High-Accuracy, Single-Base Resolution of Near-Full-Length HIV Genomes

Michael P. S. Brown[1], Melissa E. Laird[1], Lance Hepler[1], Yan Guo[1], Colleen Ludka[1], Ellen E. Paxinos[1], Richard Lempicki[2], Da Wei Huang[2], Qian Chen[2], Tomozumi Imamichi[2], Cliff Lane[3], Hiromi Imamichi[3]

[1]Pacific Biosciences, Menlo Park, CA [2]Leidos Biomedical Research, Frederick, MD [3]NIAID, NIH, DHHS, Bethesda, MD

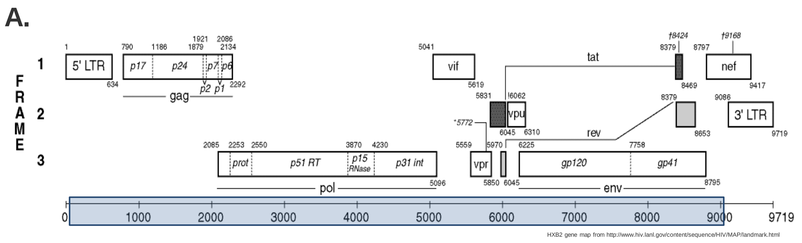
Introduction

The HIV-1 proviral reservoir is incredibly stable, even while undergoing anti-retroviral therapy, and is seen as the major barrier to HIV-1 eradication. Identifying and comprehensively characterizing this reservoir will be critical to achieving an HIV cure. Historically, this has been a tedious and labor intensive process, requiring high-replicate single-genome amplification reactions and reconstruction into full-length genomes by algorithmic imputation. Novel deep methods allow, for the first time, near-full-length HIV genomes.

Objectives

To develop single-molecule sequencing and analysis methods to determine the exact identity and relative abundances of near-full-length HIV genomes from samples containing mixtures of genomes without shearing or complex bioinformatic reconstruction.

Measuring the Diversity of Viral Infection



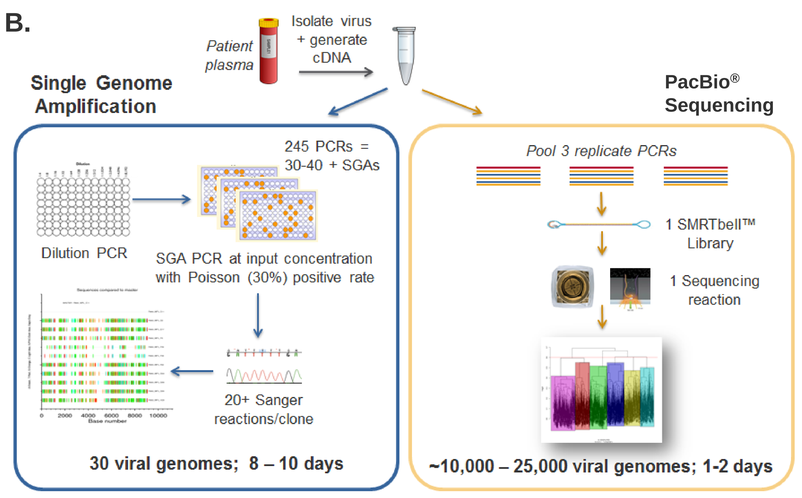


Figure 1. A) Clonal near-full-length amplicons were derived from single genome amplification (SGA) of primary proviral isolates or well-documented control strains. **B)** Viral genome SGAs have been analyzed via limiting dilution and overlapping Sanger sequencing, which then required complex bioinformatic reconstruction; PacBio® sequencing allows for pooling and sequencing of thousands of near-full-length HIV SGA genomes in one reaction.

PacBio® Sequencing of Intact HIV Genomes

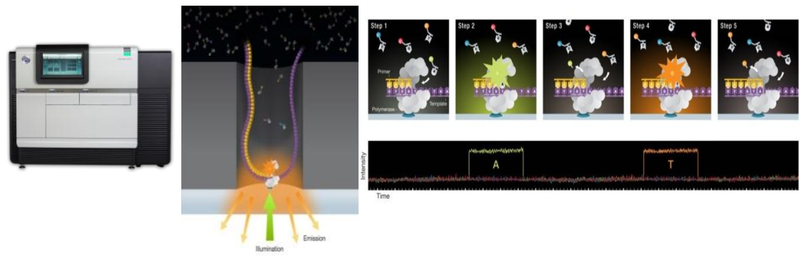
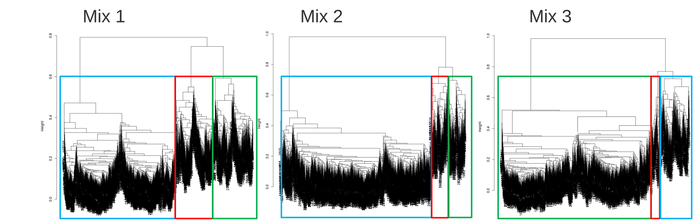


Figure 2. Clonal SMRTbell™ libraries were mixed at various abundances and sequenced as near-full-length (~9 kb) amplicons without shearing on the PacBio® RS II using P4-C2 chemistry and standard protocols.

De-convolute Mixture of HIV-1 Genomes

- Sequenced three samples containing synthetic mixtures of HIV-1 clones at different abundances



EXP (%):	60	20	20	80	10	10	80	5 15
OBS (%):	59	21	20	81	10	9	80	5 15

RESULT: Three genomes estimated in each sample with 100% consensus accuracies across the entire ~9 kb genomes and within 1% of expected abundances.

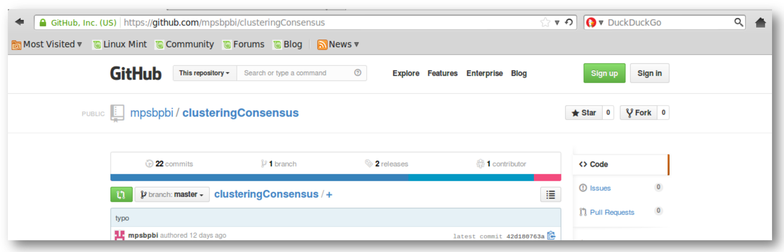
CluCon Strong Clustering Technique

- Consensus Sequence Clustering Under Binomial Bounds:
  - Generate alignment to run-specific consensus
  - Identify minor variant column regions
  - If no minor variants, return consensus
  - If divergence between read pairs is larger than expected by noise, then separate out as sub-cluster.
  - Recursively work on sub-clusters
- Clones in mixtures at ~5% divergence (or 500 positions)
- High statistical confidence in cluster consensus calls at this degree of genetic divergence.

CluCon Software

- CluCon software and data available:

  https://gibhub.com/mpsbpbi/clusteringConsensus



Single Base Resolution of Mixed Genomes

- Performed SMRT Sequencing on a synthetic mixture of 5 HIV clones that differed by only one or two nucleotides.
- Analyzed using CluCon software
- 5 strains differed by 1-2 bases (highlighted by color below)
  - Two genomes differ by one base (1, 3)
  - Two genomes differ by two bases (0, 4)
  - Single genome had no differences (2)

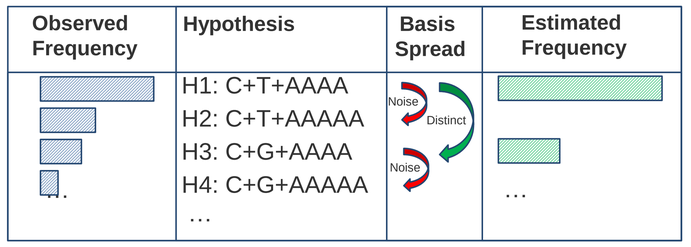
Estimated Genome Haplotypes and Relative Abundances:

```
0 'AAAT+AAAAAAG+GT+GCA+ATTTACC+ACCC+' 25.2%
1 'AATT+AAAAAAG+AT+GGA+ATTTACC+ACCC+' 21.6%
2 'AAAT+AAAAAAG+AT+GGA+ATTTACC+ACCC+' 21.0%
3 'AAAT+AGAAAAG+AT+GGA+ATTTACC+ACCC+' 19.6%
4 'AAAT+AAAAAAG+AT+GGA+ATTTTACC+ACCC+' 12.6%
```

RESULT: 100% accurate near-full-length HIV-1 genomes.

CluCon Fine Clustering Technique

- Consensus Linear System Basis Function De-convolution
- Modified CluCon Strong: call when the number of minor variant column regions is too small for binomial bounds to have power
- Tally read identities at the variant sites; putative haplotypes count
- Estimate true haplotypes from observed reads by discounting noisy haplotypes using basis functions representing sequencing noise



Conclusions

Complete characterization of near-full-length HIV-1 genomes with Single Molecule, Real-Time Sequencing

- Sanger-quality, fully phased across entire genome
- *De novo* clustering technique is reference agnostic
- Single run yields 1,000s of distinct HIV genomes
- Deconvolute samples containing complex mixtures of genomes with single base resolution
- Provides unique capacity to identify and characterize integrated, intact proviral HIV genomes

Acknowledgements. This work was funded in part through the intramural research program of the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (Bethesda, Maryland) and in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E