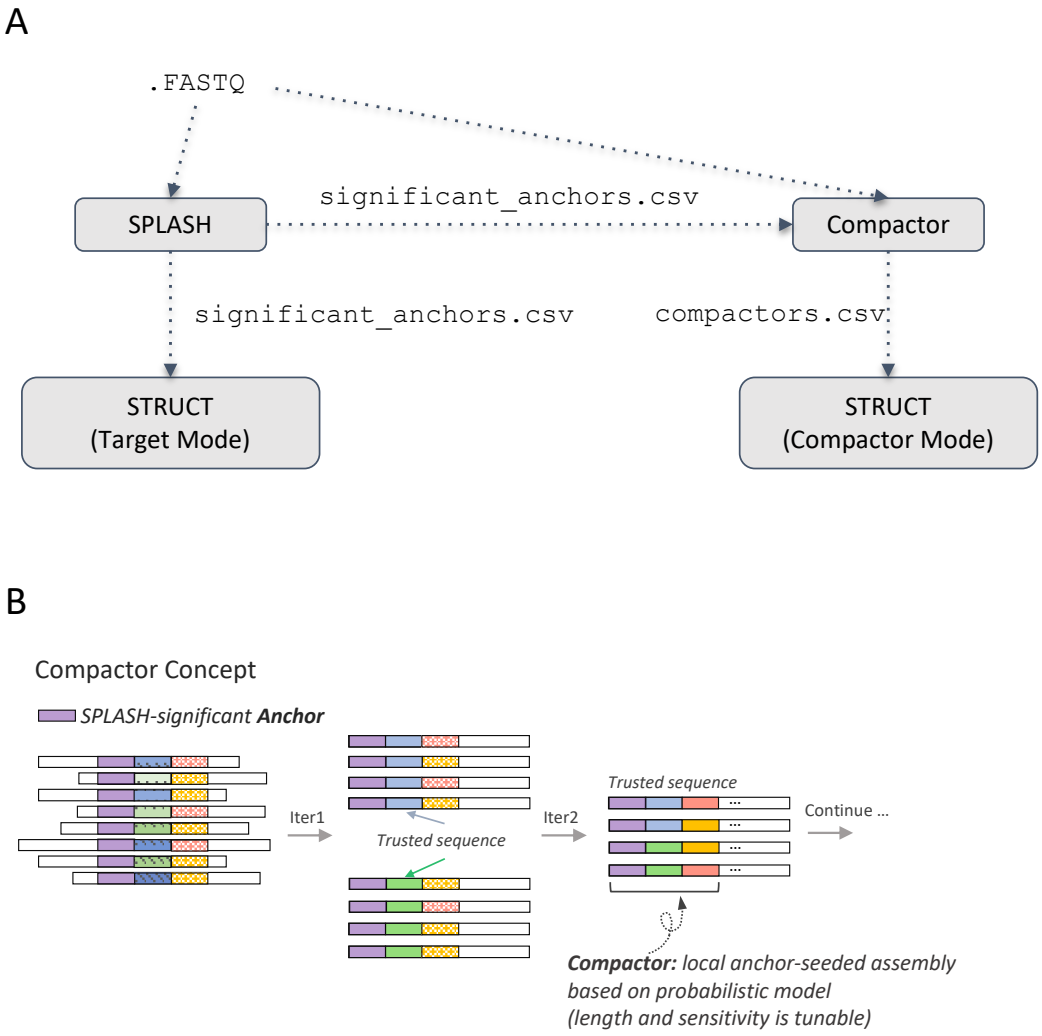
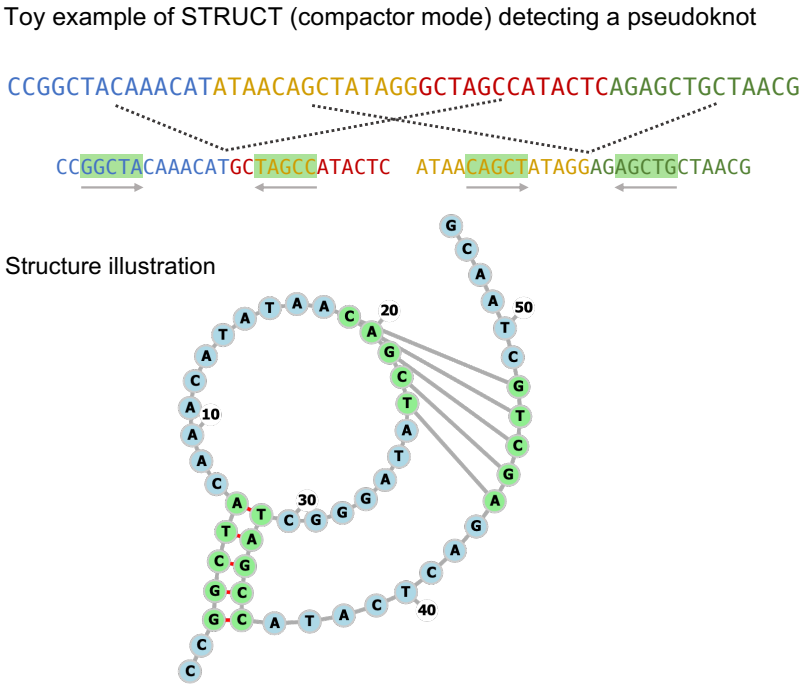
Supplementary Figures

Julie Fangran Wang, Arjun Rustagi, Julia Salzman

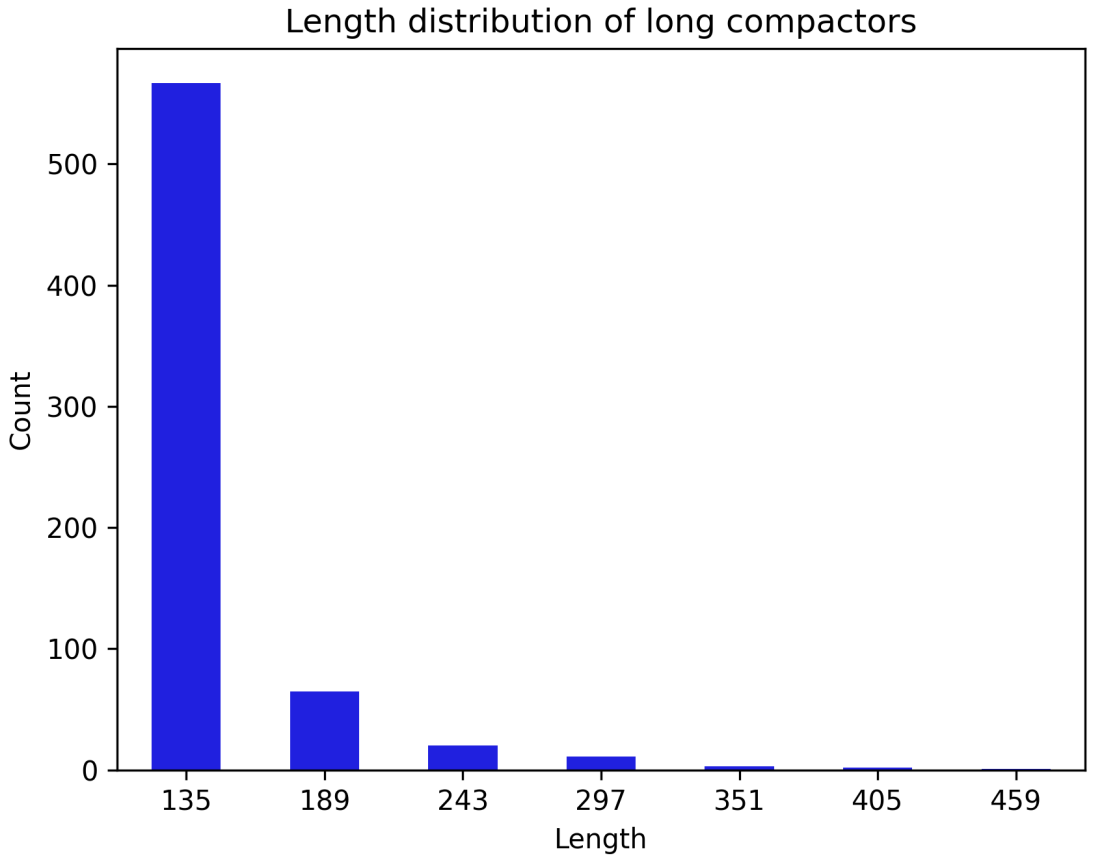
2026



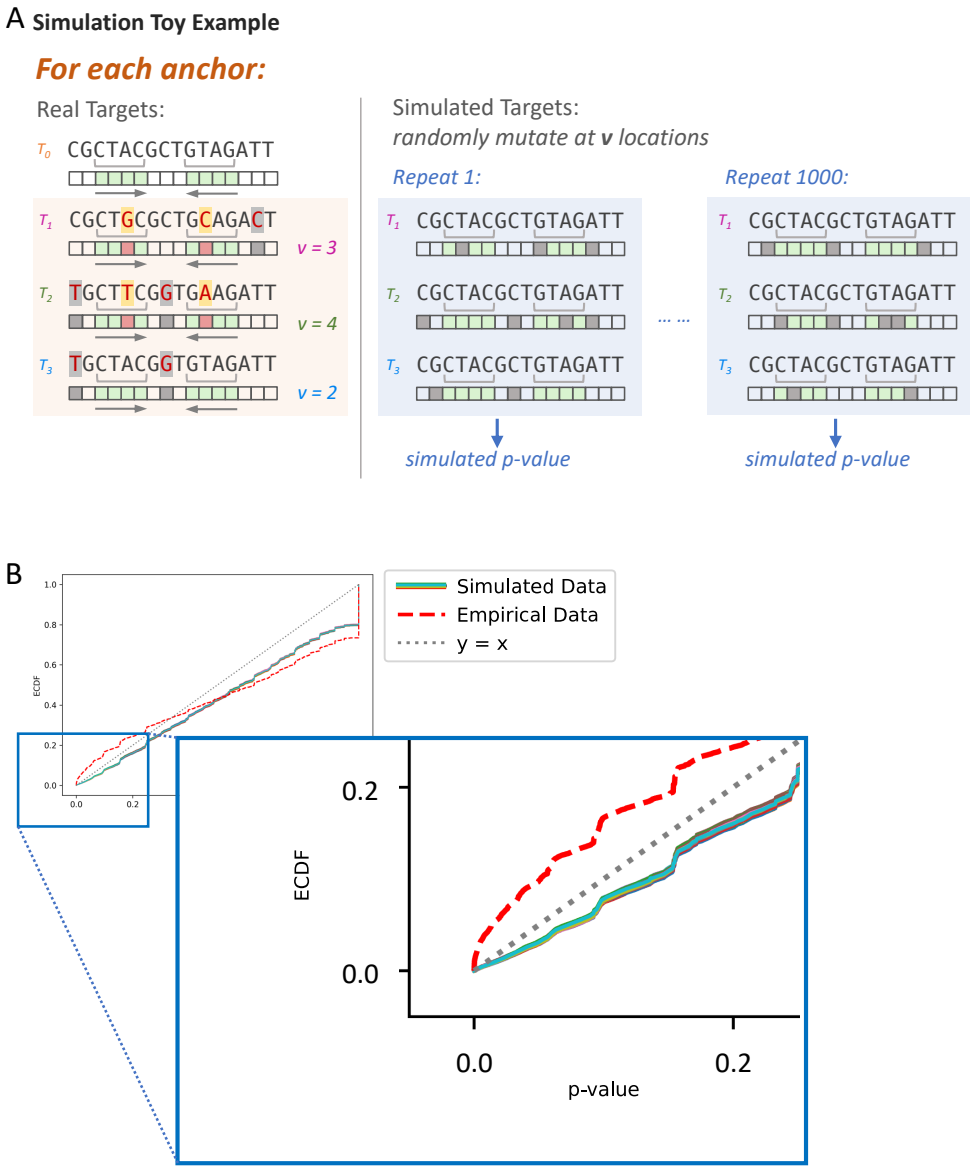
Suppl. Figure 1: STRUCT (compactor mode) workflow and compactor generations. (A) The raw sequencing data in FASTQ format is initially analyzed by SPLASH, generating an output file significant_anchors.csv that contains the SPLASH-significant anchors and their targets. Compactors are then constructed using the SPLASH-significant anchors as seeds to query the original FASTQ file, and deposited in the output file compactors.csv. The targets and compactors derived from the SPLASH-significant anchors are subsequently used as inputs for the two modes of STRUCT: Target Mode and Compactor Mode, respectively. Dotted lines represent file dependencies between the processes. (B) Compactors are local anchor-seeded assemblies of regions that vary across samples. They are generated iteratively based on a probabilistic model.



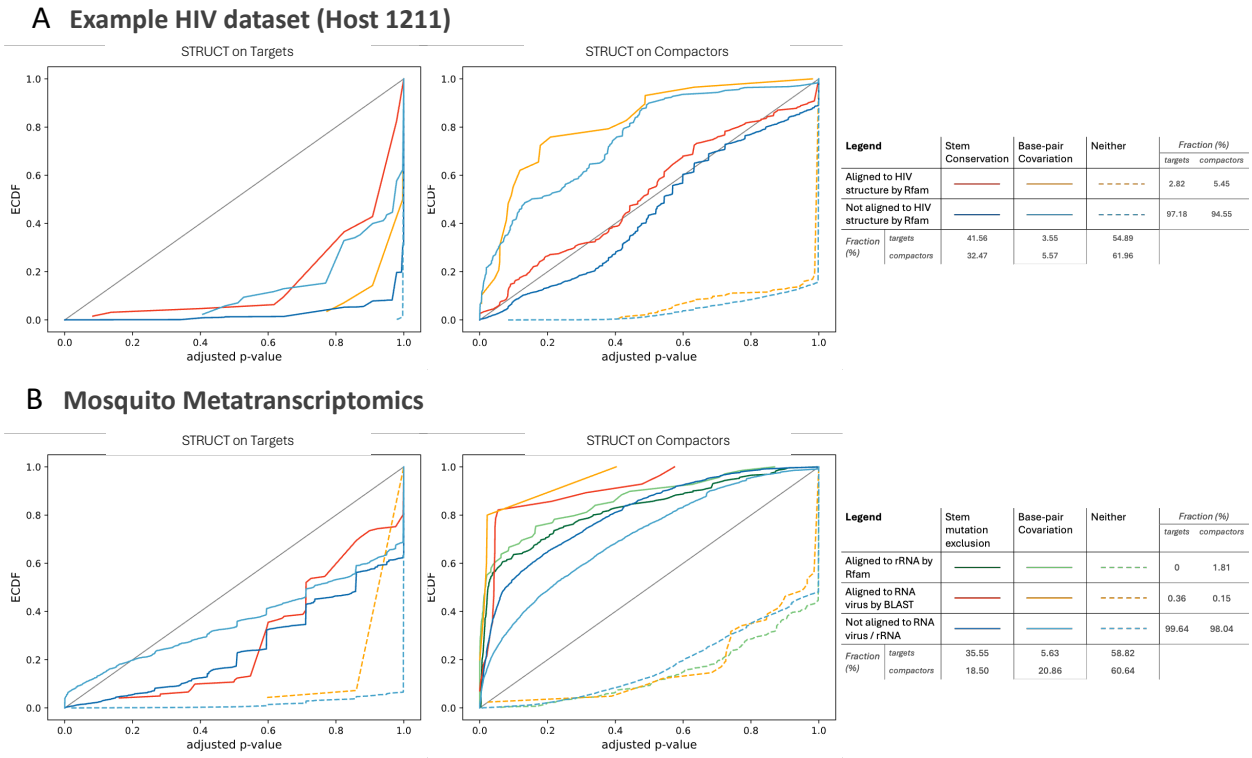
Suppl. Figure 2: Toy example of STRUCT (compactor mode) detecting a pseudoknot. In principle, STRUCT can detect pseudoknots in sequences. The toy example demonstrates that a trimmed compactor adopts the second method of split-and-concatenation, as illustrated in Figure 3. This results in two composite sequences, with compensatory stems identified in both. A forna drawing of the detected pseudoknot is provided, with nucleotides in stems highlighted in green circles and pseudoknot links represented by gray straight lines.



Suppl. Figure 3: Length distribution of extended compactors over 81 nucleotides (nts). To interpret unannotated base compactors (81 nt) using BLASTn, longer compactors with lengths capped at 1000 nucleotides are generated for the anchors of these base compactors. The length distribution of these compactors is visualized in the bar plot.



Suppl. Figure 4: Comparison of p-value distribution of STRUCT ran on empirical and simulated data. (A) A toy example illustrating the simulation procedure of test statistic, anchor score, in STRUCT (target mode). The process begins with an anchor and its associated targets, treating each as an anchor-targets simulation unit. The real dataset consists of anchors for which a stem-loop structure is detected in their base targets. For each anchor-targets simulation unit, nucleotides are randomly mutated at several positions in each target. In the example, the anchor has three targets, shown in the orange-shaded box. The green squares represent the stem locations, and the number of mismatches (v) in each target comes from the real data. For simulation, the corresponding v nucleotide positions in each target are randomly mutated (denoted as gray squares in the blue-shaded box), and a simulated p-value is computed. This process is repeated 1000 times for all anchor-targets units in the dataset. (B) Empirical cumulative distribution functions (ECDFs) of p-values (unadjusted) obtained from real data and 1000 simulated datasets. The ECDFs show unadjusted p-values obtained from the STRUCT output of a mosquito metatranscriptomics dataset (Batson et al.) alongside 1000 simulated datasets. Each data point corresponds to an anchor where a stem-loop is detected in the base compactor. The zoomed-in view highlights the enrichment of small p-values in the empirical dataset, indicating that STRUCT effectively identifies conserved stem-loops in RNA sequences.



Suppl. Figure 5: ECDF of BH-adjusted p-values in HIV and mosquito metatranscriptomics datasets. (A) and (B) show ECDF of adjusted p-values stratified across various RNA sequence categories obtained from different datasets using STRUCT on targets and compactors. For targets, each data point in the ECDF represents an anchor's adjusted p-value. For compactors, each data point in the ECDF represents an anchor's p-value for a way of split-and-concatenation of compactors. (A) For mosquito metatranscriptomics, all extendors (combined anchor and target sequences) and compactors undergo alignment with the Rfam database and a variety of viral genomes. Sequences that do not correspond with known rRNA sequences or match RNA virus genomes are categorized as "Unannotated RNA." Sometimes different extendors (or compactors) linked to a single anchor get different annotations; for instance, if four out of five extendors associated with one anchor are identified as rRNA while one remains unannotated, we apply a majority rule approach for labeling the anchor, in this case as rRNA. The sequences are also classified based on the type of structural variation they exhibit, which is detailed in Figure 1C. The categories include BPC, SVE, and sequences with "Neither" (noting no conservation nor covariation in their stem regions). Similarly, the label for each anchor is determined by the majority annotation of its data points. (B) In the dataset of HIV, all extendors and compactors are aligned to the Rfam database. Sequences are classified by whether they align to an HIV structural element and by mechanisms of conserved structures. We employed the same majority rule to label anchors.