SUPPLEMENTAL TEXT

High-throughput sequencing. Partitioned nucleic acid populations were recovered from the nitrocellulose filters by adding 500 µl of extraction buffer (8 M Urea, 10 mM EDTA, and 50 mM NaCl) to the filters, vortexing the sample, then placing the sample in a 95°C heating block for 5 min followed by additional vortexing. Recovered nucleic acid was ethanol precipitated, dried and re-suspended in 100 µl of DNase buffer (50 mM Tris-HCl pH 8, 6 mM MgCl, 10 mM CaCl), and treated with 1 unit of DNase RQ1 (Promega) at 37°C for 10 min. Samples were then extracted with an equal volume of phenol:chloroform followed by chloroform extraction. 1/8th of the recovered RNA was then reverse transcribed in transcription buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM MgCl₂, 50 mM DTT) with 500 µM of each dNTP using 2 µl of ImProm-II reverse transcriptase (Promega) to extend a complementary primer for 5 min at 23°C and 45 min at 42°C followed by 3 min at 95°C. In the first PCR amplification cDNA was amplified with primers that were complementary to part of either the 5’ or 3’ constant region on their 3’ ends and carried Illumina-specific sequences on their 5’ ends. For the primers partially complementary to the 5’ constant region, the extent of the complementarity was slightly staggered for each sample so that the Illumina sequencing reads would begin at different positions within the constant region. These variable starts provide a more uniform distribution of base calls in the early rounds of sequencing, which is critical for proper image processing. A second PCR amplification utilized primers that included indexing sequences to allow multiplexing multiple populations in a single sequencing run. The PCR products were purified on a 2% agarose gel and recovered using the QIAquick gel extraction kit (Qiagen). 5 µl of the purified samples were then diluted into 50 µl and amplified by 4 additional rounds of PCR, purified with a QIAquick PCR purification kit, and quantified by UV absorbance. The six populations described here were then pooled along with fifteen additional populations from unrelated projects at equal concentrations and assessed using a Bioanalyzer (Agilent).
Sequence analysis

In our analysis, constant regions (primer-binding sequences) were first removed from the 5’ and 3’ ends of all reads to eliminate biases introduced by the trivial nature of their conservation. Removing the constant regions resulted in trimmed reads of 68-73 nt in length. For trimmed reads that had any position with a Qphred score of 13 or less the entire read was removed from further analysis. The remaining reads were ranked according to abundance and then clustered based on sequence edit distance (Levenshtein distance). Sequences within a cluster are an edit distance of <7 from the most abundant sequence in that cluster.

For the 100 most abundant clusters, the sequences within each cluster were aligned using Mafft (1), and these alignments were used to predict basepairing patterns using RNAalifold (weight of the covariance term = 0.6, penalty for non-compatible sequences = 6) (2). For clusters with >300 non-redundant sequences, the 300 most abundant sequences within that cluster were used for these structure predictions. These alignments and the centroid structures predicted by RNAalifold were then used to generate covariance models (CMs) using Infernal (3). Infernal was then used to calibrate and search the covariance models against the population. The CMs were searched against a reduced sequence data set comprising only the most abundant sequence in each of the top 5000 clusters identified in each population. An expectation score cut-off of 0.1 was used for all CM searches. Initial CM searches were carried out based on CMs derived from the 10 most abundant clusters in the population. Groups of clusters identified by the first round of CM searches were aligned to the CM and used to generate a new refined CM. Among the 5000 top clusters, the remaining cluster that did not match these CMs after 10 rounds of CM searching and refinement were identified, and new CMs from each of the 10 most abundant members of these remaining clusters were used to seed new CM searches. This process was repeated until all of the 100 most abundant clusters had been assigned to at least one CM. The sequences from the reduced data set were aligned to their respective CMs and R2R (4) was used to visualize sequence and
structure conservation. Basepairs that were conserved from the original intra-cluster based prediction in more than 80% of the clusters identified by the CM searching were used to generate the new CMs. The new CMs were then aligned to the single cluster from which they were originally derived, and that alignment was used to seed an additional 10 rounds of CM searching and refinement to yield the final CM. To identify clusters that utilize the constant regions of the RNA sequence to form their functional structures, the two 24 nt primer-binding sequences were appended to the 5’ and 3’ ends of the 5000 sequences in the data set and the final refined CMs were used to search these ~118 nucleotide full-length sequences. The consensus descriptions shown in figure 3 were generated using R2R based on the 240s population with the primer-binding sequences appended.

Previous LTS analysis identified a dominant pseudoknot motif in HIV-1 RT aptamer populations referred to as either Tuerk-type or Family 1 pseudoknot (F1Pk) that is essentially identical to the F1Pk motif identified here (5,6). This F1Pk motif can be readily identified based on a sequence signature that is highly predictive for the presence of the motif, and therefore the F1Pk motif served as a convenient control for an initial evaluation of our covariance analysis. While the comparative sequence analysis tools used here do not explicitly predict pseudoknots, their presence is revealed by identification of the pseudoknots’ component stems which form the basis for stem specific CMs. Stem specific CM searches (each based on only one of the two component stems) independently identified groups of clusters that overlapped significantly with each other. Based on this overlap the clusters within these groups were identified as belonging to one of the three pseudoknot motifs (Supplemental Figure S1).

Multiple CM searches that began with F2Pk clusters eventually failed to identify the original cluster used to build the initial CM. This occurred because the CM searches identified several F1Pks, which during CM refinement came to define the CM consensus to the exclusion of the starting cluster. To deal with CMs initiated from these clusters, the CMs were initially searched against a data set that excluded F1Pks, and only searched against the entire 5000 clusters in the final round, after the F2Pk consensus was established.
Cell Lines and Viruses

The human cell line, 293FT (Invitrogen, Carlsbad, CA), was maintained in standard culture media containing Dulbecco’s Minimum Essential Medium (Sigma, St. Louis, MO), 10% FBS (Sigma, St. Louis, MO), 2 mM L-glutamine (Gibco, Life Technologies, Grand Island, NY), 1 mM non-essential amino acids (Gibco, Life Technologies, Grand Island, NY), 1 mM sodium pyruvate (Gibco, Life Technologies, Grand Island, NY), and 0.5 mg/mL G418 (Sigma, St. Louis, MO). Cells were maintained at 37°C in 5% carbon dioxide with splitting twice per week.

SUPPLEMENTAL REFERENCES

1. Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511-518.
2. Hofacker, I.L., Fekete, M. and Stadler, P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, **319**, 1059-1066.
3. Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335-1337.
4. Weinberg, Z. and Breaker, R.R. (2011) R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, **12**, 3.
5. Tuerk, C., MacDougal, S. and Gold, L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc Natl Acad Sci U S A*, **89**, 6988-6992.
6. Burke, D.H., Scates, L., Andrews, K. and Gold, L. (1996) Bent pseudoknots and novel RNA inhibitors of type 1 human immunodeficiency virus (HIV-1) reverse transcriptase. *J Mol Biol*, **264**, 650-666.
Supplemental Figure S1. Multiple rounds of CM searching and refinement identifies converged motifs. (A) The number of clusters found to fit the constraints of a CM (y-axis) during successive rounds of CM searching (x-axis) within the 0s population (t0). Searches that started from CMs based on the intracluster structure prediction based on the $2^{nd}$ – $7^{th}$ most abundant clusters from the 0s population are plotted. All six of these clusters have the potential to form a F1Pk, but only some of the initial structure predictions are compatible with formation of a F1Pk. CM searches based on stem 1 of the F1Pk motif (green) or stem 2 of the F2Pk (blue) are shown along with CM searches based on basepairing that is incompatible with the F1Pk structure (red). The CM searches based on either of the two component stems of the F1Pk grew faster and larger than the CMs based on basepairing that is incompatible with the F1Pk. Importantly, these CM searches found that clusters 2, 5, and 6 fit the constraints of CMs that were compatible with F1Pk. Thus the performance of CMs provides a measure of the accuracy of the initial structure prediction based on intracluster variation. (B) After multiple rounds of CM searching, the CM searches based on stem 1 show significant overlap with CM searches based on stem 2 of the F1Pk. The cluster t0.7 was found by the CM searches initiated by cluster t0.3 and pseudoknot sequence of t0.7.1 is shown with the basepairing present in the CM derived from cluster t0.3 (stem 1). The same sequence is shown with the basepairing present in the CM derived from cluster t0.7 (stem 2). These two basepairing patterns can be satisfied simultaneously resulting in the F1Pk. (C) The most abundant sequences in the $2^{nd}$ – $7^{th}$ most abundant clusters are shown aligned to structure predictions based on the initial intracluster variation within their respective clusters. The sequences that correspond to stem 1 are shown in green and sequences that correspond to stem 2 are shown in blue. Two of the predicted structures include stem 1 (green) and one of the predicted structures includes stem 2 (blue).
Supplemental Figure S2. Multiple sequences satisfy the constraints of both the F1Pk and F2Pk CMs. Two examples are shown. The two CM definitions are based on stem 1, the two versions of stem 1 used by the CMs are indicated. In both cases the CMs use a four basepair stem; however the actual stem 1 for these specific sequences is almost certainly includes the five potential basepairs. Stem 2 is also indicated, but this pairing does not explicitly contribute to the definition of the CMs.
Supplemental Figure S3. The fraction of the pre-enriched population that is captured on a nitrocellulose filter decreases following the addition of the DNA trap R1T. The fraction of $[^{32}\text{P}]$-5’end labeled material added to the filter that was recovered is plotted as a function of time since the addition of R1T. The dashed line indicates the fraction of material recovered in the NoRT population as a measure of protein-independent background binding. After 960s (this population was not sequenced) the fraction of material recovered from the filters fell to background levels.
**Supplemental Figure S4**

Simulated distributions of read counts based on 100 rounds of simulated resampling. The simulated resampling was based on randomly sampled values from Poisson distributions with average frequencies equal to the observed frequencies of sequence 240.3.1 in the 0s population (3.23%) and in the 240s population (2.07%). The number of times specific read counts were observed in the simulated resampling is displayed as histograms with a bin width of 1. *(A)* For panel A, the resampled population size was 711,100, which is equal to the size of the quality-filtered 240s HTS data set. The difference between the simulated distribution of observed reads in the 240s population (Observed Change, black) and the simulated distribution of reads that would be expected if there was no change in abundance relative to the 0s population (No change, red) is large relative to the variation of the two distributions that would be expected due to simple resampling. Thus, the HTS data set provides clear support for a real shift in the relative abundance of this aptamer (compare with **Figure 4B** in the main manuscript). *(B)* For panel B, the resampled population size was 100, which is typical of a relatively large LTS data set. In this much smaller data set, the two distributions show significant overlap, revealing that the change in abundance observed for this sequence in the HTS data set would not have been resolved by LTS.
Supplemental Figure S5. Verification of the functional significance of the (6/5)AL. (A) Truncations and variants of the (6/5)AL sequence 240.1.1. (B) Inhibition of primer extension by RT in the presence of the sequences shown in (A) and a version of truncation 1 in which all pyrimidines have their 2’OH substituted with 2’F. The faster migrating unextended primer and slower fully extended product are indicated at the top and the bottom and top of the gel respectively.