Sequencing Complex Mixtures of HIV-1 Genomes with Single-Base Resolution

Michael P. S. Brown[1], Melissa Laird[1], Lance Hepler[1], Yan Guo[1], Colleen Ludka[1], Ellen E. Paxinos[1], Richard Lempicki[2], Da Wei Huang[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

Abstract

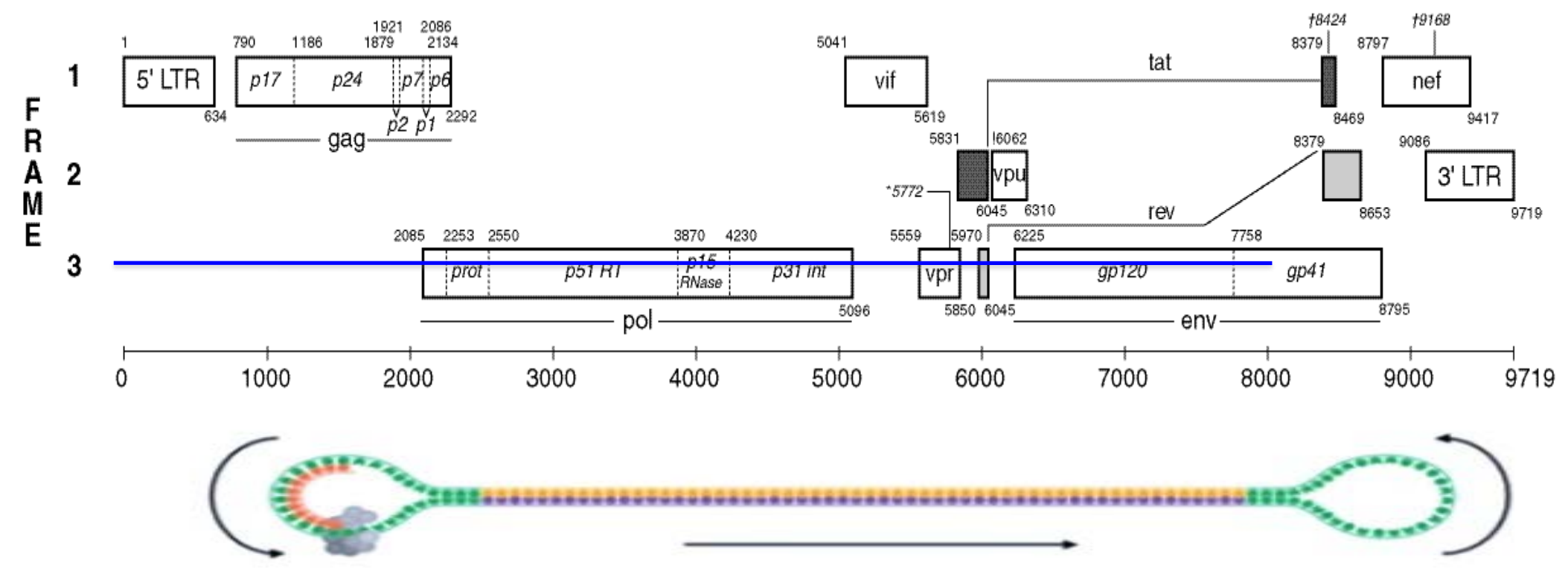
A large number of distinct HIV-1 genomes can be present in a single clinical sample from a patient chronically infected with HIV-1. We examined samples containing complex mixtures of near-full-length HIV-1 genomes. Single molecules were sequenced as near-full-length (9.6 kb) amplicons directly from PCR products without shearing.

Mathematical analysis techniques deconvolved the complex mixture of reads into estimates of distinct near-full-length viral genomes with their relative abundances. We correctly estimated the originating genomes to single-base resolution along with their relative abundances for mixtures where the truth was known exactly by independent sequencing methods. Correct estimates were made even when genomes diverged by a single base.

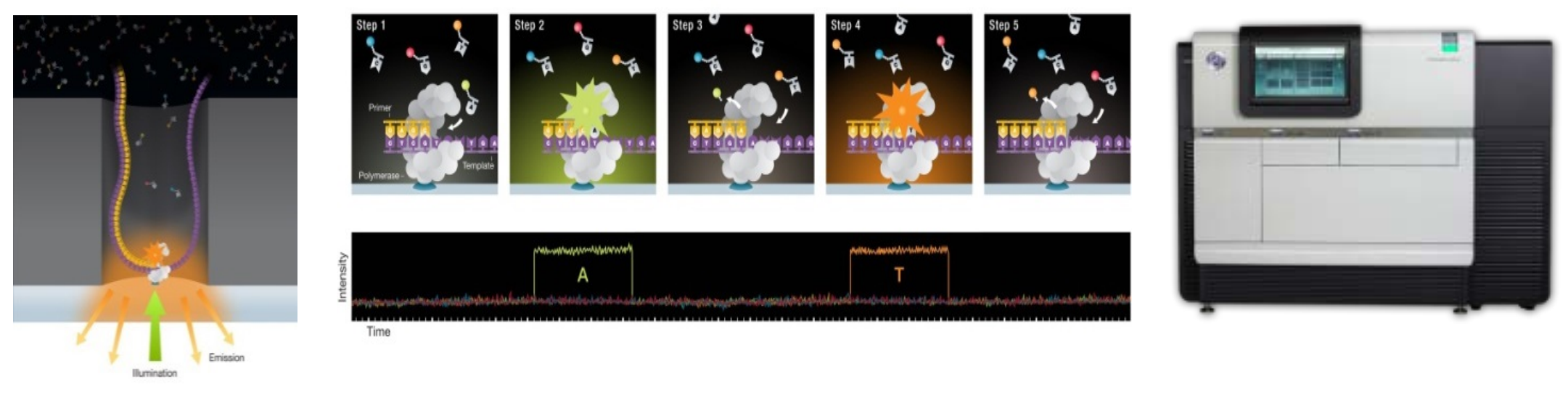
Minor abundances of 5% were reliably detected. SMRT® Sequencing data contained near-full-length continuous reads for each sample including some runs with greater than 10,000 near-full-length-genome reads in a three-hour collection time. SMRT Sequencing yields long-read sequencing results from individual DNA molecules with a rapid time-to-result. The single-molecule, full-length nature of the sequencing method allows us to estimate variant subspecies and relative abundances even from samples containing complex mixtures of genomes that differ by single bases.

These results open the possibility of cost-effective full-genome sequencing of HIV-1 in mixed populations for applications such as incorporated-HIV-1 screening. In screening, genomes can differ by one to many thousands of bases and the ability to measure them can help scientifically inform treatment strategies.

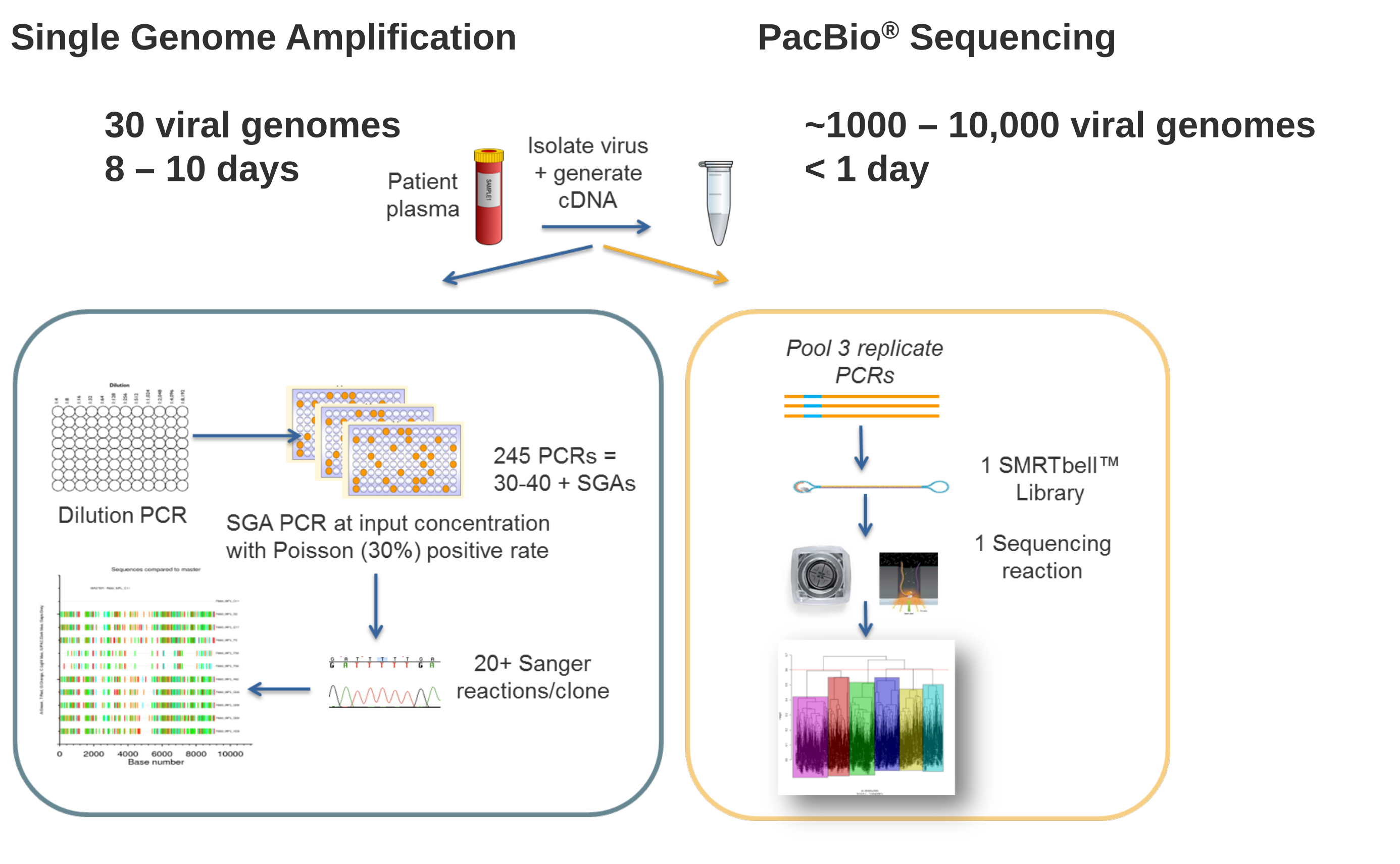
SMRT® Sequencing of Intact HIV-1 Genomes



Full Genome – 9,084 bp



Two Methods to Measure the Genetic Diversity of Viral Infection



Single Genome Amplification

30 viral genomes
8 – 10 days

PacBio® Sequencing

~1000 – 10,000 viral genomes
< 1 day

Patient plasma

Isolate virus + generate cDNA

Dilution PCR

245 PCRs = 30-40 + SGAs

SGA PCR at input concentration with Poisson (30%) positive rate

20+ Sanger reactions/clone

Pool 3 replicate PCRs

1 SMRTbell™ Library

1 Sequencing reaction

Mixtures of HIV-1: One or Two Base Differences

- Sequenced a sample containing a uniform mixture of HIV-1 clones differing by only one or two bases
- Analyzed using CluCon software
- **RESULT**: Five genomes estimated in the sample with 100% consensus accuracies across the entire ~9 kb genomes

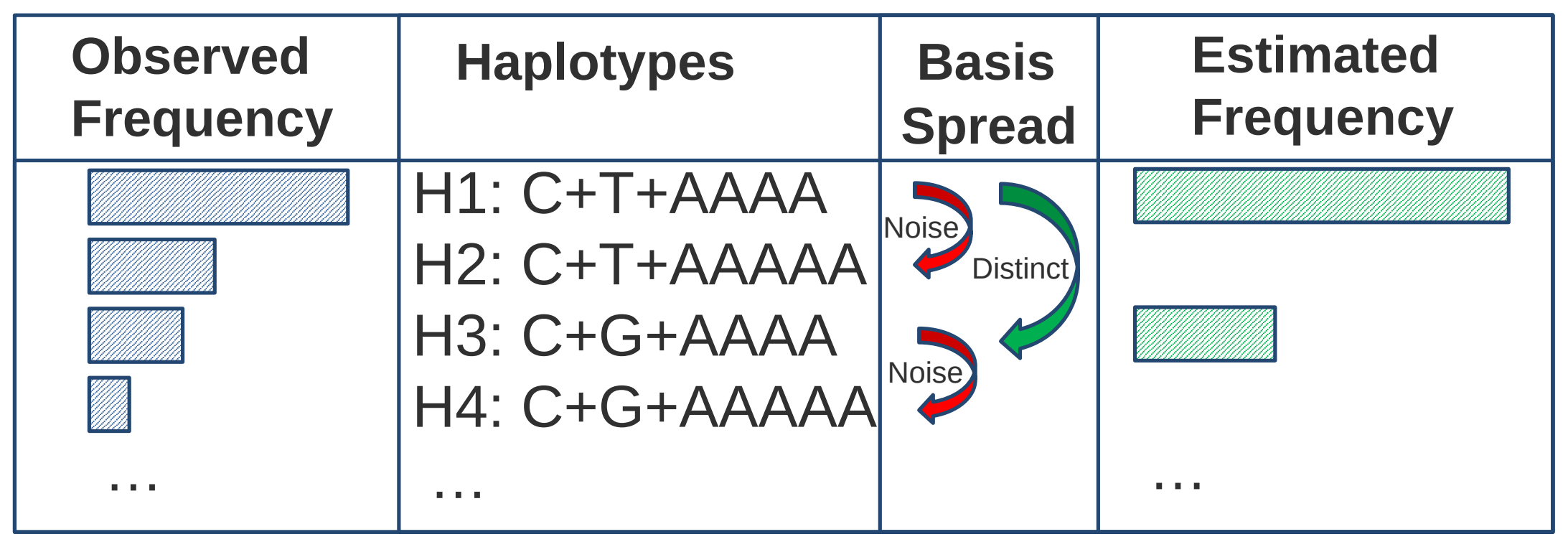
```
Estimated Genome Haplotypes:

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+ATTTTAC+ACCC+' 12.6%
```

Near-Full-Length Genomes from Mixture

- 100% correct near-full-length HIV-1 genomes with differences of five strains in the mixture highlighted by different colors
  - Two genomes differ by one base
  - Two genomes differ by two bases
  - Single genome had no differences

CluCon Fine Clustering Technique

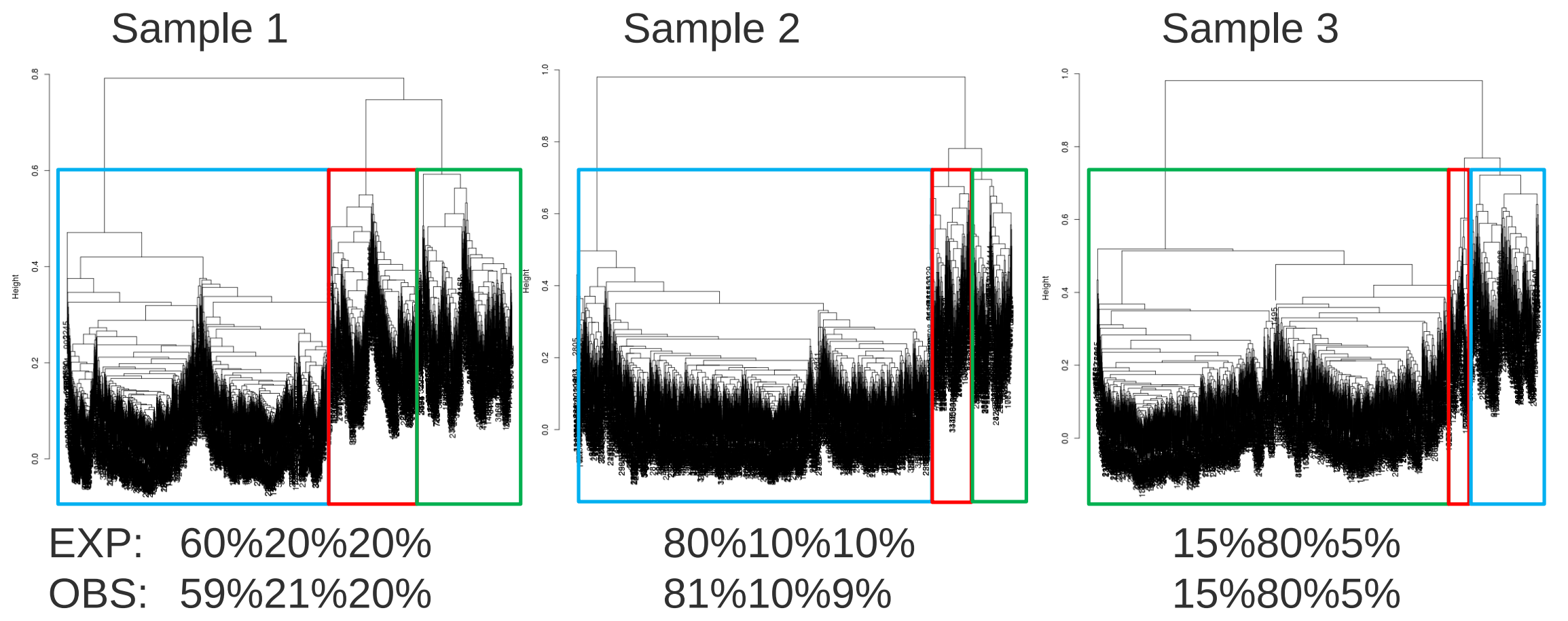
- Consensus Linear System Basis Function Deconvolution
- Identical to CluCon Strong Clustering but called when the number of minor variant column regions is too small for binomial bounds to have power
- Tally the read identities at the variant sites, the putative haplotypes counts
- Estimate true haplotypes from the observed reads by discounting noisy haplotypes using basis functions representing sequencing noise (point spread functions in signal processing)



Observed Frequency	Haplotypes	Basis Spread	Estimated Frequency
	H1: C+T+AAAA H2: C+T+AAAAA H3: C+G+AAAA H4: C+G+AAAAA …	Noise Distinct Noise	
…			…

Mixture of HIV-1 Genomes

- Sequenced three samples containing synthetic mixtures of HIV-1 clones at different abundances
- **RESULT**: Three genomes estimated in each sample with 100% consensus accuracies across the entire ~9 kb genomes and within 1% of expected abundances



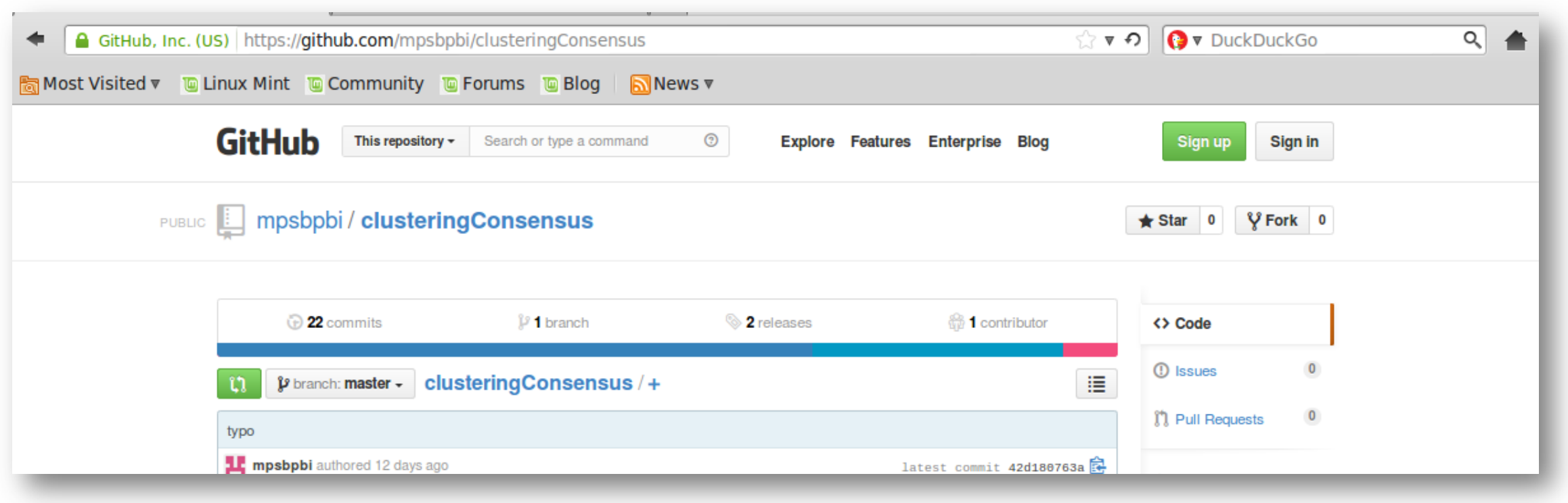
CluCon Strong Clustering Technique

- 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 subcluster.
  - Recursively work on subclusters
- Clones in mixtures are at ~5% divergence (or 500 positions)
- Mathematically almost impossible to confuse one clone for another under sequencing with these divergences

CluCon Software

- CluCon software and data available:

  https://gibhub.com/mpsbpbi/clusteringConsensus



typo

mpsbpbi authored 12 days ago

latest commit 42d180763a

Conclusions

- Complete characterization of HIV-1 genomes from single molecules
  - Sanger-quality, fully phased across entire genome
  - One SMRT Cell of sequencing
  - From samples with possibly complex mixtures of genomes even differing by single bases
- Sequencing and methods useful in other applications
  - Viral
  - Cancer
  - Metagenomics

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

For Research Use Only. Not for use in diagnostic procedures.