Rapid Sequencing of HIV-1 Genomes as Single Molecules from Simple and Complex Samples

Michael P. S. Brown[1], Yan Guo[1], Malinda Schaefer[2], William Kilembe[3], Susan Allen[2], Eric Hunter[2], and Ellen E. Paxinos[1]
[1]Pacific Biosciences®, Menlo Park, CA [2]Emory University, Atlanta, GA [3]Zambia Emory HIV Research Project, Lusaka, Zambia

Abstract

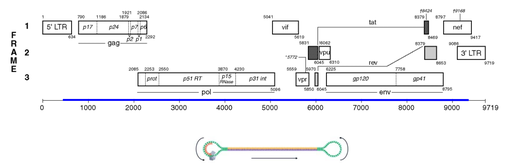
Background: To better understand the relationships among HIV-1 viruses in linked transmission pairs, we sequenced several samples representing HIV transmission pairs from the Zambia Emory HIV Research Project (Lusaka, Zambia) using Single Molecule, Real-Time (SMRT®) Sequencing.

Methods: Single molecules were sequenced as full-length (9.6 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 detail 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.

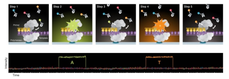
Results: SMRT Sequencing data contained multiple full-length (>9 kb) continuous reads for each sample. Simple whole-genome consensus estimates easily identified transmission pairs. Clustering of genome reads showed diversity differences between samples, allowing characterization of the quasi-species 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 subspecies and relative abundances by counting methods. The results open up the potential for reference-agnostic and cost effective full genome sequencing of HIV-1.

SMRT® Sequencing of Intact HIV-1 Genomes



Full Genome – 9,084 bp



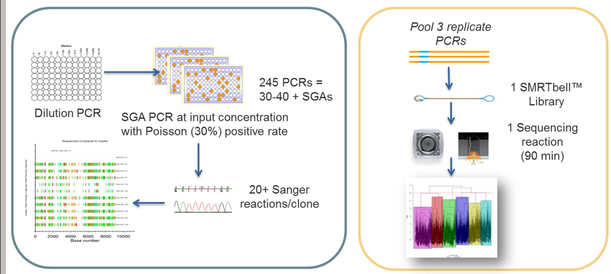
Measuring the Genetic Diversity of Viral Infection

Single Genome Amplification

30 viral genomes
8 – 10 days

PacBio Sequencing

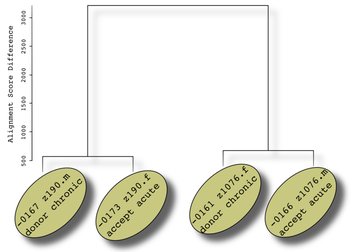
~1000 - 3000 viral genomes
< 1 day



Identify Transmission Pairs Correctly

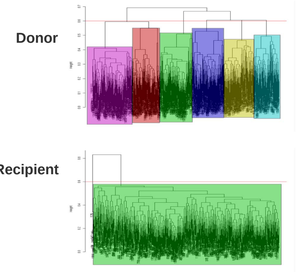
- Sequence four clinical samples from two linked transmission pairs
- Estimate a consensus genome sequence for each sample
- Pairwise align all four consensus sequences
- Result: The true transmission pairs are easily identified

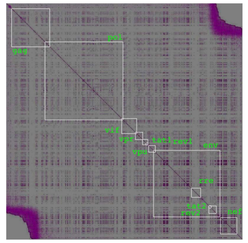
Transmission Pair Identification
by Pairwise Consensus Alignments



Genome Diversity and Pairwise Positional Dependence

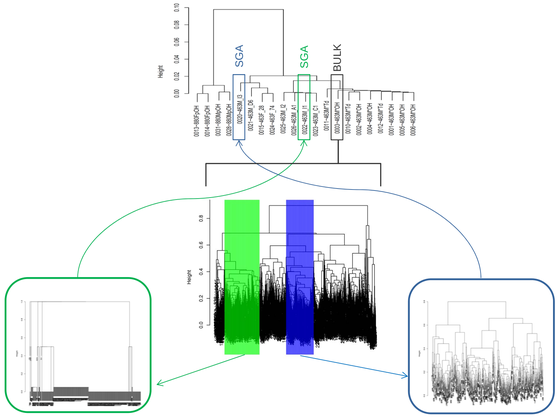
- Relate genome reads: Chronic donor has complex mixture of genomes. Acute recipient has single genome.
- Relate HIV genome positions: pairs of positions have some statistical dependence by mutual information (9k by 9k matrix).





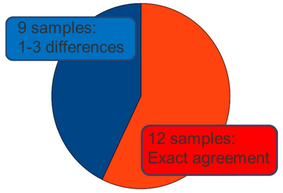
$$\sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

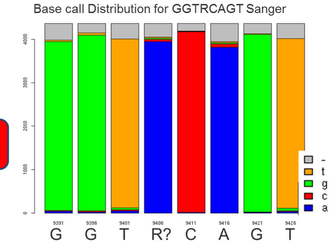
Clustering Consensus Recovers SGA Genomes



Consensus Genomes Agree Exactly with Sanger

- For 12 genomes, we estimate the **entire** HIV-1 genome **exactly** as compared to the Sanger estimates for the entire 9 kb length.
- For 9 genomes, there are 1 to 3 differences (including corrections)
  - Sanger calls ambiguous R base
  - PacBio corrects the call to an A with 3948 counts (next highest is 59 counts)





Conclusions

- Complete HIV-1 genomes from single molecules
  - Sanger-quality, fully phased across entire genome
  - One SMRT Cell of sequencing in 90 minutes
  - From samples with possibly complex mixtures of genomes
- Full genomic characterization of HIV-1 clinical transmission events
- Sequencing and methods useful in other applications
  - Viral
  - Cancer
  - Metagenomics

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.