

**Figure XX. Phylogenetic Analysis of polyA Hairpin** (A) An inferred phylogeny of representative HIV-1 strains of Group M subtypes (see methods for selection criteria). The phylogeny was built with bootstrapping (n=1000, branch width indicates branch support) a multiple sequence alignment of the viral polymerase protein sequences. The tree is annotated with a multiple sequence alignment of the polyA-HP as colored blocks where solid blocks indicate residues that are un-base paired and transparent blocks indicate residues that are base paired in the predicted secondary structure. Positional frequencies are represented as a bar graph where white space in the bars represents the gaps in the alignment. Positions marked by an asterisk are the hexameric AAUAAA polyadenylation signal. Bold-faced strain names indicate they were part of our experimental analysis, with the majority of the focus on the MAL strain (denoted by◂).(B) The consensus secondary structure of the polyA-HP as calculated by locARNA (1-3). The structure is annotated with the position specific frequency of the types of base pairs present at each position in the helical portion of the polyA-HP. Positional frequencies were determined after the predicted helical structures were aligned using a modified Needleman-Wunch algorithm to align RNA helices.

**Results**

We conducted phylogenetic analysis of the 5′ polyA-HP across HIV-1 Group M subtypes with the aim to elucidate conserved structural elements that may be suggestive of functional roles in viral replication (Figure XX. A). Analysis of the structural conservation of the polyA-HP posed the non-trivial challenge of identifying the polyA-HP in each strain since sequence data from the LANL HIV compendium (4) do not have annotations for UTR structural elements. Using a traditional sequence alignment (i.e. BLAST) with previously defined polyA-HP sequences, such as in the MAL (5) and NL4-3 (REF) strains, to identify the polyA-HP in each strain was an option; however, this approach would introduce a bias in the discovery of conserved structural elements. The resulting set of hairpin sequences would be expected to have high sequence similarity which would implicitly result in near complete conservation of the structural elements. We instead used a novel discovery pipeline that rooted the search at the hexameric AAUAAA signal to programmatically extract the most probable polyA-HP in each strain without explicitly requiring sequence similarity (Sup Figure XX, Methods). Of the ~4000 strains inputted into the pipeline, we were able to identify polyA-HP sequences for 280. Representative strains from each subtype and strains we analyzed experimentally were assembled (n=32) onto a phylogenetic tree (see Methods for representative strain selection process) (Figure XX. A).

Analysis of these representative strains showed complete sequence conservation of the 5′ and 3′ termini and the apical loop of the hairpin with majority of the sequence variation existing between positions 35 and 42 of the alignment. The predicted secondary structures, as calculated by RNAfold (6), had an average free energy of -16.55 ± 1.9 kcal/mol. The polyA-HP from each of the 32 strains was used to deduce a consensus structure calculated using the locARNA software tool (1-3) (Figure XX. B). The helical base pairs at the basal and apical regions of the helix are completely conserved. The stem had the greatest frequency of bulged residues between positions 9 and 11 on the helix (Figure XX. B). The apical loop was 100% conserved.

**Discussion**

**Methods**

**References**

1. S. Will, I. L. Joshi T Fau - Hofacker, P. F. Hofacker Il Fau - Stadler, R. Stadler Pf Fau - Backofen, R. Backofen, LocARNA-P: accurate boundary prediction and improved detection of structural RNAs.

2. S. Will, I. L. Reiche K Fau - Hofacker, P. F. Hofacker Il Fau - Stadler, R. Stadler Pf Fau - Backofen, R. Backofen, Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering.

3. M. Raden *et al.*, Freiburg RNA tools: a central online resource for RNA-focused research and teaching. *Nucleic Acids Research* **46**, W25-W29 (2018).

4. Anonymous (2016) Los Alamos HIV sequence compendium 2016.

5. J. D. Brown, Kharytonchyk, S., Chaudry, I., Iyer, A. S., Carter, H., Becker, G., Desai, Y., Glang, L., Choi, S. H., Singh, K., Lopresti, M. W., Orellana, M., Rodriguez, T., Oboh, U., Hijji, J., Ghinger, F. G., Stewart, K., Frances, D., Edwards, B., Chen, P., Case, D. A., Telesnitsky, A., Summers, M. F., Structural basis for transcriptional start site control of HIV-1 RNA fate. *Science* **368**, 413-417 (2020).

6. R. Lorenz *et al.*, ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).