Long-range HIV genotyping using viral RNA and proviral DNA for analysis of HIV drug-resistance and HIV clustering

Vlad Novitsky,[a] Melissa Zahralban-Steele,[a] Mary Fran McLane,[a] Sikhulile Moyo,[b] Erik Widenfelt,[b] Simani Gaseitsiwe,[b] Joe Makhema,[b] and M. Essex[a,b]#

Harvard School of Public Health, Boston, Massachusetts, USA[a]; Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana[b]

Running Head: Long-range HIV genotyping

#Address correspondence to M. Essex, messex@hsph.harvard.edu.

Abstract

The goal of the study was to improve the methodology of HIV genotyping for analysis of HIV drug resistance and HIV clustering. Using the protocol of Gall et al. (J Clin Microbiol 50:3838–44, 2012), we developed a robust methodology for amplification of two large fragments of viral genome covering about 80% of the unique HIV-1 genome sequence. Importantly, this method can be applied to both viral RNA and proviral DNA amplification templates, allowing genotyping in HIV-infected subjects with suppressed viral load (e.g., subjects on ART). The two amplicons cover critical regions across the HIV-1 genome (including *pol* and *env*), allowing analysis of mutations associated with resistance to protease inhibitors, reverse transcriptase inhibitors (NRTIs and NNRTIs), integrase strand transfer inhibitors, and virus entry inhibitors. The two amplicons generated span 7,124 bp, providing substantial sequence length and number of informative sites for comprehensive phylogenic analysis and greater refinement of viral linkage analyses in HIV prevention studies. The long-range HIV genotyping from proviral DNA was successful in about 90% of 212 targeted blood specimens collected in a cohort where the majority of patients had suppressed viral load, including 65% of patients with undetectable levels of HIV-1 RNA load. The generated amplicons could be sequenced by different methods, such as population Sanger sequencing, single-genome sequencing, or next-generation ultra-deep sequencing. The developed method is cost-effective – the cost of the long-range

35 HIV genotyping is under $150 per subject (by Sanger sequencing), and has the
36 potential to enable the scale-up of public health HIV prevention interventions.

Introduction

HIV genotyping is a critical tool for antiviral drug-resistance testing that has revolutionized HIV care and advanced HIV-related research. Routine antiretroviral (ARV) drug-resistance testing is useful in choosing an optimal treatment regimen and monitoring its efficiency in clinical practice (1-12). HIV genotyping has been used successfully in research on HIV transmission clusters and HIV transmission dynamics (13-35).

Initial broadly used ARV regimens included combinations of nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs). To monitor the emergence of drug-resistant mutations associated with NRTIs and NNRTIs, HIV genotyping targeted viral sequences spanning an approximately 1,000 to 1,300 bp region of the HIV-1 genome, encoding viral protease and partial reverse transcriptase (RT) using viral RNA as a template for amplification. While the RNA-based approach works well in ART-naïve individuals, it is less successful if levels of viral replication are low, such as in individuals on ART. The sequence length of traditional RNA-based HIV genotyping for drug resistance is relatively short, and does not cover the HIV-1 region encoding viral integrase, or the viral envelope, hindering analysis of drug-resistant mutations associated with integrase strand transfer inhibitors, or entry inhibitors. The global scale-up of ARV treatment and successful introduction of integrase strand transfer inhibitors and entry inhibitors into clinical trials and

Journal of Clinical Microbiology

clinical practice necessitates modification of traditional methods of HIV genotyping.

Two commercial genotyping assays, ViroSeq HIV-1 from Abbott Molecular and TruGene HIV-1 from Siemens Molecular Diagnostics, have been widely used for analysis of HIV-1-associated drug resistance. Both genotyping kits were extensively tested and validated (36-45). While the ViroSeq HIV-1 kit is still on the market, Siemens discontinued selling and supporting the TruGene HIV-1 kit in 2014. ViroSeq HIV-1 kit covers the entire protease coding region and the RT region encoding the first 320 amino acids. The TruGene HIV-1 sequences span the protease (amino acids 4 to 99) and RT (amino acids 40 to 240) coding regions. The CDC supplies WHO-designated and CDC-supported PEPFAR Genotyping Laboratories with the ATCC HIV-1 Drug Resistance Genotyping Kit (46) for drug-resistance testing. Many experienced genotyping laboratories have developed their own “in-house” amplification and sequencing protocols (11, 47-56), including identification of minor viral variants that are normally missed by commercial genotyping kits (57-61). All of these approaches generally include smaller and more restricted regions for testing HIV-1 drug resistance.

Recently, the protocol developed by Gall et al. (62) has enabled high-throughput near full-length HIV-1 genome genotyping in individuals infected with multiple HIV-1 subtypes. This method has become the cornerstone of the PANGEA-HIV Consortium (Phylogenetics and Networks for Generalized HIV Epidemics in

Africa; http://www.pangea-hiv.org/) aiming to establish worldwide scientific collaborations across phylogenetics, public health and epidemiology. The Gall protocol (62) targets viral RNA as a template for cDNA synthesis and amplification, and is very robust and reproducible when HIV-1 RNA load is high (e.g., above 10,000 cps/ml). However, specimens with levels of HIV-1 RNA below 1,000 cps/ml, or lower thresholds, present a substantial challenge, and few of those samples could be genotyped. This is consistent with the commercially available assays for HIV drug-resistance genotyping, ViroSeq and TrueGene, which are unable to genotype specimens with low or undetectable HIV-1 RNA load.

In HIV infection, proviral DNA presents an alternative template for HIV genotyping. Drug-resistance mutations detected in viral RNA from plasma and proviral DNA from PBMCs or dried blood spots (DBS) show a substantial correlation in treated patients, suggesting that either compartment is suitable for the detection of mutations as a virological guide for clinical care (63-65).

It is known that amplified HIV sequences and sequences from proviral DNA could have substantial numbers of G-to-A transitions. Such an inordinate number of identical guanine-to-adenine transitions is a retroviral signature known as hypermutation (66-69). G-to-A hypermutations produce multiple stop codons and reduce HIV replication, leading to an evolutionary dead end. It is an innate host intra-cellular defense mechanism. Host factors APOBEC3F and APOBEC3G

induce G-to-A substitutions in reverse transcribed nascent retroviral DNA (70). G-to-A hypermutations play an important role in the evolution of antiretroviral drug resistance (71, 72) and could be associated with ART failure (73). The extent of G-to-A hypermutations is not associated with levels of HIV-1 RNA (74), although hypermutations are frequent in virologic controllers (75). For sequence quality control, it is important that G-to-A hypermutations are not products of PCR amplification (76).

In this study, we present a technique for long-range HIV genotyping using proviral DNA, as well as viral RNA, as templates for amplification and sequencing. The outcome of the long-range HIV genotyping are two large fragments that span about 80% of the unique full-length HIV-1 genome sequence. The proposed technique is a modification of the method by Gall et al. (62). The key modifications include using (1) a proviral DNA template, (2) extra-round PCR, (3) selection of robust primers, and (4) modified running conditions. To illustrate the potential utility of the long-range HIV genotyping, the technique was applied to a set of specimens collected in Botswana.

Materials and Methods

Study subjects

The technique of long-range HIV genotyping was applied to specimens collected within three Botswana-Harvard AIDS Institute Partnership (BHP) studies performing viral genotyping: An HIV Prevention Program for Mochudi, Botswana (Mochudi Prevention Project, or MPP; R01 AI083036; PI: M. Essex) (34, 35), the GWAS on Determinants of HIV-1 Subtype C Infection study (RC4 AI092715; PI: M. Essex), and the Botswana Combination Prevention Project (BCPP, or Ya Tsie; U01 GH000447; PI: M. Essex) (77). All study subjects signed a consent form and donated a blood sample for viral genotyping. The first large fragment of HIV-1 genome, "Amplicon 1," was amplified and sequenced in 649 HIV-infected subjects (single sequence per subject) originating from eight geographic localities in Botswana: Digawana, Gaborone, Lobatse, Mochudi, Molapowabojang, Molepolole, Otse, and Ranaka. The second large fragment of HIV-1 genome, Amplicon 2, was amplified and sequenced in 90 subjects (work is still in progress) originating from Mochudi, Molapowabojang, Otse, and Ranaka.

A total of 212 specimens from the BCPP study were used for analysis of genotyping efficiency. These samples were collected consecutively from subjects participating in the BCPP baseline household survey (20% of households) in the first four communities, Ranaka, Digawana, Molapowabojang and Otse, from November 2013 to June 2014. Specimens from two other studies, MPP and

GWAS, represented subsets successfully amplified in the past for a shorter region of HIV-1 *env* gp120, V1C5. Due to potential selection bias, the MPP and GWAS specimens were not used in analysis of genotyping efficiency.

For sequences used in this study, the accession numbers are KR860607–KR861255 for 649 amplicon 1 sequences and KR861256–KR861345 for 90 amplicon 2 sequences.

Analyzed regions of the HIV-1 genome

The extent of HIV clustering was analyzed by using the following sub-genomic regions across the HIV-1 genome: (1) Amplicon 1 spanning the 3'-end of *gag* and almost the entire *pol* and corresponding to "amplicon 2" in Gall et al. (62), nt positions 1,486–5,058; (2) Amplicon 2 spanning *vpu*, *env*, *nef* and TATA-box in the U3 region of 3'-LTR and corresponding to "amplicon 4" in Gall et al. (62), nt positions 5,967–9,517; (3) ViroSeq – a partial *pol* sequence spanning the region encoding HIV-1 protease and the first 335 amino acids of reverse transcriptase, and corresponding to the sequence produced by ViroSeq (39, 44, 45, 78), nt positions 2,253–3,554; and (4) V1C5 – a partial *env* sequence spanning the region encoding gp120 V1C5 (34, 79, 80), nt positions 6,570–7,757. In addition, the following combinations of the sub-genomic regions included concatenated Amplicon 1 + Amplicon 2 and Amplicon 1 + V1C5. All multiple sequence codon-based alignments were generated using MUSCLE (81) in MEGA6 (82).

To prevent sample contamination, basic lab rules were enforced and included controlled flow of specimens, use of dedicated areas and equipment, proper training, and routine implementation of QA/QC program.

Analysis of drug resistance

The WHO 2009 list of mutations for surveillance of transmitted drug-resistant HIV strains was used for analysis of PI-, NRTI-, and NNRTI-associated mutations (2). The list of PI-associated mutations included 40 mutations at 18 positions across protease. The list of NRTI mutations included 34 mutations at 15 positions in RT. The list of NNRTI mutations included 19 mutations at 10 positions across RT. The IAS-USA list (2014 update) of drug-resistance mutations in HIV-1 was used for analysis of integrase strand transfer inhibitors (20 mutations at 11 positions in integrase) and entry inhibitors (10 mutations at 7 positions in gp41) (3).

APOBEC-induced hypermutations

The APOBEC-induced hypermutations were assessed by Hypermut (83) at the LANL HIV Database (http://www.hiv.lanl.gov/). The HIV-1C consensus sequence was used as a reference. Two parameters related to APOBEC-induced hypermutations were analyzed, adjusted hypermutations and hypermuation ratio. The adjusted hypermutations were expressed as a number of identified hypermuations adjusted by sequence length. The hypermutation ratio was computed as the ratio between weighted mutations (matched mutation out of potential mutations) and weighted controls (control mutations out of potential

controls), and was derived as a statistical outcome of the Hypermut package (83).

Definition of HIV cluster

The HIV cluster was defined as a viral lineage that gives rise to a monophyletic sub-tree of the overall phylogeny with strong statistical support. The bootstrapped Maximum Likelihood (ML) method (84-86) was used to determine the statistical support of clusters. The four bootstrap thresholds for identification of HIV clusters were ≥0.7, ≥0.8, ≥0.9, and =1.0. A viral lineage (group, sub-tree) with at least two viral sequences and specified statistical support was considered to be an HIV cluster. Clusters were identified using a depth-first algorithm (87, 88), a method for traversing or searching tree or graph data structures starting from the root. This approach eliminated double-counting of viral sequences in clusters when clusters had internal structure with strong support.

Confidentiality

The sharing of data, including generated HIV sequences, with the scientific community for the purpose of research is of key importance in ensuring continued progress in our understanding of how to contain the HIV epidemic. The generated HIV sequences were deposited to GenBank. The confidentiality of study subjects was protected by re-coding of deposited HIV sequences at the country level (no community or village data).

Journal of Clinical Microbiology

Phylogenetic inference

The ML tree inference was implemented in RAxML (89, 90) under the GAMMA model of rate heterogeneity. The statistical support for each node was assessed by bootstrap analysis from 100 bootstrap replicates performed with the rapid bootstrap algorithm implemented in RAxML (89). The RAxML runs were performed using RAxML ver.8.0.20 at the high-performance computing cluster Odyssey (https://rc.fas.harvard.edu/resources/odyssey-architecture/) at the Faculty of Arts and Sciences, Harvard University (https://rc.fas.harvard.edu/).

Proportion of HIV-1C sequences in clusters

To test whether the extent of HIV clustering is associated with any sub-genomic region, the proportion of clustered sequences was compared between long (Amplicon 1, Amplicon 2, concatenated Amplicons 1 + 2, and concatenated Amplicon 1 + V1C5) and short (ViroSeq and V1C5) HIV-1C sequences. The proportion of HIV sequences in clusters was estimated at the bootstrap thresholds for cluster definition from 0.7 to 1.0 under ML inference.

Statistical analysis

The HIV sequences in clusters were enumerated with PhyloPart v.2 (88) using bootstrap thresholds 0.7, 0.8, 0.9 and 1.0. All confidence intervals of estimated proportions are asymptotic 95% binomial confidence intervals (95% CI) computed with the prop.test() function in R version 3.1.2 (91). Comparisons of continuous outcomes between two groups were performed using the Wilcoxon

Rank Sum test. P-values less than 0.05 were considered statistically significant. All reported p-values are 2-sided. Proportions of viral sequences in clusters between targeted loci were compared by McNemar's test in R, and p-values less than 1.0E-04 were considered statistically significant. All plots were produced in R. All figures were finalized in Adobe Illustrator CS6.

Results

Long-range HIV genotyping

The original protocol for near full-length HIV-1 genome genotyping by amplification of four large overlapping amplicons in a single round of RT-PCR using viral RNA as a template was developed by Gall et al. (62). The protocol by Gall et al. (62) is robust and highly reproducible for samples with relatively high HIV-1 RNA. However, specimens with low, or undetectable levels of HIV-1 RNA presented a substantial challenge for amplification from viral RNA. Attempts to apply the original protocol to proviral DNA produced large number of non-specific products evident from smeared "ladders" on the electrophoretic gel (data not shown).

The modifications of the Gall et al. (62) protocol included the following steps: (1) focus on 2 (amplicons #2 and #4 in the original protocols) instead of 4 amplicons; (2) extra round of PCR: the amplified ~8.3 kb product was used as a template for

263 the second round of PCR; (3) highly-specific primers for the first round of PCR
264 and for cDNA synthesis (for viral RNA templates); and (4) modified PCR running
265 conditions.
266
267 The rationale for focusing on two instead of four amplicons was driven by a
268 balance between sequencing data and cost. Two amplicons have lengths of
269 3,574 bp and 3,550 bp (HXB2 nt length), which cumulatively covers about 80% of
270 the unique full length HIV-1 genome sequence (Figure 1). The first amplicon
271 (corresponds to amplicon #2 in Gall et al. paper (62)) spans partial *gag* at the 3'-
272 end and almost the entire *pol* (HXB2 nt positions 1,486–5,058). The second
273 amplicon (corresponds to amplicon #4 in Gall et al. paper (62)) spans *vpu*, *env*,
274 *nef* and 3'-LTR up to TATA-box in the U3 region (HXB2 nt positions 5,967–
275 9,517).
276
277 Amplification of a large fragment spanning almost the entire HIV-1 genome (Fig.
278 1; hatched bar) was introduced as the first round of PCR (RT-PCR for RNA
279 template). Primers OFM19 and SK145 (Supplementary Table S1) substantially
280 increased specificity of viral amplification. For proviral DNA template, the 1st
281 round of PCR was run with primers SK145 and OFM19. The PrimeSTAR GXL
282 DNA Polymerase (Takara; cat. #R050A) was used in 30 amplification cycles with
283 annealing temperature at 62 °C (98°C for 15 sec ⇢ 62°C for 30 sec ⇢ 68°C for 9
284 min cycling). For RNA template, cDNA synthesis with primer OFM19 was
285 followed by PCR with primers SK145 and OFM19 in a single tube RT-PCR. The

SuperScript III One-Step RT-PCR High Fidelity enzyme (Invitrogen; cat. #12574035) was used with cDNA synthesis step of incubation at 50°C for 60 min and 94°C for 2 min followed by 30 cycles of amplification in the 1st PCR round (98°C for 15 sec ⇢ 62°C for 30 sec ⇢ 68°C for 9 min cycling). In cases of specimens from subjects with low viral load, a lower annealing temperature between 58 °C and 60 °C was used in the 1st round.

The 1st round product was used as template in two separate 2nd round PCRs with specific primers (Table S1) to obtain amplicon 1 and amplicon 2 (Fig. 1; gray bars). The PrimeSTAR GXL DNA Polymerase (Takara; cat. #R050A) was used in 30 amplification cycles with annealing temperature at 62 °C (98°C for 15 sec ⇢ 62°C for 30 sec ⇢ 68°C for 4 min cycling). No additional extension step was performed at the end of the run.

After standard purification with USB ExoSAP-IT (92) (Affymetrix, cat. #782011ML), amplicon 1 was subject for direct Sanger sequencing on both strands using a total of 12 sequencing primers (Supplementary Table S2). In about 30% of cases direct sequencing of amplicon 1 failed apparently due to heterogeneity of amplified product. These cases were cloned, and Sanger sequenced on both strands. All amplicon 2 products were cloned before Sanger sequencing on both strands with a total of 12 sequencing primers (Supplementary Table S2). Direct Sanger sequencing was performed on the ABI 3730 DNA Analyzers using BigDye technology.

High diversity of HIV presents a challenge for direct Sanger sequencing. Samples collected during the early stage of HIV infection are relatively homogeneous (in case of transmission of a single HIV variant). In contrast, samples obtained from chronically infected individuals are likely to include a heterogeneous pool of viral quasispecies. High heterogeneity of viral quasispecies combined with numerous insertions and deletions (indels) could result in low quality of the directly sequenced specimens. In this case cloning may be considered, as an alternative solution to direct sequencing. If time of HIV infection is unknown, the diversity of the targeted region, or sub-region, could guide the initial sequencing strategy. The amplicon 1 spans a relatively conserved region of the HIV-1 genome. In contrast, amplicon 2 includes the most variable regions of the HIV-1 genome, with multiple indels. Our preliminary results suggest that applying cloning to about 30% of amplicon 1 sequences and to 100% of amplicon 2 sequences is the most efficient sequencing strategy to overcome the complexity of HIV quasispecies. The goal of this study was to obtain a single HIV sequence per subject. Therefore, generation of a single amplicon 1 and single amplicon 2 sequence was considered a success. If a study aimed to address multiplicity of HIV infection, or diversity of viral quasispecies, multiple sequences (e.g., 20 per targeted region per subject) could be generated by appropriate amplification methods.

Cloning was performed by PCR Cloning Kit (NEB, cat. #E1202S) using Fast-Media Amp XGal (Invivogen, cat. #fas-am-x). Ligation, transformation and plating was performed according to manufacturer's instructions. Colonies were checked for insert by EmeraldAmp GT PCR Master Mix (Takara, cat. #RR310A), and submitted to GENEWIZ (http://www.genewiz.com/public/DNA-sequencing-services.aspx) for colony sequencing. A list of sequencing primers used with clones are presented in Supplementary Table S3.

All sequence contigs were assembled by SeqScape v.2.7.

Troubleshooting

Some amplification issues during long-range HIV genotyping such as lack or insufficient amplification, over-amplified product, or presence of multiple bands could be resolved by troubleshooting. The initial amplification results could guide troubleshooting. A lack of visible bands (or weak bands) on the gel after second round PCR could be resolved by decreasing annealing temperature in the first round PCR to 58 °C, and/or increasing the number of cycles in the first round PCR to 35, or increasing the amount of RNA template (e.g., up to 5 µl). The over-amplified products could be overcome by reducing the number of cycles in the first round PCR to 25, or in the second round PCR to 20-25, or by decreasing the amount of input template. Multiple bands on the gel could be resolved by either extracting the right size band from the gel using Wizard SV Gel and PCR Clean-

Up System (Promega, cat. #A9281), or re-running the first round PCR in replicates and with serial dilutions.

HIV genotyping results

Amplicon 1 was amplified and sequenced in 649 HIV-infected subjects (single sequence per subject), while amplicon 2 was amplified and sequenced in 90 subjects.

The long-range HIV genotyping from proviral DNA was applied to 212 specimens collected from subjects participating in the BCPP baseline household survey in the first four communities, Ranaka, Digawana, Molapowabojang and Otse, from November 2013 till June 2014. The distribution of amplified and sequenced samples from proviral DNA is presented in Table 1. Amplicon 1 was successfully amplified in 89.6% (95% CI 84.5% to 93.2%) cases. Viral sequences were obtained for all amplified samples. The majority of amplified Amplicon 1 sequences, 144 of 167 (86.2%; 95% CI 79.8% to 90.9%) were obtained by direct Sanger sequencing. In 23 cases (12.1%; 95% CI 8.0% to 17.8%) amplicon 1 sequences obtained by direct Sanger sequencing had gaps that did not exceed 10% of the amplicon 1 length. Cloning followed by Sanger sequencing helped to resolve gaps in all 23 cases.

Levels and distribution of HIV-1 RNA load in amplified and non-amplified specimens from proviral DNA were of particular interest. HIV-1 RNA load data

were available for a subset of 202 HIV-positive subjects from BCPP study. The proportion of successfully amplified cases was 89.6% (95% CI from 84.3% to 93.3%; Table 1). Sequences were obtained for all amplified products. Partial sequences (less than 10% of missing data) were obtained in 11.6% (95% CI 7.5% to 17.4%) of amplified cases.

Distribution of HIV-1 RNA load among specimens amplified and sequenced from proviral DNA is presented in Figure 2. The histogram shows the distribution of HIV-1 RNA among 202 specimens with available viral load data (both amplified and failed specimens). The distribution indicates that high proportion of subjects participating in the baseline household survey in four BCPP communities had suppressed levels of HIV-1 RNA, primarily due to high proportion of HIV-infected individuals receiving ART. In fact, 71.3% (95% CI 64.4% to 77.3%) of HIV-infected subjects had HIV-1 RNA below 1,000 cps/ml including 65.3% (95% CI 58.3% to 71.8%) with undetectable HIV-1 RNA below 40 cps/ml. Distributions of HIV-1 RNA were similar among amplified (n=181) from proviral DNA and failed (n=21) specimens (Figure 2, pie charts).

Amplification and sequencing of amplicon 2 was completed for 90 subjects. Given that the first round (RT-) PCR product is used for amplification of both amplicons 1 and 2, obtaining amplicon 1 suggests a successful amplification of amplicon 2. Amplification of the overlapping product designated as “amplicon 3”

in Gall et al. paper (62) should be possible, as the first round PCR product completely covers “amplicon 3”. This strategy has not been explored yet.

Overall, the long-range HIV genotyping from proviral DNA (for amplicon 1) was successful in about 90% of targeted blood specimens collected in a cohort where majority of patients had suppressed viral load including 65% patients with undetectable HIV-1 RNA load.

Amplification from viral RNA

To assess the utility of long-range HIV genotyping for amplification and sequencing from viral RNA template, we performed a small-scale genotyping (n=32) from viral RNA in plasma (Table 2). HIV-1 RNA load was available for 31 of these samples, and was above 1,000 cps/ml in 29 cases. A subset of 23 specimens were successfully amplified and sequenced. Interestingly, two of nine specimens that failed amplification from proviral DNA (HIV-1 RNA load 1,576 cps/ml and 5,620 cps/ml), were successfully amplified from viral RNA.

Nine failed cases included one sample with unknown and eight specimens with available viral load. Among the later group, two samples had viral load below 1,000 cps/ml (181 and 497 cps/ml), 5 samples had viral load between 1,191 and 8,528 cps/ml, and one sample had viral load of 156,821 cps/ml. The later failed sample with high viral load also failed amplification from proviral DNA, apparently suggesting an intrinsic problem with mismatch of amplification primers.

Analysis of mutations associated with antiretroviral drug-resistance

Amplicon 1 covers almost the entire HIV-1 *pol* gene and allows analysis of mutations associated with antiretroviral drug-resistance to Protease inhibitors (PI), Nucleoside Reverse Transcriptase inhibitors (NRTI), non-Nucleoside Reverse Transcriptase inhibitors (NNRTI), and Integrase strand transfer inhibitors. Amplicon 2 covers the entire HIV-1 *env* gene and allows analysis of mutations associated with drug-resistance to virus entry inhibitors.

To illustrate the validity of long-range HIV genotyping for analysis of mutations associated with antiretroviral drug-resistance, we estimated drug-resistance profiles within two groups of specimens originating from MPP and BCPP studies, respectively. Amplicon 1 sets included 192 MPP sequences and 186 BCPP sequences. Amplicon 2 sets included 35 MPP and 55 BCPP sequences.

Despite relatively rare use of Protease inhibitors in Botswana, mutations associated with drug-resistance to PI were detected at five positions in Protease: D30N (5% MPP and 6% BCPP), M46I (5% MPP and 10% BCPP), G73S (10% MPP and 9% BCPP), I85V (1% MPP), and N88S (1% BCPP). The encoding analysis revealed that all 22 D30N mutations were caused by GAT (Asp) to AAT (Asn) substitution, 26 of 27 M46I mutations were due to ATG (Met) to ATA (Ile) substitution, and 35 of 36 G73S mutations were found because of GGT (Gly) to

443 AGT (Ser) substitution. Thus, it is likely that the majority of identified mutations in
444 the Protease gene were caused by G-to-A hypermutations.
445
446 NRTI and NNRTI have been part of National antiretroviral program in Botswana
447 since 2002. Viral mutations associated with drug-resistance to NRTI were found
448 at the following positions across RT: M41L (1% BCPP), D67N (1% MPP and 2%
449 BCPP), K70R (1% BCPP), K70E (1% BCPP), V75M (1% BCPP), M184V (2%
450 BCPP), M184I (16% BCPP), and T215Y (1% BCPP). Almost all M184I (60 out of
451 61) mutations were caused by ATG (Met) to ATA (Ile) substitution. Mutations to
452 NNRTI were observed at multiple RT positions and demonstrated low frequency:
453 K101E (1% MPP and 1% BCPP), K103N (1% MPP and 3% BCPP), K103S (1%
454 MPP and 1% BCPP), Y181C (1% BCPP), Y188C (1% BCPP), G190A (1%
455 BCPP), G190S (1% BCPP), G190E (1% MPP and 1% BCPP), and P225H (1%
456 BCPP).
457
458 HIV mutations associated to Integrase strand transfer inhibitors were detected at
459 three positions in Integrase: L74M (1% MPP), T97A (3% MPP and 1% BCPP),
460 and E138K (3% MPP and 6% BCPP). Mutations to entry inhibitors were found at
461 the following positions in gp41: G36S (24% BCPP) and V38M (2% BCPP). All 13
462 out of 55 G36S mutations were caused by a switch from GGT (Gly) to AGT (Ser),
463 which is a likely effect of G-to-A hypermutation.
464

G-to-A hypermutations

Presence of G-to-A hypermutations in the products amplified from proviral DNA is not surprising, as massive APOBEC-induced G-to-A transitions in retroviruses are well recognized as a key innate defense by host. Distribution of identified APOBEC-induced hypermutations in HIV-1C sequences amplified from proviral DNA is presented in Figure 3.

The sequence length among the 649 cases analyzed differed from 3,190 bp to 3,625 bp. Therefore, the number of potentially G-to-A hypermutated sites (compared to the HIV-1 subtype C consensus sequence) was adjusted for the sequence length and expressed as a proportion. Two cutoff values, 0.02 and 0.05, were used to demonstrate the proportion of viral sequences in the analyzed set with potential G-to-A hypermutations. Figures 3A and 3C demonstrate distribution of G-to-A hypermutations adjusted by sequence length for amplicons 1 (n=649) and 2 (n=90), respectively. For example, 125 of 649 amplicon 1 sequences (19.3%; 95% CI 16.3% to 22.6%) and 37 of 90 amplicon 2 sequences (41.1%; 95% CI 31.0% to 52.0%) exceeded the 0.02 level of adjusted hypermutations. Or, 23 of 649 amplicon 1 sequences (3.5%; 95% CI 2.3% to 5.4%) and 11 of 90 amplicon 2 sequences (12.2%; 95% CI 6.6% to 21.2%) were above the 0.05 level of adjusted hypermutations. Figures 3B and 3D show distribution of the hypermutation ratio estimated by Hypermut (83). Both metrics indicate presence of APOBEC-induced hypermutations among amplicon 1 and amplicon 2 sequences amplified from proviral DNA.

The majority of viral sequences with antiretroviral mutations had high rates of APOBEC-induced hypermutations, suggesting association between hypermutations and drug-resistant mutations. Distribution of APOBEC-induced hypermutations among 36 MPP sequences with drug-resistant mutations within Protease, RT and Integrase is presented in Supplementary Table S4, while hypermutations among 46 BCPP sequences with drug-resistant mutations are presented in Supplementary Table S5. It is evident that many hypermutated sequences have multiple drug-resistant mutations due to G-to-A transition.

HIV-1C sequences with identified drug-resistant mutations demonstrated high rate of APOBEC-induced hypermutations. Horizontal boxplots in Figure 3 indicate distribution of adjusted number of hypermuations (Figs. 3A and 3C) and hypermutation ratio (Figs. 3B and 3D) among viral sequences with drug-resistant mutations in relation to the distribution of hypermutation parameters in the entire set of sequences. Comparison of these distributions indicates association between APOBEC-induced hypermutations and drug-resistant mutations.

To further address how G-to-A hypermutations can affect drug-resistance mutations, we compared hypermutations between two groups, *with* and *without* M184I mutations. Individuals with M184I mutations have higher adjusted numbers of hypermutations (Figure 4A) and higher hypermutation ratios (Figure 4B). Summary statistics are presented at the bottom of Figure 4. The difference

Journal of Clinical Microbiology

was highly significant for both comparisons (p-value < 0.0001, Wilcoxon Sum Rank test).

HIV cluster analysis

To exemplify utility of the long-range HIV genotyping for analysis of HIV transmission dynamics and viral linkage, we compared extent of clustering within viral sequences generated in this study. The concatenated amplicons 1 + 2 span over about 80% of unique HIV-1 genome sequence. In this study the number of matched amplicons 1 + 2 sequences was limited to 83. The extent of HIV clustering within this small set was compared for three long loci, amplicon 1 (3,574 bp), amplicon 2 (3,550 bp), and concatenated amplicons 1 + 2 (7,124 bp), and two short loci, ViroSeq (1,263 bp) and V1C5 (1,188 bp).

The proportion of clustered HIV sequences was compatible within long loci (Table 3). For example, at bootstrap support of ≥0.80, proportion of clustered HIV sequences was 0.265, 0.289, and 0.337 for amplicon 1, amplicon 2 and concatenated amplicons 1 + 2, respectively. For short loci ViroSeq and V1C5 at the same bootstrap support of ≥0.80, proportion of HIV sequences in clusters was 0.157 and 0.145, respectively, The proportion of clustered sequences seemed to be higher for long regions than for short loci, although the difference reached significance of the 0.05 level in selected comparisons only.

A larger set of available HIV-1C sequences, n=547, included matched viral sequences for amplicon 1 and V1C5 region of gp120 generated in our previous studies (34, 35, 80). Clustering patterns were compared for two long loci, amplicon 1 and concatenated amplicon 1 + V1C5, and for two short regions across the HIV-1 genome, ViroSeq, and V1C5. Similarly to the small set of HIV sequences (n=83), proportion of clustered sequences in the large set (n=547) was higher for long loci than for short regions (Table 4).

To address whether longer loci are associated with higher extent of HIV clustering, we analyzed congruent (++ and −−) and discordant (+− and −+) clustering between different combinations of long and short HIV-1C sequences (Figure 5). At all bootstrap thresholds from 0.70 to 1.0, amplicon 1 and concatenated amplicon 1 + V1C5 demonstrated higher extent of HIV clustering than ViroSeq and V1C5 sequences (significant difference is highlighted by gray squares on background).

Estimated cost of the long-range HIV genotyping

The estimated cost for amplification and Sanger sequencing of both amplicons 1 and 2 in this study was $137.50 for proviral DNA and $139.75 for viral RNA. This includes cost of reagents, materials and disposables for nucleic acid isolation, amplification (RT-PCR and PCR), purification of amplicons, cloning up to 30% of amplicon 1 products and 100% of amplicon 2 products, and Sanger sequencing. The estimated cost does not include labor, training, supervision, or indirect costs.

Discussion

A technique for long-range HIV genotyping from both viral RNA and proviral DNA has been presented. Using proviral DNA as a template, long-range HIV genotyping was successfully performed in one of the BCPP cohorts with a high proportion of virologically suppressed individuals with the success rate at about 90%.

Both clinical trials and clinical care could benefit from routine use of long-range HIV genotyping. The proposed long-range HIV genotyping has potential to improve methodology of drug-resistance testing, broaden spectrum of monitored ARV's, and enable surveillance of transmitted drug-resistance. Mapping of HIV transmission networks performed by long-range genotyping could help reveal transmitting viral variants in Treatment-as-Prevention studies. Implementation of long-range HIV genotyping could allow greater refinement of viral linkage analyses in HIV prevention studies, and better coordination with evaluation of prevention strategies based on such interventions as behavior change, male circumcision, and treatment as prevention. The cost-effective long-range HIV genotyping technique has the potential to enable the scale-up of public health HIV prevention interventions across communities.

In this study we demonstrated that long-range HIV genotyping using proviral DNA could be successfully applied to a population with high level of ART-experienced individuals that normally presents a challenge for HIV genotyping from viral RNA. In fact, more than 70% of individuals in the BCPP cohort participating in the baseline household survey in the first four communities had HIV-1 RNA below 1,000 cps/ml including 65% with undetectable levels of HIV-1 RNA below 40 cps/ml. The ongoing scale up of National ARV programs in Africa leads to growing number of individuals with suppressed HIV-1 RNA load across communities. The presented technique of long-range HIV genotyping from proviral DNA should alleviate and enable analysis of HIV drug-resistance and HIV transmission dynamics using samples collected from individuals on ART.

The comprehensive strategy of HIV genotyping could include two steps. First, viral RNA template for amplification could be targeted, if HIV-1 RNA load is relatively high (e.g., above 1,000 cps/ml). If amplification is successful, there is no need for proviral DNA. However, if amplification from viral RNA does not work, or HIV-1 RNA load is below 1,000 cps/ml, or undetectable, using proviral DNA is a logical step toward successful HIV genotyping. A complimentary use of both viral RNA and proviral DNA templates could be an efficient and cost-effective approach for HIV genotyping.

The long-range HIV genotyping enables analysis of drug resistance (both transmitted and acquired) for all major groups of ARVs, including protease

inhibitors, NRTI, NNRTI, integrase strand transfer inhibitors, and virus entry inhibitors. A comprehensive analysis of HIV drug-resistance is feasible due to a long sequence length of generated amplicons that span the HIV-1 *pol* and *env* genes. While long-range HIV genotyping is able to identify drug-resistant mutations, it cannot distinguish transmitted and acquired drug-resistant mutations without additional information on sampling strategy and stage of HIV infection. For example, drug-resistant mutations identified in individuals during early stages of HIV infection (e.g., in seroconverters) are likely to represent transmitted drug-resistant mutations. In contrast, specimens collected in chronic HIV infection, or from individuals on ART, are more likely to be associated with acquired HIV drug resistance.

The G-to-A hypermutations observed in sequences amplified from proviral DNA and their relation to drug-resistant mutations should be interpreted cautiously in the context of specific study. Our data suggest that G-to-A hypermutations are likely to contribute to critical drug-resistant mutations, such as M184I. We recommend controlling viral sequences generated from proviral DNA for the adjusted number of hypermutations and/or hypermutation ratio using the online package Hypermut (83) at the LANL HIV Database (http://www.hiv.lanl.gov/), and the subtype consensus sequence, as a reference. Based on IQR boundaries in our data, the adjusted number of hypermutations above 2.8% (1st IQR in individuals with M184I) indicates a hypermutated sequence, and below 0.5% (3rd IQR in individuals without M184I) suggests a non-hypermutated sequence.

Whether HIV-associated drug-resistant mutations should be interpreted differentially depending on the extent of G-to-A hypermutations still needs to be addressed in future studies.

The study showed utility of the long-range genotyping for analysis of HIV transmission dynamics and HIV clustering. A higher extent of clustering for longer HIV sequences in this study corroborates well with results of our recent study (93) that used a set of near full-length HIV-1C sequences from LANL HIV Database (http://www.hiv.lanl.gov/). Longer HIV sequences are more informative for HIV cluster analysis due to a larger number of informative sites (93). The technique of long-range HIV genotyping allows using amplicon 1 and amplicon 2 sequences either separately or in concatenation for a powerful cluster analysis. The concatenated amplicons 1 and 2 span about 80% of the unique HIV-1 genome sequence, and could be considered as a cheaper alternative to near full-length HIV-1 sequencing. A combination of conserved (amplicon 1) and variable (amplicon 2) regions could help to deal with different and/or unknown stages of HIV infection in an analyzed set of viral sequences. The choice of particular bootstrap value and filtering by the threshold of pairwise distances and/or internode certainty (94, 95) could depend on specific scientific question and take into account specifics of analyzed set of sequences including sampling density (35).

In summary, the presented technique of long-range HIV genotyping using viral RNA and proviral DNA can help in analysis of HIV drug-resistance and HIV clustering in cohorts and populations on ART when amplification from viral RNA is unsuccessful due to low levels of HIV-1 RNA load.

Acknowledgements

We are very grateful to all participants of the BHP projects in Botswana. We thank CDC and the Botswana Ministry of Health for their collaboration. The Mochudi Prevention Project in Botswana was supported and funded by NIH grant R01 AI083036, *An HIV Prevention Program for Mochudi, Botswana*. The GWAS on Determinants of HIV-1 Subtype C Infection study was supported and funded by NIH grant RC4 AI092715. The Botswana Combination Prevention Project, BCPP, has been supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the United States Centers for Disease Control and Prevention under the terms of grant number U01 GH000447. We thank Lendsey Melton for excellent editorial assistance.

References:

1. **Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, Wensing AM, Richman DD.** 2013. Update of the drug resistance mutations in HIV-1: March 2013. *Top Antivir Med* **21:**6-14.
2. **Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, Vandamme AM, Sandstrom P, Boucher CA, van de Vijver D, Rhee SY, Liu TF, Pillay D, Shafer RW.** 2009. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* **4:**e4724.
3. **Wensing AM, Calvez V, Gunthard HF, Johnson VA, Paredes R, Pillay D, Shafer RW, Richman DD.** 2014. 2014 Update of the drug resistance mutations in HIV-1. *Top Antivir Med* **22:**642-650.
4. **Rojas Sanchez P, Holguin A.** 2014. Drug resistance in the HIV-1-infected paediatric population worldwide: a systematic review. *J Antimicrob Chemother* **69:**2032-2042.
5. **Ssemwanga D, Lihana RW, Ugoji C, Abimiku A, Nkengasong J, Dakum P, Ndembi N.** 2014. Update on HIV-1 Acquired and Transmitted Drug Resistance in Africa. *AIDS Rev* **17**.
6. **Smit E.** 2014. Antiviral resistance testing. *Curr Opin Infect Dis* **27:**566-572.
7. **Bhargava M, Cajas JM, Wainberg MA, Klein MB, Pant Pai N.** 2014. Do HIV-1 non-B subtypes differentially impact resistance mutations and clinical

disease progression in treated populations? Evidence from a systematic review. *J Int AIDS Soc* **17:**18944.

8. **Snedecor SJ, Sudharshan L, Nedrow K, Bhanegaonkar A, Simpson KN, Haider S, Chambers R, Craig C, Stephens J.** 2014. Burden of nonnucleoside reverse transcriptase inhibitor resistance in HIV-1-infected patients: a systematic review and meta-analysis. *AIDS Res Hum Retroviruses* **30:**753-768.

9. **Ambrosioni J, Nicolas D, Sued O, Aguero F, Manzardo C, Miro JM.** 2014. Update on antiretroviral treatment during primary HIV infection. *Expert Rev Anti Infect Ther* **12:**793-807.

10. **Iwuji CC, Orne-Gliemann J, Tanser F, Boyer S, Lessells RJ, Lert F, Imrie J, Barnighausen T, Rekacewicz C, Bazin B, Newell ML, Dabis F, Group ATS.** 2013. Evaluation of the impact of immediate versus WHO recommendations-guided antiretroviral therapy initiation on HIV incidence: the ANRS 12249 TasP (Treatment as Prevention) trial in Hlabisa sub-district, KwaZulu-Natal, South Africa: study protocol for a cluster randomised controlled trial. *Trials* **14:**230.

11. **Manasa J, Danaviah S, Pillay S, Padayachee P, Mthiyane H, Mkhize C, Lessells RJ, Seebregts C, de Wit TF, Viljoen J, Katzenstein D, De Oliveira T.** 2014. An affordable HIV-1 drug resistance monitoring method for resource limited settings. *J Vis Exp* doi:10.3791/51242.

12. **Siliciano JD, Siliciano RF.** 2013. Recent trends in HIV-1 drug resistance. *Curr Opin Virol* **3:**487-494.

13. **Brenner BG, Roger M, Moisi DD, Oliveira M, Hardy I, Turgel R, Charest H, Routy JP, Wainberg MA.** 2008. Transmission networks of drug resistance acquired in primary/early stage HIV infection. *AIDS* **22:**2509-2515.

14. **Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, Baril JG, Thomas R, Rouleau D, Bruneau J, Leblanc R, Legault M, Tremblay C, Charest H, Wainberg MA.** 2007. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis* **195:**951-959.

15. **Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, Turgel R, Charest H, Koopman J, Wainberg MA.** 2011. Transmission Clustering Drives the Onward Spread of the HIV Epidemic Among Men Who Have Sex With Men in Quebec. *J Infect Dis* **204:**1115-1119.

16. **Brenner BG, Wainberg MA.** 2013. Future of phylogeny in HIV prevention. *J Acquir Immune Defic Syndr* **63 Suppl 2:**S248-254.

17. **Leventhal GE, Gunthard HF, Bonhoeffer S, Stadler T.** 2014. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol Biol Evol* **31:**6-17.

18. **Leventhal GE, Kouyos R, Stadler T, Wyl V, Yerly S, Boni J, Cellerai C, Klimkait T, Gunthard HF, Bonhoeffer S.** 2012. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol* **8:**e1002413.

19. **Stadler T, Bonhoeffer S.** 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos Trans R Soc Lond B Biol Sci* **368:**20120198.

20. **Stadler T, Kouyos R, von Wyl V, Yerly S, Boni J, Burgisser P, Klimkait T, Joos B, Rieder P, Xie D, Gunthard HF, Drummond AJ, Bonhoeffer S.** 2012. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol* **29:**347-357.

21. **Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ.** 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A* **110:**228-233.

22. **Frost SD, Volz EM.** 2010. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci* **365:**1879-1890.

23. **Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, Koopman JS.** 2013. HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis. *PLoS Med* **10:**e1001568.

24. **Volz EM, Koelle K, Bedford T.** 2013. Viral phylodynamics. *PLoS Comput Biol* **9:**e1002947.

25. **Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SD.** 2012. Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection. *PLoS Comput Biol* **8:**e1002552.

26. **Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD.** 2009. Phylodynamics of infectious disease epidemics. *Genetics* **183:**1421-1430.

27. **Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT, Collaboration UHDR.** 2011. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J Infect Dis* **204:**1463-1469.

28. **Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, Kosakovsky Pond SL.** 2014. The global transmission network of HIV-1. *J Infect Dis* **209:**304-313.

29. **Kuhnert D, Stadler T, Vaughan TG, Drummond AJ.** 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *J R Soc Interface* **11:**20131106.

30. **Bezemer D, Faria NR, Hassan AS, Hamers RL, Mutua G, Anzala O, Mandaliya KN, Cane PA, Berkley JA, Rinke de Wit TF, Wallis CL, Graham SM, Price MA, Coutinho R, Sanders EJ.** 2013. HIV-1 transmission networks amongst men having sex with men and heterosexuals in Kenya. *AIDS Res Hum Retroviruses* doi:10.1089/AID.2013.0171.

31. **Bezemer D, Ratmann O, van Sighem A, Dutilh BE, Faria N, van den Hengel R, Gras L, Reiss P, de Wolf F, Fraser C, ATHENA observational cohort.** 2014. Ongoing HIV-1 Subtype B Transmission Networks in the Netherlands, abstr *CROI 2014*, Boston, MA,

32. **Bezemer D, van Sighem A, Lukashov VV, van der Hoek L, Back N, Schuurman R, Boucher CA, Claas EC, Boerlijst MC, Coutinho RA, de Wolf F, cohort Ao.** 2010. Transmission networks of HIV-1 among men having sex with men in the Netherlands. *AIDS* **24:**271-282.

33. **Carnegie NB, Wang R, Novitsky V, De Gruttola V.** 2014. Linkage of Viral Sequences among HIV-Infected Village Residents in Botswana: Estimation of Linkage Rates in the Presence of Missing Data. *PLoS Comput Biol* **10:**e1003430.

34. **Novitsky V, Bussmann H, Logan A, Moyo S, van Widenfelt E, Okui L, Mmalane M, Baca J, Buck L, Phillips E, Tim D, McLane MF, Lei Q, Wang R, Makhema J, Lockman S, DeGruttola V, Essex M.** 2013. Phylogenetic Relatedness of Circulating HIV-1C Variants in Mochudi, Botswana. *PLoS One* **8:**e80589.

35. **Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M.** 2014. Impact of Sampling Density on the Extent of HIV Clustering. *AIDS Res Hum Retroviruses* **30:**1226-1235.

36. **Maes B, Schrooten Y, Snoeck J, Derdelinckx I, Van Ranst M, Vandamme AM, Van Laethem K.** 2004. Performance of ViroSeq HIV-1 Genotyping System in routine practice at a Belgian clinical laboratory. *J Virol Methods* **119:**45-49.

37. **Ribas SG, Heyndrickx L, Ondoa P, Fransen K.** 2006. Performance evaluation of the two protease sequencing primers of the Trugene HIV-1 genotyping kit. *J Virol Methods* **135:**137-142.

38. **Church JD, Mwatha A, Bagenda D, Omer SB, Donnell D, Musoke P, Nakabiito C, Eure C, Bakaki P, Matovu F, Thigpen MC, Guay LA, McConnell M, Fowler MG, Jackson JB, Eshleman SH.** 2009. In utero HIV infection is associated with an increased risk of nevirapine resistance in ugandan infants who were exposed to perinatal single dose nevirapine. *AIDS Res Hum Retroviruses* **25:**673-677.

39. **Cunningham S, Ank B, Lewis D, Lu W, Wantman M, Dileanis JA, Jackson JB, Palumbo P, Krogstad P, Eshleman SH.** 2001. Performance of the applied biosystems ViroSeq human immunodeficiency virus type 1 (HIV-1) genotyping

system for sequence-based analysis of HIV-1 in pediatric plasma samples. *J Clin Microbiol* **39:**1254-1257.

40. **Eshleman SH, Crutcher G, Petrauskene O, Kunstman K, Cunningham SP, Trevino C, Davis C, Kennedy J, Fairman J, Foley B, Kop J.** 2005. Sensitivity and specificity of the ViroSeq human immunodeficiency virus type 1 (HIV-1) genotyping system for detection of HIV-1 drug resistance mutations by use of an ABI PRISM 3100 genetic analyzer. *J Clin Microbiol* **43:**813-817.

41. **Eshleman SH, Guay LA, Mwatha A, Brown ER, Cunningham SP, Musoke P, Mmiro F, Jackson JB.** 2004. Characterization of nevirapine resistance mutations in women with subtype A vs. D HIV-1 6-8 weeks after single-dose nevirapine (HIVNET 012). *J Acquir Immune Defic Syndr* **35:**126-130.

42. **Eshleman SH, Hackett J, Jr., Swanson P, Cunningham SP, Drews B, Brennan C, Devare SG, Zekeng L, Kaptue L, Marlowe N.** 2004. Performance of the Celera Diagnostics ViroSeq HIV-1 Genotyping System for sequence-based analysis of diverse human immunodeficiency virus type 1 strains. *J Clin Microbiol* **42:**2711-2717.

43. **Eshleman SH, Hoover DR, Chen S, Hudelson SE, Guay LA, Mwatha A, Fiscus SA, Mmiro F, Musoke P, Jackson JB, Kumwenda N, Taha T.** 2005. Nevirapine (NVP) resistance in women with HIV-1 subtype C, compared with subtypes A and D, after the administration of single-dose NVP. *J Infect Dis* **192:**30-36.

44. **Mracna M, Becker-Pergola G, Dileanis J, Guay LA, Cunningham S, Jackson JB, Eshleman SH.** 2001. Performance of Applied Biosystems ViroSeq HIV-1

Genotyping System for sequence-based analysis of non-subtype B human immunodeficiency virus type 1 from Uganda. *J Clin Microbiol* **39:**4323-4327.

45. **Sturmer M, Berger A, Doerr HW.** 2003. Modifications and substitutions of the RNA extraction module in the ViroSeq HIV-1 genotyping system version 2: effects on sensitivity and complexity of the assay. *J Med Virol* **71:**475-479.

46. **CDC.** 2015. ATCC HIV-1 Drug Resistance Genotyping Kit. http://www.atcc.org/products/cells_and_microorganisms/hiv-1_drug_resistance_genotyping_kit.aspx. Accessed 1/6/2015.

47. **Wallis CL, Papathanasopoulos MA, Lakhi S, Karita E, Kamali A, Kaleebu P, Sanders E, Anzala O, Bekker LG, Stevens G, de Wit TF, Stevens W.** 2010. Affordable in-house antiretroviral drug resistance assay with good performance in non-subtype B HIV-1. *J Virol Methods* **163:**505-508.

48. **Youngpairoj AS, Masciotra S, Garrido C, Zahonero N, de Mendoza C, Garcia-Lerma JG.** 2008. HIV-1 drug resistance genotyping from dried blood spots stored for 1 year at 4 degrees C. *J Antimicrob Chemother* **61:**1217-1220.

49. **Devereux HL, Youle M, Johnson MA, Loveday C.** 1999. Rapid decline in detectability of HIV-1 drug resistance mutations after stopping therapy. *Aids* **13:**F123-127.

50. **Neogi U, Sahoo PN, De Costa A, Shet A.** 2012. High viremia and low level of transmitted drug resistance in anti-retroviral therapy-naive perinatally-infected children and adolescents with HIV-1 subtype C infection. *BMC Infect Dis* **12:**317.

51. **Zhou Z, Wagar N, DeVos JR, Rottinghaus E, Diallo K, Nguyen DB, Bassey O, Ugbena R, Wadonda-Kabondo N, McConnell MS, Zulu I, Chilima B,**

Nkengasong J, Yang C. 2011. Optimization of a low cost and broadly sensitive genotyping assay for HIV-1 drug resistance surveillance and monitoring in resource-limited settings. *PLoS One* **6:**e28184.

52. **Inzaule S, Yang C, Kasembeli A, Nafisa L, Okonji J, Oyaro B, Lando R, Mills LA, Laserson K, Thomas T, Nkengasong J, Zeh C.** 2013. Field evaluation of a broadly sensitive HIV-1 in-house genotyping assay for use with both plasma and dried blood spot specimens in a resource-limited country. *J Clin Microbiol* **51:**529-539.

53. **Zhang G, Cai F, Zhou Z, DeVos J, Wagar N, Diallo K, Zulu I, Wadonda-Kabondo N, Stringer JS, Weidle PJ, Ndongmo CB, Sikazwe I, Sarr A, Kagoli M, Nkengasong J, Gao F, Yang C.** 2013. Simultaneous detection of major drug resistance mutations in the protease and reverse transcriptase genes for HIV-1 subtype C by use of a multiplex allele-specific assay. *J Clin Microbiol* **51:**3666-3674.

54. **Acharya A, Vaniawala S, Shah P, Misra RN, Wani M, Mukhopadhyaya PN.** 2014. Development, validation and clinical evaluation of a low cost in-house HIV-1 drug resistance genotyping assay for Indian patients. *PLoS One* **9:**e105790.

55. **Chen JH, Wong KH, Chan K, Lam HY, Lee SS, Li P, Lee MP, Tsang DN, Zheng BJ, Yuen KY, Yam WC.** 2007. Evaluation of an in-house genotyping resistance test for HIV-1 drug resistance interpretation and genotyping. *J Clin Virol* **39:**125-131.

56. **Steegen K, Demecheleer E, De Cabooter N, Nges D, Temmerman M, Ndumbe P, Mandaliya K, Plum J, Verhofstede C.** 2006. A sensitive in-house

RT-PCR genotyping system for combined detection of plasma HIV-1 and assessment of drug resistance. *J Virol Methods* **133:**137-145.

57. **MacLeod IJ, Rowley CF, Thior I, Wester C, Makhema J, Essex M, Lockman S.** 2010. Minor resistant variants in nevirapine-exposed infants may predict virologic failure on nevirapine-containing ART. *J Clin Virol* **48:**162-167.

58. **Rowley CF, Boutwell CL, Lee EJ, MacLeod IJ, Ribaudo HJ, Essex M, Lockman S.** 2010. Ultrasensitive detection of minor drug-resistant variants for HIV after nevirapine exposure using allele-specific PCR: clinical significance. *AIDS Res Hum Retroviruses* **26:**293-300.

59. **Rowley CF, Boutwell CL, Lockman S, Essex M.** 2008. Improvement in allele-specific PCR assay with the use of polymorphism-specific primers for the analysis of minor variant drug resistance in HIV-1 subtype C. *J Virol Methods* **149:**69-75.

60. **Rozera G, Abbate I, Bruselles A, Vlassi C, D'Offizi G, Narciso P, Chillemi G, Prosperi M, Ippolito G, Capobianchi MR.** 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* **6:**15.

61. **Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW.** 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* **17:**1195-1201.

62. **Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, Berry N, Pillay D, Kellam P.** 2012. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J Clin Microbiol* **50:**3838-3844.

63. **Devereux HL, Loveday C, Youle M, Sabin CA, Burke A, Johnson M.** 2000. Substantial correlation between HIV type 1 drug-associated resistance mutations in plasma and peripheral blood mononuclear cells in treatment-experienced patients. *AIDS Res Hum Retroviruses* **16:**1025-1030.

64. **Steegen K, Luchters S, Demecheleer E, Dauwe K, Mandaliya K, Jaoko W, Plum J, Temmerman M, Verhofstede C.** 2007. Feasibility of detecting human immunodeficiency virus type 1 drug resistance in DNA extracted from whole blood or dried blood spots. *J Clin Microbiol* **45:**3342-3351.

65. **Diallo K, Murillo WE, de Rivera IL, Albert J, Zhou Z, Nkengasong J, Zhang G, Sabatier JF, Yang C.** 2012. Comparison of HIV-1 resistance profiles in plasma RNA versus PBMC DNA in heavily treated patients in Honduras, a resource-limited country. *Int J Mol Epidemiol Genet* **3:**56-65.

66. **Vartanian JP, Henry M, Wain-Hobson S.** 2002. Sustained G-->A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. *J Gen Virol* **83:**801-805.

67. **Vartanian JP, Meyerhans A, Asjo B, Wain-Hobson S.** 1991. Selection, recombination, and G----A hypermutation of human immunodeficiency virus type 1 genomes. *J Virol* **65:**1779-1788.

68. **Vartanian JP, Meyerhans A, Sala M, Wain-Hobson S.** 1994. G-->A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. *Proc Natl Acad Sci U S A* **91:**3092-3096.

69. **Wain-Hobson S, Sonigo P, Guyader M, Gazit A, Henry M.** 1995. Erratic G-->A hypermutation within a complete caprine arthritis-encephalitis virus (CAEV) provirus. *Virology* **209:**297-303.

70. **Harris RS, Liddament MT.** 2004. Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* **4:**868-877.

71. **McCallum M, Oliveira M, Ibanescu RI, Kramer VG, Moisi D, Asahchop EL, Brenner BG, Harrigan PR, Xu H, Wainberg MA.** 2013. Basis for early and preferential selection of the E138K mutation in HIV-1 reverse transcriptase. *Antimicrob Agents Chemother* **57:**4681-4688.

72. **Fourati S, Malet I, Lambert S, Soulie C, Wirden M, Flandre P, Fofana DB, Sayon S, Simon A, Katlama C, Calvez V, Marcelin AG.** 2012. E138K and M184I mutations in HIV-1 reverse transcriptase coemerge as a result of APOBEC3 editing in the absence of drug exposure. *AIDS* **26:**1619-1624.

73. **Neogi U, Shet A, Sahoo PN, Bontell I, Ekstrand ML, Banerjea AC, Sonnerborg A.** 2013. Human APOBEC3G-mediated hypermutation is associated with antiretroviral therapy failure in HIV-1 subtype C-infected individuals. *J Int AIDS Soc* **16:**18472.

74. **Ulenga NK, Sarr AD, Hamel D, Sankale JL, Mboup S, Kanki PJ.** 2008. The level of APOBEC3G (hA3G)-related G-to-A mutations does not correlate with viral load in HIV type 1-infected individuals. *AIDS Res Hum Retroviruses* **24:**1285-1290.

930 75. **Eyzaguirre LM, Charurat M, Redfield RR, Blattner WA, Carr JK, Sajadi**
931 **MM.** 2013. Elevated hypermutation levels in HIV-1 natural viral suppressors.
932 *Virology* **443:**306-312.
933 76. **Cheynier R, Gratton S, Vartanian JP, Meyerhans A, Wain-Hobson S.** 1997.
934 G --> A hypermutation does not result from polymerase chain reaction. *AIDS Res*
935 *Hum Retroviruses* **13:**985-986.
936 77. **Wang R, Goyal R, Lei Q, Essex M, De Gruttola V.** 2014. Sample size
937 considerations in the design of cluster randomized trials of combination HIV
938 prevention. *Clin Trials* **11:**309-318.
939 78. **Eshleman SH, Jones D, Flys T, Petrauskene O, Jackson JB.** 2003. Analysis of
940 HIV-1 variants by cloning DNA generated with the ViroSeq HIV-1 Genotyping
941 System. *Biotechniques* **35:**614-618, 620, 622.
942 79. **Novitsky V, Lagakos S, Herzig M, Bonney C, Kebaabetswe L, Rossenkhan R,**
943 **Nkwe D, Margolin L, Musonda R, Moyo S, Woldegabriel E, van Widenfelt E,**
944 **Makhema J, Essex M.** 2009. Evolution of proviral gp120 over the first year of
945 HIV-1 subtype C infection. NIHMSID # 79286. *Virology* **383:**47-59. PMCID:
946 PMC2642736.
947 80. **Novitsky V, Wang R, Rossenkhan R, Moyo S, Essex M.** 2013. Intra-host
948 evolutionary rates in HIV-1C env and gag during primary infection. *Infect Genet*
949 *Evol* **19C:**361-368.
950 81. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and
951 high throughput. *Nucleic Acids Res* **32:**1792-1797.

82. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30:**2725-2729.

83. **Rose PP, Korber BT.** 2000. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. *Bioinformatics* **16:**400-401.

84. **Felsenstein J.** 1985. Confidence limits on phylogenies: an approach using a bootstrap. *Evolution* **39:**783-791.

85. **Felsenstein J.** 2004. *Inferring phylogenies*. Sinauer Associates, Inc.

86. **Nei M, Kumar S.** 2000. *Molecular evolution and phylogenetics*. Oxford University Press, New York, NY.

87. **Even S.** 2011. Graphic Algorithms, 2nd ed., p 202. *In* Even G (ed). Cambridge University Press.

88. **Prosperi MC, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, Di Giambenedetto S, Bruzzone B, Capetti A, Vivarelli A, Rusconi S, Re MC, Gismondo MR, Sighinolfi L, Gray RR, Salemi M, Zazzi M, De Luca A, group Ac.** 2011. A novel methodology for large-scale phylogeny partition. *Nat Commun* **2:**321.

89. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:**2688-2690.

90. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:**1312-1313.

91. **R Core Team.** 2014. *R: A language and environment for statistical computing*. http://www.R-project.org/. R Foundation for Statistical Computing, Vienna, Austria. .

92. **Dugan KA, Lawrence HS, Hares DR, Fisher CL, Budowle B.** 2002. An improved method for post-PCR purification for mtDNA sequence analysis. *J Forensic Sci* **47:**811-818.

93. **Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M.** 2015. Importance of Viral Sequence Length and Number of Variable and Informative Sites in Analysis of HIV Clustering. *AIDS Res Hum Retroviruses* doi:10.1089/AID.2014.0211.

94. **Salichos L, Rokas A.** 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497:**327-331.

95. **Salichos L, Stamatakis A, Rokas A.** 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* **31:**1261-1271.

96. **Kampstra P.** 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software* **28:**1-9.

991 **Figure Legends**

992

993 **Figure 1.** Overview of the long-range HIV genotyping: 1st and 2nd round products
994 are mapped against the HIV-1 genome structure. First round (RT-) PCR product
995 is shown at the bottom as a hatched bar. Second round PCR products, amplicon
996 1 and amplicon 2 are shown as gray bars.

997

998 **Figure 2.** Distribution of HIV-1 RNA load in BCPP samples (n=202) that were
999 subjects for long-range HIV genotyping using proviral DNA as a template for
1000 amplification. Histogram depicts distribution of HIV-1 RNA in all samples, n=202.
1001 The *x* axis shows HIV-1 RNA on a $\log_{10}$ scale. Two pie charts illustrate
1002 distribution of HIV-1 RNA among successfully amplified (n=181) and failed
1003 (n=21) samples. Legend at the top right outlines break down intervals of HIV-1
1004 RNA presented on pie charts.

1005

1006 **Figure 3.** Distribution of APOBEC-induced hypermutations in sequences
1007 amplified from proviral DNA (histograms). Horizontal boxplots outline distribution
1008 of APOBEC-induced hypermutations in subsets of sequences with identified
1009 drug-resistance mutations. A: Amplicon 1, n=649, distribution of hypermutations
1010 adjusted by sequence length. B: Amplicon 1, n=649, distribution of hypermutation
1011 ratio data (see Materials and Methods). C: Amplicon 2, n=90, distribution of

hypermutations adjusted by sequence length. D: Amplicon 2, n=90, distribution of hypermutation ratio data (see Materials and Methods).

Figure 4. G-to-A hypermutations in HIV-1C sequences with and without M184I mutation. Beanplots – a combination of a box plot, density plot, and a rug with ticks for each value in the middle – are shown (96). Comparison between groups was performed by Wilcoxon signed rank test. **A:** Hypermutations adjusted by sequence length. **B:** Hypermutation ratio.

Figure 5. Clustering of HIV-1C sequences by loci, n=547. Proportion of HIV-1C sequences in clusters was estimated by bootstrapped ML inference. The extent of HIV clustering was analyzed at bootstrap thresholds for cluster definition ≥0.70, ≥0.80, ≥0.90, and 1.0.The number of viral sequences found in clusters for specified locus and at specified bootstrap support was compared between loci. Four loci were used: Amplicon 1 concatenated with V1C5 region of gp120 shown as “Amp 1 + v1c5”, Amplicon 1 alone as “Amp 1”, ViroSeq sequence as “ViroSeq”, and V1C5 region of gp120 as “V1C5”. Pie charts show concordant (++ and −−) and discordant (+− and −+) clustering between specified sequence loci (the first sign corresponds to the first sequence locus listed). Cases of significantly different clustering between loci with p-values less than 1.0E-04 in McNemar’s test are highlighted by gray squares on the background.

Table 1. Summary of HIV genotyping from proviral DNA, amplicon 1 (BCPP)

Proviral DNA specimens	Total		Subset with available HIV-1 RNA load data	
	n	proportion (95% CI)	n	proportion (95% CI)
Attempted cases	212		202	
Amplified	190	0.896 (0.845−0.932)	181	0.896 (0.843−0.933)
Amplification failed*	22	0.104 (0.068−0.155)	21	0.104 (0.067−0.157)
Sequenced by direct Sanger sequencing**	167	0.879 (0.822−0.920)	160	0.884 (0.826−0.925)
Partial sequences by direct Sanger sequencing: gaps at ≤10% of sequence length; resolved by cloning**	23	0.121 (0.080−0.178)	21	0.116 (0.075−0.174)

* - proportion of failed cases is calculated from the number of attempted cases;
** - proportion of sequenced cases is calculated from the number of amplified cases.

Table 2. Summary of HIV genotyping from viral RNA, amplicon 1 (BCPP)

Viral RNA specimens	Total		Subset with available HIV-1 RNA load data	
	n	proportion (95% CI)	n	proportion (95% CI)
Attempted cases	32		31	
Amplified	23	0.719 (0.530−0.856)	23	0.742 (0.551−0.875)
Amplification failed*	9	0.281 (0.144−0.470)	8	0.258 (0.125−0.449)
Sequenced**	23	1.0 (0.822−1.0)	23	1.0 (0.822−1.0)

* - proportion of failed cases is calculated from the number of attempted cases;
** - proportion of sequenced cases is calculated from the number of amplified cases.

Table 3. Observed proportion of HIV-1C sequences in clusters, small set of sequences (n=83)

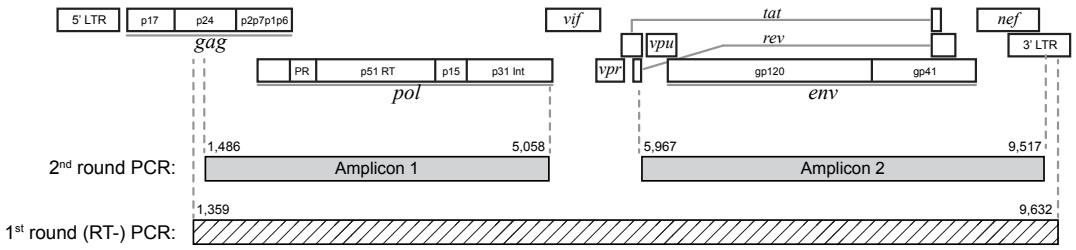
Loci	Number (proportion) of HIV-1C sequences in clusters at specified bootstrap support of splits			
	≥0.70	≥0.8	≥0.9	=1.0
Amplicon 1	32 (0.386) ◆§	22 (0.265)	19 (0.229)	11 (0.133)
Amplicon 2	32 (0.386) ◆§	24 (0.289) §	24 (0.289) ◆§	12 (0.145)
Amplicon 1 + 2	33 (0.398) ◆§	28 (0.337) ◆§	24 (0.289) ◆§	11 (0.133)
ViroSeq	11 (0.133)	13 (0.157)	11 (0.133)	7 (0.084)
V1C5	16 (0.193)	12 (0.145)	9 (0.108)	4 (0.048)

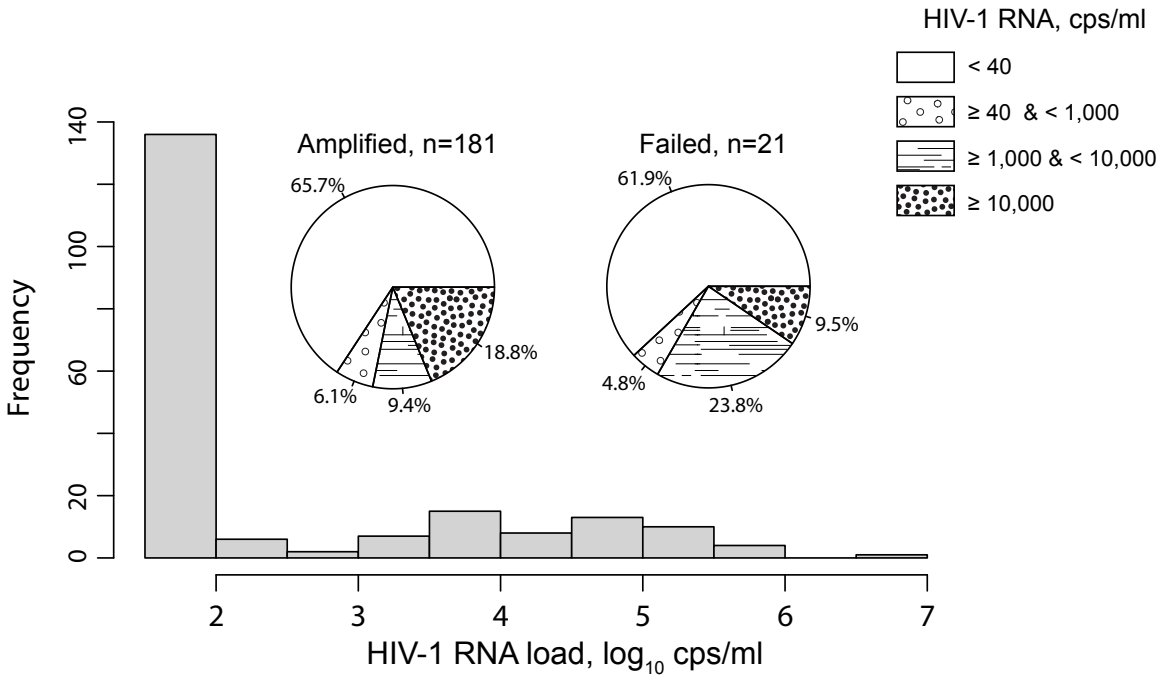
◆ - p-value < 0.05 for comparison to ViroSeq (Fisher exact test);
§ - p-value < 0.05 for comparison to V1C5 (Fisher exact test)

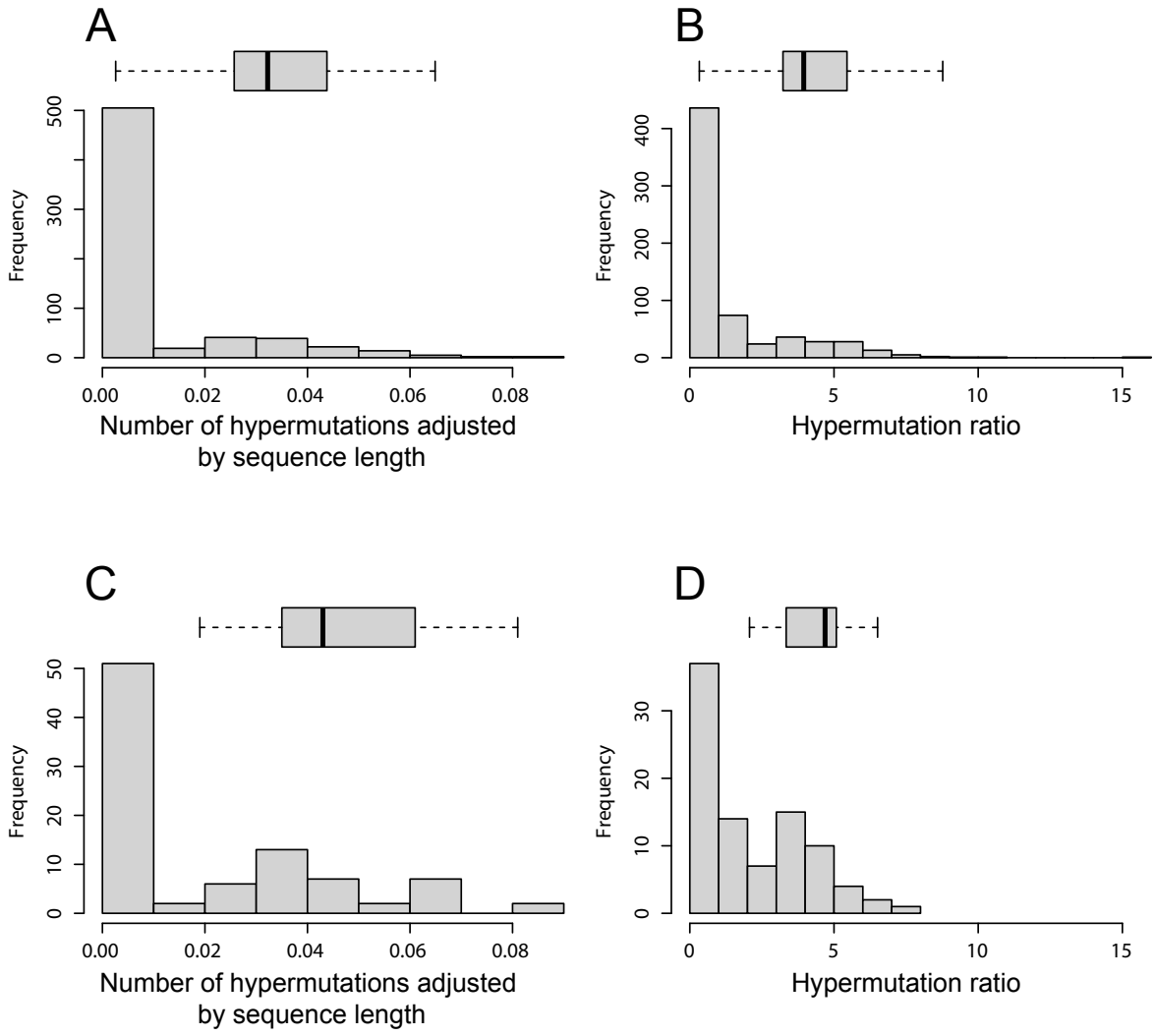
Table 4. Observed proportion of HIV-1C sequences in clusters, large set of sequences (n=547)

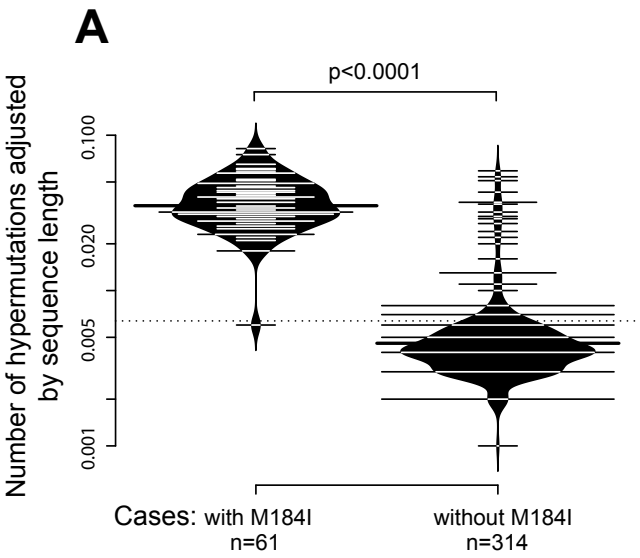
Loci	Number (proportion) of HIV-1C sequences in clusters at specified bootstrap support of splits			
	≥0.70	≥0.8	≥0.9	=1.0
Amplicon 1	251 (0.459) ◆§	215 (0.393) ◆§	159 (0.291) ◆§	88 (0.161) ◆§
Amplicon 1 + V1C5	267 (0.488) ◆§	220 (0.402) ◆§	181 (0.331) ◆§	122 (0.223) ◆§
ViroSeq	120 (0.219)	90 (0.165)	73 (0.133)	34 (0.062)
V1C5	135 (0.247)	114 (0.208)	88 (0.161)	44 (0.080)

◆ - p-value < 0.001 for comparison to ViroSeq (Fisher exact test);
§ - p-value < 0.001 for comparison to V1C5 (Fisher exact test)

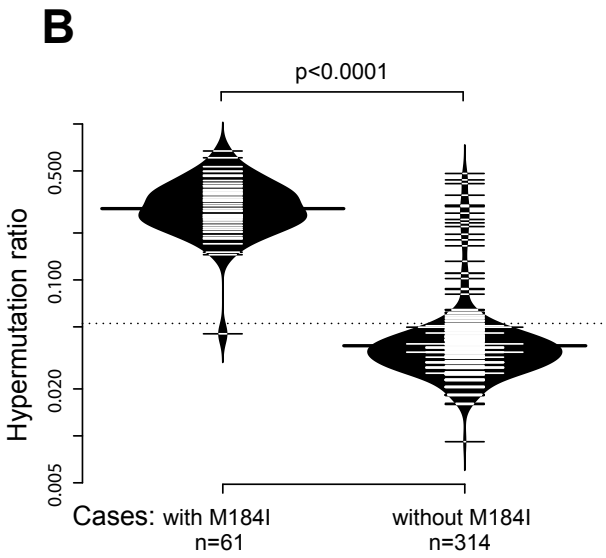




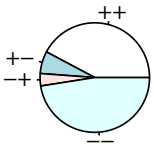
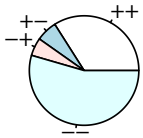
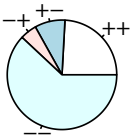
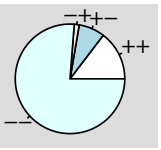
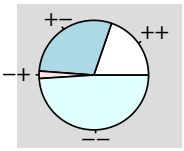
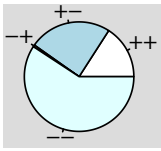
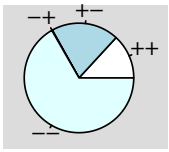
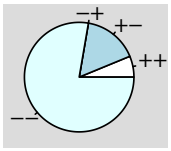
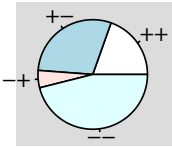
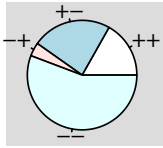
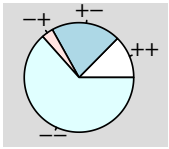
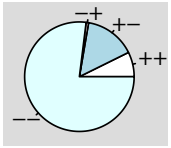
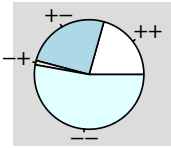
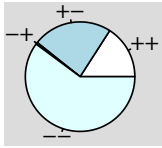
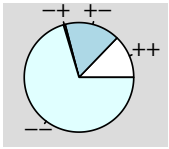
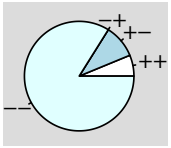
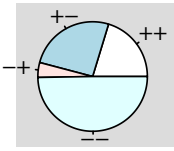
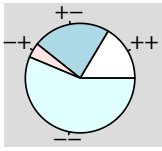
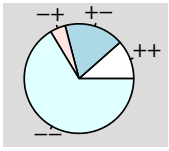
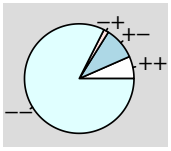




	with M184I n=61	without M184I n=314
Minimum:	0.0060	0.0010
1st Quartile:	0.0280	0.0030
Median:	0.0340	0.0040
Mean:	0.0379	0.0058
3rd Quartile:	0.0460	0.0050
Maximum:	0.0820	0.0590



	with M184I n=61	without M184I n=314
Minimum:	0.0451	0.0092
1st Quartile:	0.2330	0.0276
Median:	0.2803	0.0346
Mean:	0.3089	0.0480
3rd Quartile:	0.3724	0.0431
Maximum:	0.6711	0.4814

HIV-1C sequence loci	Bootstrap support of splits for cluster definition			
	≥0.70	**≥0.80**	**≥0.90**	**=1.0**
Amp 1 + V1C5 vs. Amp 1:				
Amp 1 + V1C5 vs. ViroSeq:				
Amp 1 + V1C5 vs. V1C5:				
Amp 1 vs. ViroSeq:				
Amp 1 vs. V1C5:				
ViroSeq vs. V1C5: