Complete HIV-1 Genomes from Single Molecules: Diversity Estimates in Two Linked Transmission Pairs Using Clustering and Mutual Information

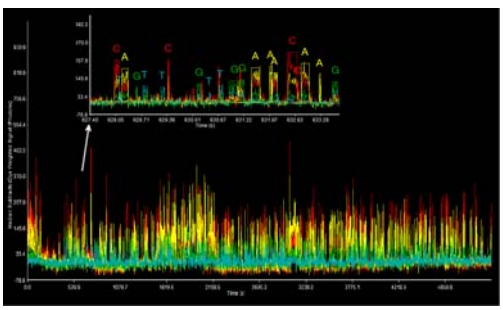
Michael P.S. Brown[1], Malinda Schaefer[2], Yan Guo[1], William Kilembe[3], Susan Allen[2], Eric Hunter[2], and Ellen Paxinos[1].

[1] Pacific Biosciences, Menlo Park, CA, USA [2]Emory University, Atlanta, GA, USA [3]Zambia Emory HIV Research Project, Lusaka, Zambia.

Abstract

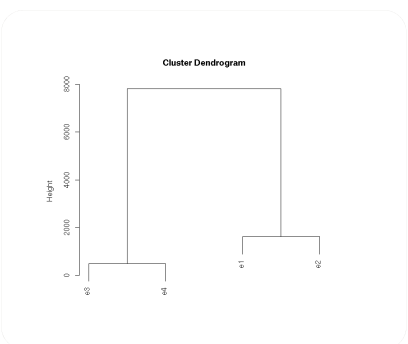
Single-Molecule, Real Time (SMRT®) sequencing yields long-read sequencing results from individual DNA molecules with a rapid time-to-result. These attributes make it a useful tool for continuous monitoring of viral populations. The single-molecule nature of our sequencing method allows us to estimate variant subspecies and relative abundances by simple counting methods. We detail mathematical techniques used in viral variant subspecies identification including clustering distance metrics and mutual information. Specifically, we sequenced full-length HIV-1 genomes from single molecules in order to better understand the relationships between the specific sequences of transmitted viruses in two linked transmission pairs.

Full-Length HIV-1 Genome in 1 Read



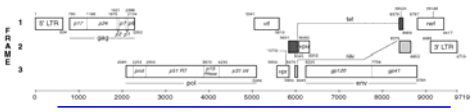
HIV Transmission Pair Identification

- Full Genome reads (9,055 bp) verify transmission pairs by simple hierarchical clustering of patient consensus sequences.
- Transmission pairs are each monophyletic, consistent with epidemiological linkage analysis.



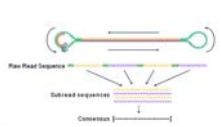
ACKNOWLEDGEMENTS: The authors gratefully acknowledge: US National Institutes of Health (R01s AI-064060 (Hunter); MH-66767 (Allen); AI-51321 (Allen); F32 AI-084409 (Schaefer)), International AIDS Vaccine Initiative, US Centers for Disease Control, Fogarty AIDS International Training in Research Program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

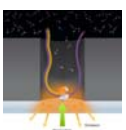
Methods

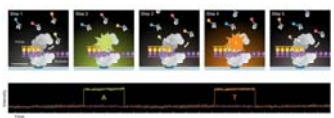


Full Genome – 9,084 bp

HXB2 gene map from http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html



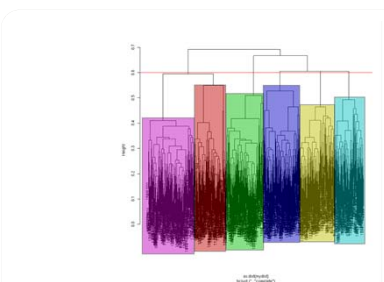




$$\sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

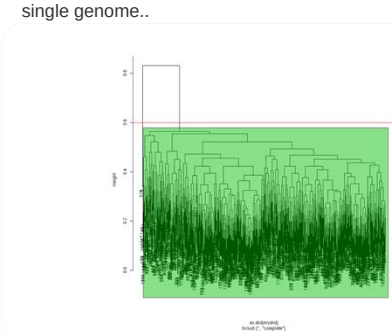
Species Diversity in Donor

- Emerging clusters indicate consistent patterns of variation and suggest variant species.



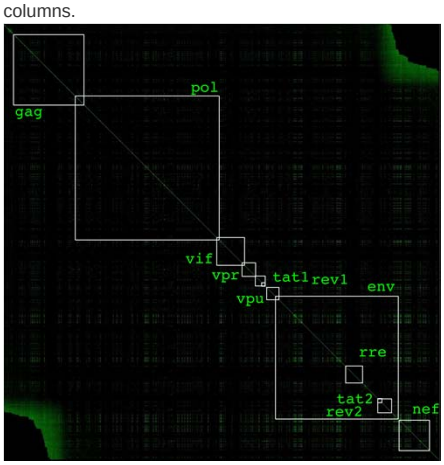
Species Diversity in the Linked Recipient

- Only a single cluster below the threshold suggesting a single genome..



HIV Genome Mutual Information

- Take single run from single patient from the Donor HIV result and compute mutual information between all pairs of columns.



Distance Metrics, Clustering, and Mutual Information

- Our simple distance metric uses the number of mismatches in aligned reads.
- Binomial bounds can be used to estimate distributions given expected error rate.
- The importance of feature selection is apparent with larger genomes where the expected noise over a large number of positions might be expected to swamp truly informative positions.
- These bounds give us guidance on appropriate thresholds however our feature selection step makes the analysis more complex.
- We use simple complete-linkage hierarchical clustering for fully-explored deterministic results. However more sophisticated techniques like Dirichlet processes could be used for clustering.
- We are exploring bounds from the distribution of mutual information to gauge significance.

Conclusions

- Near full-length HIV genomes were sequenced from clinical samples as single molecules, in single reads, using a $99 and 90-minute sequencing run.
- Simple consensus estimates correctly identified transmission paired samples.
- Clustering methods using multiple alignments and simple distance functions allowed us to estimate complex mixtures of genomes from a single sample, single prep, and single sequencing run.
- Mutual information was computed on the multiple alignment of reads. Long reads allowed us to estimate long-range interactions directly without any need to “phase” mutations.