Phylogenetic estimates of HIV-1 gp120 indel rates across the group M subtypes

Insertions and deletions (indels) in the HIV-1 envelope glycoprotein gp120 play a significant role in the evolution of HIV pathogenesis and transmission fitness. While substitution rates in HIV-1 are well characterized by phylogenetic models, there is a lack of quantitative measures of indel rates in HIV-1. Here we use a dated-tip phylogenetic analysis of gp120 sequences to estimate indel rates for 7 subtypes and CRFs of HIV-1 group M.

We obtained and processed 26,359 HIV-1 gp120 sequences from the Los Alamos National Laboratory HIV Sequence database. After filtering these sequences, we extracted the conserved and variable regions from the remaining 6,605 sequences by pairwise alignment. We used FastTree2 to reconstruct phylogenies from the alignment of concatenated conserved regions, and used least-squares dating (LSD) to rescale these trees in time. We estimated variable region indel rates by fitting a binomial-Poisson model to length discordance in sequences related by cherries.

Indel rate estimates ranged from 3e-5 to 1.5e-3/nt/year and varied significantly among variable regions and subtypes; e.g., rates were significantly lower for subtype B. Variable regions V1, V2 and V4 accumulated significantly longer indels irrespective of subtype, and we found evidence of positive selection for indels affecting N-linked glycosylation sites in V1/V2. Further, we observed that indel sequences were enriched for G and depleted for T relative to the flanking sequences.

Our results comprise the first phylogenetic measures of indel rates in HIV-1 gp120 across subtypes and variable regions, and identify novel and unexpected patterns for further investigation into HIV-1 evolution.