A Proposed HIV Reference Sample Characterized by Three Next-Generation Sequencing Platforms

Alejandro Gonzalez-Serna[1], Chanson J. Brumme[1], Winnie Dong[1], Eric Cai[1], Art F. Y. Poon[1], Kelly D. Huber[2], Nicolas Sluis-Cremer[2], Ellen E. Paxinos[3], Luke C. Swenson[1], Paul Richard Harrigan[1,4]

[1]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, CANADA, [2]Department of Medicine, Division of Infectious Diseases, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA, [3]Pacific Biosciences, Applications and Collaborations, Menlo Park, CA, USA, [4]Division of AIDS, Department of Medicine, Vancouver, BC, CANADA

BC Centre for Excellence in HIV/AIDS
603-1081 Burrard Street
Vancouver BC Canada V6Z 1Y6
Tel.: (604) 806-8281
Fax: (604) 806-8464
email: prharrigan@cfenet.ubc.ca

Background:

"Deep" sequencing allows detection of low frequency minority variants, but establishing the analytical characteristics of such assays, including the error rate and limit of detection, will require repeated sequencing of known reference samples. We present sequencing results from three next-generation platforms for one of a panel of such potential samples we will make available upon request for use as either run controls or External Quality Assessment

Methods:

Fifteen infectious viral clones, each containing a single NNRTI resistance mutation (V90I, K101E, K103N, V108I, E138A/G/K/Q, V179D, Y181C, Y188C, G190A/S, M230L and P236L) were created by site directed mutagenesis of HIV-1LAI. Mutants and wild-type viruses were pooled at equal infectivity into a single sample, "RSVP" (reference sample for validating platforms), which was extracted and subjected to RT-PCR. Two second round PCR amplicons were generated: a 1.8 kb region of pol (HXB2 2469-4295) sequenced on the PacBio® RS (v1), and 339 bp region of RT (HXB2 2813-3151) was sequenced on the Illumina MiSeq (v2 2x250 paired end) and 454 Life Sciences GS-Jr. Cross-platform variation was evaluated including amino acid (aa) frequency of the expected NNRTI mutants and error rates in the region covered by all three platforms (RT codons 96-194). In addition, serial dilutions of the 1st round PCR product were sequenced by MiSeq.

Results:

Approximately 30,000, 185,000, and 3500 sequence reads were obtained for PacBio, MiSeq and 454 respectively. When raw sequences were analyzed (no quality score filtering) marginally higher error rates were observed for PacBio compared to MiSeq and 454 (Table 1). The frequency of minor NNRTI resistant variants detected by the three platforms was similar (Figure 1). For MiSeq, discarding bases with quality below a specific cutoff (q-score 10, 15, 20) resulted in improved error rates, with q20 balancing a tradeoff between read retention and sequencing accuracy (q20 error rate 0.32%). Sequencing errors in homopolymer-rich regions and dilution effects of the 1st round PCR product on MiSeq were explored (Figures 2, 3 and 4).

Table 1. Higher median RT 96-194 sequencing error rates were observed for PacBio compared to MiSeq and 454. For MiSeq, reduced error rates were found discarding bases with quality below 20 (MiSeq Q20).

% Error	PacBio	454 GS-Jr	MiSeq Q0	MiSeq Q10	MiSeq Q15	MiSeq Q20
Mean	3.60	0.97	1.20	1.20	1.00	0.40
Median	1.76	0.20	0.69	0.69	0.57	0.32
IQR	1.2-4.5	0.11-0.36	0.44-1.56	0.44-1.56	0.41-1.10	0.24-0.43
Min - Max	0.48 -17.56	0 – 58.02	0.19 - 6.33	0.19 - 6.33	0.18 - 4.97	0.09 - 1.59

Figure 1. The frequency of minor NNRTI resistant variants detected by PacBio, 454 and MiSeq Q0 were similar. ***(Additional quality filtering steps, similar to what was performed for the MiSeq, should improve estimates for PacBio and 454. Please see companion poster by Sethuraman et al, in this session).***

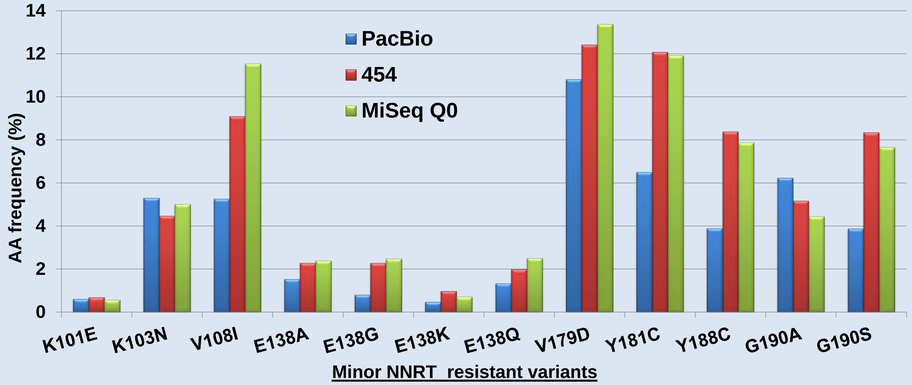


Figure 2. Sequencing errors in the form of Insertion/Deletions and AA substitutions in homopolymer-rich regions were frequently observed in 454 sequences, occasionally in PacBio, but rarely in MiSeq Q0. ***(The higher rates of AA substitution are mostly due to indel errors causing local misalignments that may be mitigated by QV-informed local re-alignment.).***

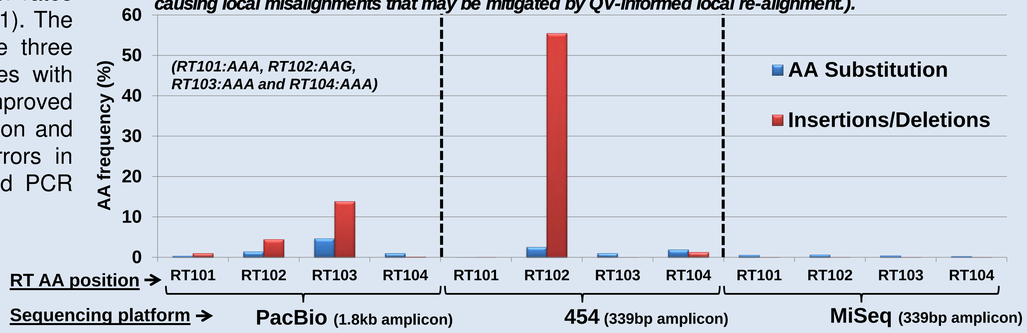


Figure 3. A) MiSeq read coverage was consistent regardless of the quality score **B)** MiSeq read coverage decreases together with the sample dilution.

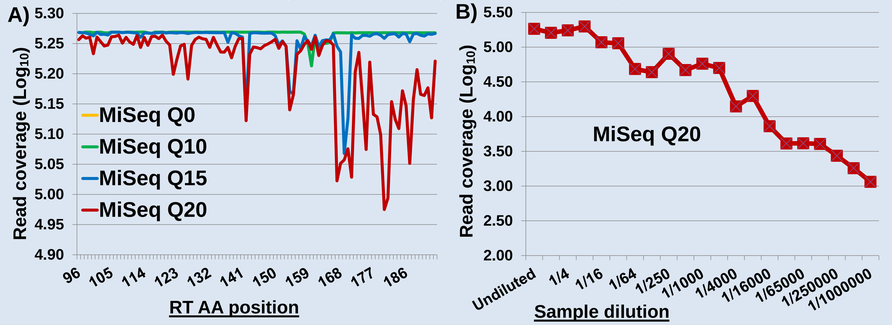
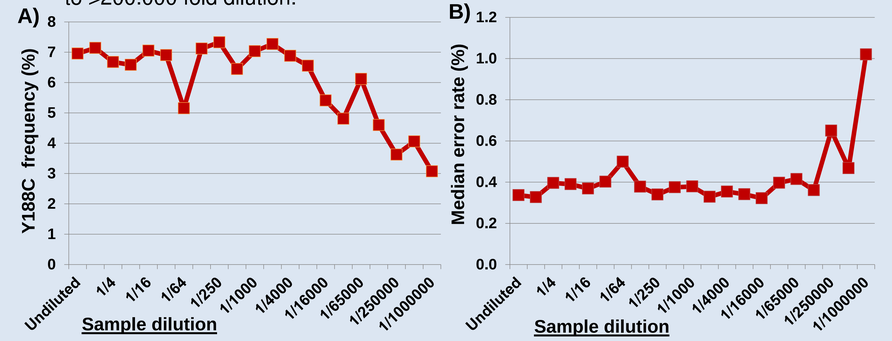


Figure 4. A) The frequency of the Y188C NNRTI resistant variant detected by MiSeq Q20 was robust up to >20.000 fold dilution (same pattern was observed for the other resistant variants). **B)** The median sequencing error rate in MiSeq Q20 was robust up to >200.000 fold dilution.



Conclusions:

We have characterized the performance of three next-generation sequencing platforms using a proposed HIV reference sample; RSVP is a mixture of protease-RT from 16 HIV-1 variants which we have a large amount and that we are making available to interested groups, along with multiple different samples, to use when validating the detection of minority NNRTI resistant variants. These are DNA and not RNA, and therefore do not reflect the inherent variability in the extraction or RT steps.

BRITISH COLUMBIA CENTRE for EXCELLENCE in HIV/AIDS

Providence HEALTH CARE
How you want to be treated.