Filtering of viral haplotypes from noisy long-read sequencing data

Seth H Borrowman[1,2], Natalie Stegman[1], Ramon Lorenzo-Redondo[1,2]

[1] Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA. [2] Center for Pathogen Genomics and Microbial Evolution, Northwestern University Institute for Global Health, Chicago, IL 60611, USA.

Background:

Retroviruses such as human immunodeficiency virus (HIV) exist as genetically diverse populations, known as *quasispecies*, that evolve rapidly in response to host pressures, antiretroviral therapy, and other selective forces. Characterizing this diversity is critical for understanding viral dynamics, immune evasion, and/or treatment resistance, but it remains technically challenging, particularly when analyzing highly variable pathogens using high-throughput long-read sequencing technologies that are prone to elevated error rates. Distinguishing mutational changes in the pathogen from read errors in the sequencing remains a large hurdle in quasispecies reconstruction.

Objective:

To develop a statistical framework applicable to nanopore long-read sequencing for distinguishing true viral haplotypes from sequencing noise and evaluate how filtering decisions impact estimates of viral diversity.

Methods:

We formalize two complementary, statistically grounded filtering strategies for viral haplotype analysis in noisy long-read data. The first, count-based filtering, uses binomial tail probability bounds to compute a minimum abundance threshold, enabling the exclusion of low-frequency haplotypes likely attributable to sequencing error. The second, variant-level filtering, applies dual binomial tests at each alignment position to identify high-confidence single-nucleotide variants (SNVs), requiring both dominant base under-dispersion and significant support for a minor allele. We implement these methods on real nanopore-derived simian immunodeficiency virus (SIV) haplotype datasets and evaluate the impact of varying thresholds on downstream quasispecies diversity metrics.

Results:

Filtering thresholds strongly affect both the size and composition of inferred haplotype populations, directly impacting downstream measures of quasispecies diversity. We demonstrate how statistically grounded filtering can be used to systematically balance sensitivity and specificity. This improves robustness in analyses of viral populations, such as diversity estimation, phylogenetics, and lineage tracking in complex viral populations.

Conclusions:

This work introduces a statistically principled and computationally efficient framework for filtering viral haplotypes in noisy long-read sequencing data. By incorporating model-based uncertainty and leveraging diversity metrics across parameter space, the approach facilitates high-resolution, reproducible inference of intra-host viral evolution. These techniques are broadly applicable across viral genomics and may enhance interpretation in viral research, clinical surveillance, and population-scale molecular epidemiology.