Efficient High-Throughput Sequencing of Variable and Near Full Length Integrated HIV-1 Proviruses

Kevin W. Joseph[1], Elias K. Halvas[1], Leah Brandt[1], Sean Patro[2], Jason Rausch[3], Abha Chopra[4], Mary Kearney[2], John Coffin[5], John W. Mellors[1]

[1]University of Pittsburgh, Department of Medicine, Division of Infectious Disease, Pittsburgh PA, 15261; [2]HIV Dynamics and Replication Program, National Cancer Institute, Frederick, MD 21702; [3]Basic Research Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702; [4]Institute for Immunology & Infectious Diseases, Murdoch University, Murdoch, Western Australia 6150; [5]Tufts University, Boston MA 02111

Abstract

Problem
- Efforts to cure HIV-1 infection will require a better understanding of the proviral reservoir that persists despite current antiretroviral therapies, but current methods are inefficient due to current technological limitations and inefficient sequencing (high % off target reads)

Objectives:
- Develop a high-throughput and efficient workflow for amplifying integrated single genome HIV proviruses

Methods:
- Single proviral template multiple displacement amplification
- Long range, variable and near-full length proviral nested PCR
- Novel Nullomer-mediated integration site nested PCR

Results:
- Median success rate of 71% for enriching integration sites from MDA wells containing single proviruses without PBS deletions (45.5%-73.9%)
- >84% of sequencing reads are utilized for both proviral and integration site consensus sequence assembly
- Amplification and sequencing of all but 69 bp of the 3' LTR of both variable and near-full-length integrated proviruses
- The validity of the assay has been confirmed by sequence identity of full-length proviruses amplified directly from PBMCs using flanking host sequences as priming sites

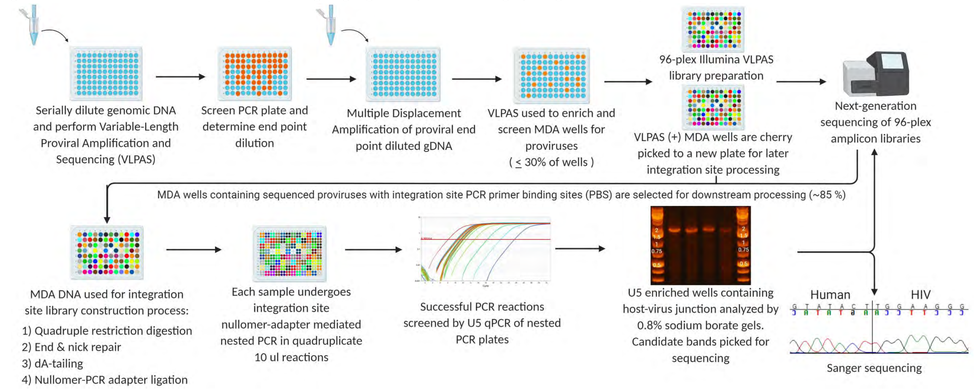
Discussion and Conclusions:
- This **I**ntegrated **P**roviral **S**equencing **A**ssay (**IPSA**) has been used to characterize intact and deleted proviruses as well as their integration sites, including replication-competent proviruses found in cell clones
- The assay can be scaled up with automation to provide in-depth characterization of proviral reservoirs that need to be targeted to achieve a cure of HIV-1 infection

Introduction

- Intact HIV proviruses integrated in the human genome allow persistence of long-term infection, despite long-term suppressive antiretroviral therapy (ART)
- However, HIV proviruses are rare in general (1 in 1,000 $CD4^+$ T cells), and their rarity makes efficient and high-throughput amplification across unknown host-virus junctions difficult for single copy targets
- Current methods are very costly and often discard more than >95% of sequencing reads as background, and also cannot pair proviruses with integration sites efficiently.
- These limitations have prevented in-depth characterization of the proviral reservoir that needs to be targeted to achieve a cure of HIV-1 infection or determine if an intervention has worked
- This improved workflow allows for enrichment and high-plexity sequencing of a single-genome, variable and near-full length HIV proviruses and their integration sites, *in which a median of 85% of sequencing reads are utilized during assembly*

Methods

Figure 1. Maco-level IPSA workflow for sequencing single-genome, near-full length, integrated HIV-1 proviruses.



What is Nullomer-mediated PCR?
- Nullomers are sequence motifs absent in a genome, which can be used to enhance specificity of linker-mediated PCR (LM-PCR)
- Reduces unintended mispriming from adapter primers in either genomes, which inhibits traditional linker-mediated PCR

Results: Virograms of Integrated Proviruses from Individual Donors

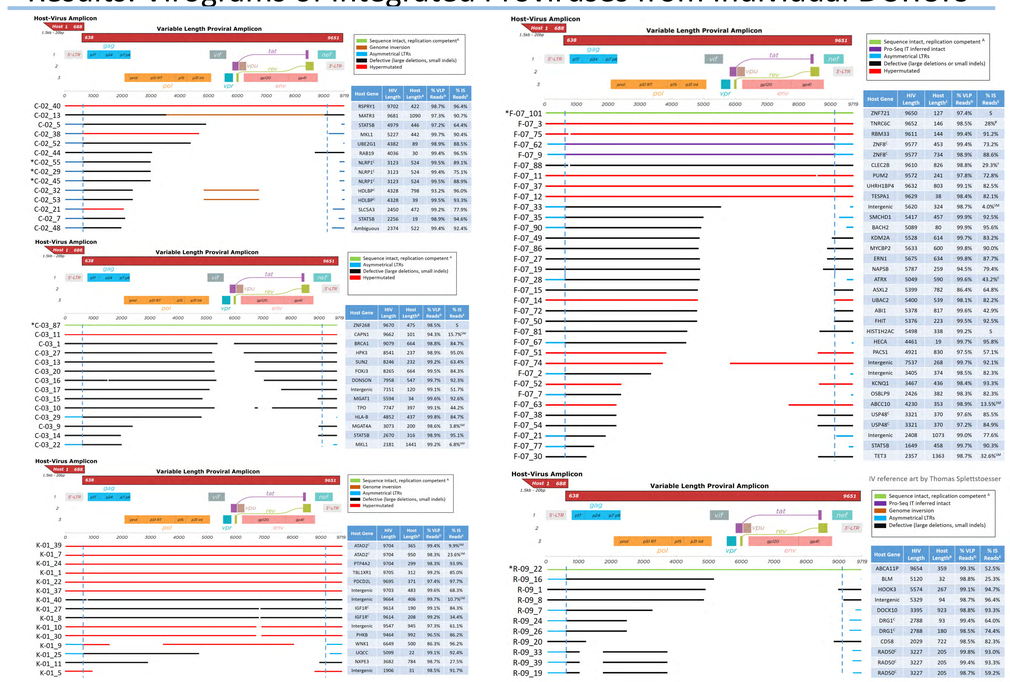


Figure 2. Donor-specific virograms depicting assembled proviruses containing matching variable length provirus and integration site amplicons. Alignment of consensus assemblies was performed using MUSCLE. Blue dashed lines denote LTR borders. *, denotes proviruses with sequence identity validated using gene-specific integration site PCR from unamplified genomic DNA. [A], replication competence determined by quantitative viral outgrowth assay (qVOA). [B], denotes length of amplified flanking host sequence. [C], denotes clonally expanded proviruses. [D], denotes the percentage of total reads utilized during assembly of variable length provirus amplicon (VLPAS). [E], denotes the percentage of total reads utilized during assembly of host-virus junction containing amplicon (IS). [S], denotes integration sites sequenced by dideoxy-sequencing only. [SM], denotes amplicons in which 10ng of a smear with no defined band by agarose gel analysis was sequenced.

Results

MDA-derived VLPAS Amplicons from Donor C-02

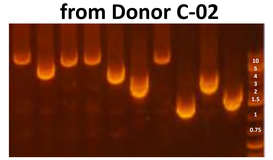


Figure 3. 0.8% sodium borate gel showing variable length proviral amplicons from separate MDA wells. Amplified proviral DNA lengths vary due to large deletions, indels, and other genomic defects, but all of the individual provirus is amplified except the *5'LTR* and 69bp of the 3'LTR *U5*.

Nullomer-mediated Integration Amplicons from Donor C-02

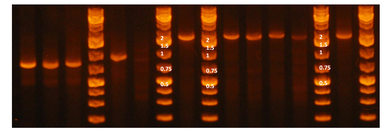


Figure 4. 0.8% sodium borate gel showing integration site amplicons. Represented above are amplicons containing enriched host-virus junctions from PCR quadruplicates of 5 different proviruses separated by GeneRuler 1kb+ ladders.

Donor ID	Number of VLPAS(+) MDA wells	Number (%) of Integration Site (IS) (+) MDA wells	Number (%) of Proviruses with Deleted IS Priming Site [A]	Number (%) of VLPAS(+) Wells Containing Mixtures [B]	Final Yield (%) of Integrated Proviral Sequences [C]	Flanking Host Sequence Length (mean, range) [D]
C-02	34	19 (55.9%)	12 (35.3%)	5 (14.7%)	14 (41.2%)	430, 19-1090
C-03	22	16 (72.7%)	1 (4.5%)	2 (9.1%)	14 (63.6%)	318, 18-664
F-07	92	68 (73.9%)	18 (19.6%)	34 (36.9%)	34 (36.9%)	495, 18-1363
K-01	28	20 (71.4%)	3 (10.7%)	5 (17.8%)	15 (53.6%)	510, 22-1137
R-09	33	15 (45.5%)	5 (15.2%)	4 (12.2%)	11 (33.3%)	328, 32-923
Median	33	19 (71.4%)	5 (15.2%)	5 (14.7%)	14 (41.2%)	430, 18-1363

Table 1. Donor-specific proviral enrichment efficiencies.
The overall percent efficiency of enriching integration sites from VLPAS(+) MDA wells across five subtype B donors was 71.4% ± 12.6%. [A] Percentage of VLPAS sequences that contain PBS deletions that would inhibit downstream integration site PCR. [B] Number of MDA wells that produced VLPAS and IS amplicons that had 5'/3'-LTR sequence identity issues. [C] Number of MDA wells that produced VLPAS and IS amplicons that did not have 5'/3'-LTR sequence identity issues. [D] Mean and range length of 5' flanking host gene sequence upstream of integration site enriched and sequenced in base pairs.

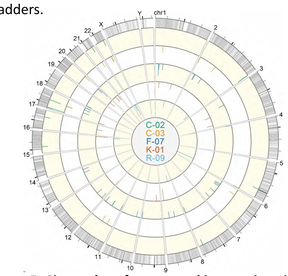


Figure 5. Circos plot of sequenced integration sites by donor. Sequencing reads containing the 5'-host-virus junction were used to extract integration sites from sequencing FASTQs for alignment against the hg38 consensus genome. Each ring depicts integration sites of sequenced proviruses from each respective donor. Length of tick marks are relative to the number of times a given integrated provirus was independently sequenced.

Discussion

- Across 5 donors, a median of 71.4% ± 12.6% of MDA wells containing proviruses had their integration sites successfully amplified and sequenced
- The very high level of specificity provided during Nullomer-mediated host-virus PCR allows high levels of multiplexing during NGS sequencing reducing the cost of sequencing
  - 96-plexing with MiSeq v2 nano kits can be used
- The ability to enrich up to 1.5kb of flanking host sequence allows for the unequivocal calling of integration sites even in ambiguous genomic regions
- Sequencing all but 69 bp of integrated proviruses allows for characterization of clonally expanded proviruses with significantly improved ability to map genomic defects (like major deletions) including those in 5' and 3' LTRs that would otherwise be unidentifiable by PCR methods requiring proviral amplification using multiple overlapping proviral amplicons

Conclusions

- This integrated proviral sequencing assay (IPSA) provides an efficient and high-throughput means of sequencing single genome integrated HIV proviruses and can be used to provide in-depth characterization of proviral reservoirs that need to be targeted to achieve a cure of HIV-1 infection
- Additionally, this method can be used to assess the impact of experimental interventions on the composition of the proviral reservoir

Acknowledgements

1. We would like to thank the study participants that made this research possible
2. Funding for this research was provided by the Internal Development Funds from the University of Pittsburgh