Medicine®

DIAGNOSTIC ACCURACY STUDY

OPEN

A Generalized Entropy Measure of Within-Host Viral Diversity for Identifying Recent HIV-1 Infections

Julia Wei Wu, ScD, Oscar Patterson-Lomba, PhD, Vladimir Novitsky, MD, PhD, and Marcello Pagano, PhD

Abstract: There is a need for incidence assays that accurately estimate HIV incidence based on cross-sectional specimens. Viral diversity-based assays have shown promises but are not particularly accurate. We hypothesize that certain viral genetic regions are more predictive of recent infection than others and aim to improve assay accuracy by using classification algorithms that focus on highly informative regions (HIRs).

We analyzed HIV *gag* sequences from a cohort in Botswana. Forty-two subjects newly infected by HIV-1 Subtype C were followed through 500 days post-seroconversion. Using sliding window analysis, we screened for genetic regions within *gag* that best differentiate recent versus chronic infections. We used both nonparametric and parametric approaches to evaluate the discriminatory abilities of sequence regions. Segmented Shannon Entropy measures of HIRs were aggregated to develop generalized entropy measures to improve prediction of recency. Using logistic regression as the basis for our classification algorithm, we evaluated the predictive power of these novel biomarkers and compared them with recently reported viral diversity measures using area under the curve (AUC) analysis.

Change of diversity over time varied across different sequence regions within *gag*. We identified the top 50% of the most informative regions by both nonparametric and parametric approaches. In both cases, HIRs were in more variable regions of *gag* and less likely in the p24 coding region. Entropy measures based on HIRs outperformed previously reported viral-diversity-based biomarkers. These methods are better suited for population-level estimation of HIV recency.

The patterns of diversification of certain regions within the *gag* gene are more predictive of recency of infection than others. We expect this result to apply in other HIV genetic regions as well. Focusing on these informative regions, our generalized entropy measure of viral diversity demonstrates the potential for improving accuracy when identifying recent HIV-1 infections.

(*Medicine* 94(42):e1865)

Abbreviations: ART = antiretroviral therapy, AUC = area under the curve, HIRs = highly informative regions, HRM = high resolution melting, IQR = interquartile range, Q10 = the tenth quantile, ROC = receiver operating characteristic, SE = segmented entropy, SGA = single genome amplification.

INTRODUCTION

Accurate HIV incidence assays are important for characterizing HIV epidemics, and for designing and assessing intervention efforts.[1,2] An incidence rate is defined as the number of new cases per population at risk in a given time period. Given the long asymptomatic period for HIV infection, diagnosis counts in standard surveillance systems cannot be used reliably for this purpose. One approach is to rely on cohort studies that follow sero-negative persons over time and document HIV acquisition. This cohort approach is not only time-consuming and expensive, but also subject to selection and follow-up biases. In many cases, high-risk persons are also those who are more likely to be lost-to-follow-up, which leads to an underestimation of the incidence. The cohort approach can also generate incidence estimates that are not generalizable to the whole population, either because the study participants have modified infection risk, or the cohort sample does not represent the population of interest.[3]

An alternative approach based on a single, cross-sectional survey can address many of these challenges. In this approach biological samples are collected in a cross-sectional survey, and host or viral biomarkers used to identify recent versus chronic HIV infections.[3] However, identifying recent infections is still a challenge.[3] In recent years, within-host viral genetic diversity measures have stood out as a promising biomarker for this purpose. The rationale is based on the observation that early in infection, within-host viral genetic diversity increases in an approximately linear fashion.[4,5] Previous studies demonstrate that the majority of HIV infections are caused by a single founder strain.[6] Over time, large quantities of distinct viral variants are generated due to rapid viral replication, frequent mutation, and recombination events.[6,7] As a result, within-host viral genetic sequences are usually homogeneous early on during infection.[8,9] Over time the viral diversity increases and stabilizes or declines in later stages of the disease.[4,9,10] This sets a biological foundation to use HIV genetic diversity as a potential biomarker to identify recent infections. The minority of HIV infections caused by multiple founder viruses presents a challenge that needs to be addressed separately.[11,12–15] Recent reports include using the fraction of ambiguous nucleotide calls obtained during population sequencing,[14] a rapid diversity assay based on high resolution melting (HRM) technology[12,16–18] as well as the quasi-

Editor: Ken Rosenthal.
Received: July 27, 2015; revised: September 25, 2015; accepted: September 28, 2015.
From the Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA (JWW); Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA (OP-L, MP); and Department of Immunology and Infectious Disease, Harvard T.H. Chan School of Public Health, Boston, MA (VN).
Correspondence: Julia Wei Wu, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA (e-mail: wew758@-mail.harvard.edu).
Supplemental Digital Content is available for this article.
Author contribution: Conceived and designed the experiments: JWW, OPL, MP; performed the experiments: JWW, OPL, MP; analyzed the data: JWW, OPL, MP; contributed reagents/materials/analysis tools: VN; wrote the paper: JWW, OPL, VN, MP.
Funding: The present work was supported by National Institutes of Health (NIH) grants R01AI097015-03 and T32AI007358-26.
Data presented and abstract published previously at the annual Conference on Retroviruses and Opportunistic Infections (CROI) 2015.
The authors have no conflicts of interest to disclose.

ISSN: 0025-7974
DOI: 10.1097/MD.0000000000001865

species sequencing-based diversity measures. Particularly worth noting are the tenth quantile (Q10) of the pairwise Hamming genetic distance proposed by Park et al,[13] the sequence clustering-based diversity assay introduced by Xia et al,[15] and the Segmented Entropy proposed by Exner and Pagano.[19] However, the ability of these biomarkers to classify recent versus nonrecent cases accurately needs improvement.

In this study, we hypothesize that the change of within-host diversity over time varies across different regions of the viral genome. As a result, certain viral genetic regions should contain stronger temporal signals, and thus be more informative for comparing recent and chronic stages of infection. Consequently, classification algorithms that focus on within-host viral diversity of highly informative genetic regions (HIRs) can display better accuracy.

We previously developed a viral diversity measure based on a modified entropy definition.[19] In the present study, we build upon this method by using a generalized entropy measure of within-host viral diversity. We first screen for genetic regions that best differentiate recent versus chronic stages of infection, to which we then apply our modified entropy measure. Descriptive analyses show that change of entropy over time varies across different gene sequence regions. As a result, some regions are more predictive of time since infection than others. To evaluate the levels of information in the different genetic regions, we use both parametric and nonparametric approaches, and construct our entropy metric giving more weight to the regions identified as highly informative. We find that this generalized version of entropy, which focuses more on highly informative regions, can outperform other previously proposed diversity measures. We also find that a combined measure of diversity that includes both the new generalized entropy and the skewness of the distribution of within-host pairwise distances further improves prediction, whereas predictors such as the Q10, and minimum or maximum of the pairwise distance distribution do not add sufficient predictive power.

MATERIALS AND METHODS

Sequence Data Sources

We analyzed HIV *gag* sequences from the Primary HIV-1 Subtype C Infection Study in Botswana, collected in the Tshedimoso study.[20–23] Subjects with acute and recent HIV infection were enrolled in a primary HIV-1 subtype C infection cohort in Botswana from April 2004 to April 2008. The primary study, the Tshedimoso study, was approved by the Institutional Review Boards in both Botswana and the USA.[20] The 42 subjects included 8 acutely infected (Fiebig stage II) and 34 recently infected (Fiebig stage IV or V) individuals.[21] Time of seroconversion (time zero) for acutely infected subjects was estimated as the midpoint between the last ELISA-negative and the first ELISA-positive test (within a week in most cases), and for recently infected subjects it was estimated by Fiebig staging. As described in our earlier paper,[24] the beginning of Fiebig stage III in HIV-1C infection coincides with the time of detectable seroconversion (time 0), the mean duration of Fiebig stage III is 3 days, for stage IV is 6 days, and for stage V is 70 days. Thus, the time from seroconversion until detection was assumed to average 6 days for subjects in stage IV (3 days of phase III and 3 days to the midpoint of phase IV), 44 days for subjects in stage V (9 days of phases III and IV and 35 days to the midpoint of phase V), and 79 days for stage V/VI (9 days of phases III and IV, and 70 days of phase V). Subjects identified within Fiebig stage VI were excluded from analyses. The cohort included 9 males and 33 females. The median age at enrollment was 27 years. All subjects were nationals of Botswana, and all infections were HIV-1 subtype C.

Subjects were followed longitudinally through at most 500 days post-seroconversion. The median follow-up period was 378 days post-seroconversion (p/s). At each time point, sequences from the core structural gene *gag* were obtained using single genome amplification (SGA) followed by direct sequencing.[21] The analyzed region of *gag* corresponded to nucleotide positions 841 to 2217 of the reference strain HXB2 (amino acids 18 to 476 in relation to the *gag* coding DNA sequence in HXB2). After aligning the sequences, hypermutants were removed from the sample using Hypermut 2.0.[25] The total length of the alignment was 1518 nucleotides. Only samples with at least 5 sequences collected were considered for analysis. The codon-based sequence alignment was performed using muscle in the MEGA4 program. The default penalties for gap opening and extension were used.

Measuring Diversity

We previously developed a Segmented Shannon Entropy measure that takes into account the length of the genetic region.[19] Briefly, suppose N viral genetic sequences have been obtained from an HIV-infected individual at time t. Suppose also that these sequences are segmented into S regions with regions indexed 1,2... S. Then, for each segment $k \in\{1,2,..,\ S\}$, the Shannon entropy of that segment at time t is determined by

$$H_t^k = -\frac{1}{\log(N)}\sum_{a=1}^{n} P_{a,t}^k \log(P_{a,t}^k) \quad (1)$$

where n is the number of distinct segment patterns (ie, regions differing by at least 1 nucleotide base) within the N sequences, and $P_{a,t}^k$ is the proportion of sequence regions in the k region with distinct pattern a at time t.

Sliding Window Analysis

Change of within-host diversity over time may vary across different gene sequence regions, resulting in certain regions being more predictive of time-since-infection than others. Focusing on these informative regions should improve prediction of recency of infection. To test this hypothesis, we first conducted an initial screening for gene sequence regions where diversity measures were consistently higher in chronic infections compared to recent infections. To that end, we used the sliding window function slide analyses in the R package SPIDER version 1.05 (http://spider.r-forge.r-project.org/). This function partitions a sequence alignment into windows (or regions) of a chosen size and performs diversity measures on each window[26,27]. We explored window sizes of 50 bp, 100 bp, 150 bp, 200 bp, and 250 bp and calculated Segmented Shannon Entropy according to Eq. (1). To visualize the development of within-host diversity over time at different nucleotide regions, we plotted entropy measures for each patient on each genetic sequence window within time periods of 0 to 3 months, 3 to 6 months, 6 to 9 months, and beyond 9 months. Here we defined a recent case as a sample collected within 6 months from seroconversion.

Screening for Highly Informative Regions (HIR)

To evaluate the discriminatory abilities of sequence regions, we developed 3 methods that relied on nonparametric and parametric approaches. Our screening strategy was to look for viral sequence regions where viral diversity consistently increased over time across individual hosts.

In Method I, for each sliding window, we first compared the Segmented Shannon Entropy in recent versus chronic stages within each individual, assigned a score of 1 (if diversity is larger in the chronic stage), 0 (if diversity is the same in both stages), or −1 (if diversity is smaller in the chronic stage). We then summed the scores over all individuals and rank the sequence windows accordingly. Sequence windows were ranked according to their overall scores.

In Method II we used the concept of *information gain*, a well-known variable segmentation procedure.[28] This method helps us evaluate how well each region (segment) of the gene splits the sample of infected people into recent and chronic. To briefly describe this method, we focused on a region k of the *gag* gene, whose predictive power we wanted to assess. We computed the entropy of the 228 samples in that genetic region, and then partitioned these 228 entropy values into a predefined number of sets. We then proceeded to compute how much information we gained subsequent to this partition. This measures how predictive of recency status region k is. Similarly, we computed the information gain for all of the other regions, having then a metric to rank them, with the most informative regions having larger information gain.

Finally, in Method III, for each sequence window, we used all the 228 person-time samples and regressed entropy measures over time points using linear mixed models and ranked the sequence windows based on P values of the temporal trend. Windows of lower P values were ranked higher. See the appendix for details on all these approaches.

Generalized Entropy Measure as a Biomarker of Recency of Infection

For each of the approaches described above, sequence regions (ie, windows) within the top 50% rankings were considered as HIRs. Using Eq. (1) we computed the Segmented Shannon Entropy measures of the HIRs and averaged them to develop combined entropy measures as biomarkers of recent infection. We used a logistic regression scheme as the basis to our classification algorithm, evaluated the predictive power of these newly developed biomarkers, and compared them against previous biomarkers using AUC of the receiver operating characteristic (ROC) analysis. To account for repeated measures in the dataset we also conducted sensitivity analysis using mix-effect logistic regression. To further improve the predictive power of the classification scheme, we also explored other diversity-related biomarkers such as skewness of the distribution of pairwise Hamming distances.

RESULTS

Forty-two subjects were observed at a total of 228 time points with a median of 11 sequences per time point (range 5, 32). Among them 90 were defined as recent infections and 106 were chronic infections. The median time since infection was 80 days (interquartile range: 44, 122 days) for the recent cases and 323 days (interquartile range IQR: 240, 415) for the chronic cases. Antiretroviral therapy (ART) was initiated in 10 of the 42 subjects within the observed period of time due to a drop in CD4 T cells. The median time of ART initiation was 316 days p/s (IQR 186−415 days p/s; range 112−491 days p/s).

Both proviral DNA and viral RNA were included in the analysis. In the preliminary analysis we compared diversity (uncorrelated p-distances with pairwise deletion of gaps) between HIV-1C *gag* quasispecies amplified from viral RNA and proviral DNA in a subset of 18 subjects including 6 acutely infected individuals and 12 recently infected individuals. A total of 27 paired time points included 14 identical sampling points and 13 cases sampled within 60 days. The range of analyzed time points spanned from 4 to 755 days p/s (median of sampling time points for viral RNA was 201 days p/s with IQR 78 to 347 days p/s; median of sampling time points for proviral DNA was 198 days p/s with IQR 104 to 347 days p/s). The HIV-1C *gag* pairwise diversity predictably increased over time and remained relatively low (medians ranged from 0–1.1%). In 11 of 27 comparisons viral RNA diversity was higher than proviral DNA, in 8 cases proviral DNA was higher than viral RNA, and in 8 cases no difference between viral RNA and proviral DNA diversity was found (Wilcoxon rank-sum test).

Visualization of the development of within-host diversity over time shows that change of diversity over time varied greatly across different gene sequence regions, with certain regions being more predictive of time since infection than others. As an illustrative example Figure 1 shows the entropy profiles of 2 patients at time periods 0 to 3 months, 3 to 6 months, 6 to 9 months, and beyond 9 months, using regions of size 50 bp. In both instances, sequence diversity between the fifth and tenth regions was more predictive of time since infection than other regions. These results suggest that indeed some genetic regions were more informative for recency prediction.

We identified highly informative regions through the 3 methods described above. Figure 2 shows the overall scores for each sequence region using the first nonparametric procedure, with window sizes of 50 bp and 100 bp. In both cases HIRs were in more variable regions (ie, regions of higher mutation rates), such as p17 and p2/p7/p1/p6, and less likely in the more conserved p24 coding region.

Park et al proposed a biomarker for recency prediction based on the Q10 of the pairwise Hamming distance distribution, which appeared to be robust to both viral subtype and multiplicity of infection.[13] We had also previously developed another biomarker based on segmented entropy (SE) measure.[19] We selected these 2 biomarkers as benchmarks for evaluating the generalized entropy measures proposed in this work.

Table 1 reports these comparisons using AUC of the ROC analysis subsequent to a logistic regression. We selected the HIRs based on the 3 algorithms described above and in the appendix. We also performed a sensitivity analysis for several window sizes. Noteworthy, we do not expect for the HIRs to be identical using different window sizes, but rather that the HIRs were consistently located around the same gene regions regardless of window sizes.

To minimize the bias caused by correlated data, we created another data set where only the first- and last-observations of the 42 individuals were used. The corresponding results are reported and compared in Table 2. We felt that this was a more appropriate data set for a proper comparison between the AUC of the different methods. In both cases, our newly developed biomarkers based on highly informative regions outperformed previously developed biomarkers, especially when sequences were segmented into regions of 150 bp and 200 bp.

Figure 3 shows the AUC plots of the best performance of the newly developed biomarkers compared to the best performance of previously developed biomarkers, using first and last observations only. Including skewness of the pairwise

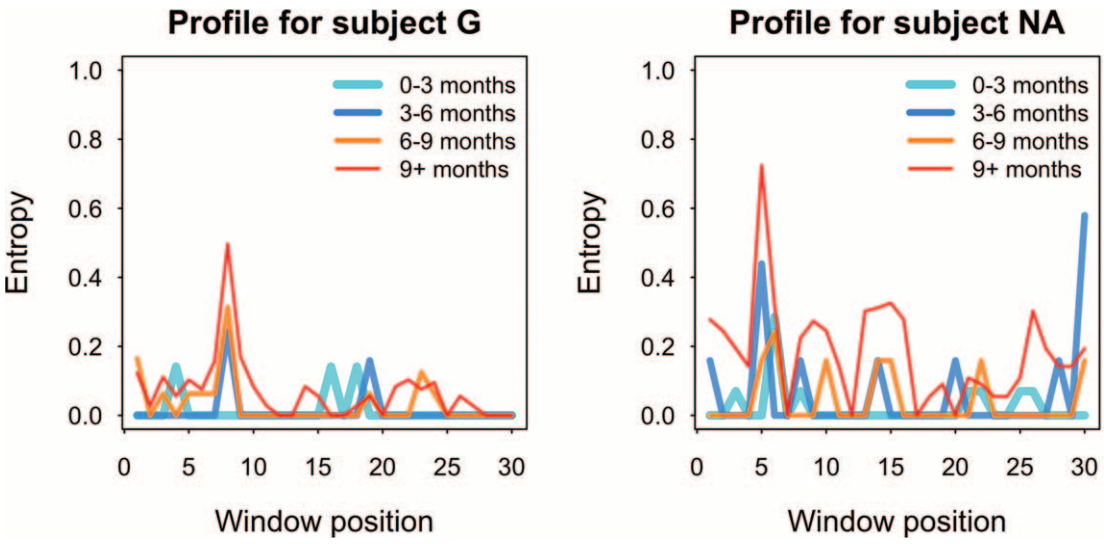

FIGURE 1. Entropy profiles of 2 patients at time periods 0 to 3 months, 3 to 6 months, 6 to 9 months, and beyond 9 months. Genetic region: HIV1-C gag gene; length: 1518 bp; window size: 50 bp. Note, for example, that in both cases, around the 5th to 10th windows the entropy increases consistently with time, whereas for regions around the 20th window, entropy does not show a systematic temporal increase.

Hamming distance distribution in the newly developed prediction model further improved the AUC up to 89%. Our Method III algorithm was significantly better than Q10 ($P = 0.01$) and SE ($P = 0.02$) based on the Delong test. Our Method I approach was significantly better than Q10 ($P = 0.05$) but the improvement over SE did not reach statistical significance ($P = 0.11$).

Noteworthy, we show the results when selecting the top 50% of the HIR. We conducted sensitivity analyses of this percentage using 2 additional cutoffs: 75% and 33%. The analysis indicates that the 50% cutoff rendered slightly better prediction power, although predictive performance was not greatly affected by the selection of cutoff within such range of values.

To better understand the potential impact of ART use on the performance of our viral diversity measure we conducted sensitivity analyses excluding the samples from ART-exposed subjects. We found that the recency prediction did not change appreciably. The same HIRs were identified, and the resulting AUCs were almost identical compared to the AUCs obtained from the full set analysis (See Figure S4 in the appendix).

DISCUSSIONS

Good estimates of HIV-1 incidence are essential for monitoring HIV transmission dynamics, designing and ascertaining the effectiveness of containment and prevention interventions, as well as informing resource allocation. Critical to this enterprise is the development of novel assays that can accurately identify recent HIV-1 cases. To further improve the predictive accuracy of existing viral-diversity-based biomarkers, we propose an approach based on differential predictability of regions across viral genetic sequences. To that end, we (1) used sliding window analysis to screen for the informative genetic region, (2) identified highly informative regions using nonparametric and parametric approaches, (3) averaged the segmented entropy measures of these highly informative regions to generate the generalized entropy biomarkers, and (4) compared the prediction power of our new biomarkers with 2 previously developed biomarkers. Our generalized entropy measure outperformed these 2 benchmarks, demonstrating the potential for improving accuracy to identify recent HIV-1 infections.

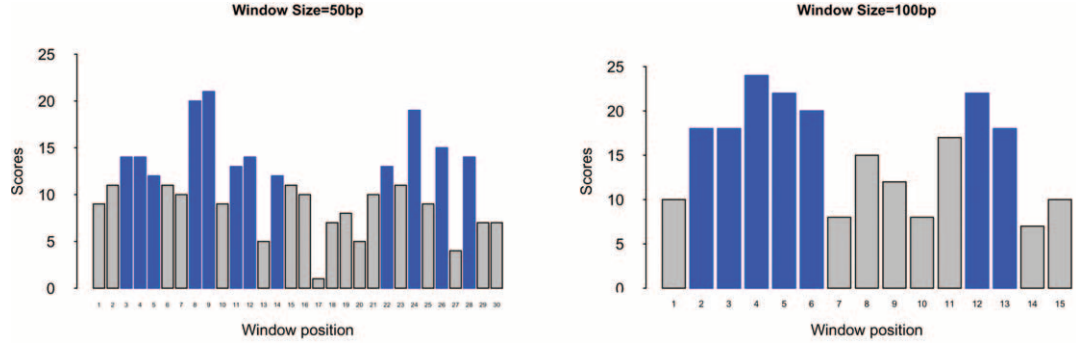

FIGURE 2. Screening for highly informative regions (HIR) using Method I. Performance scores were summed over all individuals. Genetic region: HIV1-C gag gene; length: 1518 bp; window sizes: 50 bp and 100 bp. Sequence windows' overall scores for the 2 window sizes. The blue bars are those with the highest scores, that is, the most informative regions. HIR = highly informative region.

TABLE 1. Comparing the ROC AUC Values of Classification Methods With All 228 Observations. The Best Performance of Each Method is Highlighted in Bold Type

Method/Window Sizes	50 bp	100 bp	150 bp	200 bp	250 bp
Q10 [13]	**68%**	**68%**	**68%**	**68%**	**68%**
SE [19]	66%	66%	68%	68%	**68%**
Entropy HIR Method I	69%	69%	**71%**	70%	69%
Entropy HIR Method II	66%	67%	69%	69%	**72%**
Entropy HIR Method III	68%	69%	71%	**72%**	72%

AUC = area under the curve, HIR = highly informative region, Q10 = the tenth quantile (of the pairwise Hamming genetic distance), ROC = receiver operating characteristic, SE = segmented entropy.

TABLE 2. Comparing the ROC AUC Values of Classification Methods With First and Last Observations Only. The Best Performance of Each Method is Highlighted in Bold Type

Method/Window Sizes	50 bp	100 bp	150 bp	200 bp	250 bp
Q10 [13]	**75%**	**75%**	**75%**	**75%**	**75%**
SE [19]	76%	77%	79%	80%	**80%**
Entropy HIR Method I*	80%	82%	**84%**	81%	80%
Entropy HIR Method II*	76%	78%	82%	82%	**85%**
Entropy HIR Method III*	80%	80%	85%	**86%**	84%

AUC = area under the curve, HIR = highly informative region, Q10 = the tenth quantile (of the pairwise Hamming genetic distance), ROC = receiver operating characteristic, SE = segmented entropy.
* *P* values <0.05 when compared to Q10 by the Delong test.

We show that the patterns of genetic diversification of certain sequence regions have higher predictive capacity for recent infections and that, consequently, focusing on these highly informative regions can improve predictive accuracy. Moreover, we demonstrate in several ways how to screen for highly informative genetic regions, and our procedures can be extended to other regions of the viral genome, with the potential for gaining additional information for prediction purposes. In addition, we show that our generalized entropy measure based on highly informative regions can be applied in combination with other predictive biomarkers, such as skewness measure of the pairwise Hamming distance distribution, to further improve the discriminatory ability. Moreover, this approach can be used as part of a multiple-assay algorithm and in combination with other biomarkers such as viral load. As a next step we would like to compare our generalized entropy approach to serologic biomarkers.

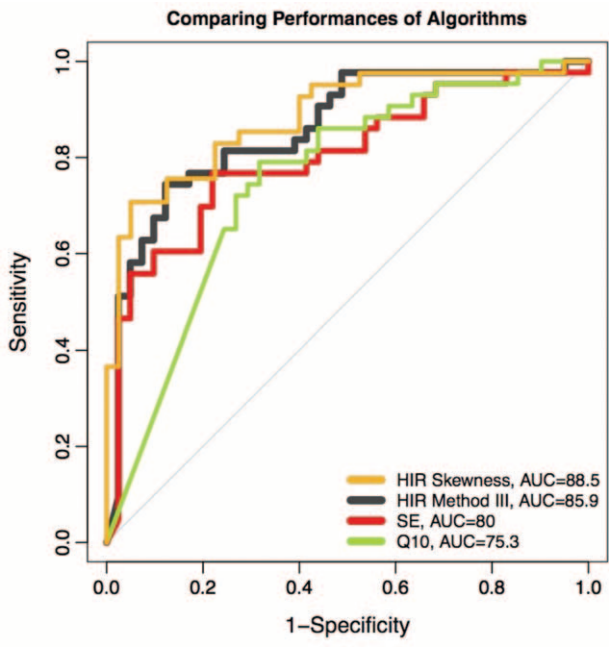

FIGURE 3. Comparing AUC plots of the best performance of the different biomarkers with first and last observations only. The newly developed biomarkers (HIR Skewness & HIR Method III) outperform existing ones (Q10 [13] and SE [19]). AUC = area under the curve, HIR = highly informative regions, Q10 = the tenth quantile (of the pairwise Hamming genetic distance), SE = segmented entropy.

We used post-seroconversion of 180 days as the cut-off for defining recent infection, but the same methods can be easily extended to different time cut-offs. The accuracy of these models might vary depending on the richness of within-host temporal signals in comparison to the ''noise'' (ie, the between-host variation) in the training data set and need further validations.

The choice of the HIV-1C *gag* gene was motivated by our previous work on the complex dynamics of selective pressure that affect viral mutations in *gag*. HIV-1 *gag* is a structural viral protein able to induce potent virus-specific T cell responses associated with control of viral replication, lower viral set point, and more favorable disease prognosis.[29–38] We have analyzed *gag* diversity and evolution in the primary[21,24,39,40] and chronic[32,33,41–43] HIV-1C infection. We addressed intra-host evolutionary rates in HIV-1C gag in primary infection[40] and demonstrated that during primary infection, the median intra-patient substitution rates within gag were 5.22E-03 (IQR 3.28E-03–7.55E-03) substitutions per site per year of infection. Viral sequences encoding partial *gag* (HXB2 nt positions 832–2217; HXB2 Gag amino acid positions 15–476) were generated in our study of primary HIV-1C infection,[20–24,40,44] and used in this study. Previously we examined the time of appearance, dominance, completeness, and loss of different types of viral mutations in *gag* soon after seroconversion,[21] timing of *gag* mutations[21] including dynamics of viral mutations at *gag* residue 242,[39] and intrahost evolutionary rates in HIV-1C *gag*.[40] In future work, we plan to extend this methodology to other genetic regions.

The results of our research provide insights into the relationship between within-host diversity and time since infection. Longitudinal quasispecies sequence data that provide valuable information on within-host viral evolution under no antiretroviral pressure, such as the data we use here, are scarce. These data were collected as part of an HIV primary infection study in Botswana.[20–23] Patients were recruited through a referral strategy of expanded Voluntary Counseling and Testing.[20] The richness of the data provided us with a rare opportunity to examine the evolution of within-host viral diversity since early infection. Prospective ''seroconverter'' cohort studies are prohibitively expensive to conduct on a large scale and have limited patient visit frequency. Previous work that aimed to describe HIV viral genetic diversity had to rely on either meta

sequence data sets downloaded from public sequence databases[45,46] or smaller follow-up studies with limited sampling frequency.[47,48] Yang[45] and Li et al[46] assessed the overall sequence variability across the viral genome within and between different HIV subtypes based on publically available sequences. Although the studies were not designed to specifically examine within-host diversity evolution and differentiability of regions in terms of recency, their observations suggesting that the level of overall genetic diversity varies greatly in different genetic regions is consistent with our findings. It is interesting to note that the less informative genetic regions we identify correspond to the more structurally conserved ''major homology region'' within gag, providing a potential biological explanation for our results. Certainly, there are other highly informative genetic regions within the whole HIV genome. Hence, applying our method to screen for additional HIRs has the potential for further improving HIV recency assays. In fact, Poon et al[49] report estimating time of infection based on phylogenetic tools and show differential predictive accuracy across different genes. Our work further illustrates that focusing on highly informative regions within any given gene has the potential to further improve prediction accuracy. Similar sequence variability patterns across HIV-1 subtypes have been observed,[45] potentially making our method generalizable to most of the circulating strains worldwide. Further studies along this line are warranted.

It is known that ART use in chronically infected persons reduces the individual's viral load levels, which might lead to false-recent classification when serological assays are used.[50] There have been concerns that ART might also reduce viral diversity, potentially resulting in false recency of some treated patients when viral diversity-based biomarkers are used. However, studies on both subtype B[51] and subtype C[52] have shown that HIV-1 population structure in ART-experienced individuals might be indistinguishable from pre-therapy samples, even following greater than 100-fold decreases in plasma HIV-1 RNA levels. To help us understand the potential impact of ART use on the performance of our viral diversity measure we conducted sensitivity analyses. In our sample, ART was initiated in 10 of 42 subjects within the observed period of time due to a drop in CD4 + T cells. We repeated the previous analyses but excluded the samples from ART-exposed subjects. We found that the recency prediction did not change significantly. The subset of individuals on ART did not have different entropy profile patterns compared to the treatment-naive individuals. Cousins et al[16] and Kouyos et al[14] also report a lack of association between ART use and viral diversity biomarkers. It is worth noting, however, that the sample size of ART-exposed person-times was rather small and the robustness of our method to ART use warrants further evaluation.

Different sequence alignment parameter settings can lead to different genetic segmentations and thus affect the determination of HIRs so we explored several alignment parameter sets. The 2 major parameters for nucleotide-based alignment that were explored were: penalty for gap opening (0–400) and penalty for gap extension (0–50). We find that in all the different alignment settings we explored, generalized-entropy-based classification algorithms outperform the benchmarks. Nonetheless, we recognize that sequence alignment parameter setting might depend on specific samples, and on the diversity of the targeted region in the HIV-1 genome. Particularly, when applied to the field, specific guidance in alignment procedures that are appropriate to the population of interest should be in place to ensure the proper usage of the assay.

We also find that evaluation of assay performances can be highly dependent on the sequence datasets used. We examined a public sequence dataset previously used in Xia et al.[15] This dataset, namely D561, represents a meta database (freely available at Los Alamos HIV public database) containing viral sequences from the *env* gene of 462 subjects (561 samples) infected with subtype B and C. Our new diversity-based biomarkers were similarly compared against the same benchmarks. The new biomarkers outperform or match the previous ones (data not shown). However it is important to note that when assessing assay performance with the public dataset (D561), all biomarkers achieved very high predictive accuracy, similar to what was reported in Xia et al.[15] Unlike the Botswana cohort data set, which is representative of potential targeted populations for cross-sectional HIV incidence estimation, the D561 is a convenience sample consisting of all available SGA sequences from the Los Alamos HIV database and it is unlikely to be representative of targeted populations of interest. Caution is needed when such data sets are used for validating assay performances. Further investigations on how the structure of data sets can impact assay's performance are being carried out and described elsewhere.

Among the limitations of our work is that, due to high between-host variation, viral-diversity-based biomarkers in general, including ours, might be unsuited for individual-level classification. Rather this type of biomarker is more suitable for population-level estimations, such as incidence estimation. Also, our method is still in the stage of proof of concept and collaborations are underway to further evaluate this approach based on larger sequence data. We are currently expanding our analysis to larger genomic regions, and formal scanning based on functional regions is underway. Additionally, single genome amplification and direct sequencing is expensive and can be impractical for initial screening. Although the screening methods we have developed and the identification process for HIRs rely on SGA sequence data, our concept can be applied to other types of genetic data. For example, our approach can have great potential with next generation sequencing data becoming more available and less expensive. Our procedures provide a tool for whole genome screening and guiding toward the optimal design of viral diversity assays.

In summary, our work shows, as a proof of concept, that focusing on highly informative viral genetic regions can improve predictive accuracy for identification of HIV recent infection. Further studies are needed to evaluate the performance of this approach across other viral genetic regions and as part of multiple-assay algorithms.

ACKNOWLEDGMENTS

We would like to thank Drs. Oliver Hofmann, Natalie Exner Dean, Hsiang-Han Chang, and George R. Seage III for helpful discussions. We thank the study participants.

REFERENCES

1. Brookmeyer R. Reconstruction and future trends of the AIDS epidemic in the United States. *Science.* 1991;253:37–42.
2. Hall HI, Song R, Rhodes P, et al. Estimation of HIV incidence in the United States. *JAMA.* 2008;300:520–529.
3. Brookmeyer R, Laeyendecker O, Donnell D, et al. Cross-sectional HIV incidence estimation in HIV prevention research. *J Acquir Immune Defic Syndr.* 2013;63(Suppl 2):S233–239.

4. 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;73:10489–10502.
5. Kearney M, Maldarelli F, Shao W, et al. Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J Virol.* 2009;83:2715–2727.
6. Keele BF. Identifying and characterizing recently transmitted viruses. *Curr Opin HIV AIDS.* 2010;5:327–334.
7. Tebit DM, Nankya I, Arts EJ, et al. HIV diversity, recombination and disease progression: how does fitness ''fit'' into the puzzle? *AIDS Rev.* 2007;9:75–87.
8. 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;105:7552–7557.
9. Delwart EL, Pan H, Sheppard HW, et al. Slower evolution of human immunodeficiency virus type 1 quasispecies during progression to AIDS. *J Virol.* 1997;71:7498–7508.
10. Learn GH, Muthui D, Brodie SJ, et al. Virus population homogenization following acute human immunodeficiency virus type 1 infection. *J Virol.* 2002;76:11953–11959.
11. 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.
12. Cousins MM, Konikoff J, Laeyendecker O, et al. HIV diversity as a biomarker for HIV incidence estimation: including a high-resolution melting diversity assay in a multiassay algorithm. *J Clin Microbiol.* 2014;52:115–121.
13. Park SY, Love TM, Nelson J, et al. Designing a genome-based HIV incidence assay with high sensitivity and specificity. *AIDS.* 2011;25:F13–19.
14. Kouyos RD, von Wyl V, Yerly S, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clin Infect Dis.* 2011;52:532–539.
15. Xia XY, Ge M, Hsi JH, et al. High-accuracy identification of incident HIV-1 infections using a sequence clustering based diversity measure. *PLoS One.* 2014;9:e100081.
16. Cousins MM, Laeyendecker O, Beauchamp G, et al. Use of a high resolution melting (HRM) assay to compare gag, pol, and env diversity in adults with different stages of HIV infection. *PLoS One.* 2011;6:e27211.
17. Cousins MM, Ou SS, Wawer MJ, et al. Comparison of a high-resolution melting assay to next-generation sequencing for analysis of HIV diversity. *J Clin Microbiol.* 2012;50:3054–3059.
18. Cousins MM, Swan D, Magaret CA, et al. Analysis of HIV using a high resolution melting (HRM) diversity assay: automation of HRM data analysis enhances the utility of the assay for analysis of HIV incidence. *PLoS One.* 2012;7:e51359.
19. Exner N, Surveillance methods for monitoring HIV incidence and drug resistance. Doctoral Dissertation, Harvard, 2014.
20. Novitsky V, Woldegabriel E, Wester C, et al. Identification of primary HIV-1C infection in Botswana. *AIDS Care.* 2008;20:806–811.
21. Novitsky V, Wang R, Margolin L, et al. Timing constraints of in vivo gag mutations during primary HIV-1 subtype C infection. *PLoS One.* 2009;4:e7727.
22. Novitsky V, Wang R, Kebaabetswe L, et al. Better control of early viral replication is associated with slower rate of elicited antiviral antibodies in the detuned enzyme immunoassay during primary HIV-1C infection. *J Acquir Immune Defic Syndr.* 2009;52:265–272.
23. Novitsky V, Woldegabriel E, Kebaabetswe L, et al. Viral load and CD4+T-cell dynamics in primary HIV-1 subtype C infection. *J Acquir Immune Defic Syndr.* 2009;50:65–76.
24. Novitsky V, Wang R, Baca J, et al. Evolutionary gamut of in vivo Gag substitutions during early HIV-1 subtype C infection. *Virology.* 2011;421:119–128.
25. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G –> A hypermutation. *Bioinformatics.* 2000;16:400–401.
26. Boyer S, Brown SD, Collins RA, et al. Sliding window analyses for optimal selection of mini-barcodes, and application to 454-pyrosequencing for specimen identification from degraded DNA. *PLoS One.* 2012;7:e38215.
27. Brown SD, Collins RA, Boyer S, et al. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol Ecol Resour.* 2012;12:562–565.
28. Provost F, Fawcett T. Data Science for Business. Sebastopol, CA: O'Reilly; 2013:http://proquest.safaribooksonline.com/9781449374273.
29. Edwards BH, Bansal A, Sabbaj S, et al. Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. *J Virol.* 2002;76:2298–2305.
30. Kiepiela P, Ngumbela K, Thobakgale C, et al. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat Med.* 2007;13:46–53.
31. Zuniga R, Lucchetti A, Galvan P, et al. Relative dominance of Gag p24-specific cytotoxic T lymphocytes is associated with human immunodeficiency virus control. *J Virol.* 2006;80:3122–3125.
32. Novitsky VA, Gilbert PB, Shea K, et al. Interactive association of proviral load and IFN-gamma-secreting T cell responses in HIV-1C infection. *Virology.* 2006;349:142–155.
33. Novitsky V, Gilbert P, Peter T, et al. Association between virus-specific T-cell responses and plasma viral load in human immunodeficiency virus type 1 subtype C infection. *J Virol.* 2003;77:882–890.
34. Serwanga J, Shafer LA, Pimego E, et al. Host HLA B*allele-associated multi-clade Gag T-cell recognition correlates with slow HIV-1 disease progression in antiretroviral therapy-naive Ugandans. *PLoS One.* 2009;4:e4188.
35. Geldmacher C, Currier JR, Herrmann E, et al. CD8 T-cell recognition of multiple epitopes within specific Gag regions is associated with maintenance of a low steady-state viremia in human immunodeficiency virus type 1-seropositive patients. *J Virol.* 2007;81:2440–2448.
36. Boaz MJ, Waters A, Murad S, et al. Presence of HIV-1 Gag-specific IFN-gamma+IL-2+ and CD28+IL-2+ CD4 T cell responses is associated with nonprogression in HIV-1 infection. *J Immunol.* 2002;169:6376–6385.
37. Eyeson J, King D, Boaz MJ, et al. Evidence for Gag p24-specific CD4 T cells with reduced susceptibility to R5 HIV-1 infection in a UK cohort of HIV-exposed-seronegative subjects. *AIDS.* 2003;17:2299–2311.
38. Ndongala ML, Peretz Y, Boulet S, et al. HIV Gag p24 specific responses secreting IFN-gamma and/or IL-2 in treatment-naive individuals in acute infection early disease (AIED) are associated with low viral load. *Clin Immunol.* 2009;131:277–287.
39. Novitsky V, Wang R, Margolin L, et al. Dynamics and timing of in vivo mutations at Gag residue 242 during primary HIV-1 subtype C infection. *Virology.* 2010;403:37–46.
40. Novitsky V, Wang R, Rossenkhan R, et al. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. *Infect Genet Evol.* 2013;19:361–368.

41. Novitsky V, Cao H, Rybak N, et al. Magnitude and frequency of cytotoxic T-lymphocyte responses: identification of immunodominant regions of human immunodeficiency virus type 1 subtype C. *J Virol.* 2002;76:10155–10168.

42. Novitsky V, Rybak N, McLane MF, et al. Identification of human immunodeficiency virus type 1 subtype C Gag-, Tat-, Rev-, and Nef-specific elispot-based cytotoxic T-lymphocyte responses for AIDS vaccine design. *J Virol.* 2001;75:9210–9228.

43. Novitsky V, Wang R, Lagakos S, et al. HIV-1 subtype C phylodynamics in the global epidemic. *Viruses*. 2010;2: 33–54.

44. Novitsky V, Wang R, Margolin L, et al. Transmission of single and multiple viral variants in primary HIV-1 subtype C infection. *PLoS One.* 2011;6:e16714.

45. Yang OO. Candidate vaccine sequences to represent intra- and inter-clade HIV-1 variation. *PLoS One*. 2009;4:e7388.

46. Li G, Piampongsant S, Faria NR, et al. An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology.* 2015;12:18.

47. Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology.* 2013;10:49.

48. Henn MR, Boutwell CL, Charlebois P, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog*. 2012;8:e1002529.

49. Poon AF, McGovern RA, Mo T, et al. Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. *AIDS.* 2011;25:2019–2026.

50. Mastro TD, Kim AA, Hallett T, et al. Estimating HIV incidence in populations using tests for recent infection: issues, challenges and the way forward. *J HIV AIDS Surveill Epidemiol.* 2010;2:1–14.

51. Kearney MF, Spindler J, Shao W, et al. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog*. 2014;10:e1004010.

52. Rossenkhan R, Novitsky V, Sebunya TK, et al. Viral diversity and diversification of major non-structural genes vif, vpr, vpu, tat exon 1 and rev exon 1 during primary HIV-1 subtype C infection. *PLoS One.* 2012;7:e35491.