Modeling of CD4 and Provirus Distributions Quantifies Undersampling Bias in Reservoir Studies

Daniel B Reeves[1,2], Hao Zhang[3], Rebecca Hoh[4], Leah H Rubin[3], Stephen J Gange[3], Nadia R Roan[4,5], Phyllis C Tien[4], Joseph B Margolick[3], Luis J Montaner[6], Michael J Peluso[4], Steven G Deeks[4], Joshua T Schiffer[1,2], Janet D Siliciano[3], Robert F Siliciano[3,7], Annukka AR Antar[3]

[1]Fred Hutchinson Cancer Center, Seattle, WA, USA, [2]University of Washington, Seattle, WA, USA, [3]Johns Hopkins University, Baltimore, MD, USA, [4]University of California San Francisco, San Francisco, CA, USA, [5]Gladstone Institutes, San Francisco, CA, USA, [6]The Wistar Institute, Philadelphia, PA, USA, [7]Howard Hughes Medical Institute, Baltimore, MD, USA

Acknowledgements. We are extremely grateful to all the generous study participants without whom none of this work would be possible.

0339

Contact:
dreeves@fredhutch.org
aantar1@jhmi.edu

Fred Hutch Cancer Center

Johns Hopkins University

BACKGROUND

- The HIV reservoir persists in CD4+ T cell clones. We hypothesize that the clonality of HIV-infected cells is similar to that of the underlying population of memory (m)CD4+ T cells.
- Open question: How deeply must we sample to uncover the dominant mechanisms sustaining the reservoir?
- Data from 4 cohorts of people with HIV on ART and without HIV and with and without CMV (**Fig 1**).

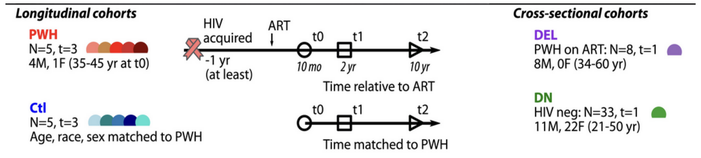


Figure 1. Cohort schematic from Reeves et al. Cell Systems 2025. 4 people living with HIV (PWH) and 4 age/race/sex matched controls sampled longitudinally at 3 time points each. 2 additional published cohorts with single time points were used for comparison.

METHODS

- Rank abundance curves and power law modeling of memory CD4+ T cell (mCD4) TCRβ repertoires and near-full length provirus sequences (mean 74 sequences/time point)

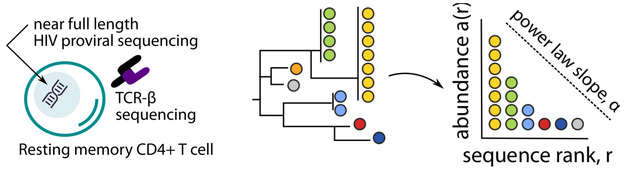


Figure 2. Sequencing TCRβ sequences and HIV proviruses in PWH. A phylogenetic tree with clonal sequences is transformed into a rank abundance distribution, and the slope of this distribution summarizes the clonality via the power law slope.

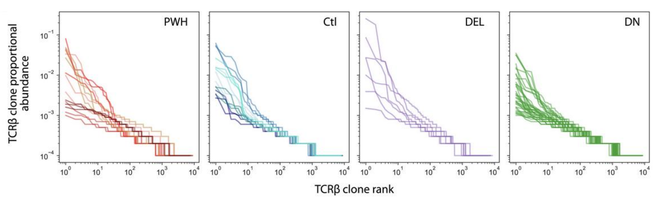


Figure 3. mCD4 TCRβ rank abundance curves for each of 4 cohorts. In longitudinal cohorts each individual has the same color for all time points.

We developed power-law models of rank abundance distributions with 1 or 2 power law slopes. Computational fitting approaches were validated on ground truth synthetic data. An objective ("loss") function

$$f_{loss} = \min_i \left(\left\{ \frac{1}{Q} \sum_q \left(\mathcal{M}_i[p(q), s] - a^{obs}(q) \right)^2 \right\}^{1/2} \right)$$

reliably reproduced simulation parameters, where $\mathcal{M}_i$ is multinomial sampling (i repeats) of the underlying distribution $p(q)$ at sample size s compared to data a^{obs}

Joint HIV and T cell ecology modeling informs the mechanisms sustaining the reservoir and the sample sizes needed to uncover them.

RESULTS

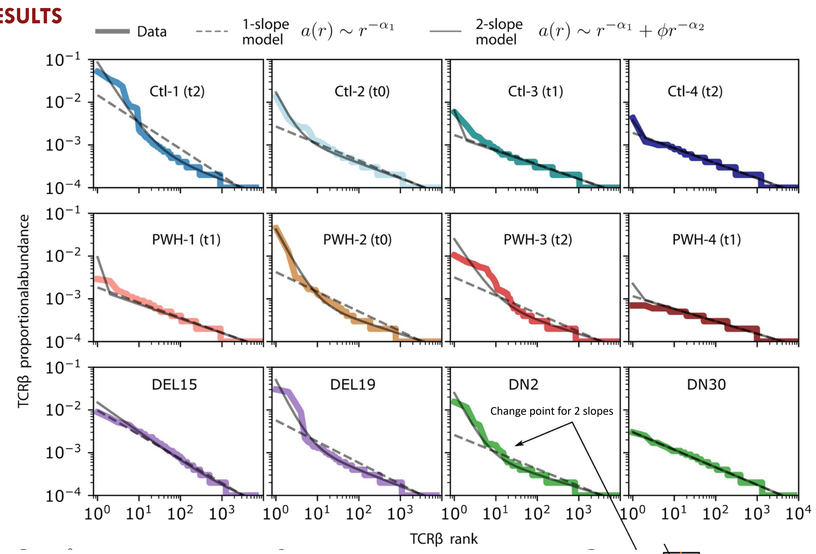


Figure 4. mCD4 TCRβ rank abundance curves (colored lines) with 1 (dashed) and 2 slope (solid) model fits. Certain individuals fit much better to a 2-slope model. The change point for 2-slope models is noted and ranges from 3-300.

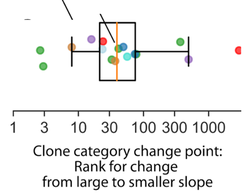
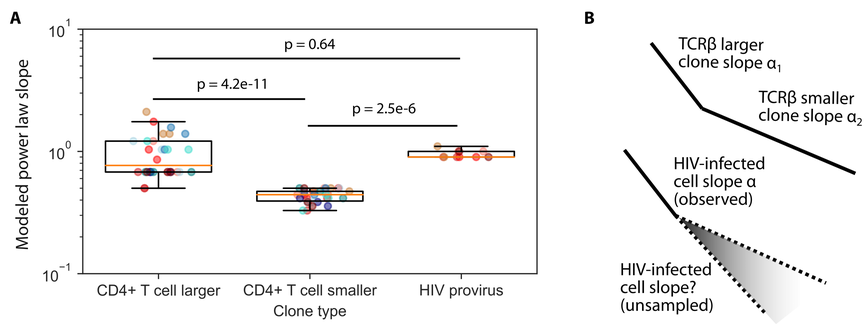


Figure 5. A) Large and small slope power law exponents from mCD4 rank abundances compared to HIV provirus rank abundance slope. B) Cartoon diagram illustrating that large CD4+ clones and HIV sequences have similar slopes, but it is unknown if HIV sequences also have a second smaller slope.



RESULTS

- mCD4 clonality was not significantly different between PWH or HIV-seronegative groups.
- The mCD4s of CMV-positive individuals fit a 2-slope model significantly better than a 1-slope model.
- The median rank differentiating the larger and smaller clone slopes in the 2-slope model was 9, meaning that the top ~9 most highly-expanded ('larger') mCD4 clones had a significantly more clonal distribution than the rest of the ('smaller') mCD4 clones.
- The clonality of the larger mCD4 clones was similar to that of HIV-infected cells from the same samples and higher than previously reported from bulk mCD4 clonality.
- Using our mathematical models, we estimated the sample size of HIV proviruses / integration sites required to observe an underlying HIV-infected clone distribution that fit a 2-slope model
  - sample sizes of 1000 proviruses would correctly identify second slope 50% of the time.

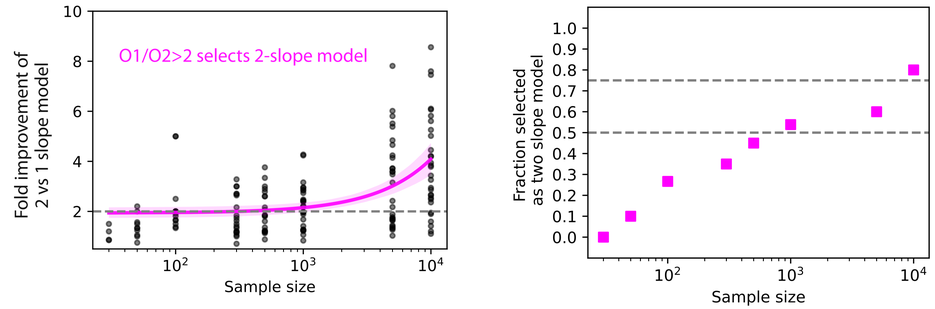


Figure 6. Using synthetic data simulated from two slope models, we varied sample size and assessed the fold-improvement of the two vs one slope model for these synthetic data. Given data simulated from a two slope model, our algorithm predicts increasing identification of a two slope model based on sample size.

CONCLUSIONS

- Studies sampling <1000 HIV proviruses or integration sites/time point are unlikely to distinguish between larger and smaller infected clones. Undersampled reservoir studies preferentially identify the largest clones. Thus, these studies may have biased the field towards inferring that the majority of HIV reservoir persistence is predominantly due to antigen-driven proliferation (if this underlies larger clone persistence) vs homeostatic proliferation of (smaller but more numerous) infected CD4s.

PLAIN LANGUAGE SUMMARY

We used computational modeling of T cell and HIV provirus genetic sequences to compare the abundance of cells that harbor the HIV reservoir (memory CD4+ T cells) vs abundace of the HIV reservoir itself. This work asks the question of whether reservoir cells are merely passengers following the normal behaviors of T cells, and sets guidelines for how we could fully test this question in experiments.