Developing High-Throughput HIV Incidence Assay with Pyrosequencing Platform

Sung Yong Park, Nolan Goeken, Hyo Jin Lee, Robert Bolan, Michael P. Dubé, Ha Youn Lee

Department of Molecular Microbiology and Immunology and Department of Medicine and Division of Infectious Diseases, Keck School of Medicine, University of Southern California, Los Angeles, California, USA; Research Health Services, Los Angeles Gay and Lesbian Center, Los Angeles, California, USA

ABSTRACT

Human immunodeficiency virus (HIV) incidence is an important measure for monitoring the epidemic and evaluating the efficacy of intervention and prevention trials. This study developed a high-throughput, single-measure incidence assay by implementing a pyrosequencing platform. We devised a signal-masking bioinformatics pipeline, which yielded a process error rate of $5.8 \times 10^{-4}$ per base. The pipeline was then applied to analyze 18,434 envelope gene segments (HXB2 7212 to 7601) obtained from 12 incident and 24 chronic patients who had documented HIV-negative and/or -positive tests. The pyrosequencing data were cross-checked by using the single-genome-amplification (SGA) method to independently obtain 302 sequences from 13 patients. Using two genomic biomarkers that probe for the presence of similar sequences, the pyrosequencing platform correctly classified all 12 incident subjects (100% sensitivity) and 23 of 24 chronic subjects (96% specificity). One misclassified subject’s chronic infection was correctly classified by conducting the same analysis with SGA data. The biomarkers were statistically associated across the two platforms, suggesting the assay’s reproducibility and robustness. Sampling simulations showed that the biomarkers were tolerant of sequencing errors and template resampling, two factors most likely to affect the accuracy of pyrosequencing results. We observed comparable biomarker scores between AIDS and non-AIDS chronic patients (multivariate analysis of variance [MANOVA], $P = 0.12$), indicating that the stage of HIV disease itself does not affect the classification scheme. The high-throughput genomic HIV incidence marks a significant step toward determining incidence from a single measure in cross-sectional surveys.

IMPORTANCE

Annual HIV incidence, the number of newly infected individuals within a year, is the key measure of monitoring the epidemic’s rise and decline. Developing reliable assays differentiating recent from chronic infections has been a long-standing quest in the HIV community. Over the past 15 years, these assays have traditionally measured various HIV-specific antibodies, but recent technological advancements have expanded the diversity of proposed accurate, user-friendly, and financially viable tools. Here we designed a high-throughput genomic HIV incidence assay based on the signature imprinted in the HIV gene sequence populations. By combining next-generation sequencing techniques with bioinformatics analysis, we demonstrated that genomic fingerprints are capable of distinguishing recently infected patients from chronically infected patients with high precision. Our high-throughput platform is expected to allow us to process many patients’ samples from a single experiment, permitting the assay to be cost-effective for routine surveillance.

HIV incidence, the number of newly infected individuals within a given amount of time, typically per year, is the key parameter of monitoring the epidemic’s rise and decline (1). It serves as a direct measure for evaluating the efficacy of HIV intervention and prevention trials and also as an objective reference for allocating HIV-related health care resources (2, 3). Therefore, accurately estimating HIV incidence is an immediate need in the community. To most efficiently address this need, such an assay should be capable of determining the stage of an infection from a single blood measure via cross-sectional sampling. However, since the first assay-based method for estimating HIV incidence was proposed in 1995 (4), numerous serologic approaches based on the characteristics of HIV-specific antibody response maturation have not been satisfactory (5–8). It has been reported that the serologic assays were overly dependent on the infecting virus subtype and displayed a substantial false-recency rate, which resulted in the overestimation of HIV incidence (2, 3, 7, 9). Overall, the sensitivity—the proportion of incident infections correctly classified as incident—was 89% and the specificity—the proportion of chronic infections correctly identified as chronic—was 87% across 13 serologic assays (9). Recently, a new limiting antigen avidity assay has demonstrated a significantly lower false-recency rate than the current standard, the BED assay (10), among individuals with AIDS (0.2% versus 2.9%) (11). The WHO has recommended that the limiting antigen avidity assays be used in environments where specimens can be further tested for HIV RNA level and the effects of antiretroviral therapy (ART) (12). Although running multiple assays in parallel or in sequence can increase precision (11, 13), the additional resources required can offset the gains in precision.
making a single-assay approach ideal for determining HIV incidence.

In pursuit of a single-assay approach, we recently proposed to turn to the other aspect of HIV infection, HIV sequence diversification over the course of infection, to develop a genomic incidence assay. We designed a genomic assay that showed high sensitivity and specificity, both over 95%, holding great promise for precise assessment of HIV incidence (14, 15). In this assay, the presence of closely related gene sequences within the HIV population of an infected individual was the signature by which incident infections were identified (14). The presence of a high proportion of similar strains indicates that a viral population has recently replicated and evolved from either a single or multiple transmitted viral variants. On the other hand, at the chronic stage of infection, viral strains within an individual mostly differ from each other as a consequence of accumulated mutations by HIV reverse transcriptase errors and immune selections. Our meta-analysis with 182 incident and 43 chronically infected subjects demonstrated that the biomarker of the 10% quantile (Q10) of the Hamming distance (HD) distribution showed not only high accuracy but also high robustness with respect to viral and host-specific factors (14). In the present study, we aimed to optimize the genomic assay for routine use, in particular, by focusing on designing a high-throughput platform.

The pyrosequencing method (16) has been utilized in HIV research to address drug resistance mutations (17), viral population diversity (18, 19), minority HIV variants (20), and immune escapes (21). A massive amount of reads can be produced from a single experiment. For example, over 700,000 reads can be obtained from a single run of a Roche GS Junior pyrosequencer. The depth of this sequencing method implies a capacity to process many patients’ samples in parallel. However, sequencing reads must be interpreted with caution because the technique is more error prone than the conventional, albeit more cost-prohibitive, Sanger sequencing technique (22). Another factor preventing pyrosequencing from recording an accurate census of intrahost viral population is sequencing resampling caused by PCR amplification prior to a sequencing run (23). To address these two issues, the present study devised a bioinformatics pipeline tolerant of pyrosequencing errors while using frequency-based biomarkers which reduce the effects of resampling. Our high-throughput platform for the genome-based assay provides a new opportunity for determining incidence from a single measure in cross-sectional surveys with high accuracy and low routine surveillance cost.

**MATERIALS AND METHODS**

Sources of blood specimens and published sequence data set. The 24 study subjects who were chronically infected were enrolled from the Rand Schrader Clinic at the University of Southern California. The study (HS-12-00121) was approved by the Institutional Review Board of the University of Southern California, and all study participants provided written informed consent at enrollment. Blood specimens were obtained from 24 patients who had a documented first positive HIV antibody test record in 12 subjects (enzymelinked immunosorbent assay [ELISA] and Western blot analysis), as listed in Table 1. Measurements of the CD4+ T cell count and HIV load (Cobas AmpliPrep/Cobas TaqMan HIV-1) were performed as part of clinical care at a mean of 36 days prior to specimen collection (Table 1). The gender and age of each chronic subject are listed in Table 1. The risk behavior of each individual was categorized as either heterosexual or men who have sex with men (MSM) as shown in Table 1. Of 24 study participants, 17 subjects had received antiretroviral therapy (ART) and 7 subjects were ART naive at the time of specimen collection (Table 1). Table 2 lists all antiviral drugs that had been used to treat each ART-experienced patient. The 12 HIV-positive patients who were at the incident stage within 1 year of the last documented negative HIV test were enrolled from the Los Angeles Gay and Lesbian Center (Table 3). The date of the last negative HIV test was documented for each individual based on a rapid-testing algorithm which was used to conduct serial tests in the following order: Orasure, Clearview, and nucleic acid amplification test (NAAT). The first HIV-positive date was documented if the result of one or more of these tests became positive. Viral load was measured by the Roche Amplicor HIV-1 Monitor Test, version 1.5. The CD4 count was measured by LabCorp test 505271, CPT code 86360. The risk group of all 12 incident patients was categorized as MSM. Most patients (10 of 12) had received ART by the time of sample collection (Table 2 and Table 3).

Additional HIV envelope sequences of 182 incident and 43 chronic patients from cohorts originating from the United States, Trinidad, South Africa, Malawi, and Canada were obtained from the published data in references 24, 25, and 26. The risk groups included 92 heterosexual transmissions, 16 MSM, 12 injection drug users, and 105 unknown. A total of 5,596 envelope segments (HXB2 7212 to 7601), previously obtained by single-genome amplification (SGA), were analyzed in this study.

**Plasma sample preparation.** The blood specimen collected from each study participant was centrifuged at 400 × g for 10 min, isolating the plasma layer. The isolated top plasma layer was transferred to a sterile 50-ml Falcon tube, which was then centrifuged at 1,200 × g for 10 min to remove any remaining cellular debris. After the second centrifugation, each 1 ml of plasma was divided into aliquots and transferred to 1.5-ml sterile tubes. The plasma aliquots were stored at −80°C.

**Amplipcr library preparation.** Viral pelleting with 2 ml of each subject’s plasma specimen was performed by first prespinning plasma in a centrifuge at 4°C and 5,300 × g for 10 min. The top layer was put into a 2-ml Sarstedt Microtube and was then centrifuged at 25,000 × g for 1 h at 4°C. To designate the pellet position, the tube was marked at the side facing away from the center of the centrifuge. After the supernatant was removed while ensuring that the marked region containing the viral pellet was not disturbed, the pellet was resuspended in 200 μl of 1× phosphate-buffered saline (PBS) buffer. Viral RNA was extracted from the plasma sample using a Purelink viral RNA extraction kit (Invitrogen) per the manufacturer’s protocol. In the final step, the concentrated HIV RNA was eluted with 50 μl sterile RNase-free water. Immediately after RNA extraction, reverse transcription-PCR (RT-PCR) of reverse transcription of viral RNA into cDNA was set up using a solution of 25 μl of viral RNA combined with 25 μl RT master mix (22 μl 2× buffer, 1 μl forward primer, 1 μl reverse primer, 1 μl Super Script III RT/Platinum Taq High Fidelity enzyme mix) using a Superscript III One Step RT-PCR kit (Invitrogen) with forward primer envB3out (5′-TAGACCGCTGGAAGCATCAGG AAG-3′) and reverse primer envB3out (5′-TTGCTACTTTGATTTGCT CCACTGT-3′). The PCR conditions were as follows: reverse transcription phase at 50°C for 30 min, 2-min hot start phase at 94°C, 25 cycles of DNA denaturing phase at 94°C for 30 s, primer annealing phase at 55°C for 20 s, and primer extension phase at 68°C for 5 min, with a final extension at 68°C for 3 min.

Subsequently, a nested PCR was conducted using 5 μl of RT-PCR product and 15 μl of master mix (11.5 μl H2O, 2 μl 10× buffer, 0.4 μl 10 mM deoxynucleoside triphosphate [dNTP], 0.6 μl 50 mM MgSO4, 0.2 μl forward primer, 0.2 μl reverse primer, 0.1 μl Invitrogen Platinum High Fidelity Taq polymerase) with one of the following three sets of primers: (i) HIAfor4 (5′-TACAAGACCGAGGAAGAAGAG GAT-3′) and HIArev4 (5′-GCCCTTCTACTTGTGCACTG-3′), and (ii) HIAfor5 (5′-TATAATGGGTTTACCCAGAC-3′) and HIArev5 (5′-GCTTTCGTTGATTTTCTGTAGGAAG-3′), and (iii) For15 (5′-CACACAGTACATGTACGCAAAC-3′) and Rev17 (5′-CTCTGGACGCTGTATAATGCCCGAC-3′). The PCR conditions consisted of 94°C for 2 min followed by 35
cycles of 94°C for 15 s, 55°C for 30 s, and 68°C for 5 min, with a final extension at 68°C for 10 min.

The PCR product was then subjected to the second nested PCR with pyrosequencing fusion primers that incorporated a 21-base-long adaptor sequence (forward primer, 5'-Adaptor-Key-MID-GCTTTTCCTACTTCCTGCCACAT-3'; reverse primer, 5'-Adaptor-Key-MID-GCTTTTCCTACTTCCTGCCACAT-3'). The PCR was performed on 5 μl of template with 35 μl master mix (28 μl H2O, 4 μl 10× buffer, 0.8 μl 10 mM dNTP, 1.2 μl 50 mM MgSO4, 0.4 μl forward primer, 0.4 μl reverse primer, 0.2 μl Invitrogen Platinum High Fidelity Tag polymerase). The PCR conditions consisted of 94°C for 4 min followed by 45 cycles of 94°C for 15 s, 50.2°C for 30 s, and 68°C for 5 min, with a final extension at 68°C for 10 min.

The amplicons were visualized on an E-gel EX gel (Life Technologies) (2%) and were gel extracted using a QIAquick gel extraction kit (Qiagen). The purification was carried out per the manufacturer’s protocol with two modifications: 30 μl of buffer EB was used to elute the PCR product instead of the suggested 50 μl, and the incubation time was increased from 1 min to 3 min in order to increase the final DNA concentration.

The purified amplicons were quantified using a Quant-iT PicoGreen double-stranded DNA (dsDNA) assay kit (Life Technologies). By serially diluting a DNA standard with 1× Tris-EDTA (TE) buffer, a standard curve between the fluorescence level and DNA concentration was obtained; in reference to the standard curve, each sample’s DNA concentration was determined in units of ng/μl. The fluorescence level was measured by the use of a SpectraMax M3 microplate reader (Molecular Devices). Each sample’s DNA concentration was converted to molecules/μl and diluted to the concentration of 2 × 108 molecules/μl first using 1× TE buffer. All of the 20-μl diluted amplicon volumes were pooled, and the pooled amplicon was diluted further to the final concentration of 50-fold by adding 10 μl of the diluted sample to 490 μl of nuclease-free water and then 12.5 μl of the diluted sample was added to an emulsion-based reaction mix. The PCR conditions of the emulsion-based clonal amplification consisted of 94°C for 4 min followed by 50 cycles of 94°C for 45 s, 58°C for 12.5 s, and 68°C for 30 s.

Emulsion-based clonal amplification was performed using a GS Junior Titanium emPCR kit, Library-A (Roche). The volume of pooled amplicons to be added to a reaction mix containing 5 million capture beads, 0.25 μl, was determined by the manufacturer’s standard protocol using 0.1 template copies per bead. Because the recommended volume of DNA library to be added ranges from 10 μl to 20 μl, the sample was diluted 50-fold by adding 10 μl of the sample to 490 μl of nuclease-free water and then 12.5 μl of the diluted sample was added to an emulsion-based reaction mix. The PCR conditions of the emulsion-based clonal amplification consisted of 94°C for 4 min followed by 50 cycles of 94°C for 45 s, and 68°C for 30 s.

The emulsion-based PCR product was purified by aspirating the emulsions using a vacuum source which collected the DNA-carrying beads. Subsequently, the beads were washed 3 times using 100% isopropyl alcohol, 100% ethanol, and enhancing buffer serially. Annealing buffer and enrichment primers were added to the washed beads, and then the preenrichment mixture was incubated at 65°C for 5 min followed by cooling on ice for 2 min. By addition of magnetic enrichment beads to the prepared preenrichment mixture, an enrichment process was conducted using a DynaBead magnetic particle concentrator (Life Technologies). Subsequently, the enrichment beads were removed using the melt solution of the Bead Recovery kit (Roche), and approximately 300,000 enriched DNA beads were collected using the GS Junior bead counter. To ensure that the beads were distributed homogeneously in the solution, the manufacturer’s protocol was modified to include a 3-s-long sonication for each washing step throughout the bead recovery, bead enrichment, and sequencing preparation phases, using a sonicator.

---

### TABLE 1 Clinical and laboratory information of 24 chronically infected subjects

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gender</th>
<th>Age (yr)</th>
<th>Risk behavior</th>
<th>Documented minimum duration of infection (days)</th>
<th>Viral load (RNA copies/ml)</th>
<th>CD4 count (cells/mm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ART naive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL7408</td>
<td>M</td>
<td>25</td>
<td>MSM</td>
<td>749</td>
<td>507</td>
<td>840</td>
</tr>
<tr>
<td>IJ7234</td>
<td>M</td>
<td>45</td>
<td>MSM</td>
<td>2,984</td>
<td>5,599</td>
<td>364</td>
</tr>
<tr>
<td>UD9992x</td>
<td>M</td>
<td>44</td>
<td>MSM</td>
<td>48</td>
<td>1,230</td>
<td>185</td>
</tr>
<tr>
<td>QZ8149</td>
<td>F</td>
<td>50</td>
<td>Heterosexual</td>
<td>3,216</td>
<td>15,000</td>
<td>542</td>
</tr>
<tr>
<td>LE2707</td>
<td>F</td>
<td>55</td>
<td>Heterosexual</td>
<td>2,407</td>
<td>1,394</td>
<td>607</td>
</tr>
<tr>
<td>NK9147x</td>
<td>F</td>
<td>60</td>
<td>Heterosexual</td>
<td>65</td>
<td>156,582</td>
<td>87</td>
</tr>
<tr>
<td>EC8287</td>
<td>M</td>
<td>30</td>
<td>MSM</td>
<td>427</td>
<td>46,398</td>
<td>591</td>
</tr>
</tbody>
</table>

ART experienced

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gender</th>
<th>Age (yr)</th>
<th>Risk behavior</th>
<th>Documented minimum duration of infection (days)</th>
<th>Viral load (RNA copies/ml)</th>
<th>CD4 count (cells/mm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CX7332x</td>
<td>M</td>
<td>67</td>
<td>Heterosexual</td>
<td>5,464</td>
<td>389,081</td>
<td>63</td>
</tr>
<tr>
<td>KI6633x</td>
<td>M</td>
<td>60</td>
<td>MSM</td>
<td>2,428</td>
<td>460,004</td>
<td>115</td>
</tr>
<tr>
<td>CF6610</td>
<td>M</td>
<td>50</td>
<td>MSM</td>
<td>3,710</td>
<td>17,001</td>
<td>634</td>
</tr>
<tr>
<td>ZI9493x</td>
<td>M</td>
<td>58</td>
<td>MSM</td>
<td>41</td>
<td>85,900</td>
<td>70</td>
</tr>
<tr>
<td>CS0442x</td>
<td>M</td>
<td>50</td>
<td>MSM</td>
<td>1,947</td>
<td>1,980</td>
<td>58</td>
</tr>
<tr>
<td>GE6444x</td>
<td>M</td>
<td>54</td>
<td>MSM</td>
<td>707</td>
<td>556</td>
<td>120</td>
</tr>
<tr>
<td>MP7985x</td>
<td>F</td>
<td>55</td>
<td>Heterosexual</td>
<td>471</td>
<td>30,262</td>
<td>411</td>
</tr>
<tr>
<td>PT8907x</td>
<td>M</td>
<td>59</td>
<td>MSM</td>
<td>1,864</td>
<td>465</td>
<td>485</td>
</tr>
<tr>
<td>JW8291x</td>
<td>M</td>
<td>56</td>
<td>Heterosexual</td>
<td>682</td>
<td>748</td>
<td>166</td>
</tr>
<tr>
<td>ST7619</td>
<td>M</td>
<td>66</td>
<td>MSM</td>
<td>2,032</td>
<td>1,365</td>
<td>635</td>
</tr>
<tr>
<td>RE1725x</td>
<td>M</td>
<td>55</td>
<td>MSM</td>
<td>2,164</td>
<td>176</td>
<td>160</td>
</tr>
<tr>
<td>EV9387x</td>
<td>M</td>
<td>56</td>
<td>Heterosexual</td>
<td>2,539</td>
<td>11,336</td>
<td>491</td>
</tr>
<tr>
<td>FX6385x</td>
<td>M</td>
<td>51</td>
<td>Heterosexual</td>
<td>2,115</td>
<td>2,690</td>
<td>434</td>
</tr>
<tr>
<td>CD8867x</td>
<td>M</td>
<td>55</td>
<td>Heterosexual</td>
<td>3,513</td>
<td>22,136</td>
<td>284</td>
</tr>
<tr>
<td>SK4851x</td>
<td>M</td>
<td>46</td>
<td>MSM</td>
<td>800</td>
<td>91,406</td>
<td>208</td>
</tr>
<tr>
<td>TP0539x</td>
<td>M</td>
<td>46</td>
<td>MSM</td>
<td>900</td>
<td>73,000</td>
<td>25</td>
</tr>
<tr>
<td>QN2681x</td>
<td>F</td>
<td>54</td>
<td>MSM</td>
<td>1,638</td>
<td>3,581,418</td>
<td>356</td>
</tr>
</tbody>
</table>

---

a M, male; F, female.

b The minimum duration of infection of each individual was based on the elapsed time between the date of first HIV-positive documentation at the Rand Schrader Clinic and the date of specimen collection.

c Subjects who had an AIDS diagnosis at the time of specimen collection.
TABLE 2 The antiviral drugs used to treat 17 patients

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic subjects</td>
<td></td>
</tr>
<tr>
<td>CX7332</td>
<td>ZDV, 3TC, NVP, D4T, TDF, FTC, EFV</td>
</tr>
<tr>
<td>KI6633</td>
<td>TDF, FTC, ATV, ZDV, EFV, 3TC</td>
</tr>
<tr>
<td>C6610</td>
<td>TDF, FTC, EFV</td>
</tr>
<tr>
<td>ZI9923</td>
<td>DRV, RTV, TDF, FTC</td>
</tr>
<tr>
<td>CS6042</td>
<td>ABC, 3TC, FTC, RTV, ZDV, IDV, NVP</td>
</tr>
<tr>
<td>GE4444</td>
<td>ATV, RTV, FTC, TDF, 3TC, ZDV, D4T</td>
</tr>
<tr>
<td>MP9985</td>
<td>ABC, 3TC, ZDV</td>
</tr>
<tr>
<td>PY9507</td>
<td>ETR, TDF, 3TC</td>
</tr>
<tr>
<td>JY8921</td>
<td>ZDV, 3TC, IDV, D4T, NFP, ETR, DRV, FTC, TDF</td>
</tr>
<tr>
<td>ST7619</td>
<td>ZDV, 3TC, ABC, TDF, D4T, DDI, SQV, RTV, NVP, EFV, NVP, APV, IDV</td>
</tr>
<tr>
<td>RE1725</td>
<td>DRV, FTC, ETR, RAL</td>
</tr>
<tr>
<td>EY9387</td>
<td>RPV, TDF, FTC, DRV, RTV, D4T, 3TC, EFV</td>
</tr>
<tr>
<td>FX6386</td>
<td>ATV, RTV, TDF, FTC</td>
</tr>
<tr>
<td>CD8867</td>
<td>3TC, TDF, NVP, ABC, APV, IDV, DRV, RTV</td>
</tr>
<tr>
<td>SK4851</td>
<td>ATV, RTV, ABC, 3TC, NVP, D4T, LPV</td>
</tr>
<tr>
<td>TP5099</td>
<td>3TC, TDF, RTV, IDV, ETR, FTC, TDF</td>
</tr>
<tr>
<td>QN2691</td>
<td>ATV, RTV, TDF, FTC, EFV, ZDV, 3TC, NVP, IDV, LPV, DDI, D4T, SQV, ABC</td>
</tr>
</tbody>
</table>

Incident subjects

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW0573</td>
<td>EVG, GS-9350, TDF, FTC</td>
</tr>
<tr>
<td>AJ9301</td>
<td>EFV, FTC, TDF</td>
</tr>
<tr>
<td>WN9852</td>
<td>RTV, FTC, DRV</td>
</tr>
<tr>
<td>DV7384</td>
<td>FTC, RPV, TDF</td>
</tr>
<tr>
<td>UW1847</td>
<td>EVG, GS-9350, TDF, FTC</td>
</tr>
<tr>
<td>VH6724</td>
<td>EVG, FTC, TDF</td>
</tr>
<tr>
<td>OM0754</td>
<td>EVG, GS-9350, TDF, FTC</td>
</tr>
<tr>
<td>OG1376</td>
<td>EVG, GS-9350, TDF, FTC</td>
</tr>
<tr>
<td>HB9168</td>
<td>EVG, GS-9350, FTC, TDF</td>
</tr>
<tr>
<td>BV2682</td>
<td>RTC, DRV, FTC</td>
</tr>
</tbody>
</table>

* 3TC, lamivudine; ABC, abacavir; APV, amprenavir; ATV, atazanavir; D4T, stavudine; DDI, didanosine; DRV, darunavir; EFV, efavirenz; ETR, etravirine; EVG, elvitegravir; FTC, fritgravida; GS-9350, cobicistat; IDV, indinavir; LPV, lopinavir; NFV, nelfinavir; NVP, nevirapine; QUAD, Strided; RAL, raltegravir; RPV, rilpivirine; RTV, ritonavir; SQV, saquinavir; TDF, tenofovir; ZDV, zidovudine.

(Branson). Finally, the enriched beads were deposited onto a single-region Pico Titer Plate and run on a Roche GS Junior Sequencer using a GS Junior Titanium PicoTiterPlate kit and a Sequence kit (Roche).

Single-genome amplification and Sanger sequencing. The SGA procedure was performed as described in references 27 and 28. Immediately after the HIV RNA extraction procedure (see above), viral cDNA was synthesized using Superscript III reverse transcriptase (Invitrogen) with primer envB5out (5'-TTGCTACTTGTGATTGCTCCATG-3') and reverse primer envB3out (5'-TTGCTACTTGTTCTCGAGATACTGCTCCCACCC-3') in a total reaction volume of 40 µl composed of 15 µl of RNA template, 11 µl first master mix (8.5 µl 10 mM dNTP, 0.5 µl primers, and 14 µl of the master mix (15.5 µl buffer, 2 µl 10× buffer, 0.4 µl 10 mM dNTP, 0.6 µl 50 mM MgSO4, 0.2 µl forward primer, 0.2 µl reverse primer, 0.1 µl Platinum High Fidelity Tag polymerase). The PCR conditions consisted of 94°C for 2 min followed by 35 cycles of 94°C for 15 s, 55°C for 30 s, and 68°C for 10 min.

Then, 2 µl of the first PCR product was subjected to the inner PCR using one of the following two sets of the primers: (i) forward primer envB5in (5'-TTGCTACTTGTGATTGCTCCATG-3') and reverse primer envB3in (5'-GTCTCGAGATACTGCTCCCACCC-3') and (ii) forward primer For15 (5'-CAGCACAGTACAATGTACACATGGAA-3') and reverse primer Rev17 (5'-CCCTGGAGCTGTATTAAATGCCTCCACAGCAGC-3'). The second PCR was performed using 2 µl of the first PCR product and 18 µl of the master mix (14.5 µl H2O, 2 µl 10× buffer, 0.4 µl 10 mM dNTP, 0.6 µl 50 mM MgSO4, 0.2 µl forward primer, 0.2 µl reverse primer, 0.1 µl Platinum High Fidelity Tag polymerase). The PCR conditions were identical to those of the first PCR except that the cycle number was increased to 45. The presence of valid amplicons was verified using 1% agarose T-gel ethidium bromide (EtBr) (Invitrogen) and a GeLogic 212 Pro Imaging System (Carestream).

After collecting valid SGA amplicons, the DNA concentration was quantified using a Quant-iT PicoGreen dsDNA assay kit (Life Technologies) (see above). Amplicons were diluted as necessary to obtain an optimal amount of DNA to undergo cycle sequencing (BigDye Terminator v3.1; Applied Biosystems). The optimal amount of DNA recommended by the manufacturer was 20 to 50 ng for a template of greater than 2,000 bp, but DNA quantities ranging from 0.65 ng to 15.3 ng were confirmed to yield valid sequencing results. Cycle sequencing was performed using 5 µl of diluted (if necessary) cDNA and 5 µl of the master mix (1.5 µl H2O, 1.5 µl 5× sequencing buffer, 1 µl 5 µM cycle sequencing primer, 1 µl BigDye Terminator ready-reaction mix). The cycle PCR conditions consisted of 96°C for 1 min followed by 25 cycles of 94°C for 10 s, 55°C for 5 s, and 60°C for 4 min. The cycle sequencing products were purified using a Qiapquick PCR Purification kit (Qiagen) and sequenced with an ABI 3730xl DNA analyzer (Applied Biosystems).

Pyrosequencing analysis pipeline. We first produced pyrosequencing reads from 7 different control HIV envelope gene segments located at.

TABLE 3 Clinical and laboratory information of 12 incident subjects

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Documented minimum duration of infection</th>
<th>Viral load (no. of RNA copies/ml)</th>
<th>CD4 count (cells/µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ART naive</td>
<td>31</td>
<td>209, 246</td>
<td>17,700</td>
<td>759</td>
</tr>
<tr>
<td>QO0353</td>
<td>28</td>
<td>35, 118</td>
<td>143,000</td>
<td>845</td>
</tr>
<tr>
<td>ART experienced</td>
<td>24</td>
<td>139, 364</td>
<td>45</td>
<td>931</td>
</tr>
<tr>
<td>AS9301</td>
<td>23</td>
<td>174, 302</td>
<td>21</td>
<td>771</td>
</tr>
<tr>
<td>WN9852</td>
<td>23</td>
<td>151, 358</td>
<td>2,430</td>
<td>560</td>
</tr>
<tr>
<td>DV5384</td>
<td>29</td>
<td>182, 298</td>
<td>44</td>
<td>598</td>
</tr>
<tr>
<td>UW1847</td>
<td>24</td>
<td>63, 222</td>
<td>46</td>
<td>576</td>
</tr>
<tr>
<td>VH6724</td>
<td>29</td>
<td>80, 309</td>
<td>546</td>
<td>538</td>
</tr>
<tr>
<td>OM0754</td>
<td>26</td>
<td>214, 340</td>
<td>20</td>
<td>605</td>
</tr>
<tr>
<td>OG1376</td>
<td>24</td>
<td>56, 300</td>
<td>45</td>
<td>694</td>
</tr>
<tr>
<td>HB9168</td>
<td>23</td>
<td>85, 235</td>
<td>13,200</td>
<td>788</td>
</tr>
<tr>
<td>BV2682</td>
<td>28</td>
<td>66, 168</td>
<td>20</td>
<td>648</td>
</tr>
</tbody>
</table>

* The minimum duration of infection of each individual was based on the elapsed time between the date of the first HIV-positive documentation at the Los Angeles Gay and Lesbian Center and the date of specimen collection. The maximum duration of infection was based on the elapsed time between the date of the last HIV-negative test at the center and the date of blood sample collection.
Needleman-Wunsch algorithm (29). As the time required for the Needleman-Wunsch algorithm scales with the square of the sequence length, we expedited the alignment by dividing each read into 4 serially aligned segments with 15-base-long overlapping regions, thereby completing the process approximately 4 times faster. Completing the alignment procedure, all the base calls in disagreement with the reference sequence were mapped and categorized as base substitutions, insertions, or deletions (Table 4). Subsequently, the data for the flow intensity and base quality score of each base call were retrieved from the pyrosequencing flowgram data in a standard flowgram format. The flow intensity and base quality scores of signals were compared with those of errors (see Fig. 4A). Most correct base calls had significantly higher base quality scores than the errors. As shown (see Fig. 4A), the frequency distribution of the signals had a sharp peak at a base quality score of 40. In contrast to the signals, the errors were primarily located at intermediate flow intensities, implying ambiguous base calls for homopolymer regions. Due to the distinctive characteristics of the signals and errors displayed, we were able to establish a boundary of signals in the plane of the base quality score and flow intensity, where 99.987% of correct base calls were retained following masking. If the flow intensity and base quality score of a given base call was located outside the masking boundary, all the reads with this putative erroneous base call were discarded. This resulted in only 5% of the correct reads being removed by masking, achieving a 5% level of significance (the ratio of removed correct reads to the total number of correct reads).

Nucleotide sequence accession numbers. The SGA sequences reported in this study were submitted to GenBank (accession numbers KF945693 to KF945994).

RESULTS

Biomarkers for genomic incidence assays. Pyrosequencing platforms are susceptible to an increased risk of replicating a given viral template throughout the course of the multiple PCRs that take place prior to a sequencing run. This template resampling introduces variations in sequence population representation which, depending on the level of resampling, diminish the assay’s

FIG 1 Design scheme for the pyrosequencing platform of the genomic incidence assay. (A and B) An example of pyrosequencing flowgram data (A) of each patient’s sample was processed by the bioinformatics pipeline based on differences in the flow intensity and base quality score between correct and incorrect base calls (B). The pipeline applied a mask to the data that retained 99.987% of all correct base calls from 13,550 control reads. Only 5% of the correct reads were removed by masking, achieving a 5% level of significance and a process error rate of 5.8 $\times 10^{-7}$ per base. (C) The processed reads were examined by a frequency-based Hamming distance (HD) distribution. The HD distributions from patient PI9189 and patient PY9507 shown here provide representative examples of incident and chronic infections, respectively. (D) To quantify the amount of closely related sequences (the signature of incident infections), two biomarkers, genome similarity index 5 (GSI5) and the 25% quantile of the HD distribution (Q25), were defined. Incident and chronic samples were predominantly segregated in the two-biomarker plane. The border line between the two stages was determined by receiver operating characteristic (ROC) analysis: samples above the border line were classified as incident infections, with all others classified as chronic infections.

the region containing the V3 and V4 loops (HXB2 7212 to 7601) to devise an analysis pipeline. To ensure that all amplicons for each control strain were amplified from a single template, two procedures were employed. First, the amplicons of each control strain were obtained through the SGA procedure. Second, the Sanger sequencing chromatogram was reviewed to verify that each control sequence’s amplicon products contained a single variant of the template. The amplicons of 7 control strains were then sequenced using a pyrosequencing platform (Roche GS Junior Sequencer). To efficiently remove erroneous reads, we developed a signal-masking algorithm (Fig. 1; see also Fig. 4). An initial preprocessing procedure removed all the reads containing ambiguous base calls and errors in the primer regions, resulting in 13,550 reads with 4,787,859 base calls (Table 4). Each remaining read was aligned to the control reference strain, which was independently sequenced by the Sanger sequencing, using the Needleman-Wunsch algorithm (29). As the time required for the Needleman-Wunsch algorithm scales with the square of the sequence length, we expedited the alignment by dividing each read into 4 serially aligned segments with 15-base-long overlapping regions, thereby completing the process approximately 4 times faster. Completing the alignment procedure, all the base calls in disagreement with the reference sequence were mapped and categorized as base substitutions, insertions, or deletions (Table 4). Subsequently, the data for the flow intensity and base quality score of each base call were retrieved from the pyrosequencing flowgram data in a standard flowgram format. The flow intensity and base quality scores of signals were compared with those of errors (see Fig. 4A). Most correct base calls had significantly higher base quality scores than the errors. As shown (see Fig. 4A), the frequency distribution of the signals had a sharp peak at a base quality score of 40. In contrast to the signals, the errors were primarily located at intermediate flow intensities, implying ambiguous base calls for homopolymer regions. Due to the distinctive characteristics of the signals and errors displayed, we were able to establish a boundary of signals in the plane of the base quality score and flow intensity, where 99.987% of correct base calls were retained following masking. If the flow intensity and base quality score of a given base call was located outside the masking boundary, all the reads with this putative erroneous base call were discarded. This resulted in only 5% of the correct reads being removed by masking, achieving a 5% level of significance (the ratio of removed correct reads to the total number of correct reads).
ability to correctly classify incident and chronic infections on a consistent basis. To avoid the complications associated with template resampling, our pyrosequencing platform redefined the Hamming distance (HD) distribution of sequences sampled from each individual as a frequency-based measure. The frequency-based HD distribution is written as follows:

\[
p(HD = d) = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} I(HD_{ij} = d)
\]

where \(p(HD = d)\) is the probability of HD being \(d\), \(f_{ij}\) is the frequency of the read \(i|j\), \(n\) is the number of distinct reads within a sampled HIV sequence population, and \(I(HD_{ij} = d)\) is an indicator that the HD between sequence \(i\) and \(j\) must be \(d\). As shown in the examples in Fig. 1C, the presence of closely related sequences as a signature of an incident infection is reflected in the HD distribution as a peak at a low-HD region whereas a broadened HD distribution indicates a chronic infection. Focusing on these tail characteristics, we propose two genomic biomarkers, the genome similarity index (GSI) and the 25% quantile of the HD distribution. The GSI for HD = \(k\) is the proportion of pairs of reads with HD equal to or lower than \(k\) among all the possible pairs of distinct reads plus the sum of the squares of the frequencies of each of the distinct reads, accounting for the amount of identical reads.

Mathematically, this can be summarized as follows:

\[
GSI_k = \sum_{d=0}^{k} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} I(HD_{ij} = d) \right)
\]

For example, \(GSI_0 = \sum_{i=1}^{n} f_{i0}^2\) quantitates the amount of identical sequence pairs and \(GSI_k = GSI_0 + \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} I(HD_{ij} = 1)\) quantitates the proportion of identical sequence pairs and sequence pairs with one nucleotide base difference (HD = 1) in a sampled HIV population, respectively. The samples of recent infections are presumed to have relatively high GSI scores since a group of similar sequences may exist within each lineage of transmitted virus at the early stage of infection. In contrast, the samples of chronic infections would presumably have relatively low GSI values reflective of an intrahost viral population that had diversified considerably throughout the course of infection. The second biomarker quantifies the HD value marking the 25% quantile of the frequency-based HD distribution (Q_{25}). These two biomarkers, while similar, give us two separate methods that can either fix the area under the HD distribution, returning an HD value, or fix the latter while returning the former, fully characterizing the tail characteristics of the HD distribution. The central signature of recent infections would presumably have relatively low GSI scores since a group of similar sequences may exist within each lineage of transmitted virus at the early stage of infection. In contrast, the samples of chronic infections would presumably have relatively low GSI scores since a group of similar sequences may exist within each lineage of transmitted virus at the early stage of infection. The second biomarker quantifies the HD value marking the 25% quantile of the frequency-based HD distribution (Q_{25}). These two biomarkers, while similar, give us two separate methods that can either fix the area under the HD distribution, returning an HD value, or fix the latter while returning the former, fully characterizing the tail characteristics of the HD distribution. The central signature of recent infections, the presence of closely related strains, manifests itself through high GSI and low Q_{25} scores, whereas the opposite tends to be the case in chronic infections, as exemplified in Fig. 1.

**Sensitivity of genomic biomarkers to sequencing errors.** We evaluated the biomarkers’ sensitivities to the sequencing error rate by conducting a simple Monte Carlo (MC) simulation. Each simulation started with 100 identical 400-base-long DNA sequences. In this case, the GSIs for these sequences should be equal to 1, as we can see from equation 2, since the HD between all pairs of the 100 identical sequences is 0. Then, we introduced random base substitutions, insertions, and deletions to these identical sequences at a given error rate, mimicking pyrosequencing errors, and monitored how the GSIs were affected. Here we repeated the procedure 100 times for each error rate and measured the GSIs by averaging the results over 100 simulation runs. As shown in

---

**TABLE 4 Summary of control data set**

<table>
<thead>
<tr>
<th>Control data set no.</th>
<th>No. of read errors</th>
<th>% of read errors</th>
<th>Process error rate (per base)</th>
<th>No. of matched process reads/total no. of process reads (%</th>
<th>% of process errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,639/2,134 (76.5)</td>
<td>9.2</td>
<td>8.8 × 10⁻⁴</td>
<td>1,564/1,880 (83.3)</td>
<td>4.9 × 10⁻⁴</td>
</tr>
<tr>
<td>2</td>
<td>1,509/1,868 (84.9)</td>
<td>8.8</td>
<td>9.9 × 10⁻⁴</td>
<td>1,187/1,456 (81.6)</td>
<td>8.8 × 10⁻⁴</td>
</tr>
<tr>
<td>3</td>
<td>1,259/1,868 (71.7)</td>
<td>8.8</td>
<td>9.9 × 10⁻⁴</td>
<td>1,259/1,922 (65.2)</td>
<td>8.8 × 10⁻⁴</td>
</tr>
<tr>
<td>4</td>
<td>1,503/1,792 (83.9)</td>
<td>8.8</td>
<td>9.9 × 10⁻⁴</td>
<td>1,238/1,450 (83.9)</td>
<td>8.8 × 10⁻⁴</td>
</tr>
<tr>
<td>5</td>
<td>1,885/2,388 (78.9)</td>
<td>8.8</td>
<td>9.9 × 10⁻⁴</td>
<td>1,238/1,450 (83.9)</td>
<td>8.8 × 10⁻⁴</td>
</tr>
</tbody>
</table>

---

**Notes:**

- Data represent the average of the total numbers of reads.
- SD (percent error rate) 1,936
- SD (percent process reads) 342
- SD (percent HD) 6.7
- SD (percent Q_{25}) 1.4

---

**a** Data represent the average of the total numbers of reads.

**b** Data represent the average of the total numbers of process reads.

**SD** 1,936
From equation 3, with our choices of $V_{RT} = 0.025 \text{ ml}$, $V_p = 2 \text{ ml}$, and $V_e = 0.05 \text{ ml}$, the number of unique templates ($U$) equals each patient’s viral load ($VL$), as presented in Table S1 in the supplemental material.

The level of resampling can be assessed as the ratio between the number of pyrosequencing reads obtained ($NR$) and the number of unique templates ($U$), as listed in Table S1 in the supplemental material. In order to evaluate the biomarkers’ sensitivities to resampling and variation in the number of unique templates, we conducted a Monte Carlo simulation which directly monitored how chronic infection signatures can be affected by replicating a limited amount of viral templates. We took the SGA sequence data set of the chronically infected subject assigned random identification code JJ7234 consisting of 20 strains (HXB2 7212 to 7601) and first randomly selected a given number of unique strains, $U$, of 20 sequences, all of which were different from one another. We then obtained 100 sequences from $U$ unique templates by allowing resampling and repeated this procedure 100 times to measure the average and standard deviation of the biomarkers (Fig. 3). As $U$ started to decrease, the two biomarkers remained relatively near baseline values, with only minor changes. The $GSI_5$ and $Q_{25}$ started to significantly deviate from their true scores at values of $U$ of around 5. This trend was invariable even following 10-
Park et al.

100-fold increases in the resampling level (Fig. 3). Our simulation outcomes clearly indicated that our biomarkers were sensitive to the initial number of unique templates and not to the level of resampling. Furthermore, once the number of unique templates was increased past 15, the changes in the biomarker scores were negligible (from 26.2 to 25.1 for Q25 and from 0.069 to 0.085 for GSI5) as U decreased from 20 to 15. This chronic sample from patient IJ7234 remained below the cutoff value in the GSI5 and Q25 range at which the number of unique templates starts to affect the two biomarkers.

Devising the analysis pipeline for pyrosequencing. Identifying two biomarkers potentially capable of differentiating incident from chronic infections while tolerating the complications of sequencing errors and template resampling solidified a fundamentally important component of a successful genomic assay, but addressing upstream pyrosequencing errors within the platform was also necessary to augment the assay’s precision. For this reason, we devised a pyrosequencing analysis pipeline. Data from 7 control sequences were produced using a next-generation pyrosequencer (Roche GS Junior). There were 13,550 reads with 4,787,859 base calls after an initial preprocessing discarding all the reads with ambiguous base calls and errors in primer regions was performed. The average number of reads per control sequence was 1,936, and the average read length was 398. Control amplicons (HXB2 7212 to 7601) were prepared through the SGA procedure (27, 28), ensuring that each set of control templates originated from a single copy of viral cDNA. To identify errors, the reads were then aligned to the reference sequence, which was independently sequenced using the Sanger method (Applied Biosystems 3730xl). Prior to being processed using our analysis pipeline, the average read error rate was $1.1 \times 10^{-3}$ per base. Errors consisted of 44.8% point mutations, 35.1% insertions, and 20.1% deletions (Table 4). On average, 76% of the reads perfectly matched each control sequence.

We developed an analysis pipeline which removed read errors by employing a signal-masking algorithm. Figure 4A plots the distribution of the flow intensity and base quality scores of the correct base calls from the 7 control data sets. As expected, the majority of these signals had high base quality scores, peaking at 40. On the other end of the spectrum, errors had lower base quality scores than signals (Fig. 4B). Errors also showed a tendency to be located at intermediate flow intensities, denoting that the main source of base-calling errors was ambiguous homopolymer lengths (Fig. 4B). By utilizing this difference in the distributions of the flow intensity and base quality scores between signals and errors, our analysis pipeline masked the signals, as shown in Fig. 4B. Here, a boundary masking 99.987% of signals was determined to achieve a 5% level of significance (the ratio of removed correct reads to the total number of correct reads) and all the reads with any base calls outside the masking region were discarded. Overall, the masking pipeline discarded 14% of the total reads we had obtained from the 7 control data sets, and only 5% of total correct reads were removed. We found that the masking pipeline significantly reduced the process error rate from $1.1 \times 10^{-3}$ to $5.8 \times 10^{-4}$ per base (Table 4). This process error implied that the probability of observing a single base error in a single read was around 0.2. The proportion of the reads that perfectly
matched each control sequence was increased from 76% to 84% on average. We concluded that the masking pipeline is an effective way of processing deep-sequencing flowgram data by balancing the process error rate with the amount of correct reads retained.

Pyrosequencing data of 12 incident and 24 chronic patients. To obtain pyrosequencing data sets from both incident and chronic subjects, 12 incident and 24 chronic patients were recruited from the Los Angeles Gay and Lesbian Center and the University of Southern California’s Rand Schrader Clinic, respectively (see Materials and Methods for more details). Documented dates of HIV-negative tests and the first HIV-positive test were available for all 12 incident patients who had been infected for less than 1 year at the time of specimen collection, enabling us to obtain the minimum and maximum durations of infection for each subject, as listed in Table 3. The minimum duration of infection ranged from 35 to 214 days with mean of 121.2 (±64.1) days and the maximum duration from 118 to 364 days with mean of 271.7 (±75.9) days for the 12 incident patients. Viral loads of the 12 incident patients ranged from 20 to 143,000 RNA copies/ml, and CD4⁺ T cell counts were distributed in the range of 538 to 931 cells/mm³. Most (10 of 12) of the incident subjects were ART experienced at the time of sample collection (Table 3).

Each of the 24 chronic subjects had a documented HIV-positive date as listed in Table 1. The 24 subjects had a viral load ranging from 176 to 3,581,418 RNA copies/ml and a CD4⁺ T cell count of from 25 to 840 cells/mm³. The documented minimum duration of infection for the three study participants, UD9992, NK9147, and ZI9923, was less than 1 year, but these subjects were considered chronically infected since they had an AIDS diagnosis (CD4⁺ T cell count of less than 200 cells/mm³) at the specimen collection date (Table 1). Seven more subjects were at the AIDS stage upon sample collection as seen in Table 1. Of 10 AIDS-diagnosed subjects, two individuals, KI6633 and ZI9923, were symptomatic upon sample collection. Among the chronic subjects, 17 had received ART and 7 were ART naive at the time of specimen collection (Table 1). Table 2 lists the antiviral drugs used to treat each infected individual.

The HIV envelope gene segments located at the region containing the V3 and V4 loops (HXB2 7212 to 7601) obtained from two groups of study participants were sequenced by a pyrosequencer and processed by our masking pipeline, yielding a total of 18,434 reads (see Materials and Methods; see also Table S1 in the supplemental material). This genomic region has been chosen because the V3 and V4 loops are among the most variable regions of the envelope gene while also being enclosed by conserved primer-targeted regions. After being filtered through the masking pipeline, the reads were subjected to pairwise alignment and measured with the two biomarkers GSI₃ and Q₂₅. As expected, a high proportion of similar sequences was present in all 12 incident subjects, resulting in high GSI₃ and low Q₂₅ scores (Fig. 5A). In contrast, sampled HIV sequence strains from most chronic subjects—with the exception of subject CF6610—increasingly differed from one another, showing low GSI₃ and high Q₂₅ scores (Fig. 5A).

Noting this tendency for incident cases to be located in the upper left portion of the Q₂₅-GSI₃ plane and chronic cases to be located in the lower right, we objectively determined a cutoff line that would place incident cases above it and chronic cases below it by conducting a receiver operating characteristic (ROC) analysis. First, by adjusting the slope of the cutoff line, we obtained a ROC curve (sensitivity versus 1 − specificity) and measured the area under the ROC curve for each slope of the cutoff line. Next, we determined the range of the cutoff line slopes that would maximize the area under the ROC curve. Then, an isocost line maximizing the sum of the sensitivity and specificity values designated the y intercept of the cutoff line. All potential cutoff lines sharing the designated isocost line were examined to obtain the widest possible range of border lines that achieved the maximum accu-
racy. In this manner, any line within the region marked by the two border lines shown in Fig. 5A resulted in 100% sensitivity and 95.8% specificity. All 12 incident subjects and 23 of 24 chronically infected individuals (except for subject CF6610, whose biomarkers were located above the cutoff line) were correctly classified by the cutoff lines.

A multivariate analysis of variance (MANOVA) was performed with the stage of HIV disease as an independent factor and the two biomarkers as a set of variables. There was no statistically significant difference in the biomarkers between 10 AIDS patients and 14 non-AIDS patients ($P = 0.12$). The sequences from the AIDS patients indeed not only demonstrated the genetic signature of chronic infections but also were indistinguishable from the signatures found in chronically infected individuals who had not developed AIDS.

We monitored the performance of the assay using various read lengths by measuring the biomarkers after segmenting the pyrosequencing reads. The minimum read length necessary to maintain the assay’s high precision (100% sensitivity and 95.8% specificity) was around 250 bases (see Table S2 in the supplemental material). With 100-base-long envelope gene segments (HXB2 7212 to 7311), three chronic subjects were misclassified as incident infections, marking the specificity as 87.5%. The specificity of the assay was 96.6%. On the other hand, within the samples analyzed by the pyrosequencing platform, only 2/12 (17%) incident subjects had been infected for less than 6 months upon sample collection (see Table 3), resulting in a lower density of subjects with Q 25 values. All the replicates were located below the cutoff lines shown in Fig. 5B, which yielded 97% sensitivity and 100% specificity. While our pyrosequencing platform and SGA platform were found to distinguish incident cases from chronic ones in a highly accurate manner, the two platforms had different cutoff values that marked the thresholds between incident and chronic cases, which can be attributed to a combination of the following two factors. First, Fiebig laboratory staging (30, 31), which is based on the orderly appearance of viral RNA, antigens, and HIV-specific antibodies in plasma, estimated an infection duration of less than 6 months for around 92% of the 182 incident subjects previously examined by the SGA platform (24–26). This corroborates the observed high density at Q 25 and GSI5 = 1; 62.6% of the 182 incident subjects had these biomarker values. On the other hand, within the samples analyzed by the pyrosequencing platform, only 2/12 (17%) incident subjects had been infected for less than 6 months upon sample collection (see Table 3), resulting in a lower density of subjects with Q 25 = 0 and GSI5 = 1. Second, sequencing errors and template resampling might have affected the ROC analysis and GSI5 values, though our simulation studies demonstrated that the two biomarkers are relatively tolerant of these complications. For samples adjacent to the boundaries between the incident and chronic cases, small perturbations caused by sequencing errors and template resampling might disproportionately impact cutoff lines. With more data sets, we would be better able to determine incidence across, in addition to within, different sequencing platforms.

**FIG 5** The genomic incidence assay’s classification scheme. (A) Biomarker score distributions of HIV envelope gene segments (HXB2 8212 to 7601) obtained from all 12 incident infections and 24 chronic infections by the pyrosequencing platform. The dotted lines represent the boundaries between incident (red) and chronic (blue) infections obtained by the ROC analysis; any line enclosed within the region marked by two dotted lines discriminates between incident and chronic infections with the maximum sum of sensitivity and specificity. Each cutoff line places incident infections above it and chronic infections below it. The classification achieved 100% sensitivity and 96% specificity. (B) Applying the same classification scheme to the same region of HIV envelope gene segments of 182 incident and 43 chronic infections, which were obtained by SGA, resulted in 97% sensitivity and 100% specificity. The SGA data sets were collected from references 24, 25, and 26.

**Testing genomic biomarkers with preexisting sequencing data.** To establish a point of comparison for the pyrosequencing platform designed in this study, we cross-referenced the previously published sequence data sets which were obtained through single-genome amplification and Sanger sequencing (see Materials and Methods for more details). These sequences were obtained from 182 infected individuals who were at the incident stage and 43 individuals who were at the chronic stage. We observed distributions similar to those determined with our pyrosequencing analysis, marked by incident infections clustered in high-GSI5 and low-Q 25 regions and chronic infections located at comparatively lower GSI5 and higher Q 25 regions. The ROC analysis determined the cutoff lines shown in Fig. 5B, which yielded 97% sensitivity and 100% specificity. While our pyrosequencing platform and SGA platform were found to distinguish incident cases from chronic ones in a highly accurate manner, the two platforms had different cutoff values that marked the thresholds between incident and chronic cases, which can be attributed to a combination of the following two factors. First, Fiebig laboratory staging (30, 31), which is based on the orderly appearance of viral RNA, antigens, and HIV-specific antibodies in plasma, estimated an infection duration of less than 6 months for around 92% of the 182 incident subjects previously examined by the SGA platform (24–26). This corroborates the observed high density at Q 25 = 0 and GSI5 = 1; 62.6% of the 182 incident subjects had these biomarker values. On the other hand, within the samples analyzed by the pyrosequencing platform, only 2/12 (17%) incident subjects had been infected for less than 6 months upon sample collection (see Table 3), resulting in a lower density of subjects with Q 25 = 0 and GSI5 = 1. Second, sequencing errors and template resampling might have affected the ROC analysis and GSI5 values, though our simulation studies demonstrated that the two biomarkers are relatively tolerant of these complications. For samples adjacent to the boundaries between the incident and chronic cases, small perturbations caused by sequencing errors and template resampling might disproportionately impact cutoff lines. With more data sets, we would be better able to determine incidence across, in addition to within, different sequencing platforms.

**Replicate analysis on pyrosequencing data.** We inquired into the reproducibility of our data by selecting two chronic samples (JW8291 and TP0539) adjacent to the border lines and producing three replicates of pyrosequencing data sets for each of the two samples. By conducting separate RT-PCRs and PCRs for each replicate, three independent amplicon libraries for both samples were prepared. Following this reproduction, the six data points were replotted in the two-biomarker plane. As shown in Fig. 6, all three replicates for both patients showed comparable GSI5 values and small shifts in Q 25 values. All the replicates were located below the cutoff lines which were determined with the pyrosequencing data set as shown in Fig. 5A and therefore were correctly classified as
chronic infections, indicating the robustness of our pyrosequencing workflow. This observed consistency within same-patient samples suggests that sequencing errors and template resampling in our pyrosequencing platform occur in a controlled manner, rendering a genomic incidence assay more readily usable by avoiding the cost associated with running multiple assays or confirming patients’ results.

**Single-genome amplification and Sanger sequencing.** To cross-validate outcomes from the pyrosequencing platform, we independently produced over 300 envelope gene segments (HXB2 7212 to 7601) using the method of SGA from 3 incident individuals (PI9189, WN9587, and VH8724) and 10 chronic samples (PL7408, IJ7234, UD9992, QZ8149, LE2707, CX7332, NK9147, EC8287, KI6633, and CF6610). Note that we examined the same region of the gene to allow a direct comparison between the pyrosequencing and SGA platforms. The number of sequences obtained by the SGA for each individual is presented in Table S3 in the supplemental material. The SGA optimized the dilution of HIV cDNA prior to PCR amplification to ensure that all amplicons originated from a single copy of cDNA, providing a precise representation of each HIV strain (27, 28).

We found that each of the 3 incident and 10 chronic subjects examined by the SGA displayed the signature of corresponding stage of infection, as evidenced by each sample’s location relative to the biomarker plane’s cutoff line. Notably, subject CF6610, misclassified by the pyrosequencing platform, was correctly classified as having a long-standing infection. As shown in Fig. 7, the HIV population complexity revealed by the pyrosequencing data was found to be consistent with the cross-analyzed SGA data. We found a statistically significant correlation between the biomarker GSI5 obtained by the pyrosequencing and that obtained by the SGA platform (F-statistics, \( r = 0.88 \), \( F = 35.9 \), and \( P = 9.0 \times 10^{-7} \)). Likewise, a statistically significant correlation was measured between the biomarker Q25 values obtained by the two methods (\( r = 0.63 \), \( F = 7.06 \), and \( P = 0.022 \)). The concordance between the pyrosequencing platform and the SGA method suggested that high throughput and high accuracy can be achieved simultaneously in a pyrosequencing setting for genome-based incidence assays.

**DISCUSSION**

This study has designed a pyrosequencing platform to parallelize genomic HIV incidence assays in pursuit of a single-assay approach scalable for use in routine cross-sectional surveys. The two most important concerns to be addressed in the creation of this platform were potential complications caused by inflated sequencing error rates present in next-generation pyrosequencing as well as those caused by viral template resampling. To permit the application of deep-sequencing techniques for accurately identifying genomic signatures, we devised a bioinformatics pipeline using 7 control data sets consisting of over 4.7 million bases. Our pipeline utilized the difference in the distributions of the flow intensity and base quality scores between signals and errors found in the control data. It masked correct base calls in the plane of the flow intensity and base quality score and removed the reads outside the mask containing putative erroneous base calls. In this way, we were able to reduce the process error rate to \( 5.8 \times 10^{-4} \) per
was assessed and compared with the number of pyrosequencing
individual's maximum number of initial viral templates prior to PCRs
the viral sequence population. To examine this aspect, each indi-
SGA, as evidenced by the statistically significant association
offered a consistent standard for comparisons between reads
bution's tail characteristics. The frequency-based biomarkers
were able to define biomarkers based on unique read frequency
of the reads. Resampling of a given viral population alters the
absolute read number-based HD distribution by generating iden-
were more than the maximum numbers of initial templates were sam-
predominant factor for inducing a bias in the biomarkers. Simulations
allow us to decipher the genomic signatures that distinguish in-
incident from chronic infections, even in the presence of pyrosequencing control reads.

The second design component of our platform was creating biomarkers tolerant of sequencing errors, which allows robust
classification of incident and chronic infections in the presence of nucleotide insertions, deletions, and substitutions introduced by
the pyrosequencer. For example, genomic biomarkers must have
the ability to repeatedly classify identical sequences as similar to
one another even in the event that sequencing errors artificially
increase the differences between sequences. To test the sensitivity
of our biomarkers, a Monte Carlo simulation was performed
wherein both of our biomarkers, GSI5, and Q25, remained at their
original values after a population consisting of 100 identical
strains was subjected to artificial sequencing errors at rates around
two times higher than the process error we determined for our
pyrosequencing platform. This robustness of the two biomarkers
allows us to decipher the genomic signatures that distinguish in-
cident from chronic infections, even in the presence of pyrose-
quencing errors.

The present study examined viral resampling in the pyro-
sequencing platform from multiple angles. To create a platform that
is robust with respect to resampling, the HD distribution must
incorporate relative frequencies rather than the absolute number
of the reads. Resampling of a given viral population alters the
absolute read number-based HD distribution by generating iden-
tical copies of each read, pushing the distribution toward uniform-
ity (HD = 0). For example, the previous biomarker used in
genomic incidence assays (14), the 10% quantile of the HD distribution (Q10), decreases following an increase in the propor-
tion of identical sequence pairs, an artifact caused by resam-
pling. By measuring a frequency-based HD distribution, we
were able to define biomarkers based on unique read frequency
rather than each read's absolute copy number. This allows us to
avoid inaccuracy arising from template resampling in the py-
rosequencing procedure and to precisely define the HD distri-
bution's tail characteristics. The frequency-based biomarkers
offered a consistent standard for comparisons between reads
obtained from pyrosequencing and those obtained through SGA, as evidenced by the statistically significant association
between biomarker scores measured by the two platforms.

Even when using a frequency-based measure, an extremely low
level of initial viral templates results in an incomplete portrait of
the viral sequence population. To examine this aspect, each indi-
vidual's maximum number of initial viral templates prior to PCRs
was assessed and compared with the number of pyrosequencing
reads obtained from each sample. The maximum number of ini-
tial viral templates was estimated to range from 20 to 3.6 million
copies, and the pyrosequencing read number ranged from 37 to
4,150 across 36 patients. The level of resampling, defined as the
ratio between these two measures, ranged from 0.0001- to 37.5-
fold. In an attempt to quantify the effects of a variety of resampling
levels on the genomic biomarkers, we performed Monte Carlo
simulations. According to the simulations, the number of unique
templates, not the level of resampling, was found to be a predom-
inant factor for inducing a bias in the biomarkers. Simulations
predicted that biomarkers would remain effective at relatively low
levels of initial templates before losing reliability as the copy num-
ber fell below 10. The majority of our patients' samples have rel-
atively high viral loads and, therefore, initial template copy num-
bbers and would remain unaffected by resampling effects even if
fewer than the maximum numbers of initial templates were sam-
ped due to experimental limiting factors such as RNA extraction
efficiency and PCR primer mismatch. In the case of three incident
subjects (AS9301, OM0754, and BV0468) with a viral load near 20
copies/ml, experimental factors might have resulted in problem-
tically low levels of initial template being sampled. Taking the
data together, the genetic structure within most HIV patients' 
viral populations would be robustly represented using our plat-
form regardless of the amount of resampling taking place, with the
initial template amount being a limiting factor.

AIDS patients displayed GSI5, and Q25 signatures similar to
those obtained from non-AIDS patients with chronic infections.
It has been reported that intrahost viral diversity declines at the end-
point of disease (33), which may affect genomic biomarkers in the
same way that reduced immune responses impact the accuracy of
serologic assays at the AIDS stage (9). Notwithstanding this con-
cern, comparability of the biomarker scores from AIDS and non-
AIDS chronic patients suggests that genetic diversity may not de-
cline to a level comparable to that seen with incident infections.
Therefore, the AIDS subgroup which contributes significantly to
serologic assays’ false-recency rates may not affect our genomic
assay to the same extent.

GSI5 and Q25 also meet a significant benchmark necessary for a
successful genomic incidence assay: the ability to distinguish re-
cent infections with multiple founder viruses from chronic infec-
tions. When multiple variants were transmitted, the overall ge-
etic diversity at the early stage of infection was demonstrated to
be comparable to the diversity at the chronic stage of infection (24,
34). Therefore, genetic diversity was shown not to be an ideal
measure for distinguishing recent infections from long-standing
ones based on this criterion (14, 34). On the other hand, our
biomarkers attempted to identify the signature of recent evolution
among the descendants of each transmitted virus. The transmitted
variants in multiple-founder infections can differ considerably,
resulting in high levels of genetic diversity typically associated with
chronic infection. However, descendants of each transmitted
founder virus should be closely related since the shorter time
frame following infection would limit the accumulation of muta-
tions within a lineage. An overall measure of genetic diversity ac-
counts for distances among the founder viruses and obscures the
recent evolution signature, nullifying the predictive power to dis-
criminate early infections with multiple founders from chronic
infections. On the other hand, GSI5, and Q25 filter within-lineage
signatures by measuring the low end of the HD distribution,
which explains a remarkably high accuracy (96% sensitivity) clas-
sifying 182 incident subjects, including around 80 subjects whose
infections had originated from multiple transmitted variants (24–26).

The pyrosequencing platform of the genomic incidence assay showed 100% sensitivity, correctly classifying all 12 incident subjects, and 96% specificity, correctly classifying 23 of 24 chronically infected subjects. One chronic subject (CF6610), who was incorrectly classified by the pyrosequencing platform, did show the signature of chronic infection following SGA and Sanger sequencing. Since the subject’s viral load was 17,001 copies/ml, it is unlikely that a lack of unique viral templates played a role in this discrepancy. One possible explanation for this discrepancy is that biased sampling of the viral population resulting from differing amounts of primer mismatch to each viral strain may have been involved (23). Observing biomarker values that placed this subject’s sample above the GSI and Q25 cutoff line could be attributed to the possibility that a limited viral population was sequenced by this type of differential sampling. However, analyzing data from a randomly chosen subset consisting of 3 incident and 10 chronic subjects, which included the misclassified subject, CF6610, demonstrated statistically significant associations between the biomarker scores obtained by pyrosequencing platforms and those obtained from independent SGA measures (Fig. 7). Combining genetic biomarkers resistant to sequencing errors and template resampling with a bioinformatics pipeline capable of correcting intrinsic pyrosequencing errors enabled our platform to achieve a level of accuracy comparable to that of SGA platforms. The pyrosequencing platform’s adherence to the results obtained from the more expensive and time-consuming SGA procedure indicates the possibility of directly applying the genomic assay in high-throughput settings.

Our genomic incidence assay, which correctly classified incident infections spanning the period from 6 months to 1 year posttransmission, is likely to have a longer window period than conventional serologic assays. A window period covering the entire incident phase is advantageous since a much smaller sample size is needed to estimate the rate of incidence (2). According to the results of a meta-analysis of existing assays, no prior assays had obtained a window period greater than approximately 6 months (9). An extended detection window is capable of detecting early infections while maintaining a low false-negative rate for infections present for longer than 6 months. In addition, the genomic incidence assay’s ability to detect even the preseroconversion period (Fiebig stages I and II), a time during which serologic assays are incapable of detecting HIV-specific antibodies, underscores its potential increase in precision. The pyrosequencing and SGA platforms distinguished incident from chronic infections with similar levels of accuracy. However, the results were not completely analogous due to the slightly different incident infection–chronic infection boundaries. These differences presumably originate from intrinsic cohort imbalances, such as different sample sizes and disproportionate amounts of acute-phase samples, AIDS cases, and ART-experienced patients, between the data sets the two platforms produced. Indeed, when the platforms analyzed the same patients’ samples, the biomarkers were statistically significantly correlated, which bodes well for the genomic incidence assay’s reproducibility and robustness. Collecting more data sets across populations would be required to objectively identify incidence across, in addition to within, different sequencing platforms, and this would permit standardization of the assay.

We designed a single-assay-based, high-throughput platform for the genomic HIV incidence assay that, with careful planning, can increase its applicability for routine use in cross-sectional settings. Ensuring that laboratories from around the world can carry out this assay and attain standardized results will require more efforts for reducing cost and streamlining workflow. In addition to our platform’s utility as a stand-alone assay capable of processing many samples in parallel, the sequencing data it generates can be combined with drug resistance surveys, as suggested in reference 34. The genomic assay is expected to serve as a reference to evaluate different approaches for estimating HIV incidence due to its precision, robustness, and replicability. This report marks a significant step toward realizing the viability of genomic assays for determining HIV incidence through cross-sectional sampling.

ACKNOWLEDGMENTS

This work was supported by NIH grants R01 AI095066 and AI083115. We thank the study participants at the Rand Schrader Clinic and the Los Angeles Gay and Lesbian Center and Luis Mendez, Bartolo Santos, Ruben Rivera, and Risa Flynn for support in the recruitment of study participants. We thank Jesus Salazar-Gonzalez, Eun-Young Kim, and Steven Wolinsky for providing help in sequencing methods. We also thank the members of the DNA sequencing core laboratory at the University of Alabama at Birmingham.

S.Y.P. and H.Y.L. designed the deep sequencing platform, formulated the bioinformatics pipeline, analyzed experimental data, and performed Monte Carlo simulations. M.P.D. and R.B. designed the study cohort. S.Y.P., H.Y.L., and H.J.L. conducted sequencing experiments. H.Y.L. and N.G. performed statistical testings. All of the authors participated in the writing of the manuscript.

REFERENCES

18. Bimber
17. 2990 Park et al.
14. 20. Cousins


