Background: To better understand the relationships among HIV-1 viruses in linked transmission pairs, we sequenced several samples representing HIV transmission pairs from the Zambia Emory HIV Research Project (Lusaka, Zambia) using Single Molecule, Real-Time (SMRT®) Sequencing.

Methods: Single molecules were sequenced as full-length (9.4 kb) amplicons directly from PCR products without shearing. This resulted in multiple, fully-phased, complete HIV-1 genomes for each patient. We examined Single Genome Amplification (SGA) prepped samples, as well as samples containing complex mixtures of genomes. We detailed mathematical techniques used in viral variant subspecies identification, including clustering distance metrics and mutual information, which were used to derive multiple de novo full-length genome sequences for each patient. Whole genome consensus estimates for each sample were made. Genome reads were clustered using a simple distance metric on aligned reads. Appropriate thresholds were chosen to yield distinct clusters of HIV-1 genomes within samples. Mutual information between columns in the genome alignments was used to measure dependence. In silico mixtures of reads from the SGA samples were made to simulate samples containing exactly controlled complex mixtures of genomes and our clustering methods were applied to these complex mixtures.

Results: SMRT Sequencing data contained multiple full-length (>9 kb) continuous reads for each sample. Simple whole-genome consensus estimates easily identified transmission pairs. Clustering of genome reads showed diversity differences between samples, allowing characterization of the quasi-species diversity comprising the patient viral populations across the full genome. Mutual information identified possible dependencies of different positions across the full HIV-1 genome. The SGA consensus genomes agreed with prior Sanger sequencing. Our clustering methods correctly segregated reads to their correct originating genome for the synthetic SGA mixtures.

Conclusions: SMRT Sequencing yields long-read sequencing results from individual DNA molecules with a rapid time-to-result. These attributes make it a useful tool for continuous monitoring of viral populations. The single-molecule nature of the sequencing method allows us to estimate variant subpopulations and relative abundances by counting methods. The results open up the potential for reference-agnostic and cost-effective full genome sequencing of HIV-1.

ACKNOWLEDGEMENTS: The authors gratefully acknowledge: US National Institutes of Health (R01s AI-064060 (Hunter); MH-66767 (Allen); AI-51321 (Allen); F32 AI-084409 (Schaefer)), International AIDS Vaccine Initiative, US Centers for Disease Control, Fogarty AIDS International Training in Research Program FIC 2D43 TW001042, Social & Behavioral and Virology Cores of the Emory Center for AIDS Research through P30 AI050409.