Chinese CDC 2018

Using NGS to investigate within-host HIV evolution and latency

Art Poon

Department of Pathology & Laboratory Medicine
Western University

Background

Next-generation sequencing and HIV

- HIV-1 rapidly accumulates genetic variation within each host
- NGS is useful for measuring this diversity
- Applications:
    1. To measure the frequency of different variants (mutations).
    2. To detect rare (clinically significant) variants.
    3. Others?

Background

Aligning NGS from HIV samples

- The diversity of HIV makes these samples difficult to process.
- *de novo* assembly can be successful for a consensus, but we lose variation.
- Others have tried to develop *haplotype reconstruction methods* with limited success (Schirmer *et al*. 2012).
- Continuum of variation, frequent recombination makes this difficult!

Background

Reference mapping

- Easier than *de novo* assembly.
- Build index of reference genome.
- Index stores the location of the fragment in reference genome.
- Fragment sequence serves as "address" for rapid look-up of location.

Background

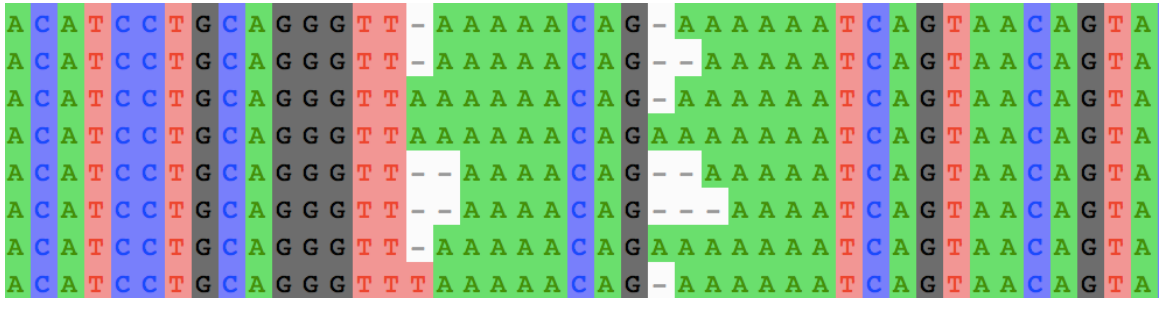
BC Centre for Excellence in HIV/AIDS

- Responsible for HIV-1 genotyping for Canada (except Quebec).
- ~2,000 samples per year.
- Extensive experience with 454 for HIV genotyping, coreceptor tropism testing.
- Nearly 50 publications in the literature.

Background

Problems with the 454

- High error rates in homopolymers (single nucleotide repeats).
- Some homopolymers associated with HIV drug resistance (e.g., K103N) — genome is ~40% A!
- Could not be resolved with bioinformatics — eventually, this platform was discontinued in 2013.



Background

Transition to the MiSeq

- For these reasons, many HIV groups transitioning to Illumina MiSeq.
- Shorter read lengths, no homopolymer issue.
- Many more reads per run (lower cost per base).
- The BC-CFE lab needed a new pipeline for MiSeq data.

MiCall bowtie2

- An open-source mapper that is tolerant of mutations (github.com/BenLangmead/bowtie2)

```
10000 reads; of these:
10000 (100.00%) were paired; of these:
  10000 (100.00%) aligned concordantly 0 times
  0 (0.00%) aligned concordantly exactly 1 time
  0 (0.00%) aligned concordantly >1 times
  ----
  10000 pairs aligned concordantly 0 times; of these:
    0 (0.00%) aligned discordantly 1 time
  ----
  10000 pairs aligned 0 times concordantly or discordantly; of
    20000 mates make up the pairs; of these:
      19979 (99.89%) aligned 0 times
      21 (0.10%) aligned exactly 1 time
      0 (0.00%) aligned >1 times
0.10% overall alignment rate
```

MiCall

Iterative remapping

- Any reads that mapped to initial (seed) reference were used to "evolve" the reference with Python script.

```
10000 reads; of these:
  10000 (100.00%) were paired; of these:
    403 (4.03%) aligned concordantly 0 times
    9597 (95.97%) aligned concordantly exactly 1 time
    0 (0.00%) aligned concordantly >1 times
    ----
    403 pairs aligned 0 times concordantly or discordantly; of these:
      806 mates make up the pairs; of these:
        749 (92.93%) aligned 0 times
        57 (7.07%) aligned exactly 1 time
        0 (0.00%) aligned >1 times
96.25% overall alignment rate
```

MiCall

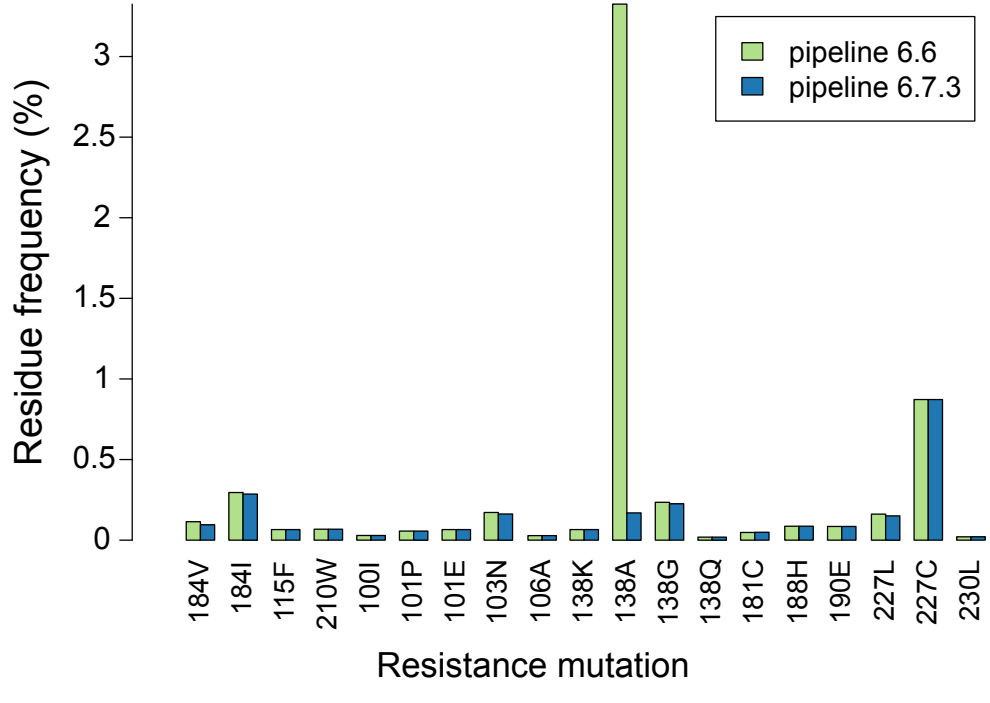
Processing *bowtie2* outputs

- Paired-end reads are merged and discordant calls in overlapping regions are resolved
- Group aligned reads by unique sequence (reduce file size ~half)
- Outputs coverage, nucleotide and amino acid frequencies, and alignment.
- Released this pipeline as open-source, github.com/cfe-lab/MiCall

MiCall

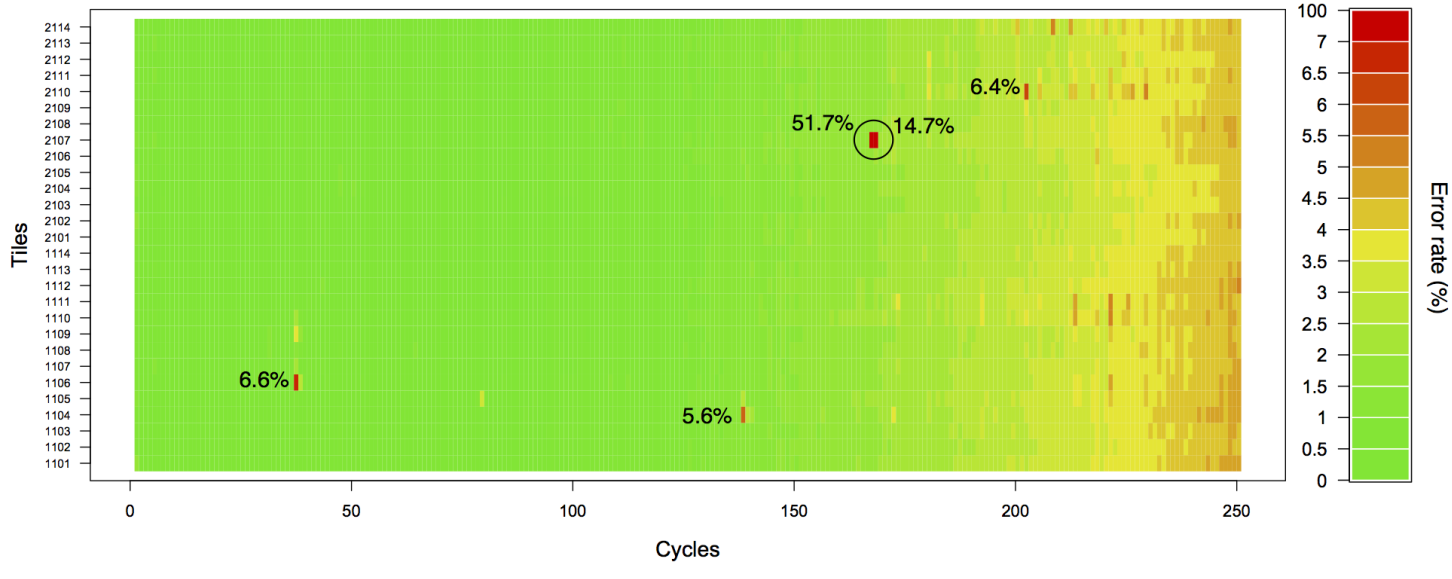
Bad tile-cycle combinations

- Every HIV RT sample in this run had ~3% E138A
- Quality scores for these bases were normal.



Bad tile-cycle combinations

- An Illumina system generates a set of "InterOp" files for every run.
- The file `ErrorMetricsOut.bin` reports ϕX174 error rates.
- The 3% E138A was due to one bad tile-cycle combo - this was not reflected in base quality scores!



MiCall

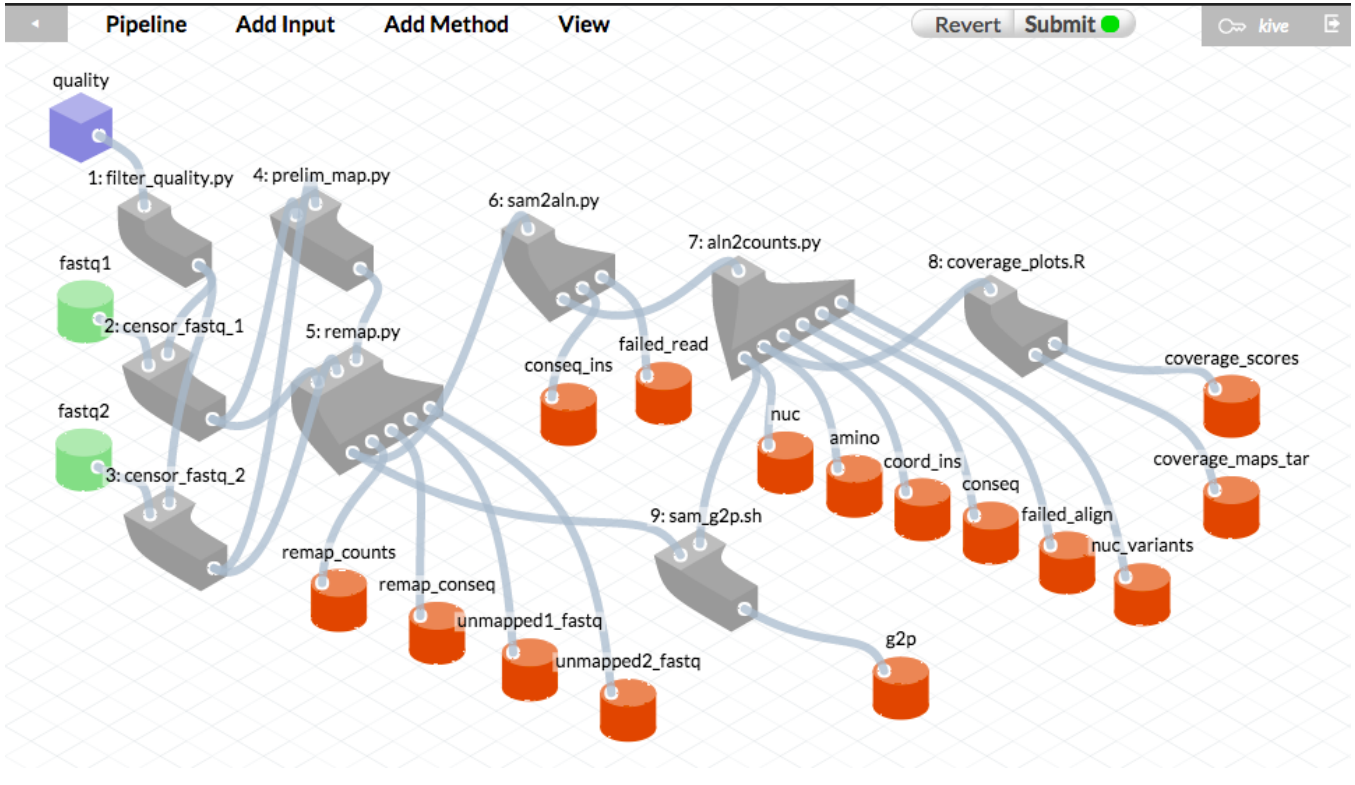
Bad tile-cycle combinations

- We modified MiCall to read this InterOp file and throw out any bad tile-cycle combinations (version 6.7.3+).
- This removed E138A from all samples in that run.
- This affects "amplicon" runs the most, because a given cycle represents the same nucleotide in the affected reads.

MiCall

Version tracking

- CFE lab stores a "fingerprint" of every data input and output handled by MiCall using an in-house software called Kive (github.com/cfe-lab/Kive).



MiCall

Open-source licensing

- Released source code under AGPL-3.0 license at github.com/cfe-lab/MiCall
- Free to use, modify and redistribute with attribution.
- Developing a more user friendly version at github.com/PoonLab/MiCall-Lite

Outline

3 applications of NGS

1. Finding new HIV drug resistance polymorphisms
2. Measuring the latent reservoir
3. Dating the latent reservoir

Drug resistance
INSTIs

- Integrase strand transfer inhibitors.
- Difficult for HIV to evolve resistance, used for salvage therapy.
- Postdoc **Mariano Avino** with **Joint Clinical Research Centre** (JCRC) in Uganda and **Case Western Reserve University**.
- Examine virological failure in non-B patients failing INSTI-based regimens.

JCRC

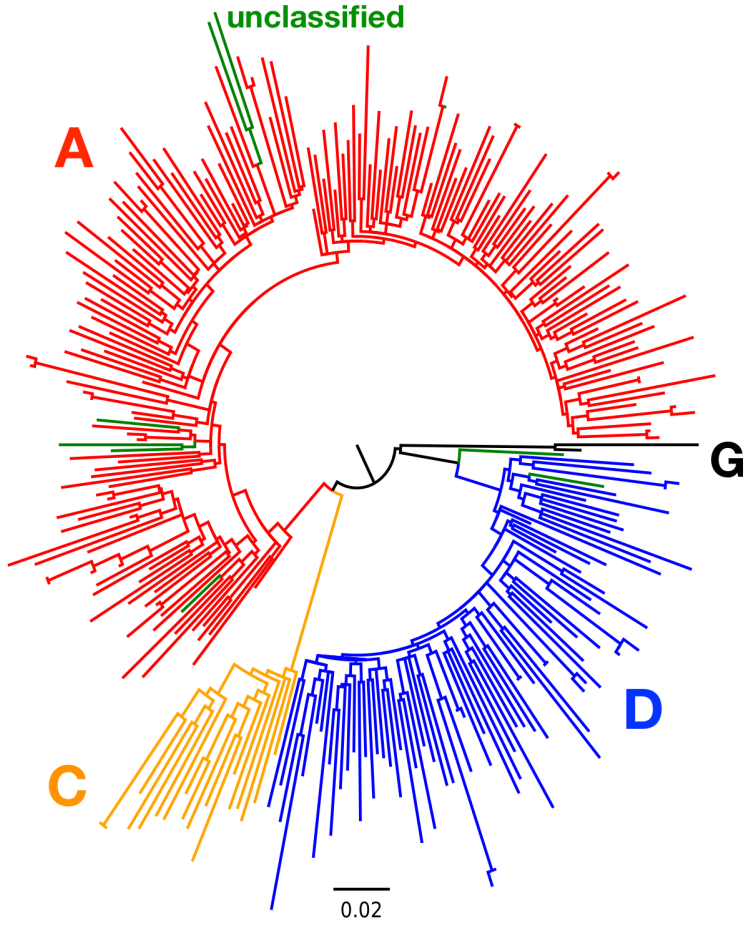
CASE WESTERN RESERVE UNIVERSITY
EST. 1826

Drug resistance
Data collection

- Retrieved archived plasma samples from *n*=382 JCRC patients with non-B infections
  - 85 treatment naive
  - 129 first-line treatment failure
  - 116 second-line treatment failure
  - 53 failure raltegravir (RAL)-based regimen
- Sequenced two overlapping regions of HIV integrase on MiSeq

Drug resistance

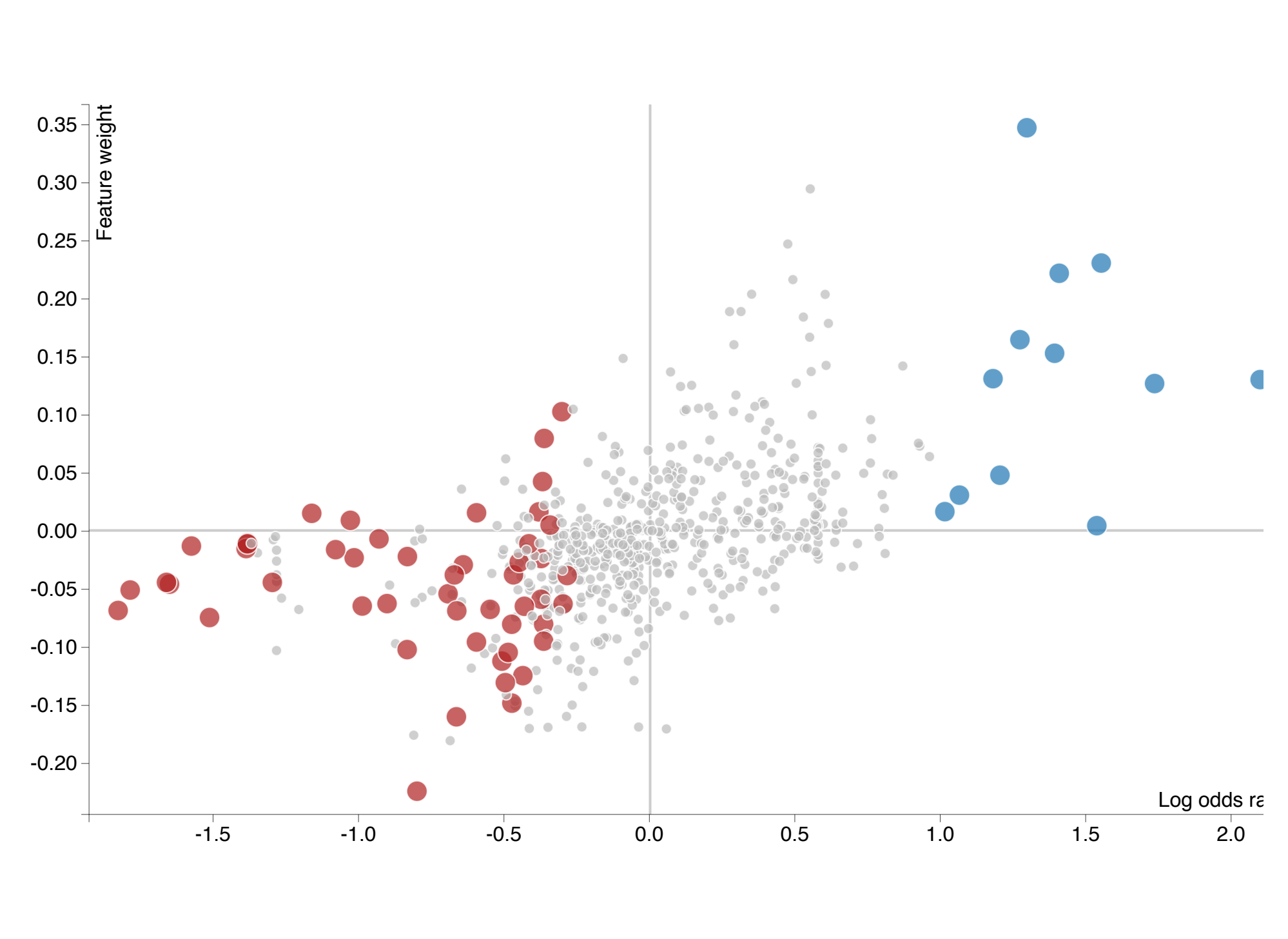
HIV-1 subtype classifications



Maximum likelihood tree (PhyML) using NGS consensus sequences.

Drug resistance
SVM classification

- Encoded NGS data as a **large binary matrix** where 1 indicates an amino acid frequency above some threshold.
- Since there are many more variables (amino acids) than observations (samples), we used a **support vector machine** (SVM) that defines a model using data points instead of variables.
- Trained and validated SVM to classify RAL-exposed samples from naive samples and other treatment failures.


Feature weight
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Log odds ra

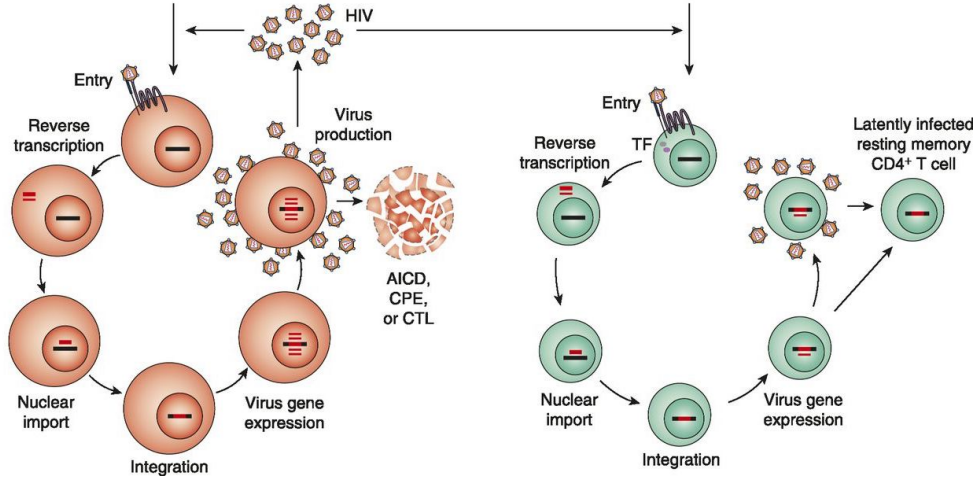
Concluding remarks

- We find both known and potentially novel mutations associated with INSTI resistance.
- NGS is useful because, in resource-limited settings, a patient may have stopped a failing treatment long before they are able to visit a clinic.
- Drug resistance mutations may have reverted to low frequencies.

HIV latency

What is the latent reservoir?

- HIV DNA becomes integrated into the host genome.
- A small fraction of infected cells enter an inactive state.
- Long-lived reservoir invisible to immune system and drug treatment; reseeds the infection.

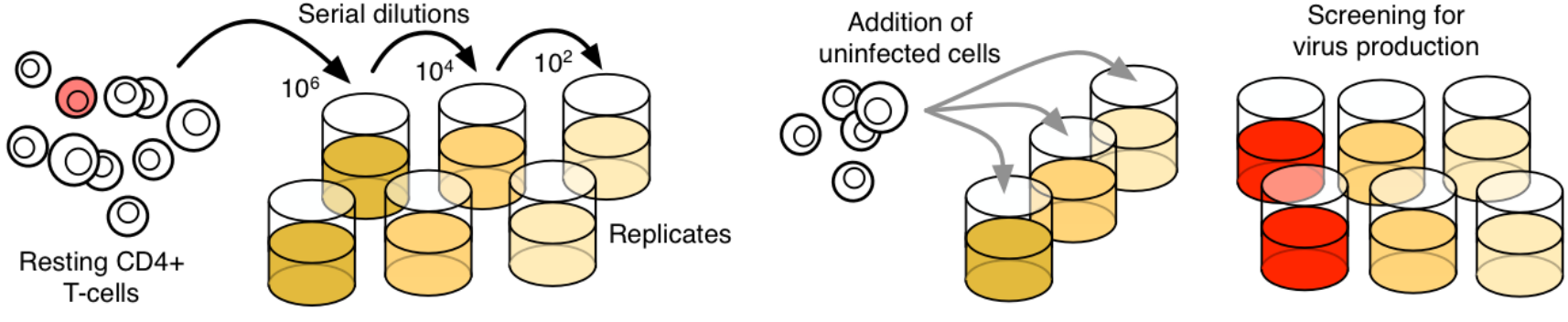


AJ Murray *et al.* (2016) J Immunol 197:407.

HIV latency

How large is the reservoir?

- The reservoir is largely composed of resting CD4+ T cells.
- We measure the latent reservoir by the number of infected units per million (IUPM) cells.
- Use a limiting dilution assay to estimate the IUPM.



HIV latency

Estimating IUPM from NGS data

- A well tests posiitve if there is one or many infected cells.
- Detecting multiple HIV variants in a well by NGS may improve estimates.
- We developed a new Bayesian method to use NGS to estimate IUPM
- Applied to samples collected from Rakai, Uganda, by Johns Hopkins Medicine and NIAID (US).

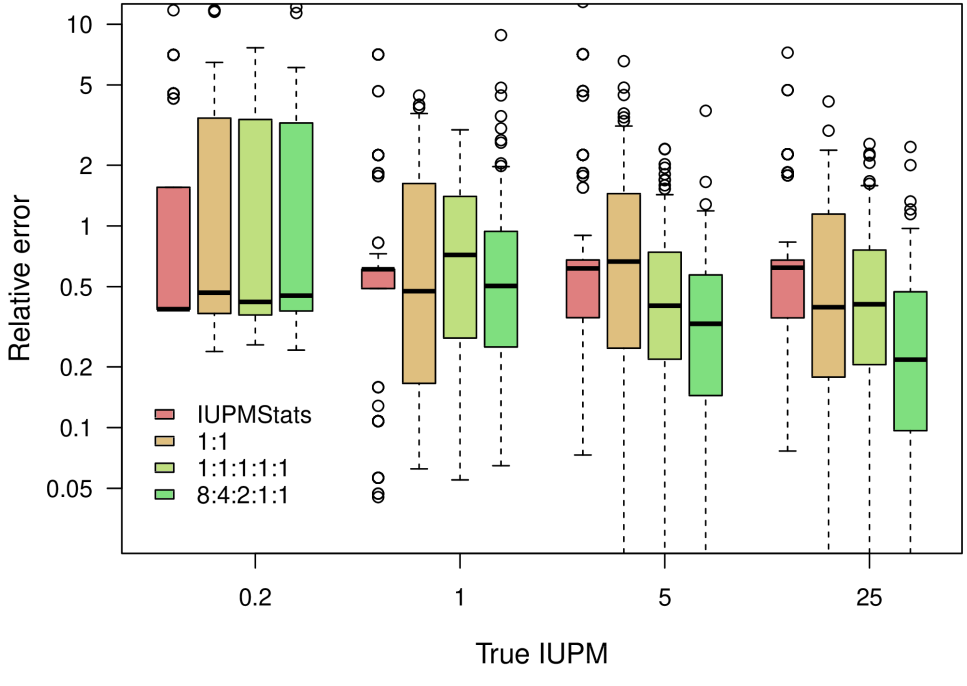
Dr. Jessica Prodger

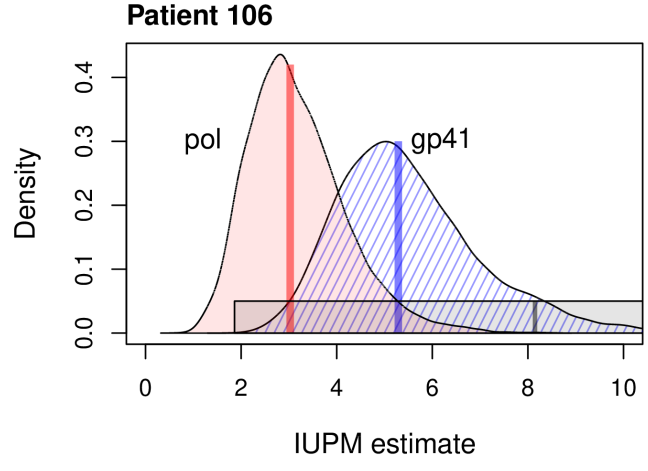
JOHNS HOPKINS
SCHOOL of MEDICINE

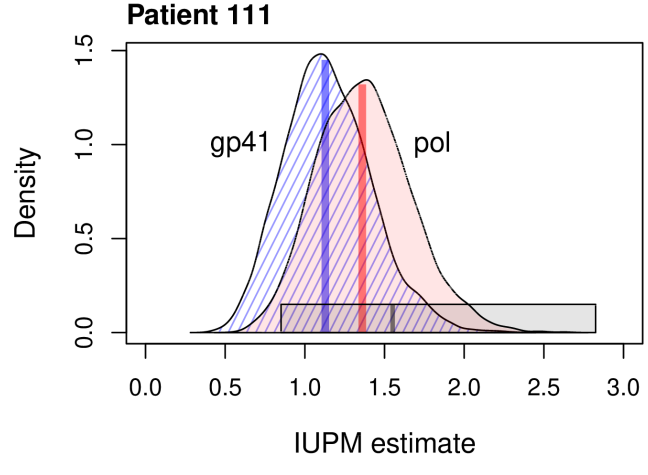
NIH
NIAID

HIV latency
Results

Simulations show greater accuracy with increasing true IUPM.







Poon *et al.* (2018) *Quantitation of the latent HIV-1 reservoir from the sequence diversity in viral outgrowth assays.* Retrovirology 15: 47.

HIV latency

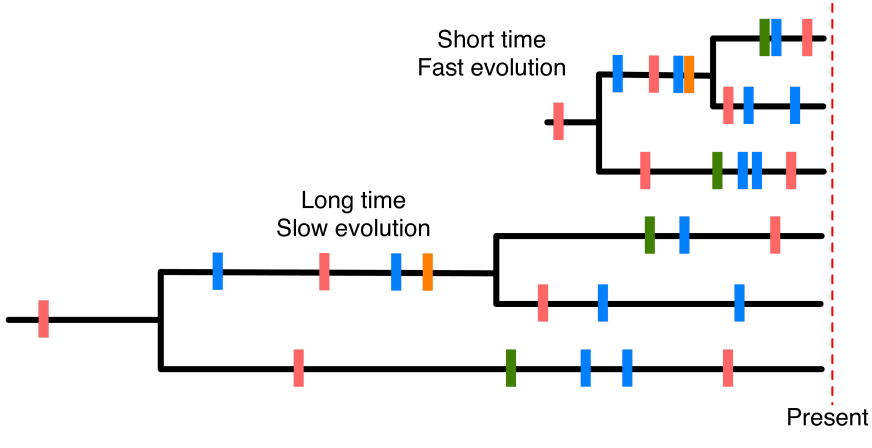
How does the reservoir persist?

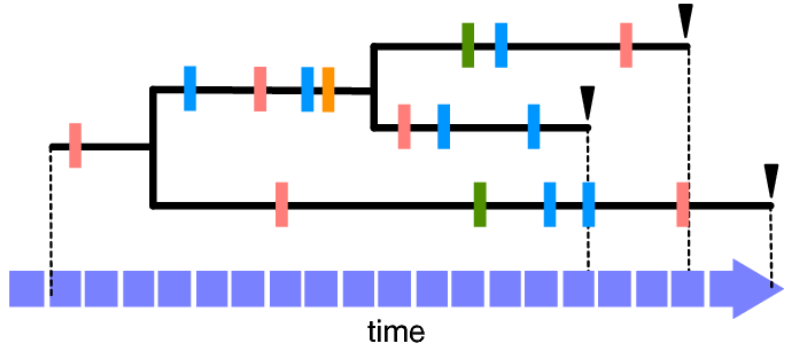
- Half-life estimates for latent reservoir range from 4 to 12 years.
- Viral rebound within days of treatment interruption implies constant high rate of reactivation.
- Hypotheses:
    1. Growth of latently infected T-cells (clonal expansion).
    2. Low-level replication of HIV in drug sanctuaries.

HIV latency

Dated-tip phylogenies

- We can build a phylogeny relating copies of HIV within a single host.
- Without other information, time and the rate of evolutiona are confounded (left).
- We can use sample collection dates to "pin" tips to a timeline, and rescale the tree in time.

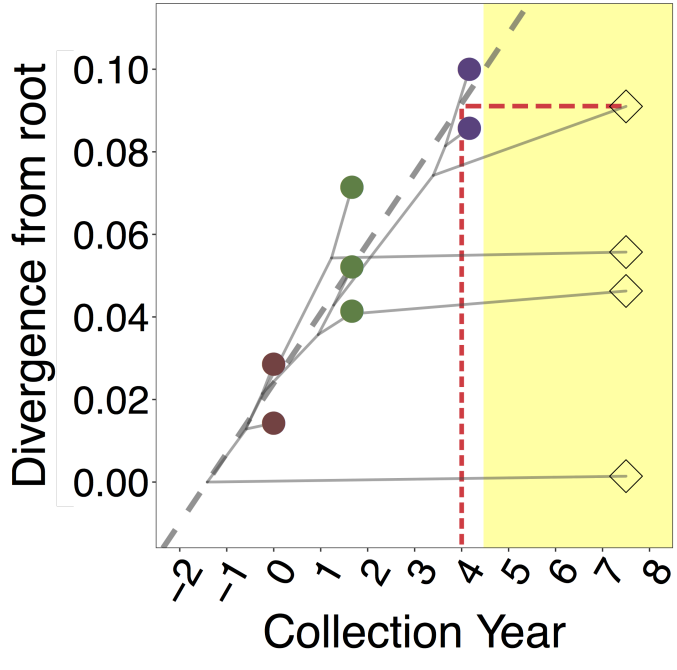




HIV latency

Dating HIV in the reservoir

- When HIV integrates into the host cell genome, its evolution is effectively frozen.
- If evolution is sufficiently "clock-like" (constant rate of evolution), then we can extrapolate when HIV DNA became latent.
- Requires that we estimate the **root** — the earliest point in time in the phylogeny.



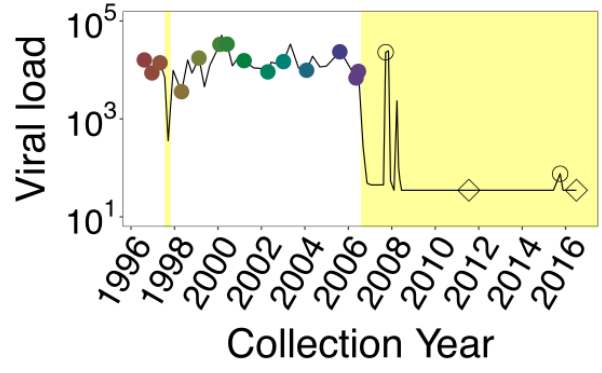
HIV latency
Data collection

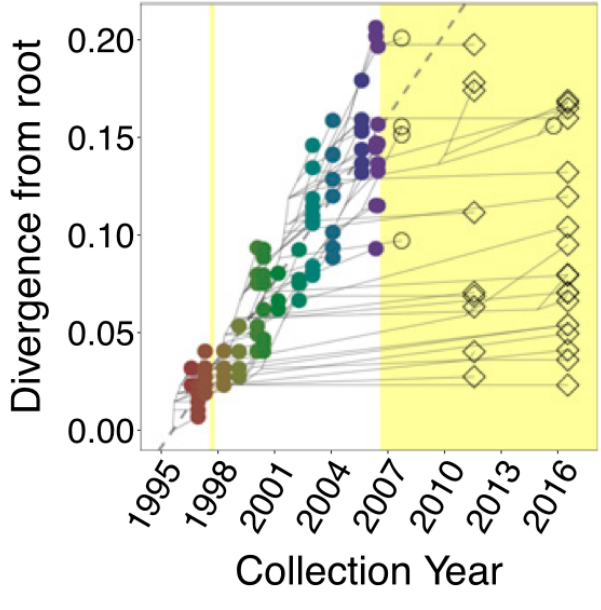
- Collaboration with Dr. Zabrina Brumme and Dr. Jeff Joy at BC Centre for Excellence (CFE) in HIV/AIDS.
- Obtained **pre-therapy** samples of HIV plasma RNA from 2 patients from CFE archive.
- Sampled post-treatment HIV DNA from same patients.
- Applied our method to date these reservoir sequences.

BRITISH COLUMBIA
CENTRE *for* EXCELLENCE
in HIV/AIDS

HIV latency
Participant 1

- P1 was diagnosed with HIV-1 on August 1996.
- Did not achieve viral suppression until August 2006.
- Viral rebound following unsuccessful regimen change (Fall 2017).
- Sequenced post-treatment HIV RNA (○) and DNA (◇).
- HIV DNA dated far deeper than post-treatment RNA.

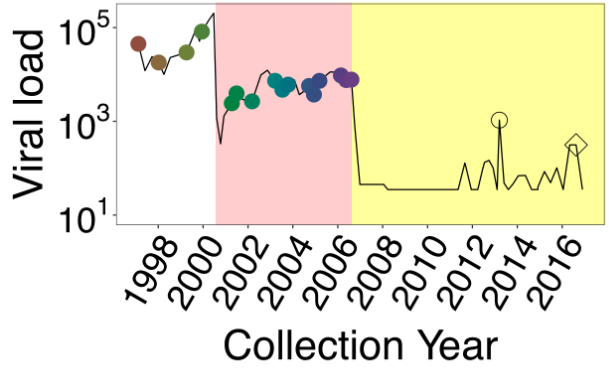


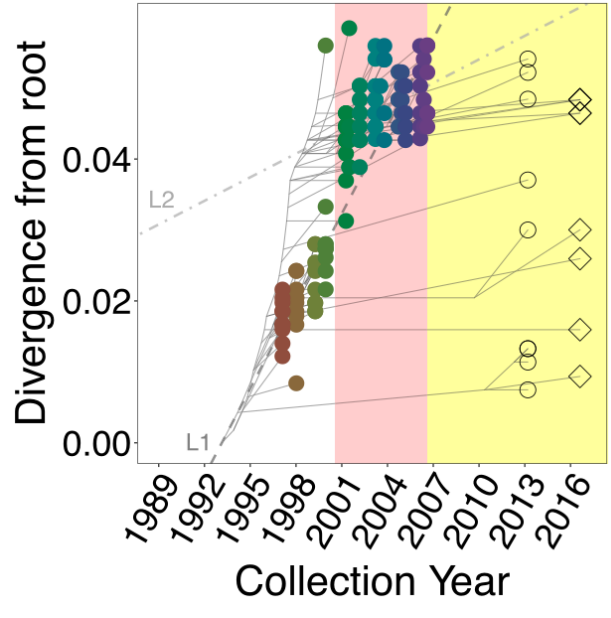


BR Jones *et al.* (2018) *Phylogenetic approach to recover integration dates of latent HIV sequences within-host.* Proc Natl Acad Sci USA.

HIV latency
Participant 2

- P2 was diagnosed with HIV-1 on April 1995.
- Initiated dual therapy in July 2000.
- Switched to combination therapy (cART) August 2006 and achieved viral suppression.
- 5 sequences dated to time of diagnosis, before earliest sample.
- Reservoir harbours replication-competent HIV from early stage of infection.





BR Jones *et al.* (2018) *Phylogenetic approach to recover integration dates of latent HIV sequences within-host.* Proc Natl Acad Sci USA.

PNAS

Phylogenetic approach to recover integration dates of latent HIV sequences within-host

Bradley R. Jones[a], Natalie N. Kinloch[b], Joshua Horacsek[a], Bruce Ganase[a], Marianne Harris[a], P. Richard Harrigan[c], R. Brad Jones[d], Mark A. Brockman[a,b], Jeffrey B. Joy[a,c,1,2], Art F. Y. Poon[e,1,2], and Zabrina L. Brumme[a,b,1,2]

[a]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada V6Z 1Y6; [b]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; [c]Department of Medicine, University of British Columbia, Vancouver, BC, Canada V5Z 1M9; [d]Department of Microbiology, Immunology and Tropical Medicine, George Washington University, Washington, DC 20037; and [e]Department of Pathology and Laboratory Medicine, University of Western Ontario, London, ON, Canada N6A 5C1

Edited by Robert F. Siliciano, Johns Hopkins University School of Medicine, Baltimore, MD, and approved August 9, 2018 (received for review February 2, 2018)

Acknowledgements

- The development and validation of the MiCall pipeline was made possible by the members of the BC Centre for Excellence in HIV/AIDS Molecular Laboratory.
- These studies were also made possible by collaborations with JCRC Uganda, Case Western Reserve University, Johns Hopkins School of Medicine, and the National Institutes of Allergy and Infectious Diseases (NIH).
- And by the donation by study participants of samples for research purposes.

Thanks!


Ontario Genomics


GenomeCanada


NIH
NIAID


CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada


NSERC
CRSNG