clarify this signature by quantifying the tail characteristics of the HD distribution; we measure the 10% quantile for HD, Q_{10} , i.e., the HD value dividing the HD distribution into 10% below it and 90% above it. Figure 3B highlights the difference between the distribution of the Q_{10} statistics for the 182 incident infection samples and that for the 43 chronic samples. Here the 182 incident

patients include both 102 single founder and 80 multiple founder cases. The incident Q_{10} distribution (red in Figure 3B), which includes both single and multiple founder infections, is visibly disparate from the chronic distribution (blue in Figure 3B).

The clear difference between the incident and chronic Q_{10} distributions led us to devise a binary classification test to identify samples from incident infections as being significantly different from the population of chronic infections. If Q_{10} is greater than the cut-off value $Q_{10}^{\text{cut-off}}$, the sample is judged as a chronic infection and otherwise the sample is scored as an incident infection. As with most binary classifications, there is a trade-off between specificity and sensitivity that is controlled by the choice of the threshold. The cut-off value of $Q_{10}^{\text{cut-off}}$ is objectively determined from an analysis of the receiver operating characteristic (ROC) curve [22]. The isocost line designates $Q_{10}^{\text{cut-off}}=7$ as the optimal value. Whereas simple measures of viral diversity and variance fail to discriminate chronic samples from incident ones, the binary classification test statistically differentiates the two groups. All of the 43 chronic subjects showed Q_{10} values greater than the threshold of 7, indicating a specificity of 100%. Only 5 out of 182 incident subjects had Q_{10} values greater than the threshold; the measured sensitivity is 97.3% and the majority of the 5 misclassified subjects was infected through intravenous drug use. These high levels of sensitivity and specificity convincingly suggest the possibility of using the tail characteristics of the HD distribution as a biomarker for identification of incident infections. Measuring HD distribution is advantageous because it can be determined from a single blood draw from each individual in cross-sectional blood surveys.

Our biomarker is robust under changes of viral specific and host specific factors such as the viral subtype, the viral load of subjects, and the length and location of the sampled envelope gene sequences. Figure 4A shows the ROC curve of the Q_{10} distributions when we exclude the dataset of subtype C infections. The area under the ROC curve with subtype B infections only remained the same as that with both subtype B and C infections, 0.998, implying that the biomarker is not sensitive to the clade of the viral strains. The ROC curve with only subtype B infections provides a sensitivity of 95.6% and a specificity of 100% with $Q_{10}^{\text{cut-off}}=7$. This is presumably because the dynamics of early HIV diversification is not greatly affected by viral subtype. In contrast, the existing serologic assays have significantly different window periods of incident infections among subtype B and other subtypes [11]. Little association is observed between the biomarker and the viral load. Figure 4B shows the scatter plot of Q_{10} values and viral loads measured from both incident and chronic subjects. The correlation coefficients were -0.04 for incident subjects and 0.13 for chronic subjects, suggesting that the biomarker is not sensitive to each patient's viral load.

We also find that the sensitivity and specificity of the assay remain very high under the changes of either the length or region of the envelope gene. While the changes in the incident Q_{10} distribution by varying the length of *env* are minor, the mean of the chronic Q_{10} distribution decreases substantially as the length of *env* decreases (Figure 4C). Despite this dependence, we stress that the sensitivity and specificity are markedly high, 95.1% or greater regardless of whether we consider 500, 1000, 2000 base long *env* segments or full *env*, as we control $Q_{10}^{\text{cut-off}}$ values objectively based on the ROC curve analysis (see Figure 5A). Our analysis suggests that read lengths of HIV *env* as short as 500 bases do not affect the accuracy of the assay.

Chronic Q_{10} distributions show a considerable amount of variation with the choice of the location within *env*. As Figure 4D displays, the 500 base long segment of *env* encompassing the major portion of the V3 loop, HXB2 7125-7624, shows the greatest mean of Q_{10} and the

segment of HXB2 7625-8124 shows the smallest mean. We postulate that differing observed distributions imply that, in chronic infections, purifying selections keep certain sections of *env* quite conserved despite a long period of infection. The presence of purifying selections in chronic infection has been reported [23]. However, the impact of purifying selection does not appear to be strong enough to weaken the signature of chronic infection. The power of discrimination even in the least sensitive region (HXB2 7625-8124) is comparable to the power of the entire *env*; the sensitivity is 98.4% and the specificity is 97.7% with the optimal $Q_{10}^{\text{cut-off}}=0$; the 10% quantile of the HD distributions of 179 out of the 182 incident subjects was 0 but only a single chronic subject had a 10% quantile value of 0. We conclude that our HIV incidence assay is robust to changes of the length and location of HIV *env* as summarized in Figure 5.

DISCUSSION

We find that sequence similarity used as a biomarker has high specificity and sensitivity and is robust to viral and host specific factors such as the clade of the viral strain, viral load, and the length and location of sequences in the HIV envelope gene. We found that a simple measure of the viral diversity or variance failed to distinguish chronically infected individuals from those infected with multiple founder viruses but who are at an early stage. This is due to the fact that distinct founder strains in multivariant transmissions caused increased HD diversity and variance (see Figure 2 and Figure 3A). On the other hand, there still exists a tangible number of very closely related sequences within each lineage of the founder virus at incident stage, which yields lower Q_{10} values than do individuals in chronic stage. Consequently, the 10% quantile of the HD distribution, instead of the mean or variance of the HD distribution, was found to be a robust measure for distinguishing incident infections, including multi-variant transmissions, from chronic infections.

One foreseeable issue for the development of a genome-based HIV incidence assay is the decline in viral sequence diversity that occurs during the later stages of infection [24, 25]. This common phenomenon of diversity decline as the end point disease is approached implies that we cannot exclude the possibility that a sequencing based assay might identify some subjects with late infection as having an incident infection. Sequences collected from patients with end point disease would be essential to examine this possibility at the population level. However, such late stage patients should be identifiable based on clinical criteria by introducing additional measures such as CD4+ T cell count.

The datasets used in the present study were obtained by single genome amplification and sequencing which conventionally samples less than 100 sequences. On the other hand, deep sequencing [26] is capable of producing more than 10,000 reads from a single blood sample. The estimation of tail characteristics of a distribution such as Q_{10} requires a substantially greater sample size than the estimation of central characteristics such as the mean or median. One of the limitations of the current deep sequencing platforms is that a relatively short read length (400-600 base long) is produced in comparison to single genome amplification (SGA) and Sanger sequencing. Our analysis suggests that short read lengths do not affect the accuracy of the assay, implying that data from deep sequencing methods could also be used to develop the assay. Caution would be needed in terms of sequencing errors and re-sampling issues of deep sequencing methods for future implementation.

Our study has demonstrated that a sequencing based HIV incidence assay could be a very powerful tool for identifying incident infections in a highly accurate manner. The rapid decrease in the cost of DNA sequencing over decades suggests the potential of our assay to be cost-effective and widely adopted in clinical practice. More empirical data will be essential to further validate this new paradigm by examining the sensitivity of the assay to

logistical factors. Our assay is advantageous as it is expected to designate incident infection or chronic infection from a single blood draw measure, not necessarily combining different measures and adopting multi-assay algorithms.

Acknowledgments

We thank Dr. George Shaw, Dr. Stuart Z. Shapiro, and Dr. Steven Wolinsky for critical review of the manuscript.

Financial support: This work was supported by NIH grants R01 AI083115 and AI095066 (to HYL), and R01 RR06555 and R37 AI28433 (to ASP) as well as NSF grant PHY05-51164.

REFERENCES

1. Brookmeyer R. Reconstruction and future trends of the AIDS epidemic in the United States. *Science*. 1991; 253:37–42. [PubMed: 2063206]
2. Burton DR, Desrosiers RC, Doms RW, et al. HIV vaccine design and the neutralizing antibody problem. *Nat Immunol*. 2004; 5:233–6. [PubMed: 14985706]
3. McGowan I. Microbicides: a new frontier in HIV prevention. *Biologicals*. 2006; 34:241–55. [PubMed: 17097303]
4. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. *PLoS Med*. 2005; 2:e298. [PubMed: 16231970]
5. Fiebig EW, Wright DJ, Rawal BD, et al. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS*. 2003; 17:1871–9. [PubMed: 12960819]
6. Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA*. 1998; 280:42–8. [PubMed: 9660362]
7. Kothe D, Byers RH, Caudill SP, et al. Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence. *J Acquir Immune Defic Syndr*. 2003; 33:625–34. [PubMed: 12902808]
8. Hargrove JW, Humphrey JH, Mutasa K, et al. Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. *AIDS*. 2008; 22:511–8. [PubMed: 18301064]
9. Chawla A, Murphy G, Donnelly C, et al. 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*. 2007; 45:415–20. [PubMed: 17151211]
10. Thomas HL, Wilson S, O'Toole CM, et al. Differential maturation of avidity of IgG antibodies to gp41, p24 and p17 following infection with HIV-1. *Clin Exp Immunol*. 1996; 103:185–91. [PubMed: 8565298]
11. Busch MP, Pilcher CD, Mastro TD, et al. Beyond detuning: 10 years of progress and new challenges in the development and application of assays for HIV incidence estimation. *AIDS*. 2010; 24:2763–71. [PubMed: 20975514]
12. Guy R, Gold J, Calleja JM, 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; 9:747–59. [PubMed: 19926035]
13. Keele BF, Giorgi EE, Salazar-Gonzalez JF, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A*. 2008; 105:7552–7. [PubMed: 18490657]
14. Abrahams MR, Anderson JA, Giorgi EE, et al. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J Virol*. 2009; 83:3556–67. [PubMed: 19193811]
15. Lee HY, Giorgi EE, Keele BF, et al. Modeling sequence evolution in acute HIV-1 infection. *J Theor Biol*. 2009

16. Li H, Bar KJ, Wang S, et al. High multiplicity infection by HIV-1 in men who have sex with men. *PLoS Pathog.* 2010; 6:e1000890. [PubMed: 20485520]
17. Bar KJ, Li H, Chamberland A, et al. Wide variation in the multiplicity of HIV-1 infection among injection drug users. *J Virol.* 2010; 84:6241–7. [PubMed: 20375173]
18. Wolinsky SM, Korber BT, Neumann AU, et al. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science.* 1996; 272:537–42. [PubMed: 8614801]
19. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978; 8:283–298. [PubMed: 112681]
20. 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; 43:406–13. [PubMed: 15635002]
21. Salazar-Gonzalez JF, Bailes E, Pham KT, et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol.* 2008; 82:3952–70. [PubMed: 18256145]
22. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press; New York: 2003.
23. Edwards CT, Holmes EC, Pybus OG, et al. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics.* 2006; 174:1441–53. [PubMed: 16951087]
24. Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol.* 1999; 73:10489–502. [PubMed: 10559367]
25. Lee HY, Perelson AS, Park SC, Leitner T. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput Biol.* 2008; 4:e1000240. [PubMed: 19079613]
26. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2009; 11:31–46. [PubMed: 19997069]

Figure 1.

A. Typical plots of viral load (red) and antibody titer (blue) following HIV transmission. Incident infection denotes the first one year of infection approximately. **B.** Schematic representations of HIV genomic populations at viral transmission, incident stage, and chronic stage. The first row represents a typical diversification pattern when an infection originates from a single founder and the second row represents when an infection starts from three founder strains.

Figure 2.

The *env* diversity (A) and variance (B) of 102 acutely infected subjects (red) with a single strain infection, 80 acutely infected subjects with multiple strain transmission (pink), and 43 chronically infected subjects (blue). Each horizontal black line denotes the median of each group and each box plots the first and third quartiles.

Figure 3.

A. The HD distribution of the sampled sequences from two incident patients, ACT54869022 in Ref. [17] (top left) and 703010228 in Ref. [14] (top right) and two chronic subjects in Ref. [13], SMRE4166 (bottom left) and SHKE4761 (bottom right). The red dashed line indicates the position of computed 10% quantile for the HD distribution, Q_{10} , for each subject. **B.** The distribution of the statistic Q_{10} for the samples of 182 incident infections with a smoothed approximation shown in red. Overlaid in blue is the smoothed estimate of the distribution of Q_{10} calculated from 43 samples of chronic infections. The incident Q_{10} distribution includes both 102 single and 80 multiple founder infections. **C.** The computed ROC curve for the binary classification test based on the incident and chronic Q_{10} distributions presented in **B.**

Figure 4.

A. Dependence of the ROC curve on the subtype of infection. The blue line represents the original ROC curve with the samples from both subtype B and C infections. The red line represents the ROC curve when we exclude 69 incident samples with subtype C infections.

B. The scatter plot of Q_{10} and viral load measured from incident (red) and chronic (blue) subjects. **C.** Dependence of the Q_{10} distribution on the length of the genome. The three overlapping red lines denote the Q_{10} distributions for the *env* genes of 182 incident infections with sequence lengths of 500, 1000, and 2000 bases. Here the three lines are indistinct since the three incident Q_{10} distributions are very close to each other. The three blue curves represent the Q_{10} distributions for the *env* genes of 43 chronic infections with sequence lengths of 500, 1000, and 2000 bases. **D.** Dependence of the Q_{10} distribution on the location of 500 base long *env* segment. For 43 chronic samples the Q_{10} distribution is shown in blue; the segment of *env* gene HXB2 7125-7624 showed the greatest mean of Q_{10} and the segment of HXB2 7625-8124 showed the smallest mean of Q_{10} . The two overlapping red curves denote Q_{10} distributions of the 182 incident samples at these two regions. The two red lines are visibly indistinct as the incident Q_{10} distributions of the two regions are extremely close to each other.

Figure 5.

A. The optimal cut-off value for the 10% quantile, $Q_{10}^{\text{cut-off}}$, of the binary classification test for each length and placement of the viral segments. The starting position of each segment is referenced to the HXB2 strain. As the length of envelope gene shortens the cut-off value decreases. **B.** The sensitivity (red) and specificity (blue) of the binary classification test for each viral segment.