Supplementary analysis 1:
Euclidean distance in the RESS

As shown in Fig. S1, Euclidean distance did not adequately distinguish between pairs of random sequence and pairs of structurally-related sequence. In fact, random sequence pairs were the closest together by this measure. One possible explanation for this is that the random sequences all scored very low for all CMs, and thus are essentially clustered around the origin in the RESS. Using a correlation-based distance measure, such as the Spearman distance, allowed us to identify RNAs with similar score patterns, which we found to be better suited for pulling out groups of related structure.

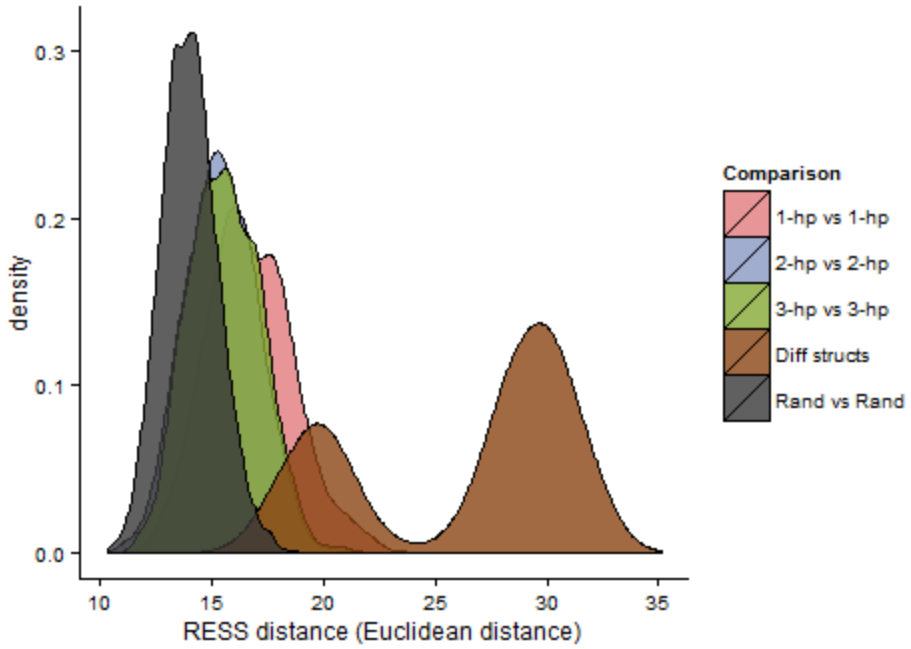

Figure S1: Similar to Fig. 2D in paper, but using Euclidean distance instead of Spearman distance.

Supplemental analysis 2:
Use of sliding windows

Since using a sliding window may result in some loss of sensitivity, we first tested this approach using datasets consisting of 50 sequences from the 2-hp synthetic structure family, each embedded in 10-500nt of random sequence on each end. We generated five versions of this dataset with different random embeddings of the sequences and then created two sliding window datasets for each: a 50nt window with a 35nt slide, and a 150nt window with a 105nt slide. The large slide (70% of the window size) was used to decrease the occurrence of clusters that can form due to highly overlapping sequence. When we clustered these datasets with NoFold, we recovered an average of 33.4±5.1 of the 2-hp structures in the 50nt window set (0.67±0.10 sensitivity) and 34.6±5.5 in the 150nt window set (0.69±0.11 sensitivity). Compared to the 0.88 sensitivity obtained from clustering the non-embedded 2-hp sequences (Table 2 in paper, middle row), this indicates that we can still maintain reasonable sensitivity with a sliding window approach.

Supplemental analsysis 3:
Non-canonical translation initiation sites

Background

Two recent studies utilized ribosome profiling in combination with harringtonine (Ingolia, Lareau, & Weissman, 2011) or lactimidomycin (Lee, Liu, Huang, Shen, & Qian, 2012) treatment to capture the locations of initiating ribosomes across the entire mouse and human transcriptomes. One interesting finding of these studies was that translation initiation at non-AUG start codons accounts for ~50% of initiation sites. Although initiation at near-AUG codons (e.g. CUG, GUG) in good Kozak context is thought to be possible through wobble base pairing of the methionine tRNA (Peabody, 1989), it is unknown whether this mechanism can support initiation at completely non-canonical sites (i.e. those with more than two base differences from AUG). Previously, certain IRES (Jan, Kinzy, & Sarnow, 2003; Kanamori & Nakashima, 2001) and hairpin structures (Ash et al., 2013; Mori et al., 2013; Zu et al., 2011) have been shown to facilitate initiation at non-canonical codons, suggesting that RNA structures may play a central role in this phenomenon. We therefore asked whether structure motifs might be found near the newly discovered non-canonical initiation sites and help explain their occurrence.

Results and Discussion

To identify motifs in the proximity of non-canonical translation initiation sites (ncTIS), we created two datasets consisting of the sequence immediately upstream of each ncTIS identified in humans by Lee et al. (2012). We restricted our analysis to ncTIS with start codons that were two or more bases different than AUG and to the first 50nt and 150nt upstream of the ncTIS. Although most known viral IRES are larger than these window sizes, putative cellular IRES often contain smaller structural modules that can promote initiation individually (Stoneley & Willis, 2004), and a recent study found several sequences smaller than 100nt that had IRES activity in humans (Wellensiek et al., 2013), suggesting these window sizes are reasonable. We first scanned the sequences directly against the Rfam database using cmsearch to determine if any sequences contained instances of known initiation-promoting structures. The only relevant hit was a sequence derived from heat shock protein 70 kDa (Hsp70) which matched, unsurprisingly, to the Hsp70 IRES CM (RF00495). We turned next to novel motif analysis. NoFold identified 33 50nt motifs and nine 150nt motifs, all of which were found to be significantly enriched upstream of ncTIS relative to non-ncTIS positions in the same transcripts (Fisher's exact test, FDR-corrected $p < 0.05$; Table S1). We note that since there was a difference in GC content between the background and test sequences (54-55% and 48-49% GC, respectively), we cannot fully rule out an effect of this difference on the enrichment test. Furthermore, it is possible that some of the sequences near the true ncTIS may be more closely related to each other than expected by chance, due to considerations such as ancient gene duplication events. Several motifs were annotated as having IRESs among their top ten CMs, although the average Z-scores for these IRESs were generally < 3. One exception to this was motif M23 (Fig. S2A), which scored highly against the Hsp70 IRES ($Z = 3.8$) and is predicted to be highly stable (minimum free energy (MFE) = -30.77). The Hsp70 sequence mentioned earlier was not in this cluster and therefore did not contribute to the high Hsp70 IRES score. This motif also had two tRNA-like CMs among its top ten (TLS-PK3, $Z = 2.8$; TLS-PK2, $Z = 2.7$), which is interesting because tRNA-mimicry is a mechanism used by certain IRES to achieve initiation at non-canonical codons in the absence of tRNA-Met (Jan et al., 2003).

The most widespread motif that we identified was found to occur in seven sequences, six belonging to histone subunit H4 genes and one belonging to heat shock protein 60 (Hsp60). This motif is predicted to be a small hairpin (Fig. S2B), and appears to occur in a similar location in each H4 gene (~50-80nt from the 5' transcript end). Interestingly, H4 transcripts were recently shown in mouse to use an unusual mechanism for translational initiation that involves loading of ribosomes independently of the 5' cap (Martin et al., 2011). This process is thought to depend on two RNA structures, one that recruits the cap binding protein eIF4E and another that may help position the ribosome over the initiation site in a manner similar to an IRES. It has not yet been investigated whether this mechanism supports initiation at non-canonical initiation codons. We note that NoFold also found several motifs within other histone gene families, including H2 and H3, but that most of these motifs were more conserved in primary sequence than in structure.

Of the 42 motifs found upstream of human ncTIS, none occurred in more than seven sequences, indicating that no single structure accounts for a large portion of human non-canonical initiation. Given that very little sequence and structural similarity has been found between different IRES in the past, this is not entirely unexpected. Additionally, a possible complicating factor in this analysis is that initiation-promoting motifs do not necessarily occur immediately upstream of the ncTIS. Some IRES are located more distally from the start codon and interact with the initiation site by pseudoknot formation (Kanamori & Nakashima, 2001). This makes it difficult to find motifs specifically involved in non-canonical initiation, since an ncTIS must be linked to the distal motif either by using pseudoknot prediction, which is computationally intensive for long sequences, or using direct experimental probing. Due to this, we expect that our analysis of only the regions upstream of ncTIS is an underestimation of the motifs involved in non-canonical initiation.

Methods

The transcript positions of ncTIS in human were obtained from (Lee et al., 2012). Codons were defined as ncTIS if they were neither AUG nor near-AUG codons but showed translation initiation through ribosome profiling analysis. Since multiple mapping of non-unique ribosome footprints was allowed in the original dataset, we removed any ncTIS that was surrounded by >20nt of sequence that was exactly identical to any other ncTIS. Such ncTIS mostly fell within repetitive elements. We extracted 50nt/150nt upstream of each remaining ncTIS to create separate 50nt and 150nt datasets. Within a dataset, extracted sequences were allowed to overlap by no more than 50% of their length (this occurred when there were two ncTIS close together within the same transcript). If such an overlap occurred, only the first sequence was kept. If the full sequence length could not be extracted due to an ncTIS falling too close to the 5' end of the transcript, the 5' end was buffered with random sequence. A background database for the enrichment analysis was created from 50nt/150nt upstream of random transcript locations that were not within 25nt of an ncTIS. The GC content of the background sequences was 49% and 48%, respectively, which was lower than the GC content of the test sequences (54% and 55%). Only transcripts that had observed expression in the ribosome profiling experiment were used to obtain background sequences. Enrichment testing was performed as described in the previous section. The search for known structures was carried out using all 1,973 CMs and the cmsearch module of Infernal (v.1.0.2) with options "--ga --toponly", which reports only matches that exceed the "gathering threshold" defined for each CM.

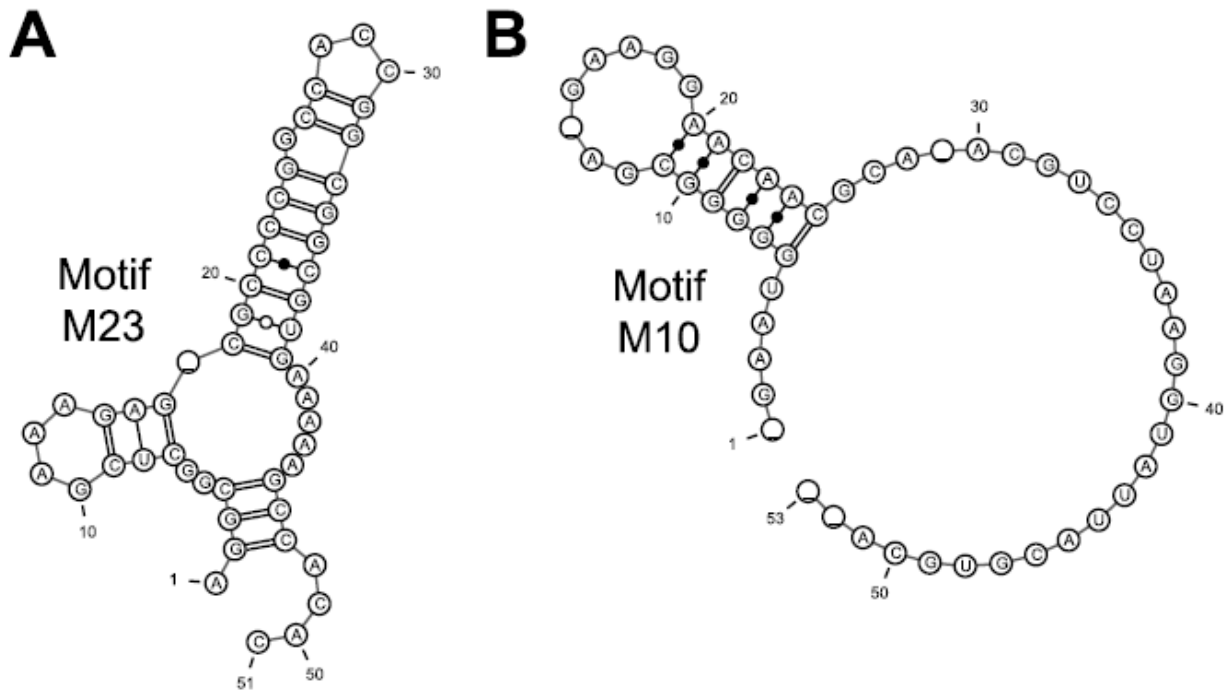

Figure S2. Consensus structures of motifs enriched upstream of non-canonical translation initiation sites. (A) A structure with possible similarity to elements of the IRES_Hsp70 CM and two tRNA-like hairpins. (B) A small hairpin found in six Histone 4 genes and Hsp60.

Table 3. Summary of motifs identified in the ncTIS datasets.

Dataset	#Seqs	Window size	#Windows	# Motifs				
				≥ 3 seq	≥ 5 seq	≥ 10 seq	Enriched	SCI > 0.5
Non-canonical translation initiation sites	-	50 nt	782	33	3	0	33	13
		150 nt	632	9	3	0	9	3

≥ 3 seq, ≥ 5seq, ≥ 10 seq indicates the number motifs found in at least 3, 5, or 10 different sequence windows, respectively.
Enriched motifs had $p < 0.05$ after FDR correction.

References

Ash, P. E. A., Bieniek, K. F., Gendron, T. F., Caulfield, T., Lin, W.-L., DeJesus-Hernandez, M., … Petrucelli, L. (2013). Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. *Neuron*, *77*(4), 1–8. doi:10.1016/j.neuron.2013.02.004

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. doi:10.1016/j.cell.2011.10.002

Jan, E., Kinzy, T., & Sarnow, P. (2003). Divergent tRNA-like element supports initiation, elongation, and termination of protein biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15410–15415.

Kanamori, Y., & Nakashima, N. (2001). A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation. *RNA*, *7*(2), 266–274.

Lee, S., Liu, B., Huang, S.-X., Shen, B., & Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), 1–9. doi:10.1073/pnas.1207846109

Martin, F., Barends, S., Jaeger, S., Schaeffer, L., Prongidi-Fix, L., & Eriani, G. (2011). Cap-assisted internal initiation of translation of histone H4. *Molecular Cell*, *41*(2), 197–209. doi:10.1016/j.molcel.2010.12.019

Mori, K., Weng, S.-M., Arzberger, T., May, S., Rentzsch, K., Kremmer, E., … Edbauer, D. (2013). The C9orf72 GGGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS. *Science (New York, N.Y.)*, *339*(6125), 1335–1338. doi:10.1126/science.1232927

Peabody, D. (1989). Translation initiation at non-AUG triplets in mammalian cells. *Journal of Biological Chemistry*, *264*(9), 5031–5035.

Stoneley, M., & Willis, A. E. (2004). Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. *Oncogene*, *23*(18), 3200–7. doi:10.1038/sj.onc.1207551

Wellensiek, B. P., Larsen, A. C., Stephens, B., Kukurba, K., Waern, K., Briones, N., … Chaput, J. C. (2013). Genome-wide profiling of human cap-independent translation-enhancing elements. *Nature Methods*, *10*(8), 3–8. doi:10.1038/nmeth.2522

Zu, T., Gibbens, B., Doty, N. S., Gomes-Pereira, M., Huguet, A., Stone, M. D., … Ranum, L. P. W. (2011). Non-ATG-initiated translation directed by microsatellite expansions.

Proceedings of the National Academy of Sciences of the United States of America, *108*(1), 260–5. doi:10.1073/pnas.1013343108