MAJOR ARTICLE

Genetic Diversity as a Marker for Timing Infection in HIV-Infected Patients: Evaluation of a 6-Month Window and Comparison With BED

Manon Ragonnet-Cronin,[1,2,a,b] Stéphane Aris-Brosou,[2,a] Isabelle Joanisse,[1] Harriet Merks,[1] Dominic Vallée,[1] Kyna Caminiti,[1] Michael Rekart,[3] Mel Krajden,[3] Darrel Cook,[3] John Kim,[1] Laurie Malloch,[1] Paul Sandstrom,[1] and James Brooks[1,4]

[1]National HIV and Retrovirology Laboratories, Public Health Agency of Canada, and [4]Department of Medicine and [2]Department of Biology, University of Ottawa, Ottawa, and [3]BC Centre for Disease Control, Vancouver, Canada

Background. It has been reported that the increase in human immunodeficiency virus (HIV) sequence diversity in drug resistance surveillance specimens may be used to classify the duration of HIV infection as <1 or >1 year. We describe a mixed base classifier (MBC) optimized to categorize the duration of subtype B infections as <6 or >6 months on the basis of sequences for drug resistance surveillance specimens and compared MBC findings with those of serologic methods.

Methods. The behavior of the MBC was examined across a range of thresholds for calling mixed bases. MBC performance was then evaluated using either complete *pol* sequences or sites reflecting evolutionary pressures (HLA selection sites, sites that increased in entropy over the course of infection, and codon positions).

Results. The MBC performance was optimal when secondary peaks on the sequencing chromatogram accounted for at least 15% of the area of primary peaks. A cutoff of <0.45% mixed bases in the *pol* region best identified recent infections (sensitivity = 82.7%, specificity = 78.8%), with improvement achieved by analyzing only sites that increased in entropy.

Conclusions. In an extended data set of 1354 specimens classified by BED, the optimized MBC performed significantly better than a simple MBC (agreement, 68.98% vs 67.13%). If further validated, the MBC may prove beneficial for detecting recent infection and estimating the incidence of HIV infection.

The early symptoms of human immunodeficiency virus (HIV) infection may not prompt patients to seek medical attention. Consequently, most HIV infections are diagnosed among people who are chronically infected [1], with the date of seroconversion remaining unknown. Nevertheless, correct identification of recent HIV infections (RHIs), defined as infections <6 months old, is critical to public health for several reasons. First, estimates of RHIs in a specified population over a defined period are needed to assess the incidence of HIV infection. This information is necessary to determine trends in the HIV epidemic and to evaluate the success of prevention strategies. Second, contact tracing of RHI patients is critical because it is believed that high rates of onward HIV transmission are associated with elevated viral loads during early infection [2]. Third, because viral strains in RHI patients are closely related to the virus population found in the transmitting partner, identification of RHI offers unique opportunities to improve our understanding of the biology of HIV transmission [3]. However, current methods used to tease apart recent from established infections have severe shortcomings [4].

Received 22 November 2011; accepted 27 February 2012.

Presented in part: 18th Conference on Retroviruses and Opportunistic Infections, Boston, Massachusetts, February 2011, Abstract 1059; 5th Public Health Agency of Canada Science and Research Forum, Gatineau, Canada, March 2011.

[a]M.R.-C. and S.A.-B. contributed equally to the study.

[b]Present affiliation: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom.

Correspondence: James Brooks, MD, National HIV and Retrovirology Laboratories, Public Health Agency of Canada, Ottawa K1A 0K9, Canada (james.brooks@phac-aspc.gc.ca).

The Journal of Infectious Diseases 2012;206:756–64

Published by Oxford University Press on behalf of the Infectious Diseases Society of America 2012.

DOI: 10.1093/infdis/jis411

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

Incident infections may be identified during the course of routine diagnostic testing. At the simplest level, discordant results of serologic tests (ie, a negative result followed by a positive result) over a brief and defined period accurately identifies RHI. Alternatively, during the first 2–4 weeks of infection and before seroconversion, the detection of viral capsid (p24) or HIV RNA/DNA in a serology-negative specimen is evidence of RHI [5, 6]. However, these methods will only identify incident infections in the minority of patients who are either frequently tested or who present for care during acute infection.

Beyond the 2–4-week window, HIV-specific immunoglobulin G (IgG) appears as a result of the host adaptive immune response. This appearance is the basis of serologic assays used to identify RHI. The "detuned" approach consists in retesting HIV-positive samples with a less sensitive diagnostic test. Samples negative with the less sensitive assay are classified as RHI [7]. The Calypte BED-CEIA is a capture enzyme immunoassay that estimates the time since seroconversion by measuring the level of HIV-specific IgG relative to the total level of IgG [8]. Alternatively, the avidity of HIV antibodies, a proxy for the maturity of the antibody response, can be used to identify RHI [9]. In Canada, national HIV surveillance relies on the BED assay, which distinguishes infections with a duration of <155 days from those with a duration of >155 days. However, the assay has a number of limitations. First, HIV-specific antibody levels can decrease during infection, either as patients progress to AIDS or when viral load is suppressed [10]. As a result of this decrease, BED misclassifies such infections as recent. Second, the recommended cutoff varies between subtypes, from 155 days for subtype B to 360 days for subtype C [4, 11]. Thus, the application of BED to a population with a multitude of viral subtypes can be challenging.

An alternative approach to determining the stage of infection has been to measure the genetic diversity of HIV within an individual [12]. Previous work showed that sexually transmitted HIV infection is established by only a limited number of viruses [13, 14], with viral genetic diversity increasing subsequently [15, 16].The proportion of "mixtures" during population-based sequencing, which reflect viral population polymorphisms, may be used as a proxy for within-patient genetic diversity. In their pioneering work, Kouyos et al determined that a mixed-base cutoff of $\gamma_{360} = 0.5\%$ achieved a sensitivity of 86.8% and a specificity of 70% in predicting infections with a duration of <1 year [12]. Here, given the existing body of work associated with BED use, we examined whether the classifier would provide concordant classifications at the 155-day BED cutoff. In addition, because the number of mixed bases in a nucleotide sequence implicitly depends on the threshold for calling mixed bases, we tested different thresholds with the classifier. Finally, because selective pressure on the *pol* gene is not evenly distributed, we evaluated whether mixed bases in subsections of *pol* provided greater resolution of recent versus established infections.

METHODS

Study Populations

The Canadian HIV Strain and Drug Resistance Surveillance Program (SDR) tracks subtype and transmitted drug resistance among patients in Canada who have newly diagnosed HIV infection and are antiretroviral (ART) naive [17]. Remnant HIV diagnostic specimens, along with basic epidemiological data, including exposure category, are sent to the National HIV and Retrovirology Laboratories (NHRL) in Ottawa, Canada, for genotyping. Before specimens are genotyped, they are carefully pedigreed to ensure they originate from the first diagnosis of HIV in that patient. Here, 1450 samples received from western Canada between 2002 and 2008 were analyzed. The SDR is approved by the Health Canada Research Ethics Board. Additional external sequences originating from patients with RHI were provided by the Jewish General Hospital in Montreal, Canada, the BC Centre for Excellence in Vancouver, Canada, and the National Cancer Institute in Bethesda, Maryland. Sequences from patients with chronic infection were obtained from published work [18]. All sequences originated from ART-naive patients.

Stage of Infection

Three data sets were assembled (Supplementary Figure 1), with 2 containing sequences associated with a well-known infection duration and 1 containing sequences associated with an unknown infection duration. Specimens were classified as recent if viral p24 antigen was detected in the sample in the absence of antibodies or if the patient tested negative for HIV within the 155 days prior to diagnosis. Specimens were classified as established if they were collected at least 155 days after the initial HIV diagnosis.

The NHRL training data set consisted of 96 specimens, of which 66 were from recent infections and 30 were from established infections. An additional 237 sequences (96 recent and 141 established) from other laboratories or the literature [18] were added to the NHRL training data set to form the full training data set. In total, the full training data set contained 333 sequences corresponding to 162 recent and 171 established infections.

The NHRL sequences for which the duration of infection was unknown were classified according to results of BED (performed according to manufacturer instructions) as having a duration of <155 days or >155 days [8].This third independent data set of 1354 specimens (440 recent and 914 established) is denoted hereafter as the BED data set. Other than the 171 sequences from established infections in the training data sets, all samples originated from patients with newly diagnosed infection.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

Amplification and Sequencing

Viral RNA was extracted from 0.5-mL serum samples, and *pol* was reverse transcribed, amplified, and Sanger sequenced as previously described [19]. The lower limit of sensitivity was 1000 copies/mL. Contigs were assembled and aligned to the National Center for Biotechnology Information HIV type 1 reference genome (accession number: NC_001802), using BioEdit 7.0.9 [20]. The obtained *pol* sequences covered 1305 base pairs, including the protease gene (bases 1–297) and a portion of the reverse transcriptase gene (bases 1–1008). Sequences from samples collected before 2005 were already in GenBank [19] and all new sequences were deposited (accession numbers: HM468499-HM468608,HM468610-HM468617,HM468619-HM468740-HM468750,HM468752-HM468766,HM468768-HM468813,HM468815,HM468817-HM468821,HM468823, HM468825-HM468853,HM468855-HM468870,HM468872, HM468874,HM468875,HM468877-HM468906,HM468908-HM468914,HM468916-HM468936,HM468938-HM468959, HM468964-HM468975,HM468978-HM468982,HM468984-HM468992,HM468994-HM468999,HM469001,HM469003-HM469045,HM469047-HM469064,HM469066,HM469068-HM469077,HM469080,HM469082-HM469088,HM469090-HM469103,HM469105-HM469108,HM469110-HM469144, HM469146-HM469179,HM469181-HM469198,HM469200-HM469213,HM469215-HM469229,HM469231-HM469240, HM469263,HM469271,HM469273,HM469276,HM469278, HM469282,HM469283,HM469285-HM469302,HM469304-HM469306,HM469309,HM469310,HM469312-HM469315, HM469319-HM469321,HM469325,HM469327-HM469329, HM469331-HM469335,HM469337-HM469342,HM469344-HM469353,HM469355,HM469356,HM469359-HM469361, HM469362,HM469364-HM469366,JQ674753-JQ675288. Subtype was determined using the REGA HIV-1 Subtyping Tool 2.0 [21]. Because subtype B predominates in the Canadian epidemic, we limited our analysis to this subtype.

Mixed Base Calls

Mixed bases were identified on sequencing chromatograms at nucleotide positions where a second trace, representing a different base, was present above a threshold τ percentage area of the dominant, primary peak. For example, at a threshold $\tau = 15\%$, the secondary peak must represent at least 15% of the primary peak for a mixed base to be called. We varied τ in SeqScape [22] from 5% to 45% for all samples sequenced in our laboratory, creating a new sequence alignment each time, and counted mixed bases with DAMBE [23].

Receiver Operator Characteristic (ROC) Analysis

ROC analysis was used to assess whether sequences could be classified as recent or established on the basis of numbers of mixed bases. We name this method the mixed base classifier (MBC). Under the hypothesis that numbers of mixed bases increase with duration of infection, we first determined with the NHRL training data set the threshold τ that maximized area under the ROC curve (AUC). We then used the full training data set to search for the mixed base cutoff γ_{155} that maximized sensitivity and specificity in distinguishing recent from long-term infections at 155 days.

Sequence Subsets

The following 11 subsets were created from the full training data set to contain only sites potentially informative for the MBC: (1) sites where numbers of mixed bases increased in established infections (True +); sites where entropy was increased at the (2) nucleotide level (ΔE + Nuc) or (3) amino acid level (ΔE + AA); sites where the entropy increase was significant at the (4) nucleotide level (ΔE + Nuc.Sig)or (5) amino acid level (ΔE + AA.Sig) (Figure 1); (6) sites that were associated with HLA (HLA +); (7) sites that were under positive selection (Pos); (8) sites that were HLA associated and

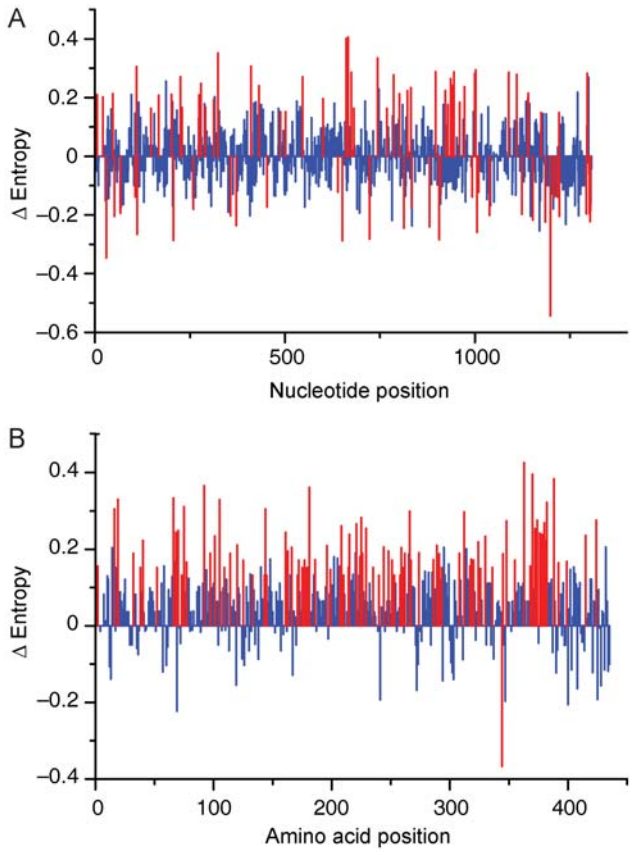

Figure 1. Site-specific differences in entropy (ΔEntropy). Site-specific ΔEntropy values were calculated between recent and established infections in the full training data set, both in nucleotide (*A*) and in amino acid (*B*) sequences. Site-specific ΔEntropy values are color-coded: red represents significant differences, and blue represents nonsignificant differences.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

Table 1. Alignment Sizes for the Categories of Sites Evaluated With the Mixed Base Classifier at the Amino Acid (AA) and Nucleotide (Nuc) Levels

Alignment name	Full Sequence	True +	ΔE + Nuc	ΔE + Nuc.Sig	ΔE + AA	ΔE + AA. Sig	HLA +	Pos	HLA + Pos
s_{AA}	435				311	104	78	26	85
s_{nuc}	1305	551	356	50	933	312	234	78	255

Four entropy-based measures are listed: sites that increased in entropy in nucleotide (ΔE + Nuc) and AA sequences (ΔE + AA), and sites that significantly increased in entropy in nucleotide (ΔE + Nuc.Sig) and AA sequences (ΔE + AA.Sig); as well as sites previously shown to be HLA associated (HLA +); sites inferred under positive selection (Pos); and a combination of both (HLA + Pos +).

under positive selection (HLA + pos); and (9–11) sites at codon positions 1–3. Table 1 lists the resulting alignment sizes (for positions, see Supplementary Data).

Shannon entropy measures the variability at each position. Site-specific differences in entropy between recent and established infections were calculated using the Shannon entropy tool (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html); significance was assessed by a randomization test (α = 0.05). HLA-associated sites were described previously [24]. Sites under positive selection were inferred with SLAC [25], under the GTR substitution model.

Mixed bases were counted in each subset. ROC analysis, which plots the true-positive rate of the new test (sensitivity) against its true negative rate (calculated as 1 – specificity), was used to identify the mixed base cutoff γ_{155} that best separates recent from established infections on the basis of the AUC: the best classifier has the highest AUC. This analysis was performed on each subset, and the best classifier was tested in the BED data set. As viral diversity may depend on transmission route [12], we further compared mixed base frequencies between patients infected through sexual exposure and those infected through blood exposure. All statistical analyses were performed using SPSS [26].

RESULTS

Mixed Base Thresholds τ of 15% and 25% Both Perform Well in the MBC

To test whether the MBC could classify infections as <155 days or >155 days old as accurately as it classifies them as <1 year or >1 year old [12], we first assessed the effect of varying the threshold τ for calling mixed bases. In the NHRL training data set, we counted mixed bases at mixture thresholds (τ) of 5%, 15%, 25%, 35%, and 45% of the dominant peak. The AUC for different thresholds τ ranged from 0.702 (95% confidence interval [CI], .597–.806) to 0.802 (95% CI, .706–.897). AUC was highest at thresholds τ = 15% and τ = 25% (Figure 2*A*), with τ = 15% yielding a sensitivity and specificity of 78.8% and 80%, respectively. At τ = 15%, only a small proportion of sequences (14.6%) remained misclassified.

Baseline Performance of MBC on the Full Training Data Set

In the full training data set, chromatograms were not available to vary the threshold τ, but consistent with our methods, τ was known to lie between 15% and 25%. Mixed base frequency distribution differed significantly between recent sequences and established sequences (on average, 0.22% and 1.28%, respectively; $P < .001$ by the Kolmogorov-Smirnov [KS] test). ROC analysis indicated that a cutoff γ_{155} = 0.45% provided the highest sensitivity and specificity (77.8% and 81.9%, respectively), such that sequences containing <0.45% mixed bases should be classified as recent. ROC analysis was repeated after excluding samples from the NHRL training data set, leaving 96 recent and 141 established samples. In this data set, γ_{155} = 0.46% provided the highest sensitivity and specificity (81.3% and 83%, respectively).

An Entropy-Based Approach Can Increase Sensitivity and Specificity of the MBC

Performance of the MBC was further increased in 4 of the subsets: True + , ΔE + AA, ΔE + AA.Sig, and ΔE + Nuc (Table 2). Of these, the ΔE + Nuc sites, which included only 27.28% of the sequenced nucleotide positions, most improved MBC performance, with sensitivity and specificity increasing to 85.2% and 83.5%, respectively, at a mixed base cutoff γ_{155} = 0.82%. The negative predictive value increased from 0.82 to 0.86, and the positive predictive value increased from 0.80 to 0.83 as compared to use of full sequences. The improvement from using only ΔE + Nuc sites appeared highly significant: when 356 sites were randomly sampled without replacement from the full training data set 10 times (among only *pol* variable sites), all resulting ROC curves had lower AUCs, with nonoverlapping CIs (Figure 2*B*).

HLA-Associated Sites and Sites Under Positive Selection Are Not Sufficient to Predict RHI

Because HIV diversifies under selective pressure early in infection [27, 28], we hypothesized that HLA + and Pos sites might increase the predictive power of the MBC. Neither of these categories alone improved the MBC as compared to the full sequence (Table 2), although Pos sites yielded a surprisingly high AUC of 0.802, despite covering only 6.4% of the

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

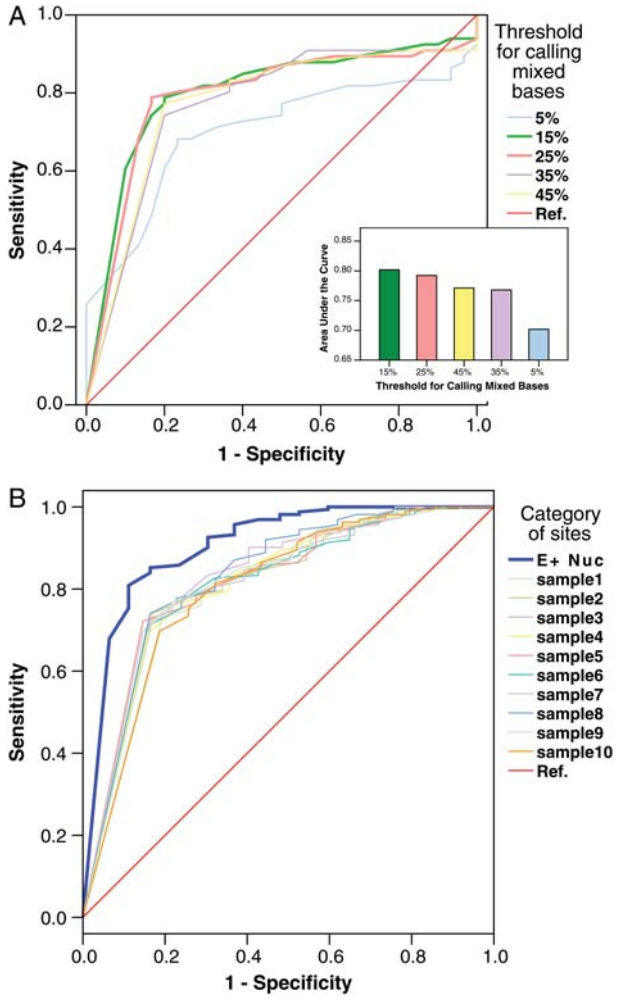

Figure 2. Receiver operating characteristic (ROC) curves for different alignments. *A*, The threshold τ for calling mixed bases was varied from 5 to 45% in SeqScape to generate a new alignment each time. For each threshold τ, the mixed base classifier (MBC) was evaluated through ROC analysis. Mixed base thresholds τ of 15% and 25% both perform well in the MBC. *B*, The performance of the 356 ΔE + Nuc sites was compared to the performance of 356 sites sampled randomly from the variable sites in pol. Confidence intervals for the area under the curve for ΔE + Nucsites and sampled sites were nonoverlapping.

alignment. Combination of HLA + and Pos sites increased performance beyond either category alone (AUC = 0.858) but not beyond use of the full sequences. Although both HLA + and Pos sites associated weakly with ΔE + Nuc ($P = .013$ and $P = .037$, respectively, by the Fisher exact test), these sites were insufficient for optimal MBC performance.

All 3 Codon Positions Are Informative for the MBC

Codon redundancy results in greater variability in the third codon position; however, restricting analysis to individual codon positions did not improve MBC performance (Figure 3). The best performance was achieved using third codon positions (AUC = 0.872), but there was little difference between codon positions, suggesting that mixed bases resulting from both synonymous and nonsynonymous mutations contribute information for the timing of infection classification using MBC.

Evaluation of the Concordance Between MBC and BED Results

Finally, the MBC was compared to BED in the BED data set. Mixed bases in the BED data set were called at a threshold $\tau = 15\%$. At this threshold, the average proportion of mixed bases was 0.852% per site for established infections and 0.346% for recent infections ($P < .001$ by the KS test). Using full sequences and $\gamma_{155} = 0.45\%$, MBC and BED agreed on the classification of 909 of 1354 samples (67.13%). Of 445 discordant samples, three-fourths were considered established by BED but recent by MBC. When the ΔE + Nuc sites determined in the full training data set at $\gamma_{155} = 0.82\%$ were used, agreement between BED and MBC increased slightly, to 68.98% (Figure 4). The highest proportion of discordant results remained for specimens classified as established by BED but as recent by MBC.

No Difference in Mixed Base Frequencies Between Exposure Categories

Of 1354 specimens in the BED data set, exposure could be categorized as sexual for 368 samples and as blood related for 803 samples. No difference in mixed base frequencies was observed between the 2 groups (0.701% vs 0.692%; $P > .1$ by the KS test).

DISCUSSION

We showed here that, in a precisely timed cohort, the stage of HIV infection can be inferred from the proportion of mixed bases identified during population-based sequencing of the *pol* region, consistent with previous findings [12, 29–31]. In addition, we improved the MBC and distinguished infections with a duration of <155 days from those with a duration of >155 days, as opposed to the 1-year threshold used by Kouyos et al [12]. This recalibration is crucial because both historical and current estimates of incidence in many Canadian [32] and international studies [33] use the 155-day limit to identify RHI. Because the risk of HIV transmission is elevated early in infection because of high viral loads [2], the ability to classify infections as <155 days old can offer greater value to understanding the dynamics of HIV transmission.

The universal application of MBC depends on calibrating sequencing instruments and base-callers across laboratories [12]. There is no clear standard for the threshold at which mixed bases are called, and both intralaboratory and interlaboratory variability is exacerbated by the manual review of sequences accepted as a component of the genotyping process.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

Table 2. Mixed Base Classifier Performance of Each Category of Sites in the Full Training Data Set

Alignment name	Full Sequence	True + [a]	ΔE + Nuc[a]	ΔE + Nuc. Sig	ΔE + AA.[a]	ΔE + AA. Sig[a]	HLA +	Pos	HLA + Pos +
Best cutoff γ_{155}(%)	0.45	0.90	0.82	1.00	0.50	0.54	0.71	0.64	1.10
AUC	0.878	0.899	0.906	0.843	0.890	0.899	0.849	0.802	0.858
Sensitivity	0.827	0.846	0.852	0.870	0.852	0.889	0.796	0.809	0.870
Specificity	0.788	0.747	0.835	0.759	0.776	0.782	0.788	0.747	0.712

Four entropy-based measures are listed: sites that increased in entropy in nucleotide (ΔE + Nuc) and amino acid sequences (ΔE + AA), and sites that significantly increased in entropy in nucleotide (ΔE + Nuc.Sig) and amino acid sequences (ΔE + AA.Sig); as well as sites previously shown to be HLA associated (HLA +); sites inferred under positive selection (Pos); and a combination of both (HLA + Pos +).

Abbreviation: AUC, area under the curve.

[a] Categories of sites that outperformed the full sequence.

Here we showed that mixed base thresholds of 15% and 25%, available in SeqScape [22] and Trugene, perform equally well in the MBC. Because commonly used thresholds lie somewhere between these values, most HIV *pol* sequences generated should be appropriate for use with MBC, even in the absence of sequencing chromatograms. Moreover, the performance of MBC was not decreased in the full training data set, where the threshold varied between 15% and 25%, as compared to the NHRL training data set, where τ was fixed for all sequences. Taken together, our results indicate that it may not matter whether the same threshold is used across all sequences analyzed, as long as it falls within this range.

We next examined whether certain sites could offer greater discriminatory power to distinguish recent from established infections. Our analysis revealed that, although the third codon position contained most of the information relevant to the MBC, the predictive power of the MBC was reduced when using individual codon positions. Similarly, although HLA-associated sites are known to diversify rapidly after transmission [27, 28], the exclusive use of these sites did not improve the MBC. It is possible that, although variability at HLA sites initially increases under adaptive immune pressure, continued selective pressure ultimately results in the fixation of better adapted variants [34], reducing numbers of mixed bases at HLA sites. Another possibility is that, because HLA adaptation occurs early in infection [35], mixed bases are already present at HLA sites in RHI, confounding the MBC. However, the performance of MBC improved significantly by focusing on the 25% of sites that increased in entropy (Figure 2*B*).

Our results add to a growing body of evidence that the MBC may be considered as a potential addition to the public health toolbox for the identification of RHI. A significant advantage of the MBC is that no additional laboratory work is required beyond the genotyping performed for clinical or

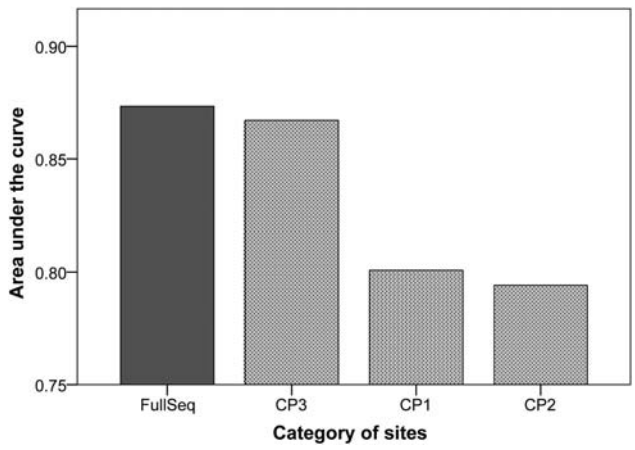

Figure 3. Mixed base classifier (MBC) area under the curve performance at individual codon positions. Gray indicates calculations based on the full training data set, and hatched gray indicates separate analysis of codon positions 1, 2, and 3 (CP1, CP2, and CP3, respectively).

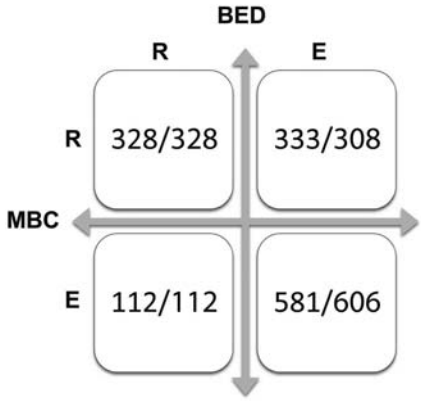

Figure 4. Performance of the mixed base classifier (MBC) in the BED data set. Of 1354 samples, BED testing classified 440 as recent (R) and 914 as established (E). By using full sequences and the γ_{155} = 0.45% cutoff, the MBC correctly classified 328 of the recent sequences and 581 of the established sequences. When the ΔE + Nuc sites determined in the full training data set and its γ_{155} = 0.82% cutoff was applied, the MBC continued to correctly classify 328 of the recent sequences, but the number of correctly classified established sequences increased to 606.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

surveillance purposes. A freely available MBC has been implemented in DAMBE [23], and the cutoff γ_{155} identified in this study can be directly applied to existing *pol* sequences generated for determining subtype and identifying drug resistant mutations. Application of MBC to existing longitudinal sequencing databases may create new potential models of HIV transmission patterns. Indeed, even if our analysis is entirely cross-sectional, we still found significantly higher frequencies of mixed bases among infections with a duration of >155 days ($P < .001$). One remaining point is the less-than-perfect concordance (70%) between MBC and the BED. The high frequency of mixed bases observed in sequences classified as recent by BED is consistent with the assay's tendency to overclassify infections as recent [36]. This result may also highlight the fact that high viral diversity is already present in some patients in primary HIV infection, possibly because of the rapid proliferation of viral variants prior to immune selection [37–39]. However, we found the greatest discordance among established infections that had fewer mixed bases. This latter finding may result from the influence of sequences from much-longer-term infections, in which viral diversity decreases. As estimates of BED accuracy vary widely [40, 41], it is possible that the lack of observed concordance on a per-specimen basis reflects the imperfect performance of BED in predicting RHI. Unfortunately, our early infections were serology negative, and determining concordance of the BED and MBC on these specimens is impossible because the BED requires positive serologic findings. This being said, the widespread reporting of BED results in the scientific literature means that reconciling BED results with MBC findings remains a critical, albeit challenging, endeavor. Contrary to Kouyos et al [12], we did not find that viral diversity depended on mode of HIV acquisition.

Another limitation is that the sequences used in this study represent a heterogeneous group of early or late infections. The late infection sequences were obtained from individuals infected anywhere from 156 days to >3 years before sampling. Although all patients in the BED data set had newly diagnosed infection, the stage of infection varied. It is unlikely that any sequences were derived from patients with AIDS because AIDS cases are so infrequent in Canada and mostly occur among patients in whom HIV infection was previously diagnosed [42]. Meanwhile, the early infection group consisted of specimens from patients identified as p24 positive before seroconversion, as well as from individuals classified as having RHI on the basis of discordant results of HIV tests performed up to 155 days apart. This inclusion of sequences from infections prior to seroconversion, when genetic diversity is known to be low, may influence results by favoring the hypothesized association.

We further highlight that the results from this analysis are only applicable to ART-naive sequences, such as those used in this study. Just as patterns of mutation differ in treated patients, suppressed viral loads will make it harder to amplify enough copies to obtain an accurate representation of viral diversity. While samples collected through the SDR exclusively originate from ART-naive patients with a recent diagnosis, equivalent programs in other countries, such as the United Kingdom, continue to collect samples from patients beyond this period.

To fully validate the MBC, it would be necessary to have access to data sets of precisely timed and longitudinal samples, but unfortunately very few such data sets exist to date. Before results can be extrapolated, further research should focus on evaluating the MBC on alternative data sets, such as those involving ART-treated subjects, elite controllers, and patients with AIDS, all of which have presented problems for antibody-based incidence assays [4].

In conclusion, we found highly significant differences in mixed base frequencies between samples associated with infection durations of <6 months or >6 months and have identified sites in *pol* at which mixed bases most correlate with the time since infection. We argue that the MBC is a significant addition to the tool kit for detecting RHI. In particular, it might be used in combination with other tools, such as the recent infection testing algorithm, which uses results from laboratory assays, as well as clinical information [43]. While future work should validate our list of entropy-identified, highly informative sites and should determine the functional significance of those that were not HLA associated, longitudinal analyses of patients with known HLA types might better determine whether mixed bases do indeed appear at HLA-associated sites. This step will likely be facilitated by the replacement of Sanger sequencing by next-generation sequencing, which is better suited to quantifying population genetic diversity.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Acknowledgments. We thank Dr Xuhua Xia, for implementing a mixed base counter in DAMBE; Drs Bluma Brenner, Frank Maldarelli, and Richard Harrigan, for kindly sharing sequences from recent infections; and Anna van Weringh, for constructive comments and support. We are also grateful to the National Laboratory for HIV Reference Services at the Public Health Agency, for BED testing, and to Mark Gilbert from the BC Centre for Disease Control, for sharing specimens.

Financial support. This work was supported by the Canada Memorial Foundation (to M.R.C.), the Natural Sciences Research Council of Canada (to S.A.B.), the Canada Foundation for Innovation (to S.A.B.), and the Federal Initiative on HIV/AIDS in Canada (to J.B.).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

References

1. Public Health Agency of Canada. HIV and AIDS in Canada. Surveillance Report to December 31, 2008. **2009**.
2. Pilcher CD, Joaki G, Hoffman IF, et al. Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection. AIDS **2007**; 21:1723–30.
3. Cohen MS, Gay CL, Busch MP, Hecht FM. The detection of acute HIV infection. J Infect Dis **2010**; 202(Suppl 2):S270–7.
4. Murphy G, Parry JV. Assays for the detection of recent infections with human immunodeficiency virus type 1. Euro Surveill **2008**; 13.
5. Fiscus SA, Pilcher CD, Miller WC, et al. Rapid, real-time detection of acute HIV infection in patients in Africa. J Infect Dis **2007**; 195:416–24.
6. Pilcher CD, Fiscus SA, Nguyen TQ, et al. Detection of acute infections during HIV testing in North Carolina. N Engl J Med **2005**; 352:1873–83.
7. 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.
8. Parekh BS, Kennedy MS, Dobbs T, et al. 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 **2002**; 18:295–307.
9. Suligoi B, Massi M, Galli C, et al. Identifying recent HIV infections using the avidity index and an automated enzyme immunoassay. J Acquir Immune Defic Syndr **2003**; 32:424–8.
10. Fisher M, Pao D, Murphy G, et al. Serological testing algorithm shows rising HIV incidence in a UK cohort of men who have sex with men: 10 years application. AIDS **2007**; 21:2309–14.
11. Parekh BS, Hanson DL, Hargrove J, et al. Determination of mean recency period for estimation of HIV type 1 incidence with the BED-capture EIA in persons infected with diverse subtypes. AIDS Res Hum Retroviruses **2011**; 27:265–73.
12. Kouyos RD, von WV, 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**; 52:532–9.
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.
14. Zhu T, Mo H, Wang N, et al. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science **1993**; 261:1179–81.
15. 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.
16. Troyer RM, Collins KR, Abraha A, et al. Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression. J Virol **2005**; 79:9006–18.
17. Jayaraman GC, Archibald CP, Lior L, Sutherland D. Integrating laboratory and epidemiological techniques for population-based surveillance of HIV strains and drug resistance in Canada. Can J Infect Dis **2000**; 11:74–80.
18. Brumme ZL, John M, Carlson JM, et al. HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. PLoS One **2009**; 4:e6687.
19. Ragonnet-Cronin M, Ofner-Agostini M, Merks H, et al. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. J Acquir Immune Defic Syndr **2010**; 55:102–8.
20. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser **1999**; 41:95–8.
21. de Oliveira T, Deforche K, Cassol S, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics **2005**; 21:3797–800.
22. SeqScape [computer program]. Life Technologies Corporation, Carlsbad, California. Version 2.7. **2009**.
23. Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. J Hered **2001**; 92:371–3.
24. Brumme ZL, John M, Carlson JM, et al. HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. PLoS One **2009**; 4(8).
25. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics **2005**; 21:676–9.
26. SPSS for Windows [computer program]. Version 18.0.0. Chicago: SPSS, **2009**.
27. Liu Y, McNevin JP, Holte S, McElrath MJ, Mullins JI. Dynamics of viral evolution and CTL responses in HIV-1 infection. PLoS One **2011**; 6:e15639.
28. Price DA, Goulder PJ, Klenerman P, et al. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc Natl Acad Sci USA **1997**; 94:1890–5.
29. Andersson E, Shao W, Bontell I, et al. Evaluation of a new subtype independent bioinformatics algorithm to detect recent HIV-1 infection [abstract 1056]. In: Program and Abstracts of the 18th Conference on Retroviruses and Opportunistic Infections (Boston). Alexandria, Virginia: CROI, **2011**:502.
30. Wilson E, Shao W, Brooks J, et al. New bioinformatic algorithm to identify recent HIV-1 infection [abstract 1057]. In: Program and abstracts of the 18th Conference on Retroviruses and Opportunistic Infections (Boston). Alexandria, Virginia: CROI, **2011**:503.
31. Ambrose J, Foster G, Chaytor S, Booth C, Geretti AM. Population sequence nucleotide ambiguity as a measure of HIV-1 infection length [abstract 1058]. In: Program and abstracts of the 18th Conference on Retroviruses and Opportunistic Infections (Boston). Alexandria, Virginia: CROI, **2011**:503.
32. Boulos D, Yan P, Schanzer D, Remis R, Archibald C. Estimates of HIV Prevalence and Incidence in Canada, 2005. 1 August **2006**.
33. Brown AE, Gifford RJ, Clewley JP, et al. Phylogenetic reconstruction of transmission events from individuals with acute HIV infection: toward more rigorous epidemiological definitions. J Infect Dis **2009**; 199:427–31.
34. Herbeck JT, Rolland M, Liu Y, et al. Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. J Virol **2011**.
35. Wood N, Bhattacharya T, Keele BF, et al. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog **2009**; 5:e1000414.
36. Niccolai LM, Verevochkin SV, Toussova OV, et al. Estimates of HIV incidence among drug users in St. Petersburg, Russia: continued growth of a rapidly expanding epidemic. Eur J Public Health **2010**.
37. Sagar M, Kirkegaard E, Long EM, et al. Human immunodeficiency virus type 1 (HIV-1) diversity at time of infection is not restricted to certain risk groups or specific HIV-1 subtypes. J Virol **2004**; 78:7279–83.
38. Haaland RE, Hawkins PA, Salazar-Gonzalez J, et al. Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. PLoS Pathog **2009**; 5:e1000274.
39. 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.
40. Le Vu S, Meyer L, Cazein F, et al. Performance of an immunoassay at detecting recent infection among reported HIV diagnoses. AIDS **2009**; 23:1773–9.
41. Truong HM, Kellogg T, Louie B, Klausner J, Dilley J, McFarland W. Recent HIV-1 infection detection: comparison of incidence estimates derived by laboratory assays and repeat testing data. J Acquir Immune Defic Syndr **2009**; 51:502–5.
42. Public Health Agency Canada. Canada's report on HIV/AIDS 2005. Surveillance and Risk Assessment Division, Centre for Infectious Disease Prevention and Control, **2005**.
43. Gill N, Delpech V, Smith R, et al. National public health monitoring of incident HIV-1 infections and primary drug resistance. UK: Health Protection Agency, **2008**.
44. UNAIDS. UNAIDS Reference Group on estimates, modelling and projections-statement on the use of the BED assay for the estimation of HIV-1 incidence for surveillance or epidemic monitoring. Wkly Epidemiol Rec **2006**; 81:40.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012

45. Hargrove JW, Humphrey JH, Mutasa K, et al. Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. AIDS **2008**; 22:511–8.
46. Brooks J, Woods C, Merks H, Wynhoven B, Hall TA, Sandstrom P. Evaluation of an automated sequence analysis tool to standardize HIV genotyping results. In: Program and abstracts of the 18th Annual Canadian Conference on HIV/AIDS Research (Vancouver). Ottawa, Canada: CAHR, **2009**.
47. Korber BT, Foley BT, Kuiken CL, Pillai SK, Sodroski JG. Numbering positions in HIV relative to HXB2CG. Report no. III. Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos Laboratory, **1998**.

Downloaded from http://jid.oxfordjournals.org/ at University of Ottawa on September 5, 2012