Benchmarking Rao's Q as a Reproducible, Quantitative, Evolution-Aware Metric of Viral α-diversity for Metagenomic Data

Florencia Martino[1], Kakhangchung Panmei[1], David L. Thomas[1], Abraham J. Kandathil[1], Dylan Duchen[2], Steven J. Clipman[1]

[1]Johns Hopkins University School of Medicine, Baltimore, MD, USA

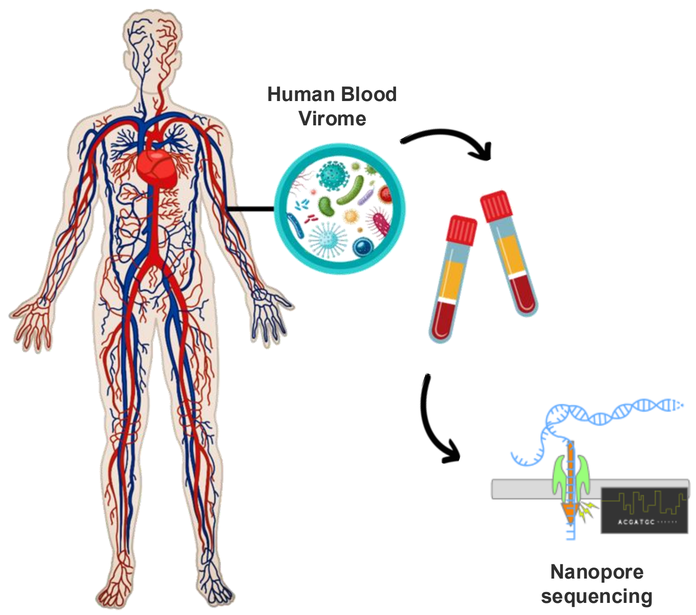
[2]Yale School of Medicine, Department of Pathology, New Haven, CT

JOHNS HOPKINS SCHOOL of MEDICINE

fmarti34@jh.edu

BACKGROUND

- Within-sample viral diversity is a biomarker of immune status and infection risk.
- Viral genomes are highly similar and recombinogenic, causing metagenomic reads to cross-map between related lineages.
- Classical α-diversity metrics (Shannon, Simpson, Hill) ignore phylogeny and inflate diversity under near-clonal expansions.
- Faith's PD incorporates phylogeny but ignores abundance.
- Viral metagenomes require a metric that integrates both evolutionary distance and relative abundance.



$$Q = \sum_i \sum_j d_{ij} p_i p_j$$

p_i: abundance of lineage i.
d_{ij}: evolutionary (patristic) distance between lineages i and j.

Q **= Rao's Q**
Computes the abundance-weighted average patristic distance among all lineage pairs.

MATERIALS AND METHODS

Reference construction:
687 *Anelloviridae* genomes + M13mp18; clustered at 70% ORF1 identity (MMseqs2); ORF1 alignment (MAFFT), ML tree (RAxML-NG, 1500 bootstraps)

Reads and abundance estimation:
Nanopore reads mapped with Minimap2 (≥ 85% ORF1 coverage, MAPQ ≥ 10, depth ≥ 5X). Per-sample normalized abundances derived from mapped coverage.

Diversity indices:
Shannon, Simpson, Hill numbers, Faith's PD, and Rao's Q computed from relative abundances and patristic distances.

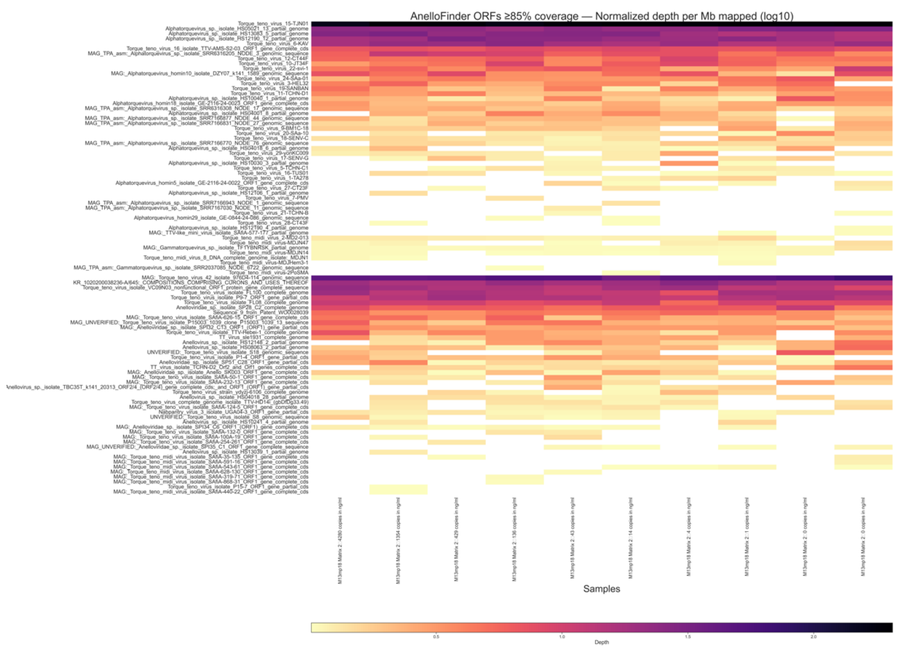
Experimental conditions

1. Controlled plasma viromes with serial dilutions of M13mp18.
2. In silico perturbations: phylogenetic distance tests, cross-mapping simulations, tree-collapse.
3. Technical reproducibility across seven independent ONT runs.
4. Clinical plasma viromes (n = 92).

RESULTS

Cross-mapping obscures lineage structure:

- Closely related lineages share homology → reads redistribute.
- White gaps reflect local coverage loss, not true absence.
- Abundance-only metrics misinterpret these artifacts as biological change.

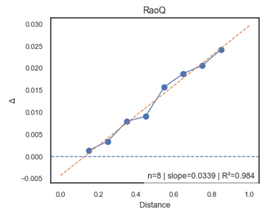


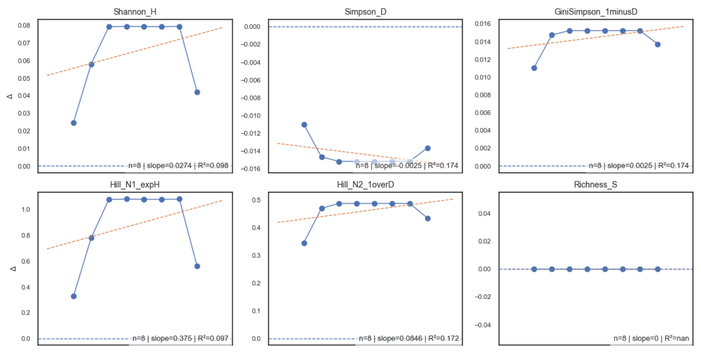
1) Biological scaling

In controlled plasma viromes (M13mp18), Rao's Q responds linearly to controlled lineage depletion.

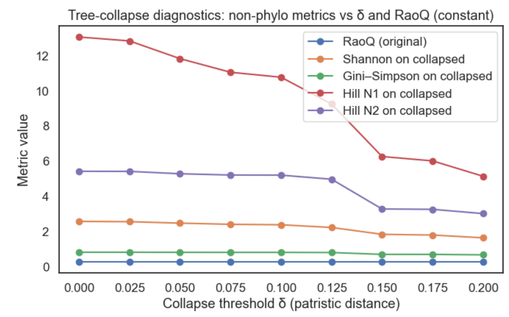
2) Rao's Q captures evolutionary dispersion:

Sensitivity to evolutionary distance: Rao's Q increases with phylogenetic divergence; classical α-diversity metrics remain insensitive.





Tree collapse test
Collapsing phylogeny makes Shannon/Hill converge toward Rao's Q.

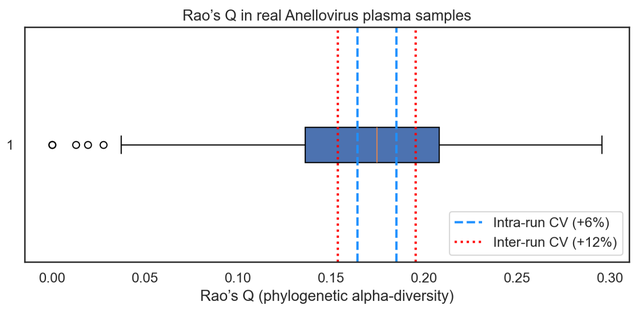


3) Technical reproducibility across ONT runs:

Within-run CV = 0.026 (±6%).
Between-run CV = 0.118 (±12%).
Negative-control dilution series show bounded variability (CV ≤ 0.05).

4) Real samples: Human Plasma Viromes (n=92)

Rao's Q ranged 0 - 0.3 (median = 0.17), above technical error.



CONCLUSIONS

- Rao's Q quantifies α-diversity as evolutionary dispersion and integrates abundance and phylogeny, resists cross-mapping artifacts, behaves linearly with evolutionary distance, and has reproducible error bounds.
- Enables robust diversity benchmarking in viral metagenomes.

Acknowledgments

- Study participants who made this research possible.
- Funding from the National Institutes of Health, USA (DP2DA056130; R01DA058567)