Identification of HIV superinfection in seroconcordant couples in Rakai, Uganda using next generation deep sequencing

Andrew D Redd1, Aleisha Collinson-Streng1, Craig Martens2, Stacy Ricklefs2, Caroline E Mullis3, Jordyn Manucci3, Aaron AR Tobian3, Ethan J. Selig3, Oliver Laeyendecker1,3, Nelson Sewankambo4,7, Ronald H Gray6, David Serwadda4,8, Maria J Wawer6, Stephen F. Porcella2, and Thomas C Quinn1,3*, on behalf of the Rakai Health Sciences Program

1-Laboratory of Immunoregulation, DIR, NIAID, NIH, Baltimore MD, USA
2- Genomics Unit, Research Technologies Section, Rocky Mountain Laboratories, DIR, NIAID, NIH, Hamilton, MT, USA
3-Johns Hopkins Medical Institute, Johns Hopkins University, Baltimore MD, USA
4-Rakai Health Sciences Program, Kalisizo, Uganda
5- Uganda Virus Research Institute, Entebbe, Uganda
6- Bloomberg School of Public Health, Johns Hopkins University, Baltimore MD, USA
7- School of Medicine, Makerere University, Kampala, Uganda
8- School of Public Health, Makerere University, Kampala, Uganda

* - Corresponding author: Thomas C Quinn, MD, MSc

Johns Hopkins University School of Medicine
Rangos Building, Room 530
855 N Wolfe St
Baltimore MD, 21205

tquinn2@jhmi.edu

Word count: Abstract: 210, Manuscript: 4024

Running title: HIV superinfection in Rakai Uganda
Abstract:

HIV superinfection, which occurs when a previously infected individual acquires a new distinct HIV strain, has been described in a number of populations. Previous methods to detect superinfection have involved a combination of labor-intensive assays with varying success. We designed and tested a next generation sequencing (NGS) protocol to identify HIV superinfection by targeting two regions of the HIV viral genome, p24 and gp41. The method was validated by mixing control samples infected with HIV subtypes A or D at different ratios to determine the inter- and intra-subtype sensitivity by NGS. This amplicon-based NGS protocol was able to consistently identify distinct inter-subtype strains at ratios of 1%, and intra-subtype variants at 5%. Using stored samples from the Rakai Community Cohort Study (RCCS) in Uganda, eleven individuals who were HIV-seroconcordant but virally unlinked from their spouses were then tested with this method to detect superinfection between 2002-2005. Two female cases of HIV inter-subtype superinfection (18.2%) were identified. These results are consistent with other African studies, and support the hypothesis that HIV superinfection occurs at a relatively high rate. Our results indicate that NGS can be used for detection of HIV superinfection within large cohorts, which could assist in determining the incidence and the epidemiologic, virologic, and immunological correlates of this phenomenon.
Introduction:

HIV superinfection occurs when a known HIV-infected individual is subsequently infected with a new phylogenetically distinct viral strain or strains. The first documented cases of HIV superinfection were found in individuals with varying modes of transmission and included inter- and intra-subtype cases (1, 9, 17). Subsequently, multiple studies have documented superinfection in small populations of high risk individuals (2, 3, 7, 8, 11, 13, 17, 20, 22, 23, 26, 29). The rate of HIV superinfection in these high-risk groups was relatively frequent, and was comparable to the incidence rate in similar populations from the same regions, especially if multiple viral genes were examined (3, 13, 14, 21, 24). In contrast, other researchers have found no evidence of superinfection in large scale population studies (6, 15).

One possible reason for this discrepancy may be due to differences in techniques and criteria used to identify superinfection (16). Initial studies designed to examine the frequency of superinfection utilized heteroduplex mobility assays (HMA) or multiregion hybridization assays (MHA) followed by selective clonal analysis of those samples that demonstrated the presence of new viral variants (3, 11, 15). MHA screening is limited in that it only can identify inter-subtype superinfection while possibly missing intra-subtype superinfection. Although HMA is sensitive enough to detect samples with greater than 1.5% differences in pairwise distance, it is susceptible to false positives due to the presence of insertions or deletions (16). Additionally, both the HMA and MHA methods require verification using in-depth cloning and Sanger-sequencing (13, 16).

The sensitivity of these screening/cloning techniques is dependent on the number of clones amplified and the number of genes examined (12-14). To detect a minor variant approaching 1% over 100 clones would need to be examined per sample, preferably from multiple PCR reactions to increase the amount and diversity of viral strains sequenced (14, 16). With the need to
examine multiple regions of the viral genome to ensure accurate phylotyping and identification of superinfecting strains, in-depth cloning and Sanger sequencing is prohibitively labor intensive for large scale studies (12, 14, 16).

Newly developed next generation sequencing (NGS) techniques provide unprecedented sequencing depth, offer the ability to multiplex samples, are quicker, more cost-effective, and less labor intensive than cloning and Sanger-based sequencing (12). Using several genomic targets and high sequence volume, NGS should be able to distinguish minor variants that arose spontaneously either through recombination, within-host viral evolution, or from newly introduced strains or subtypes (18, 30).

We designed and tested an NGS protocol and sequence analysis pipeline, which focuses on amplification and sequencing of the p24 region of the viral capsid and the gp41 region of the viral envelope. These genomic regions were chosen for examination because they are relatively genetically stable, of sufficient length, suitable for phylotyping, were previously used in PCR-cloning and Sanger sequencing studies, and are not high in polymeric regions. We further tested the protocol in 11 individuals from virally unlinked HIV seroconcordant couples from Rakai District, Uganda to detect the occurrence of HIV superinfection.

Materials and Methods:

Ethics Statement

All subjects provided written informed consent for their samples to be stored and used for future unspecified HIV related research. The study was approved by the Science and Ethics Committee of the Uganda Virus Research Institute, the Western Institutional Review Board and the Committee on Human Research at Johns Hopkins Bloomberg School of Public Health.
Study Population and Subjects

Serum samples were retrospectively selected from individuals in the Rakai Community Cohort Study (RCCS), a rural, community-based open cohort consisting of persons aged 15-49 years in Rakai District, southwestern Uganda (27). Since 1994, interviews and venous blood samples have been obtained annually from approximately 14,000 consenting adults living in fifty villages. As part of the routine interview, consenting individuals in stable sexual partnerships are linked as couples.

Control serum samples were selected from HIV-infected individuals who were previously identified as being infected with either subtype A (n=4) or D (n=6) in the 2002 community survey. Identification of subtypes was performed by Sanger-sequencing of cloned PCR products of the p24 and gp41 target regions.

Using stored sera from 2002, we identified 18 HIV-infected individuals in 9 HIV-seroconcordant couples whose viruses were phylogenetically unlinked to their partner’s virus, as determined by previous Sanger-sequencing for either the gp41 or p24 regions (4). The individual’s samples were labeled with their gender, couple number, and year of sample draw (ex. Female_1_C1_2002). Of the 18 individuals, eleven had serum samples available in 2005, and these were examined for HIV superinfection in this population. Four of the eleven individuals were from two couples (couples 1 and 2) where both members had serum samples available from 2002 and 2005; however, for this analysis each individual was analyzed independently. The remaining seven individuals only had serum samples available in 2002, but were included in this study to examine for the source of any new superinfecting HIV strains found in their partner’s 2005 samples.

Viral RNA extraction, cDNA synthesis, and PCR target amplification
Viral RNA was extracted from 140 µL of serum using a QIAmp Viral RNA Mini Kit (Qiagen, Valencia CA), and eluted into 50 µL of Qiagen Buffer AVE. For each genomic target region (p24 and gp41), two 50 µL RT-PCR reactions were performed simultaneously to maximize the amount and diversity of viral RNA genomes amplified per sample. For the gp41 region, each 50 µL RT-PCR reaction was performed using a 40µL master mix composed of 20µL of ddH₂O, 10µL of 5x buffer, 3µL of dNTPs, and 2µL of Enzyme Mix from the Qiagen OneStep RT-PCR Kit. One microliter of RNase Inhibitor was also added, along with 2 µL of 20µM dilutions of both the forward primer (GP50F1-HXB2 nt 7691→7720) and the reverse primer (GP41R1-HXB2 nt 8347←8374)(Appendix of methods). This master mix was combined with 10µL of purified viral RNA and incubated 30 minutes at 50°C, and 15 minutes at 94 °C for RT extension. PCR was then performed at 35 cycles of 30 seconds at 94°C, 35 seconds at 53.5°C, and 90 seconds at 72°C, followed by 72°C for 10 minutes. For the p24 region, the 50µL RT and PCR reactions were carried out using the same master mix as described above with one exception; forward and reverse primers specific for the p24 target were used and designated as G00 (HXB2 nt 764→782) and G01 (HXB2 nt 2264←2281), respectively (Appendix of methods). For two samples that did not amplify the p24 region during the initial PCR, a reformulated 40µL master mix containing 20µL of ddH₂O, 10µL of 5x Buffer, 3µL of dNTPs, and 2µL of Enzyme Mix from the Qiagen OneStep RT-PCR Kit as well as 2µL of MgCl₂, 1µL of RNase Inhibitor, and 1.5 µL of 20µM dilutions of both the forward primer and reverse primers was used. The two samples were pooled to maximize the depth of detection, and 10 ul of this pool was used in a nested 100 µL PCR reaction using primer sets for gp41 (E55-primer set with fourteen 454 bar-coded variations (MID1-MID14)) or p24 (G100-primer set with fourteen 454 barcoded variations (MID1-MID14)) (Appendix of methods)(Roche Inc, Branford, CT). Briefly,
each nested PCR reaction for gp41 or p24 contained 90 µL of master mix composed of 50.4 µL ddH₂O, 10 µL 10x Reaction Buffer, 20 µL MgCl₂, 3 µL dNTPs, and 0.6 µL HotStarTaq DNA Polymerase (Qiagen, Valencia, CA) as well as 3 µL of the forward and reverse E55-primers or 3 µL of the forward and reverse G100-primer set both at 20 µM final concentration for both regions (Appendix of methods). PCR amplification conditions for the 100 µL nested reactions were identical to the first round PCR conditions described above. Successful single band amplification of gp41 or p24 target products was verified by agarose gel electrophoresis.

Serum HIV-1 RNA concentrations (viral loads) were determined by the Ampli cor v1.5 (Roche Diagnostics; Basil Switzerland)

**Generation of control samples for Inter- and Intra-subtype NGS threshold detection**

Control serum samples from HIV infected individuals, previously identified via Sanger-sequencing of PCR fragments as being infected with either subtype A (n=4) or D (n=6) were used to determine the assay’s limit of detection. Phylogenetically unlinked viral isolates were mixed in inter- and intra-subtype experiments for each viral target region.

For the p24 region, viral extracts from two HIV subtype A infected control individuals (A₁ & A₂) and four subtype D infected individuals (D₁-D₄) were amplified separately in the first round PCR reaction. Aliquots were collected and set aside for pure sample analysis, while aliquots of each control sample were also mixed at a variety of ratios. The ratios tested for the p24 target region were A₂:D₁ 50:50; A₂:D₁ 95:5; A₂:D₁ 99:1; D₃:A₂ 99.9:0.1; A₁:A₂ 95:5; D₁:D₂ 95:5, and D₃:D₄ 95:5. Nested PCR reactions were performed on these samples as described above.

For the gp41 region, viral extracts from two HIV subtype A infected individuals (A₃ & A₄) and two subtype D infected control individuals (D₅ & D₆) were amplified separately in the
first round PCR reaction. Aliquots of the first round PCR were collected and set aside for pure sample analysis, while aliquots of each were mixed at a variety of ratios. The ratios tested for the gp41 target region were $A_4:D_5$ 50:50; $A_4:D_5$ 95:5; $A_4:D_5$ 99:1; $A_4:D_5$ 99.9:0.1; $A_4:A_2$ 95:5; and $D_3:D_6$ 95:5. Nested PCR reactions were performed on these samples as described above.

**PCR Product Purification**

The Amplicon Library Preparation Method was performed as recommended by the manufacturer (Roche, Branford, CT) and all PCR products were purified with the following minor alterations. In an effort to eliminate the capture of primers, the bead to target ratio was reduced by incubating 30ul of AMPure Beads XP (Agencourt, Beckman Coulter Genomics Danvers, MA) with 25uL of PCR product diluted in 25ul of water. Purified PCR products were quantified using PicoGreen (Invitrogen, Carlsbad, CA) and each template was diluted to a $1 \times 10^9$ molecules/ul stock. The amplicon pools were made by combining 5uL of each diluted barcoded template to make a final $1 \times 10^9$ molecules/ul stock containing 14 barcoded amplicons.

**DNA sequencing**

Preparation of templated beads for NGS followed the emPCR Method Manual-Lib-L-MV (Roche Branford, CT). The $1 \times 10^9$ molecules/ul library pools were diluted to $1 \times 10^5$ molecules/ul for a target addition of 0.175 copies per bead to the DNA Capture Beads. The Live Amplification Mix was based on the reagent volumes for Paired End libraries to reduce the amount of Amplification Primer in the reactions and thereby reduce the bead signal intensity during sequencing. Enriched DNA Capture Beads were sequenced on the 454 (Roche, Branford, CT) per the manufacturer’s instructions using a 4-region gasket when indicated.

**Sequence Segregation**
Sequencing results were analyzed using the GS Amplicon Variant Analyzer version 2.5 (Roche, Branford, CT). All sequence reads were compared and similar sequences were combined into a single consensus sequence. Generated consensus sequences that were within 10 bases from both ends of the amplicon and comprised of a cluster of 10 individual, near-identical sequences or more were determined using the Roche Amplicon software and were classified as being consensus sequences of HIV variants. These consensus sequences were used for subsequent phylogenetic analysis.

**Phylogenetic analysis**

Consensus sequences, subtype reference sequences, and a selection of subtype reference sequences collected from Rakai (Appendix of methods) were aligned using Clustal W(25). Phylogenetic trees were generated by the Neighbor-Joining method(19). Statistical support for a specific clade in each phylogeny was obtained by bootstrapping (1000 replicates). The NGS consensus sequences for gp41 and p24 have been submitted to Genbank accession numbers XXXX, and are also available upon request (aredd2@jhmi.edu).

**HIV Superinfection definition and analysis**

HIV superinfection was defined in an individual whose 2005 serum sample demonstrated two or more distinct consensus sequences forming a monophyletic cluster that was phylogenetically unlinked from the individual’s entire consensus sequences in the 2002 sample. In order to be considered a superinfection, the genetic distance of the new monophyletic cluster from the closest related viral sequences found at the earlier timepoint had to be \( \geq 0.55\% \) per year for the p24 region, or \( \geq 0.98\% \) per year for the gp41 region for subtype D and \( \geq 0.59\% \) per year for the p24 region, or \( \geq 0.72\% \) per year for the gp41 region for subtype D, which is equal to the mean plus twice the standard deviation of the intra-person viral divergence or evolutionary rate.
of each HIV-1 subtype in Rakai Uganda (data not shown). All newly identified consensus sequences were phylogenetically compared to the most prominent strains of the other barcoded samples within NGS runs to search for micro-contamination, misclassification or sequencing errors. If instances of these errors were found, these consensus sequences were eliminated. For further verification, newly identified superinfecting viral strain sequences were translated and analyzed in order to check that a functional protein sequence was encoded in the sequence. Newly discovered superinfecting consensus sequences within an individual were compared phylogenetically to their partner’s consensus viral sequences in order to determine if the partner was the source of the new superinfecting virus.

**Results:**

*Genomic target regions, sequencing depth and consensus sequence analysis*

The p24 and gp41 regions of the viral genome were chosen for NGS because they are located at opposing ends of the HIV genome and are two of the more conserved areas of the genome. Previous research has indicated that the sensitivity of NGS for HIV quasispecies detection is 0.1% (30). Therefore, estimating an approximate read volume of 10,000 reads per sample, a cut-off of 10 similar reads, as determined by the Roche segregation software, was selected to qualify as a consensus sequence for further analysis. A cut-off of five sequences was also examined and found to not affect the findings and the overall sensitivity of the assay(12). However, when the consensus cut-off was dropped to two similar sequences, small amounts of micro-contaminating sequences reflecting the inherent error rate for the technology were discovered. Therefore, for the purposes of this study, ten reads or more was the threshold for quality consensus viral sequences (Fig. appendix 1).
P24 Inter- and intra-subtype analysis

Previous Sanger-sequencing of PCR fragments of the p24 region identified two subtype A (A₁, A₂) and four subtype D (D₁, D₂, D₃, D₄) samples used in this analysis (Table 1)(5). In order to test the intra- and inter-subtype viral population sensitivity of our NGS protocol, first round PCR products targeting the p24 region from these subtype A and subtype D samples were mixed in varying ratios, amplified, and sequenced on the Roche 454 as described above (Table 1, Fig. 1, 2A-D, & appendix 2A-C). In order to exclude cross contamination or poor quality reads, consensus read data sets for all mixtures were merged and the resulting trees constructed (Fig. 1D). These data demonstrate that reads specific for the mixed ratio samples are segregating properly to their respective branch locations for the components of the mixture, and that the NGS protocol provides good depth and quality sequence sorting during phylogenetic analysis (Fig. 1). The ratios of A₂:D₁ 95:5 and 99:1 were examined to determine if NGS would provide adequate depth and representation of the subtypes at these ratios (Fig. 2A&B). The lower frequency of minor variant (D₁ in both cases) was adequately represented in both trees although with a slight decrease in the number of consensus reads in the 99:1 ratio (Fig. 2B).

To further test the sensitivity of this assay, we analyzed a mixture of D₃ to A₂ at a ratio of 99.9:0.1. When we merged this ratio data with the control datasets (D₃ and A₂) the minor variant (A₂) did not appear in the data (Fig. appendix 2C). These results suggest that for the p24 target, an inter-subtype ratio at or less than 0.1% cannot be reliably identified by this NGS protocol.

In order to test the protocol for its ability to adequately sequence and separate related subtypes the following ratios; A₁:A₂ 95:5, D₁:D₂ 95:5 and D₃:D₄ 95:5 were tested (Table 1 and Fig. 2C&D, appendix 2A&B). The minor viral variant population in the A₁:A₂ 95:5 ratio (A₂) was identified as 14.5% of the total number of consensus sequences (Table 1 and Fig. 2C).
D1:D2 95:5 ratio sample did not appear to adequately amplify the minor variant (D2) when the

data were merged with the datasets for D1 and D2 (Table 1 and Fig. appendix 2B). This suggests

a lower limit for D1 vs D2 related intra-subtype identification for the p24 target. To determine if

this lack of detection or amplification of D2 was unique to the D1 vs D2 ratio of 95:5, this test

was repeated using the ratio D3:D4 at 95:5. In this test, the minor variant (D4) was identified in

25% of the total number of consensus sequences (Table 1 and Fig. 2D). It was found that the

consensus sequences that were expanded from the minor variant (D3) corresponded to the most

prominent subtype sequences present in the pure sample for D3 (Fig. appendix 2A).

gp41 Inter- and intra-subtype analysis

Due to limited amounts of viral RNA available for samples A1, A2, and D1-D4 different

control samples were used to test the minor intra- and inter-subtype viral population sensitivity

of our NGS protocol of the gp41 region (A3, A4, D5, D6) (Table 1 and Fig. 3). The majority of

the p24 NGS reactions were performed on a full 454 slide with 14 different bar-coded samples,

whereas the gp41 test samples were run on a slide that had been divided into four quadrants. The

reason for this change was to increase the sample throughput per run resulting in a lower read

volume per bar-coded sample (Table 1).

NGS analysis of all four inter-subtype mixtures (A vs D) for the gp41 region

demonstrated detectable consensus sequences of the minor variant (Table 1 and Fig. 3A-D).

However, in the case of the 99.9:0.1 mixture only one consensus sequence from the minority

variant subtype was amplified (Table 1 and Fig. 3D). While the sensitivity for minor viral

variants was increased for gp41 relative to the results for p24, the lack of two or more distinct

consensus sequences means that this would not qualify as a superinfecting viral species

according to the parameters described above.
NGS analysis of the two intra-subtype comparisons (A3 vs A4, or D5 vs D6) at 95:5 ratios demonstrated that in a merged data format, the minor variants (A4 and D5) were detected (Table 1 and Fig. appendix 3A&B). These data also demonstrated that the A3 individual, who previously was identified by PCR-cloning and Sanger-sequencing analysis as being infected with only subtype A, was in fact infected with two distinct variants which coincided with both subtype A and D (Fig. appendix 3A).

**HIV superinfection in Rakai Uganda**

Eleven HIV-infected individuals for whom serum samples were collected at 2002 and 2005 were evaluated at both p24 and gp41 for evidence of HIV superinfection (Table 2). In addition, for each individual, their partner’s sample from 2002, or in the case of two couples (C_1 and C_2) from 2002 and 2005, were amplified and sequenced by NGS to examine if superinfecting strains discovered in 2005 originated from their partner (Table 2). Serum HIV viral loads were calculated for each sample tested (Table 2). Each member was treated independently in this analysis.

Using NGS, two of the eleven individuals (18.2%) had evidence of HIV superinfection in their 2005 sera (Table 2 and Fig. 4&5). The first case of superinfection was documented in female_C1, who was infected in 2002 with a viral population that grouped with subtype D in the p24 region and with subtypes D and C in the gp41 region (Table 2 and Fig. 4A & appendix 4). In 2005 she had multiple consensus sequences in the p24 target region which grouped with subtype A, indicating a superinfection of a new HIV species (Fig. 4B). NGS analysis of her male partner (male_C1) demonstrated that he was infected with an apparent D/C recombinant strain that linked with his female partner’s viral strains in both regions in 2002 and 2005 when
examined in a merged phylogenetic tree (merged data not shown), indicating that she was
superinfected by another source (Table 2).

The second case of superinfection was observed in female_C3 who was initially infected
with HIV subtype D in both genomic regions (Table 2 and Fig. 5A & appendix 5). In her 2005
sample, she had acquired a new viral strain in the p24 region with multiple consensus sequences
that clustered with subtype A (Fig. 5B). Her partner, male_C3, was infected in 2002 with a dual
population of viruses that clustered with subtype D and C in the gp41 region, and subtype D in
the p24 region (Table 2). Merged phylogenetic tree analysis demonstrated that her
superinfecting strain was not found in her partner suggesting she was superinfected by another
source (merged data not shown). No other cases of superinfection were observed in the
remaining nine individuals during merged and unmerged phylogenetic tree analysis (Table 2).

Discussion:
Identification of HIV superinfection in the past has been accomplished using a variety of
screening techniques in conjunction with labor intensive cloning or single genome amplification
(3, 6, 11-13, 21). This has led to a significant amount of variability in the estimated rates of HIV
superinfection (3, 6, 8, 21). The data presented here describe a new NGS protocol to identify
HIV superinfection with relatively high inter- and intra-subtype sensitivity. The consensus of ten
repeated sequences was chosen since it was approximately 1/1000 of the estimated total reads,
and appeared to be an appropriate cut-off to identify inter- and intra-subtype minor variants
while avoiding data artifacts. Using mixtures of HIV-infected samples containing subtypes A
and D, the predominant viral species found in Uganda, the assay’s inter-subtype sensitivity in
both the p24 and gp41 target regions was determined to be at least 1%. Minor viral strains were
found at lower levels (0.1%) in the gp41 region, but not consistently or at high enough consensus
counts to lower the threshold of detection for the protocol. Intra-subtype sensitivity was
approximately 5%, although intra-subtype detection within the subtype A mixtures seemed more
robust than for the subtype D samples. We hypothesize that primer specificity and target
sequence variation may be driving some of these differences and is a limitation of our protocol.

The NGS protocol was able to identify two cases of HIV superinfection in women from
11 individuals who were members of virally unlinked concordantly infected couples. In both
cases, the superinfecting strain was HIV subtype A, which has been shown to be more infectious
than subtype D (10). In addition, both women’s viral loads increased during the period. None of
the superinfecting strains were detected in the women’s male partners suggesting that the
superinfecting strain was acquired from another source. It is possible that the new strains found
in these two individuals were present in the earlier timepoints at levels that were too low to be
detected in our assay. However, according to the data from our mixture analysis the levels in the
first timepoint would most likely be less than 1%, and therefore we feel these events should be
classified as superinfections. The relatively high proportion of superinfected individuals in our
population agrees with other studies of high risk individuals in Africa (13, 14). However, given
the small number of individuals examined, further investigation is needed to estimate the rate
and correlates of superinfection in the Rakai population. In addition, the individuals in this study
were selected based upon a high likelihood of superinfection since they were initially virally
unlinked from their partners, and therefore may not represent the natural rate of superinfection in
the larger HIV-infected population.

NGS is substantially easier and more cost effective compared to previous methods used
to detect superinfection, particularly for screening large numbers of subjects (12, 28). It should
be noted that NGS protocols like ours require specialized equipment that somewhat limits their utility in resource poor settings. The data presented here demonstrate that HIV superinfection can be detected in an accurate and sensitive manner, in a high-throughput environment, and suggest that future studies examining HIV superinfection rates in large cohorts should utilize these types of deep sequencing techniques. The ability to rapidly determine the nature and extent of HIV superinfection could have profound influence on studies of HIV disease, therapeutic interventions, transmission of potential drug resistance, and viral evolution in the population.

Acknowledgements:

The authors would like to thank all the participants of the Rakai cohort, and the staff of the Rakai health science program. We would like to especially thank Susanna Lamers for her assistance in sequence submission. All subjects provided written informed consent for their samples to be stored and used for future HIV related research. The study was approved by the Science and Ethics Committee of the Uganda Virus Research Institute, the Western Institutional Review Board and the Committee on Human Research at Johns Hopkins Bloomberg School of Public Health. There are no conflicting interests or any of the study authors. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was supported in part by funding from the Division of Intramural Research, NIAID, NIH; NIAID (grants R01 A134826 and R01 A134265); NICHD (grant 5P30HD06826); the World Bank STI Project, Uganda; the Henry M. Jackson Foundation; the Fogarty Foundation (grant 5D43TW00010); and the Bill and Melinda Gates Institute for Population and Reproductive Health at JHU.
References


pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. Retrovirology 6:15.


Figure Legends

Figure 1: Mixture analysis of inter-subtype detection by NGS. Neighbor-joining trees of p24 next-generation consensus sequences ($\geq$10 identical reads) of control samples of $A_2$ (A, blue), subtypes $D_1$ (B, green), a mixture of $A_2$ and $D_1$ at a 50:50 ratio (C, red), and a merged tree of all three sample runs (D) are shown. The trees are constructed with a selection of subtype reference sequences and random sequences from individuals in Rakai shown in black. Brackets demonstrate the source of different clades within the merged trees. Bootstrap values higher than 80% are shown for non-merged trees (1000 replicates).

Figure 2: Inter- and intra-subtype detection of the p24 region by NGS. Neighbor-joining trees of p24 next-generation consensus sequences ($\geq$10 identical reads) of inter-subtype mixtures of $A_2$ and $D_1$ (red) at 95:5 (A) and 99:1 (B) ratios are shown. Intra-subtype mixtures of $A_2$ and $A_1$ (C, blue), and $D_3$ and $D_4$ (D, green) at the ratio of 95:5 are shown. The trees are constructed with a selection of subtype reference sequences and random sequences from individuals in Rakai shown in black. Brackets demonstrate the source of different clades within the merged trees. Bootstrap values higher than 80% are shown (1000 replicates).

Figure 3: Inter-subtype detection of the gp41 region by NGS. Neighbor-joining trees of gp41 next-generation consensus sequences ($\geq$10 identical reads) of inter-subtype mixtures of $A_4$ and $D_5$ (red) at 50:50 (A), 95:5 (B), 99:1 (C), and 99.9:0.1 (D) ratios are shown. The trees are constructed with a selection of subtype reference sequences and random sequences from individuals in Rakai shown in black. Brackets demonstrate the source of different clades within the merged trees. Bootstrap values higher than 80% are shown (1000 replicates).
Figure 4: Detection of HIV superinfection in p24 region. Neighbor-joining trees of HIV p24 next-generation consensus sequences (≥10 identical reads) from female_C1 (green) in 2002 (A) and 2005 (B) are shown. The trees are constructed with a selection of subtype reference sequences and random sequences from individuals in Rakai shown in black. Superinfecting strains are shown with a circle. Brackets demonstrate the individual’s HIV subtypes within the trees. Bootstrap values higher than 80% are shown (1000 replicates).

Figure 5: Detection of HIV superinfection in p24 region. Neighbor-joining trees of HIV p24 next-generation consensus sequences (≥10 identical reads) from female_C3 (blue) in 2002 (A) and 2005 (B) are shown. The trees are constructed with a selection of subtype reference sequences and random sequences from individuals in Rakai shown in black. Superinfecting strains are shown with a circle. Brackets demonstrate the individual’s HIV subtypes within the trees. Bootstrap values higher than 80% are shown (1000 replicates).
Table 1: Sequence read totals and consensus distribution for pure subtype samples and mixture analysis

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Region</th>
<th>Total reads</th>
<th>Total consensus (≥10)</th>
<th>Plate Divisions</th>
<th>Minor variant consensus (≥10)</th>
<th>% minor variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>p24</td>
<td>20808</td>
<td>162</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>A₂</td>
<td>p24</td>
<td>21222</td>
<td>104</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>D₁</td>
<td>p24</td>
<td>19596</td>
<td>163</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>D₂</td>
<td>p24</td>
<td>17446</td>
<td>139</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>D₃</td>
<td>p24</td>
<td>13137</td>
<td>73</td>
<td>4</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>D₄</td>
<td>p24</td>
<td>11442</td>
<td>81</td>
<td>4</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>A₂:D₁ 50:50</td>
<td>p24</td>
<td>18702</td>
<td>146</td>
<td>1</td>
<td>D₁/115</td>
<td>78.8</td>
</tr>
<tr>
<td>A₂:D₁ 95:5</td>
<td>p24</td>
<td>18694</td>
<td>133</td>
<td>1</td>
<td>D₁/86</td>
<td>64.7</td>
</tr>
<tr>
<td>A₂:D₁ 99:1</td>
<td>p24</td>
<td>19216</td>
<td>117</td>
<td>1</td>
<td>D₁/63</td>
<td>53.8</td>
</tr>
<tr>
<td>D₃:A₂ 99.9:0.1</td>
<td>p24</td>
<td>11299</td>
<td>70</td>
<td>4</td>
<td>A₂/0</td>
<td>0</td>
</tr>
<tr>
<td>A₁:A₂ 95:5</td>
<td>p24</td>
<td>18855</td>
<td>138</td>
<td>1</td>
<td>A₂/20</td>
<td>14.5</td>
</tr>
<tr>
<td>D₁:D₃ 95:5</td>
<td>p24</td>
<td>17783</td>
<td>167</td>
<td>1</td>
<td>D₃/0</td>
<td>0</td>
</tr>
<tr>
<td>D₃:D₄ 95:5</td>
<td>p24</td>
<td>7461</td>
<td>44</td>
<td>4</td>
<td>D₃/11</td>
<td>25</td>
</tr>
<tr>
<td>A₁</td>
<td>gp41</td>
<td>9328</td>
<td>55</td>
<td>4</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>A₄</td>
<td>gp41</td>
<td>7678</td>
<td>58</td>
<td>4</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>D₃</td>
<td>gp41</td>
<td>8461</td>
<td>40</td>
<td>4</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>D₆</td>
<td>gp41</td>
<td>6310</td>
<td>47</td>
<td>4</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>A₄:D₃ 50:50</td>
<td>gp41</td>
<td>7229</td>
<td>58</td>
<td>4</td>
<td>D₃/19</td>
<td>32.8</td>
</tr>
<tr>
<td>A₁:D₉ 95:5</td>
<td>gp41</td>
<td>7561</td>
<td>51</td>
<td>4</td>
<td>D₉/4</td>
<td>7.8</td>
</tr>
<tr>
<td>A₁:D₉ 99:1</td>
<td>gp41</td>
<td>7990</td>
<td>64</td>
<td>4</td>
<td>D₉/4</td>
<td>6.3</td>
</tr>
<tr>
<td>A₁:D₉ 99.9:0.1</td>
<td>gp41</td>
<td>7137</td>
<td>54</td>
<td>4</td>
<td>D₉/1</td>
<td>1.9</td>
</tr>
<tr>
<td>A₁:A₄ 95:5</td>
<td>gp41</td>
<td>9531</td>
<td>59</td>
<td>4</td>
<td>A₄/14</td>
<td>23.7</td>
</tr>
<tr>
<td>D₅:D₆ 95:5</td>
<td>gp41</td>
<td>7532</td>
<td>59</td>
<td>4</td>
<td>D₅/2</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Table 2: Subject viral loads, sequence read totals, and consensus subtype distribution

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Region</th>
<th>p24 Consensus subtype</th>
<th>gp41 Consensus subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Viral Load</td>
<td>Total reads</td>
<td>Consensus (≥10)</td>
</tr>
<tr>
<td>Female_C1_02</td>
<td>4.43</td>
<td>18186</td>
<td>91</td>
</tr>
<tr>
<td>Female_C1_05</td>
<td>5.31</td>
<td>17874</td>
<td>153</td>
</tr>
<tr>
<td>Male_C1_02</td>
<td>3.03</td>
<td>16378</td>
<td>138</td>
</tr>
<tr>
<td>Male_C1_05</td>
<td>4.52</td>
<td>19554</td>
<td>148</td>
</tr>
<tr>
<td>Female_C2_02</td>
<td>4.94</td>
<td>10921</td>
<td>72</td>
</tr>
<tr>
<td>Female_C2_05</td>
<td>5.27</td>
<td>13331</td>
<td>102</td>
</tr>
<tr>
<td>Male_C2_02</td>
<td>5.35</td>
<td>12651</td>
<td>16</td>
</tr>
<tr>
<td>Male_C2_05</td>
<td>6.30</td>
<td>12076</td>
<td>99</td>
</tr>
<tr>
<td>Female_C3_02</td>
<td>4.83</td>
<td>8119</td>
<td>50</td>
</tr>
<tr>
<td>Female_C3_05</td>
<td>5.07</td>
<td>6980</td>
<td>39</td>
</tr>
<tr>
<td>Male_C3_02</td>
<td>5.30</td>
<td>11779</td>
<td>93</td>
</tr>
<tr>
<td>Female_C4_02</td>
<td>4.08</td>
<td>13525</td>
<td>93</td>
</tr>
<tr>
<td>Female_C4_05</td>
<td>4.30</td>
<td>15274</td>
<td>95</td>
</tr>
<tr>
<td>Male_C4_02</td>
<td>5.04</td>
<td>17534</td>
<td>134</td>
</tr>
<tr>
<td>Male_C5_02</td>
<td>4.86</td>
<td>16915</td>
<td>105</td>
</tr>
<tr>
<td>Male_C5_05</td>
<td>5.66</td>
<td>15038</td>
<td>109</td>
</tr>
<tr>
<td>Female_C5_02</td>
<td>&lt;2.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Female_C6_02</td>
<td>4.10</td>
<td>20767</td>
<td>110</td>
</tr>
<tr>
<td>Female_C6_05</td>
<td>3.80</td>
<td>17158</td>
<td>93</td>
</tr>
<tr>
<td>Male_C6_02</td>
<td>5.69</td>
<td>19648</td>
<td>89</td>
</tr>
<tr>
<td>Female_C7_02</td>
<td>4.13</td>
<td>14882</td>
<td>133</td>
</tr>
<tr>
<td>Female_C7_05</td>
<td>4.77</td>
<td>12154</td>
<td>102</td>
</tr>
<tr>
<td>Male_C7_02</td>
<td>&lt;2.6</td>
<td>12799</td>
<td>90</td>
</tr>
<tr>
<td>Male_C8_02</td>
<td>4.33</td>
<td>7971</td>
<td>60</td>
</tr>
<tr>
<td>Male_C8_05</td>
<td>5.10</td>
<td>8261</td>
<td>67</td>
</tr>
<tr>
<td>Female_C8_02</td>
<td>5.31</td>
<td>8251</td>
<td>39</td>
</tr>
<tr>
<td>Male_C9_02</td>
<td>4.70</td>
<td>10655</td>
<td>62</td>
</tr>
<tr>
<td>Male_C9_05</td>
<td>4.88</td>
<td>10158</td>
<td>64</td>
</tr>
<tr>
<td>Female_C9_02</td>
<td>4.73</td>
<td>9927</td>
<td>55</td>
</tr>
</tbody>
</table>

*Samples with superinfecting strains are shown in bold. Corresponding partner read totals and consensus sequence are indicated for the individual’s samples which are labeled with their gender, couple number, and year of sample draw. Viral loads are indicated as the log_{10} value.
Figure 1:

A) Pure A_2

B) Pure D_1

C) A_2:D_1 50:50

D) Merged
Figure 2:

A)  

B)  

C)  

D)
Figure 3:

A) $A_4:D_5$ 50:50

B) $A_4:D_5$ 95:5

C) $A_4:D_5$ 99:1

D) $A_4:D_5$ 99.9:0.1
Figure 4:
A)  

B)
Figure 5:

A) B)

D

99

97

99 1

1

81

99
ERRATUM

Identification of HIV Superinfection in Seroconcordant Couples in Rakai, Uganda, by Use of Next-Generation Deep Sequencing

Andrew D. Redd, Aleisha Collinson-Streng, Craig Martens, Stacy Ricklefs, Caroline E. Mullis, Jordyn Manucci, Aaron A. R. Tobian, Ethan J. Selig, Oliver Laeyendecker, Nelson Sewankambo, Ronald H. Gray, David Serwadda, Maria J. Wawer, Stephen F. Porcella, and Thomas C. Quinn on behalf of the Rakai Health Sciences Program

Laboratory of Immunoregulation, DIR, NIAID, NIH, Baltimore, Maryland; Genomics Unit, Research Technologies Section, Rocky Mountain Laboratories, DIR, NIAID, NIH, Hamilton, Montana; Johns Hopkins Medical Institute, Johns Hopkins University, Baltimore, Maryland; Rakai Health Sciences Program, Kalisizo, Uganda; Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland; School of Medicine, Makerere University, Kampala, Uganda; and School of Public Health, Makerere University, Kampala, Uganda

Volume 49, no. 8, p. 2859–2867, 2011. Supplemental file 2: For Gp41-Forward Primer Sequences-E55, the sequences for MID1 to MID14 were incorrect. A revised supplemental file has been posted.