Functional and structural basis of extreme conservation in vertebrate 5′ untranslated regions

Gun Woo Byeon1,2, Elif Sarinay Cenik1,2,4, Lihua Jiang1, Hua Tang1, Rhiju Das3 and Maria Barna1,2,✉

The lack of knowledge about extreme conservation in genomes remains a major gap in our understanding of the evolution of gene regulation. Here, we reveal an unexpected role of extremely conserved 5′ untranslated regions (UTRs) in noncanonical translational regulation that is linked to the emergence of essential developmental features in vertebrate species. Endogenous deletion of conserved elements within these 5′ UTRs decreased gene expression, and extremely conserved 5′ UTRs possess cis-regulatory elements that promote cell-type-specific regulation of translation. We further developed in-cell mutate-and-map (icM2), a new methodology that maps RNA structure inside cells. Using icM2, we determined that an extremely conserved 5′ UTR encodes multiple alternative structures and that each single nucleotide within the conserved element maintains the balance of alternative structures important to control the dynamic range of protein expression. These results explain how extreme sequence conservation can lead to RNA-level biological functions encoded in the untranslated regions of vertebrate genomes.

One of the most fascinating findings from the comparative analysis of vertebrate genomes is the existence of extreme sequence conservation in noncoding regions, at levels often greater than in coding regions with perfectly invariant polypeptides1–11. These regions are undergoing strong purifying selection in humans and are not merely mutational cold spots12. However, the fundamental problem that was initially raised a decade ago still remains unsolved: why does such extreme conservation arise during evolution, and what are the functional roles for such sequences in the genome?

To date, efforts to understand the phenomenon of extreme conservation have focused heavily on intergenic sequences, suggesting possible roles for these elements as transcriptional enhancers13–15. However, early in vivo knockout studies paradoxically yielded viable mice lacking grossly deleterious phenotypes, raising uncertainties about the relevance and contribution of highly conserved elements to organismal development16,17. It is only more recently that mice with loss of single or pairwise deletions of ultraconserved enhancer elements have been shown to produce more subtle developmental phenotypes due to the impact on the transcription of neighboring genes18,19.

However, beyond its importance in transcriptional regulation, the biological meaning of extreme conservation in post-transcriptional regulation remains largely unknown. While a few examples, such as the functional roles for ultraconserved regions transcribed as long noncoding RNAs or alternatively spliced poison cassette exons, have been described20–21, RNA-level mechanisms for extreme conservation have not been widely explored. The observation of extreme sequence conservation across extended stretches of 5′ UTRs suggests the presence of specialized translational cis-regulatory elements. In a paradigmatic example, the Hoxa9 5′ UTR contains an ~650-nucleotide extremely conserved region that mediates noncanonical translation initiation through a structured internal ribosome entry site (IRES)-like RNA element22. Knockout of an ~150-base pair (bp) functional element within this conserved region in mice results in diminished spatiotemporal Hoxa9 protein expression and a pronounced axial skeleton phenotype leading to a homeotic transformation, demonstrating how 5′ UTR RNA sequences important for specialized translational regulation in the developing embryo can undergo extraordinary negative selection.

We were thus inspired to ask if there could be a broader, systematic trend for extreme conservation, to reveal currently unknown translational regulatory sequences, and, conversely, if such regulatory sequences could help to explain the functional basis of extreme noncoding conservation in messenger RNAs.

Results

Hyperconserved 5′ UTRs in vertebrate genomes. To address the function of extreme noncoding conservation for mRNA 5′ UTRs, we used the conservation pattern of the aforementioned Hoxa9 5′ UTR as our archetype in selecting a set of other 5′ UTRs in the genome. The length of the extremely conserved stretch in Hoxa9 5′ UTR is ~650 nucleotides; the size of the functional element within the conserved stretch is around 350 nucleotides22. We used PhastCons with a log of the odds score (LOD) minimum of 500, which marked large blocks of extremely conserved sequences throughout the genome that are 100 nucleotides long, on average22. Using mouse RefSeq gene annotations, we intersected mouse 5′ UTRs with the LOD ≥ 500 PhastCons elements (representing the top 8.25% of all PhastCons elements identified in the genome), requiring at least 250 nucleotides overlap. This resulted in a set of 589 5′ UTRs for 499 genes (Fig. 1a and Supplementary Table 1). The median nucleotide identity between mouse and human genomes in the conserved regions in the selected 5′ UTRs is 92.3% (80% identity at 5th percentile). The average total length and the average number of nucleotides overlapping PhastCons elements for these 589 5′ UTRs are 674 and 389 nucleotides, respectively, and they tend to be found more frequently closer to the start codon than to the 5′ end (Extended Data Fig. 1a–c). For the remainder of the text, we will refer to these 589 5′ UTRs as hyperconserved 5′ UTRs (h5UTRs) and the LOD ≥ 500...
Fig. 1 | Hyperconserved 5′ UTRs in vertebrate genomes. a, Schematic illustrating selection of hyperconserved vertebrate 5′ UTRs. We begin with 60-way multiple species alignment of vertebrate genomes, its per-nucleotide PhastCons probabilities and conserved element prediction tracks. High-scoring (LOD ≥ 500) PhastCons elements are overlapped with RefSeq annotated mouse 5′ UTRs. We define those with overlap ≥ 250 nucleotides to be hyperconserved (also see Supplementary Table 1). b, Distributions of cross-tissue transcriptome–proteome correlations (GTEx Consortium data across 32 human tissues) for all genes, genes with h5UTRs or genes with size-matched nonconserved 5′ UTRs. Indicated P values are from two-sided Wilcoxon rank-sum tests for cross-tissue correlation values between h5UTR genes and all genes, or between h5UTR genes and size-matched nonconserved controls. c, Scatter plot illustrating the term enrichment strategy and criteria. The x axis and y axis plot the expected (E) and the observed (O) number of genes for each term. Blue dashed line indicates the minimum observed/expected (O/E) ratio cut-off of 3. Green line indicates expected and observed counts where two-tailed Fisher’s test P value (P_f) is estimated to have FDR = 0.05. Neighbor-weighted test P value (P_fw) ≤ 0.05 is further used as an additional cut-off. The final set of enriched terms passing the filter is colored by P_f and sized by P_fw. d, Visualization of representative gene ontology terms significantly enriched for the h5UTRs according to criteria in c. A number of genes mapping to each term are also displayed (also see Supplementary Table 2).
PhastCons elements within the h5UTRs as 5' UTR hyperconserved elements (HCEs).

We next asked if h5UTRs are more likely to be discordant in their mRNA:protein expression levels, which would suggest post-transcriptional regulation. Using the GTEx Consortium tissue-specific transcriptomics and proteomics dataset, we determined if genes with h5UTRs have a different distribution of per-gene cross-tissue correlations in mRNA versus protein levels, compared to genes with similarly sized, nonconserved (defined as no overlap with LOD ≥ 500 PhastCons elements) 5' UTRs25,28. For 181 h5UTR genes, both RNA and protein expression were detectable in at least ten tissues and the h5UTRs were annotated in both human and mouse RefSeq databases. Compared to all genes or to size-matched nonconserved controls, we observe significantly lower (Wilcoxon rank-sum test \( P = 0.0013, P = 0.0017, \) respectively) cross-tissue correlations (Pearson) for h5UTR genes (Fig. 1b). We also compared cross-tissue correlations of h5UTR genes with RNA variance-matched nonconserved controls to eliminate a model in

**Fig. 2 | Hyperconserved 5' UTRs impact translation efficiency.** a, Schematic of experimental design for testing the impact of hyperconserved 5' UTRs on translation of coding genes. Shift in the distribution of the mRNAs across sucrose gradient fractions towards the right (heavier polysomes) indicates more average ribosome loading and higher translation efficiency (TE), while shift towards the left indicates lower translation efficiency. b–f, Polysome profiles of wild type versus hyperconserved element (HCE) knockout cells for Chrd1 (b), Dlx1 (c), Gdf5 (d), Sema3a (e) and Zfx (f). Distribution of mRNAs across sucrose gradient fractions are plotted. The y axis (the line) plots the mean percentage mRNA for each fraction. Error bars indicate standard error. Asterisk indicates two-sided t-test \( P \leq 0.05 \) for each fraction between the knockout and the wild type, \( n = 4 \). Indicated \( P \) value is calculated by Fisher's method across all fractions.
which h5UTRs impact the correlations only through a different dynamic range of variation in RNA expression. The correlations were still lower for the h5UTR group (P = 0.03) (Extended Data Fig. 1d). Alternative 5’ UTR isoforms are also more frequently annotated for genes with h5UTRs than for all genes or nonconserved controls (Extended Data Fig. 1e). In summary, protein levels of h5UTR genes, as a group, are more difficult to predict with RNA levels alone than those of nonconserved 5’ UTR genes, suggesting that extreme sequence conservation in the 5’ UTR may be due to tissue-specific post-transcriptional control.

To describe the potential biological functions of genes with h5UTRs, we surveyed gene ontology (GO) terms enriched in the h5UTR gene set. To ensure the specificity of the enrichment, we also analyzed a length-matched set of nonconserved 5’ UTR genes, which did not yield any enriched term (Extended Data Fig. 1f). h5UTR GO terms highlighted genes critical for vertebrate embryonic developmental processes (Fig. 1c,d and Supplementary Table 2). For example, h5UTR genes are involved in morphogenesis of major tissues and organs, especially the nervous system. Genes that are part of signaling pathways involving the molecules WNT, retinoic acid, GABA, FGF, activin, BMP, PDGF, Notch, VEGF, hedgehog or Semaphorins are also abundantly present. We also note the genes involved in epigenetics, such as chromatin remodeling and histone acetylation. Additionally, when we intersected known disease-associated variants with h5UTRs, we identified five potentially interesting associations, which suggest that these regions may also play a functional role in disease (Supplementary Table 3)35. Overall, these annotation enrichments suggest that h5UTRs may play an important role in the post-transcriptional control of core embryonic developmental regulators.

Hyperconserved 5’ UTRs impact translation efficiency. To address experimentally whether the h5UTRs could impact the translational efficiency of mRNAs, we chose five candidates (Chrdl1, Gdf5, Dlx1, Sema3a and Zfx) that function in contexts where spatiotemporal expression patterns are important for embryonic development. Chrdl1 is a bone morphogenetic protein (BMP) antagonist with numerous functional roles in cell differentiation and synapse plasticity, and is implicated in multiple neurological disorders30–35. Gdf5 is a transforming growth factor (TGF)-β family protein with roles in skeletal and nervous system development36–39. Dlx1 is a homeobox transcription factor that has critical roles in craniofacial patterning, as well as in the differentiation and survival of neurons in the brain40,41. Sema3a is a semaphorin family protein that is secreted and functions as a guidance cue for axons and vasculatures42–46. Zfx is an X-linked transcription factor protein that regulates self-renewal of embryonic and hematopoietic stem cells47.

To examine the contribution of h5UTRs, we introduced deletions into the 5’ UTRs of Chrdl1, Gdf5, Dlx1, Sema3a and Zfx using pairs of CRISPR–Cas9 single guide RNAs (sgRNAs) targeting segments ranging between 50 and 200 nucleotides within the HCEs (Supplementary Figs. 2–6 and Supplementary Table 4). We used either mouse embryonic stem cells (mESCs), mESCs treated with retinoic acid to promote differentiation or NIH3T3 cells, reflecting the cell types and conditions where these transcripts are expressed for analysis of translation (Supplementary Fig. 1a). Polyoma profiling allows quantification of translational efficiency independent from effects on transcript levels (Fig. 2a). The mRNAs that are more highly translated are expected to be present in heavier polysomes as they are bound by more ribosomes. As expected, global translation levels displayed no difference between the wild types and CRISPR–Cas9-mediated HCE knockout cells (Supplementary Fig. 1b–f). However, we observe that, for all five candidates tested, the deletion mutants exhibited a shift in the distribution of the targeted mRNA species from the heavier polysomes into the lighter polysomes, indicating a decrease in translation efficiency (Fig. 2b–f). These findings suggest that h5UTRs may frequently harbor uncharacterized, additional cis-enhancers of translation initiation.

Noncanonical translation enhancer in hyperconserved 5’ UTRs. There has been growing evidence for the importance of less understood, alternative mechanisms of initiation independent of the cap-eIF4E interaction, which have the potential for enhancing transcript-specific regulation of gene expression34,48–53. For example, it has been estimated that 5–10% of cellular mRNAs may undergo cap-independent translation35. The HCE of Hoxa9 contains a functional RNA element previously shown to direct translation initiation in a cap-independent manner, which is required for proper embryonic development35. Therefore, we asked whether other h5UTRs can similarly activate noncanonical translation initiation.

To test this hypothesis, we performed a large-scale reporter assay to measure the levels of noncanonical translation initiation from the h5UTRs. We synthesized and cloned a library of 253 full-length h5UTRs into a bicistronic reporter construct, containing two reporter genes, Renilla and firefly luciferase, that are transcribed as one mRNA. The first cistron, Renilla luciferase, is positioned immediately downstream of the promoter and is translated by cap-dependent translation. The second cistron, firefly luciferase reporter ratios of each truncation relative to its full-length wild type. Error bars indicate standard error of the log2 luciferase activity ratios. Bars numbers to the left of the bars indicate exact P values for each comparison versus the full length.
We noticed two groups of reporter activities distributed in a bimodal distribution for each of the cell types (Extended Data Fig. 2a). Within the higher group, we found the three positive controls that we included in the reporter assays, which all promote cap-independent translation: hepatitis C virus (HCV) IRES, encephalomyocarditis virus (EMCV) IRES and the Hoxa9 5’UTR. The ‘empty’ negative control reporter activity is found in the lower group, near its median. Thus, the lower component appears to represent the background noise level present in our reporter assays.
Using mixture modeling of the bimodal distribution, we estimated the false discovery rate (FDR) for each tested 5'UTR as the probability that the reporter activity of the tested 5'UTR could have come from the lower noise group. Using the maximum reporter activity across all six assayed cell types, we estimated that the proportion of the tested 5'UTRs with noncanonical initiation activity is 33%. At 10% FDR, we are able to identify 90 5'UTRs with high noncanonical translation activity in at least one cell type (Fig. 3a).

**Fig. 4 | Cellular remodeling hyperconserved 5' UTR RNA structures.** a, Schematic of identifying RNA structures under cellular remodeling in 5'UTRs. Multiplexed, targeted DMS chemical probing of 69 5'UTRs inside cells from their endogenous mRNAs is performed following ATP depletion treatment to stop RNA helicase activity. b, Heat map of correlation (Pearson’s) matrix across replicate samples for untreated, ATP depleted and in vitro refolded samples (three each). The correlation values are calculated from a vector of normalized accessibility values for all nucleotides passing per-amplicon reproducibility cut-off. c, Manhattan plot of differential accessibility tests in 11-nucleotide overlapping windows across the 5'UTRs. The y axis indicates $-\log_{10}$ (KS test $P$ value) for each window along 69 5'UTRs in the x axis. Dashed line indicates the $P$ value cut-off at which permutation FDR is at 5%. d, Enlarged view of differential accessibilities along the Csde1 5'UTR from positions 190 to 386. Top plot shows $-\log_{10}$ ($P$ value) for each window. Highlighted boxes mark significantly different windows, above the dashed line indicating 5% FDR. Middle plot shows differential accessibility on the y axis, where greater than zero indicates increased accessibility upon ATP depletion and less than zero indicates decreased accessibility. Bottom plot shows differential accessibility for in-cell versus in vitro refolded RNA. Error bars in each plot show standard error, $n=3$. 

---

**ARTICLES**

Using mixture modeling of the bimodal distribution, we estimated the false discovery rate (FDR) for each tested 5'UTR as the probability that the reporter activity of the tested 5'UTR could have come from the lower noise group. Using the maximum reporter activity across all six assayed cell types, we estimated that the proportion of the tested 5'UTRs with noncanonical initiation activity is 33%. At 10% FDR, we are able to identify 90 5'UTRs with high noncanonical translation activity in at least one cell type (Fig. 3a).
truncations in the two reporters transfected to 10T1/2 cells, we observed that at least one truncation significantly reduced noncanonical initiation, as well as the total translational levels (Fig. 3d and Extended Data Fig. 3b). The trend for truncations to frequently reduce total translation activity is notable, since cap-dependent translation initiation typically increases in efficiency when the 5′ UTR is shortened. We asked if this is more generally true in a larger set of 38 h5UTRs by comparing the total translation directed by the full length versus only the first 300 nucleotides of the h5UTR, without a large proportion of the HCE in each h5UTR. Of the 38 h5UTRs, 20 decrease significantly in the shorter truncated 5′ UTR relative to the full length, while only 5 increase significantly (Fig. 3e). In contrast, truncating long, nonconserved 5′ UTRs does not show the same trend for decreased translation (Extended Data Fig. 3c,d). Furthermore, there is no correlation between the change in the density of upstream AUGs and the change in reporter activities (Extended Data Fig. 3e). Taken together, noncanonical translation enhancer elements in h5UTRs widely impact total translation efficiency in physiological cellular conditions, suggesting that h5UTR genes may be translated via more specialized initiation mechanisms that utilize evolutionarily constrained, sequence-specific cis-regulatory features.

Cellular remodeling of hyperconserved 5′ UTR RNA structures.

Higher order structures are inherent features of RNA molecules that underpin their biochemical function. The majority of previous covariation-based predictions of RNA structures in vertebrates occur in ‘moderately’ conserved regions of the genome and miss the HCEs\(^6\). This is because covariation analysis requires not only sufficient conservation for alignment, but also sufficient variation for statistical power\(^6\). Since extreme conservation limits the extent to which covariation signals can be informative, addressing this question currently requires additional experimental data.

We postulated that specific regions of mRNA that display localized sensitivity in their structures to active cellular remodeling by RNA helicases could potentially lead us to functionally relevant structures within h5UTRs that guide translation initiation. To obtain high coverage accessibility data for a large number of h5UTRs, we initially performed a highly multiplexed amplicon sequencing adaptation of dimethyl sulfate (DMS) mutational profiling\(^6\). DMS profiling was performed in mESCs under the conditions of no treatment or depletion of ATP to eliminate helicase activity.

**Fig. 5 | icM\(^2\) reveals structured elements in the hyperconserved Csd67 5′ UTR.** a. Schematic of localized perturbation patterns that may be observed in M\(^2\) data. Here, the mutant does not disrupt the overall structure and ‘releases’ its base-pairing partner. This results in an increase of chemical accessibility relative to other nucleotides. b. Schematic of global rearrangement patterns that may be observed in M\(^2\) data. Here, multiple conformations of the RNA molecule are present together in an ensemble at nonnegligible relative proportions. Mutations can shift this balance, such that one structural state is favored over the other. In this case, M\(^2\) reveals large-scale accessibility perturbations across a longer stretch of the RNA molecule. Multiple mutations often impact the relative proportions in similar ways, which manifests as correlated accessibility changes in the M\(^2\) data matrix. c. Schematic of the icM\(^2\) method. Mutagenesis library of the target RNA of interest is first generated using error-prone PCR followed by cloning into an expression vector. The cells are transfected with the library and treated with DMS. Total RNAs are extracted. Read-through reverse transcription encodes DMS-modified nucleotides as mutations on the cDNA, which are read out by high-throughput sequencing. Correlated mutations in sequencing reads are then quantified and the resultant covariation matrix is analyzed for signature perturbation patterns. d. Heat map of icM\(^2\) accessibility matrix for Csd67 5′ UTR from position 190 to 386. For each row, the chemical mapping profile of a single-nucleotide variant of the RNA is plotted across the columns, where the colors indicate z-scaled accessibility change values from the wild-type RNA. One-dimensional data from each mutant are vertically stacked to display a two-dimensional matrix. White boxes mark the two regions (A; positions 334–363 and B; positions 215–315) that display strong perturbation signals, which reveal their structures. e. A structure model (structure W) of region A. Bases colored in red indicate mutations with accessibility changes observed in icM\(^2\) data that are consistent with the model. f. Scatter plot showing correlations of per-nucleotide accessibility changes between each mutant versus the ‘wild type’ (wild-type accessibilities are not directly measured, but mean accessibilities of 10 lowest variable mutants are used as a close approximation) on the x axis and nucleotide positions along the x axis. P indicates two-sided Wilcoxon rank-sum test P value for the difference in distributions of correlations between region B versus other nucleotides. g. Multiple species alignment for Csd67 5′ UTR from position 125 to 548. For each row, the sequence alignment of a species is plotted across the columns, where the colors indicate match/substitution/insertion/deletion at each nucleotide. The alignment positions are relative to the mouse sequence. The top row is the mouse alignment, colored separately from other rows as a reference to identify the identity of the bases in each position in the multiple species alignment.
We successfully profiled 161 tiling amplicons of 250 nucleotides in size across 69 endogenously expressed h5UTRs. We identified 140 11-nucleotide windows over 20 h5UTRs that were significantly different (FDR ≤ 0.05) between ATP depletion and no treatment (Fig. 4c and Supplementary Table 6). One known source of RNA structure remodeling in the cell is ribosome unwinding of mRNAs during translation, and thus the presence of upstream open reading frames (uORFs) may lead to differential accessibilities.

**Figure a:** Localized perturbations

**Figure b:** Global rearrangements

**Figure c:** Mutagenesis library

**Figure d:** Cds1: 190–386

**Figure e:** Supported by mutations

**Figure f:** Region B

**Figure g:** Mutations in various species

**Figure h:** Covariation matrix
Csd1 5′ UTR encodes alternative functional RNA structures. As a model to investigate cellular RNA structure and its remodeling in HCEs, we sought to further characterize the helicase-sensitive structures in the Csd1 5′ UTR. In particular, we developed in-cell mutate-and-map (icM²), a powerful methodology that enables application of the M² strategy, wherein systematic mutagenesis of RNA is coupled with chemical mapping to generate accessibility profiles for every mutated nucleotide, inside the native cellular context (Fig. 5a–c). In icM², the target sequence of interest is mutagenized using error-prone PCR, cloned as a pool into an expression plasmid and transfected into cells. Following the treatment of transfected cells with DMS, total RNAs are extracted and subjected to read-through reverse transcription, where modified nucleotides are misincorporated as mutations on the complementary DNA, which are amplified and sequenced. Correlated mutations in sequencing reads are then quantified, and the resultant covariation matrix is analyzed for signature perturbation patterns. icM² is particularly suited for analysis of h5UTRs, as it directly addresses what RNA structural changes occur if each of the extremely conserved nucleotides is mutated during evolution.

We applied icM² in three windows tiling across the Csd1 5′ UTR in mESCs. We observed strong perturbation signals in the 215–365 positions along the 5′ UTR, where we had originally observed large differential accessibilities in response to elimination of RNA helicase unwinding activities (Fig. 5d). The visualization of the icM² accessibility matrix immediately highlighted two subregions. The first region is around positions 334–363 (region A), where short-range localized perturbations indicated the presence of a small stem loop motif. Here, the data corresponded well to the expected accessibility changes for the lowest free energy structure (structure W) predicted for the region (Fig. 5e). The second region is around positions 215–315 (region B), where correlated global perturbations across a long stretch of about 100 nucleotides indicated the presence of multiple conformations. Remarkably, these correlated global perturbations occur for almost every mutation across the 100-nucleotide stretch, revealing the strong sensitivity of the ensemble state to the precise sequence identity of each base. This is highlighted by the correlation of per-nucleotide accessibility changes between each mutant versus the ‘wild type’ (Fig. 5f; Wilcoxon rank-sum test P = 0.0015 for mutants in region B versus other mutants). Therefore, at least two conformational states exist whose relative proportions inside the cell are affected by a mutation in almost any of the extremely conserved nucleotides. In addition, we observe the strongest conservation signal of the Csd1 5′ UTR in region B, where, amongst placental mammals, there is near-perfect sequence identity (Fig. 5g). These results suggest a structural explanation for why such extreme conservation levels may be required. Furthermore, examining the conservation levels and ATP-dependent accessibility profiles across all other h5UTRs reveals that the average per-nucleotide conservation levels in significantly differential accessibility regions (FDR ≤ 0.05) display exceedingly high conservation levels compared to the rest of the RNA (Extended Data Fig. 5a,b). Thus, encoding of actively remodeled cellular RNA structures may be a broadly occurring phenomenon associated with the extreme conservation levels in h5UTRs.

We next asked what candidate structures might explain the observed alternative states of the ensemble in region B. We used the average accessibility change profiles for the two clusters as two separate constraints for RNA folding (Fig. 6a). Constraining by the cluster 1 average accessibility profile revealed a well-defined conformation (structure X) disrupted by cluster 1 mutants in the 5′ UTR upstream of the luciferase reporter. Plotted are the mean; error bars indicate standard error. The bars and labels along x axis are colored according to whether they are wild type, cluster 1 mutants or cluster 2 mutants. Dashed line indicates the wild-type luciferase reporter level. Asterisk indicates a significant difference between each mutant and wild type (two-sided t-test P ≤ 0.05, n = 3).
helices and stabilized by cluster 2 mutants in the loops (Fig. 6b). Constraining by the cluster 2 profile resulted in a higher entropy fold, which was nevertheless readily visualizable by two representative medoid conformations (structures Y and Z; Fig. 6c,d). To estimate the relative mixing ratios of these structures, we chose to apply the RNA ensemble extraction from footprinting insights technique (REEFFIT)⁹. For the wild-type sequence, REEFFIT yielded proportions of 67±9%:10±4%:23±9% for the representative structures X, Y and Z, respectively (Fig. 6b–d). It also predicted how these proportions are expected to change across the individual mutants, adding quantitative estimates to our initially qualitative observations of alternative structural states. For example, cluster 1 mutants disrupt structure X to favor structures Y and Z, changing the relative proportions of X:Y:Z to 30%:13%:57% on average, while cluster 2 mutants act in the opposite direction, shifting the proportions to 91%:6%:3% (Fig. 6e). We further discovered that the accessibility change profiles of the two clusters of mutants are closely correlated with helicase-dependent accessibility change (Fig. 6). This observation suggests that elimination of RNA helicase unwinding activity decreases the proportion of structure X in the cell and does so to
increase the fraction of the alternative structures Y and Z. Notably, structure X has multiple long stems (positions 232–282); that is, the helicase activity promotes a low free energy structure and potentially may act as a chaperone\textsuperscript{11}. It is formally possible for other direct contacts on the exact methylation sites of the nucleotides, such as a direct RBP interaction on the base-pairing face, to produce localized ‘footprints’ on the accessibility profiles; however, this would not drastically impact our model. Taken together, we propose three candidate conformations to account for our icM\textsuperscript{2} signal observed in region B of the Csde1 5′ UTR and hypothesize that the cell is actively expending energy to maintain the precise relative balance of these conformations in the cellular structural ensemble.

In our initial one-dimensional DMS profiling analysis of h5UTRs in mESCs, we had observed that the accessibility profiles of many RNAs refolded in vitro were discordant from those of RNAs in cells (Fig. 4d and Extended Data Fig. 5c). To further expand on these differences and to actually compare in-cell versus in vitro RNA structures, we also performed in vitro M\textsuperscript{0} on Csde1 5′ UTR. We observed a strikingly different accessibility matrix (Extended Data Fig. 6a,b). These results highlight the importance of resolving flexible conformations that can occur uniquely under cellular conditions.

Lastly, we asked whether such a shift in the balance of structural conformations has a functional consequence on the translation of the downstream gene. We performed luciferase reporter assays with mutant Csde1 5′ UTRs carrying a number of substitutions from each of the two clusters, which are predicted to change the relative proportions. We selected four different nucleotide positions from cluster 1 and three from cluster 2, hypothesizing that similar patterns of expression level changes may be observed among each cluster. We observed that all cluster 1 mutants decreased firefly luciferase activities by 15–20% compared to the wild-type 5′ UTR (Fig. 6g). In contrast, cluster 2 mutants increased the reporter activities by 5–15%. When the three individual single mutations from cluster 2 are combined, the effect size is increased to about 50%. The dynamic range of final protein levels can thus be tuned according to the relative proportions of the multiple conformations along the RNA structural landscape of the Csde1 5′ UTR. Together, these results suggest that the exact proportions and properties of the RNA structural ensemble are a critical functional requirement under negative selection in hyperconserved vertebrate 5′ UTRs.

**Discussion**

Extreme sequence conservation has long been observed in non-coding regions of vertebrate genomes, yet our current functional knowledge of these elements falls short in explaining why and how such conservation levels exist. Here, we uncover a functional role for hyperconserved 5′ UTRs in regulation of translation and report their unexpected enrichment in noncanonical initiation sites, particularly within those transcripts critical for development in vertebrate species. We speculate that there may potentially be many different types of unknown noncanonical mechanisms that are adopted by these 5′ UTRs and that further investigations may identify new classes of RNA elements that accommodate more specialized mechanisms of translational control. The activities of h5UTRs may vary across cells and tissues, which may result in differential translatability of these mRNAs.

A crucial component of decoding cis-regulatory features at the level of RNAs is the determination of their higher order structure beyond the primary sequence. To this end, we developed a technique, icM\textsuperscript{2}, to examine the RNA structural ensemble within cells. We found that cells precisely tune protein expression levels by remodeling the hyperconserved Csde1 5′ UTR to maintain the relative proportions of multiple functional conformations. While icM\textsuperscript{2} revealed a highly dense array of mutations that disrupt such an actively enforced balance of dynamic structures in the Csde1 5′ UTR across a ~100-nucleotide-long stretch, the same mutations are negatively selected against in nature across vertebrate species. This suggests that selective pressures for translational regulation can lead to extreme sequence constraints when an ensemble of multiple functional conformations must be encoded over a single stable structure to ensure a dynamic range of translational outputs.

The observation that regions of h5UTRs under helicase-dependent structural remodeling in general display the highest conservation levels further suggests that a similar phenomenon could extend more broadly to other h5UTRs and may, at least in part, explain extreme conservation in 5′ UTRs at the level of RNA. Flexible structural states can potentially endow multiple functional states in regulatory elements that respond to environmental or cellular cues. Most current genome-wide efforts to identify functional structures have focused on single stable conformations. Our results underscore the necessity of the ensemble perspective of RNA structure in understanding the cellular activities of regulatory RNAs and the potential utility of extreme conservation in detecting such dynamically structured elements in the untranslated regions of mRNAs. We envision that hyperconserved 5′ UTRs will aid the discovery of functional RNA structures in vertebrate genomes and advance our broader understanding of post-transcriptional gene regulation in development, disease and evolution.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00830-1.

Received: 10 July 2020; Accepted: 26 February 2021; Published online: 5 April 2021

**References**

1. Dermitzakis, E. T., Reymond, A. & Antonarakis, S. E. Conserved non-genic sequences – an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6**, 151–157 (2005).
2. Harmston, N., Barcisz, A. & Lenhard, B. The mystery of extreme non-coding conservation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**, 20130021 (2013).
3. Halligan, D. L. et al. Positive and negative selection in murine ultraconserved noncoding elements. *Mol. Biol. Evol.* **28**, 2651–2660 (2011).
4. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
5. Dimitrieva, S. & Bucher, P. Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics* **28**, 1395–1401 (2012).
6. Boffelli, D., Norega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* **5**, 456–465 (2004).
7. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
8. Sandelin, A. et al. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99 (2004).
9. de la Calle-Mustienes, E. et al. A functional survey of the enhancer activity of hyperconserved non-coding sequences – an unexpected feature of mammalian genomes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**, 20130021 (2013).
10. Sakuraba, Y. et al. Identification and characterization of new long conserved non-coding sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
11. Richtsmeier, J. T. et al. Comparison of human chromosome 21 conserved nongenic elements. *Nat. Rev. Genet.* **19**, 703–712 (2008).
12. Muramatsu, K. et al. Comparison of regulatory elements conserved between human and mouse genomes. *Genome Res.* **15**, 1061–1072 (2005).
13. Deloukas, P. et al. Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mamm. Genome* **19**, 703–712 (2008).
14. Dermitzakis, E. T. et al. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
15. Katzman, S. et al. Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
16. Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
17. Vissel, A. et al. Ultracore is a suitable small subset of extremely constrained developmental enhancers. *Nature* **40**, 158–160 (2005).
18. Vissel, A. et al. A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895–908 (2013).
19. Mihovilovic, M. et al. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**, e234 (2007).
111–114 (2015).

41. Lee, A. S. Y., Kranzusch, P. J. & Cate, J. H. D. eIF3 targets cell-proliferation
47. Galan-Caridad, J. M. et al. Zfx controls the self-renewal of embryonic and
42. Polleux, F., Morrow, T. & Ghosh, A. Semaphorin 3A is a chemoattractant for
40. Panganiban, G. & Rubenstein, J. L. R. Developmental functions of the
39. Buxton, P., Edwards, C., Archer, C. W. & Francis-West, P. Growth/
38. Wu, H., Li, J., Xu, D., Zhang, Q. & Cui, T. Growth differentiation factor 5
36. Osório, C. et al. Growth differentiation factor 5 is a key physiological
35. Pei, Y.-F. et al. Hypermethylation of the CHRDL1 promoter induces
34. Liu, T. et al. Chordin-like 1 improves osteogenesis of bone marrow
33. Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric
29. Steri, M., Idda, M. L., Whalen, M. B. & Orrù, V. Genetic variants in mRNA
28. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects
27. Jiang, L. et al. A quantitative proteome map of the human body.
24. Liz, J. et al. Regulation of pri-miRNA processing by a long noncoding RNA
23. Calin, G. A. et al. Ultraconserved regions encoding ncRNAs are altered in
269–283.e19 (2020).
26. Elatmani, H. et al. The RNA-binding protein Unr prevents mouse embryonic
25. Xue, S. et al. RNA regulons in Hox 5' UTRs confer ribosome specificity to
gene regulation. Nature 517, 33–38 (2015).
24. Calin, G. A. et al. Ultraconserved regions encoding ncRNAs are altered in
human leukemias and carcinomas. Cancer Cell 12, 215–229 (2007).
23. Depew, M. J., Simpson, C. A., Morasso, M. & Rubenstein, J. L. R. Reassessing
the Dlx code: the genetic regulation of branchial arch skeletal pattern and
Distal-less/Dlx homeobox genes. Development 129, 592 (2018).
22. Thomas, J. D. et al. RNA isoform screens uncover the essentiality and
tumor-suppressor activity of ultraconserved poison exons. Nat. Genet. 52, 84–94 (2020).
21. Ni, J. Z. et al. Ultraconserved elements are associated with homeostatic
control of splicing regulators by alternative splicing and nonsense-mediated
decay. Genes Dev. 21, 708–718 (2007).
20. Larea, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive
splicing of SR genes associated with highly conserved and
ultraconserved DNA elements. Nature 446, 926–929 (2007).
19. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness
in mammalian development. Nature 554, 239–243 (2018).
18. Dickel, D. E. et al. Ultraconserved enhancers are required for normal
development. Cell 162, 491–499.e15 (2015).
17. McLean, C. & Bejerano, G. Dispensability of mammalian DNA. Genome Res. 18,
1743–1751 (2008).
16. Polleux, F., Morrow, T. & Ghosh, A. Semaphorin 3A is a chemoattractant for
hippocampal neurons during Alzheimer's disease.
15. Vazquez, K. et al. Gene- and species-specific HoX mRNA translation by
ribosomal expansion segments. Mol Cell 80, 980–995.e13 (2020).
14. Aide, B., Schepens, B. et al. A role for hnRNP C1/C2 and Unr in internal initiation of
translational reprogramming and human diseases.
13. El-Maghrabi, R. et al. RNA-mediated translational repression and human diseases.
12. Bosisio, D. et al. NF-kappaB transcription factor controls the establishment of sympathetic innervation.
11. Johansson, P. et al. Noncoding RNA in dissecting the genetic basis of human developmental disorders.
10. Wurth, L. et al. UNR/CSDE1 drives a post-transcriptional program to
control of Cbfa1/Runx2 expression. J. Cell. Biochem. 88, 493–503 (2003).
9. Jang, G. M. et al. Structurally distinct elements mediate internal ribosome entry
within the 5′-noncoding region of a voltage-gated potassium channel
premRNA. EMBO J. 27, 207–214 (2008).
8. Schepens, B. et al. A role for hnRNP C1/C2 and Unr in internal initiation of
translational reprogramming and human diseases.
7. Guo, H. et al. Disruptive variants of CSDE1 associate with autism and
interfere with neuronal development and synaptic transmission. Sci. Adv. 5, eaax2166 (2019).
6. Bosisio, D. et al. RNA-mediated translational repression and human diseases.
5. Moore, K. S. et al. Csd1 binds transcripts involved in protein homeostasis and
controls their expression in an erythroid cell line. Sci. Rep. 8, 2628 (2018).
4. Vazquez, K. et al. Gene- and species-specific HoX mRNA translation by
ribosomal expansion segments. Mol Cell 80, 980–995.e13 (2020).
3. Wurth, L. et al. UNR/CSDE1 drives a post-transcriptional program to
control of Cbfa1/Runx2 expression. J. Cell. Biochem. 88, 493–503 (2003).
2. Schepens, B. et al. A role for hnRNP C1/C2 and Unr in internal initiation of
translational reprogramming and human diseases.
80. Saltel, F. et al. Unr defines a novel class of nucleoplasmic reticulum involved in mRNA translation. *J. Cell Sci.* **130**, 1796–1808 (2017).
81. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
82. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.* **3**, 954–962 (2011).
83. Cordero, P. & Das, R. Rich RNA structure landscapes revealed by mutate-and-map analysis. *PLoS Comput. Biol.* **11**, e1004473 (2015).
84. Bhaskaran, H. & Russell, R. Kinetic redistribution of native and misfolded RNAs by a DEAD-box chaperone. *Nature* **449**, 1014–1018 (2007).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods

Data sources. See Supplementary Notes for the publicly available data used in this study.

h5'UTR definition. Sixty-way vertebrate PhastCons elements were downloaded from the UCSC mouse genome database, and elements with LOD ≥ 500 were subsetted. For each mouse (1612) transcript record, the total number of 5' UTRs, CDS, and 3' UTR nucleotides overlapping LOD ≥ 500 elements were calculated (Supplementary Table 1). The 5' UTRs with ≥ 250 nucleotide overlap were labeled hyperconserved. See Supplementary Notes for the full description.

Transcriptome–proteome correlations. Cross-tissue tandem mass tag mass spectrometry data and matching RNA-seq data were obtained from GTEx quantitative proteomics analysis of 32 human tissues from 14 individuals. Pearson’s correlation coefficient was calculated between per-tissue medians of RNA expression and per-tissue medians of protein expression. See Supplementary Notes for the full description of data-processing steps.

Term enrichment analysis. GO term enrichment analysis was performed using topGO (v2.38.11). GO-term gene mappings were obtained from the Bioconductor annotation package org.Mm.eg.db. Mammalian phenotype ontology term enrichment analysis was performed using MouseMine©. See Supplementary Notes for the full description.

CRISPR knockouts. sgRNAs were designed using CRISPOR©. The sgRNA sequences were synthesized as single-stranded (ss) DNA oligonucleotides and were cloned into the BbsI-digested expression plasmid bearing both sgRNA scaffold backbone and Cas9 nuclease, pX330-U6-Chimeric-CCB-BbsI-HspCas9. For HCE knockout in mESCs (Chd1l, Ddx1, Sema3a and Zfx), ~0.5 x 10^5 cells were plated on a 6-well plate. For HCE knockouts in 3T3Cs (Gdf5), the transfection and selection were performed using the same methods, but the cells were plated at limiting dilution of 0.5 cells per well into a 96-well plate for expansion and split for genotyping. Cells in the genotyping plate were lysed by removing the media, adding DMSO (5% of total media) and heating at 95 °C for 10 min. After cooling to room temperature, 500 μl of 500 mM Tris–HCl pH 8.0 was added to neutralize and a 1:10 dilution was taken for genotyping PCR. The genotyping PCR reaction was as follows: 1X MyTaq HS Red Mix (Meridian Bioscience, catalog no. BIO-25047), 300 nM forward primer, 300 nM reverse primer, 1 μl of 1:100 diluted crude lysate in 10 μl total reaction volume. Cycling conditions were: 95 °C, 3 min initial denaturation, followed by 30 cycles of 95 °C for 15 s, 65 °C for 15 s, 72 °C for 30 s. Clones with expected shorter amplicons were further expanded. DNA from expanded clones was isolated with Wizard Genomic DNA Purification kit (Promega, catalog np. A1120). The genotyping PCR reaction from expanded clones was as follows: 0.02 μl Kapa HiFi HotStart polymerase (Roche, catalog no. 12023871), 1X Kapa HiFi HotStart buffer, 300 nM forward primer, 300 nM reverse primer, 10 ng genomic DNA in 20 μl. Cycling conditions were: 95 °C for 3 min initial denaturation, followed by 30 cycles of 90 °C for 30 s, 65 °C for 30 s, 72 °C for 30 s. The amplicons were sequenced at QIAGEN. Cell culture. See Supplementary Notes for the description of cell culture conditions.

Mouse husbandry. All animal work was reviewed and approved by the Stanford Administrative Panel on Laboratory Animal Care (APLAC). The Stanford APLAC is accredited by the American Association for the Accreditation of Laboratory Animal Care. All mice used in the study were housed at the Research Animal Facility and at the SIM-1 Barrier Facility at Stanford University. All mice used for experiments were between two and six months old. All animal studies were performed in accordance with Stanford University Animal Care and Use guidelines.

Polysome profiling. Cells were collected 2 min after replacing media with cycloheximide (MilliporeSigma, catalog no. C6786-1G) containing media at 100 μg/ml. Approximately 10 x 10^6 cells were resuspended in 400 μl of the following lysis buffer on ice for 30 min, vortexing every 10 min: 25 mM Tris–HCl pH 7.5, 150 mM NaCl, 15 mM MgCl₂, 1 mM DTT, 8% glycerol, 1% Triton X-100, 100 μg/ml cycloheximide, 0.2 μl ml⁻¹ Superase-In RNase inhibitor (ThermoFisher Scientific, catalog no. AM2694), 1X Halt protease inhibitor cocktail (ThermoFisher Scientific, catalog no. 78430), 0.02 μl ml⁻¹ TURBO DNase (ThermoFisher Scientific, catalog no. AM2238). Nuclei were removed by two-step centrifuging, first at 1,300 g for 5 min and then at 10,000 g for 5 min, taking the supernatants from each. A 25%–50% sucrose gradient was prepared in 13.2-ml ultracentrifuge tubes (Beckman Coulter, catalog no. 331372) using BioSpin Gradient Master with the following sucrose gradient: 25 or 50% sucrose (w/v), 25 mM Tris–HCl pH 7.5, 150 mM NaCl, 15 mM MgCl₂, 1 mM DTT, 100 μg/ml cycloheximide. The lysate was layered onto the sucrose gradient and ultracentrifuged on a Beckman Coulter SW-41T rotor at 40,000 r.p.m. for 150 min at 4 °C. The gradient was density fractionated using Brandel BR-188 into 16x 750-μl fractions. In vitro transcribed spike-in luciferase RNA (50 μg) was added to each fraction. A 700-μl portion of each fraction was mixed with 100 μl of 10% SDS, 200 μl of 1.5 mM sodium acetate and 900 μl of acid phenol–chloroform pH 4.5 (ThermoFisher Scientific, catalog no. AM9720), heated at 65°C for 5 min and centrifuged at 20,000 g for 15 min at 4 °C for phase separation. A 600-μl aqueous phase was mixed with 600 μl of 100% ethanol and RNA was purified on silica columns (Zymo, catalog no. R1013). For each fraction, up to 8 μg RNA was DNAseI-treated at 37°C for 30 min using 0.2 U ml⁻¹ TURBO DNase with 1 μl ml⁻¹ Superase-In in 30 μl and purified again on a silica column. RNA (100 ng) was reverse transcribed using iScript reverse transcription (Bio-Rad, catalog no. 1708890) in 10-μl reactions. qPCR was performed using SsoAdvanced Universal SYBR Green Supermix (Bio-Rad, catalog no. 1725270) with 2 μl of 1:4 diluted reverse transcription reaction and primer pairs targeting the HCE knockout h5UTR genes or the spike-in (see Supplementary Table 9 for primer sequences).

Ct values were normalized to Ct values of spike-in luciferase and plotted as proportions across the 16 fractions. The f-statistic and P value were calculated for significance means between different genotypes. The p-value compares an independent culture (p genotype), sucrose gradient fractionation and qPCR quantification. Fisher combined P value was calculated for no difference across all fractions.

Reporter constructs for luciferase assays. Bicistronic reporter gateway plasmid, pRF_D2, was constructed from prF vector, which has SV40 promoter and two reporter genes, Renilla luciferase and firefly luciferase, with multiple cloning sites in between them (pSV40). Gateway cassette A (ThermoFisher Scientific, catalog no. 11828029) was inserted in between Renilla and firefly luciferases, replacing the cloning sites using two EcoRI sites.

RNA normalizing reporter gateway plasmid, pRF_D1, was constructed from prF vector by replacing the cloning sites with HCV IRES and inserting gateway cassette A in between AvrII and EcoRV sites upstream of Renilla luciferase.

Downstream of the Renilla luciferase was the firefly luciferase and the HCV IRES between them, such that the HCV IRES-translated downstream firefly luciferase normalizes for differences in RNA levels to enable measurement of translation efficiency.

Full-length h5UTRs were synthesized and cloned into pENTR1A (ThermoFisher Scientific, catalog no. A10462) by SGI (sequences in Supplementary Table 1). Truncation variants were either synthesized or cloned by PCR from synthesized full-length sequences into pENTR1A vectors. Gateway LR Cloning System (ThermoFisher Scientific, catalog no. 11791020) was used to recombine the full-length or truncation variants into either prF_gwy or prF_D1 vectors.

Mutant Cadet 5' UTRs were cloned by Gibson assembly reaction (NEB, catalog no. E2621S) using mutation-containing ssDNA templates with homology arms and two upstream/downstream fragments (sequences in Supplementary Table 9). Full-length wild-type and mutant 5' UTRs were inserted into pG LUC (Promega, catalog no. E1751) plasmid between EcoRI and Ncol sites upstream of firefly luciferase gene. The in vitro transcription template was amplified with T7 promoter sequence containing primer and in vitro transcribed using T7 RNA polymerase (NEB, catalog no. E2040S). The in vitro transcription RNAs were capped using Vaccinia virus capping enzyme (Cellscript, catalog no. C-SCCE0625) and polyA tailed using polyA polymerase (Cellscript, catalog no. P-CPAP104H).

Reporter transfection for luciferase assays. For DNA transfections, 200 ng of plasmid DNA was transfected to cells plated on 96-well plates. For 1OT12 cells, mESCs, NSCs, Irbm mesenchyme culture and embryoid bodies, 0.5 μl of Lipofectamine 2000 (ThermoFisher Scientific, catalog no. 11668030) was used per well. Foot Purus, 0.2 μl of ViVecf (Promega, catalog no. E9491) was used per well. Cells were incubated for 4 h with transfection reagent, DNA in OptiMEM media (ThermoFisher Scientific, catalog no. 31985062). Cells were then washed with PBS, and the medium was changed back to regular growth media.

For RNA transfection, 200 ng of firefly luciferase RNA and 10 ng of Renilla luciferase RNA was transfected to cells plated on 96-well plates. Lipofectamine 2000 (0.5 μl) was used per well.

Luciferase assays. For DNA transfections, cells were lysed using Passive Lysis Buffer (Promega, catalog no. E1941) for 30 min at room temperature, 48 hours after transfection. For RNA transfections, cells were lysed 6 hours after transfection. Firefly and Renilla luciferase values were read using Dual-Glo Luciferase Assay System (Promega, catalog no. E2920) for >96 samples or Dual-Luciferase Reporter Assay System (Promega, catalog no. E9110) for fewer samples, on a Promega GloMax-Multi plate reader. In all experiments, log ratios of the two luciferase activities were taken for each well.
For bidirectional reporter assays across multiple cell types, the data were quantile normalized across all replicate samples. The normalized rMxEM function from R package mixtools (v=1.2.0) was used for mixture modeling of maximum replicates-average value estimates. False discovery rate versus the whole amplicon. False discovery rate was estimated by the Benjamini–Hochberg procedure. See Supplementary Notes for the full description.

In-cell DMS probing following ATP depletion and multiplexed mutational profiling. One hundred h5UTRs were chosen for amplicon sequencing on the basis of coverage profiles from ENCODE E14 mESC RNA-seq data. See Supplementary Notes for the description of amplicon sequencing primer design and pooling. For ATP depletion, 10 × 10^6 mESCs were incubated for 10 min in ATP depletion media: DMEM without glucose (ThermoFisher Scientific, catalog no. A1443001), 10 mM 2-deoxy-α-glucose (2DG, MilliporeSigma, catalog no. 25972), 3.5 ml of mESC media all containing 10 mM 2DG and 10 mM NaNO_3. Bicine (1 mM; MilliporeSigma, catalog no. B3876) titrated to pH 8.5 at 25 °C was added to resuspended cells (200 mM final bicine concentration). A portion of 500 μl of 16% dimethyl sulfate (MilliporeSigma, catalog no. D186309) in ethanol was added to pH 8.5 final concentration. Cells were incubated for 30 min at 37 °C. Ice-cold 30% BME (2.5 ml; MilliporeSigma, catalog no. M3148) in ethanol was added to quench the reaction. This DMS modification protocol was adapted from protocol and data reported in ref. 70). Following centrifugation to remove the supernatant, the cells were lysed in Trizol (ThermoFisher Scientific, catalog no. AM2696) in 60 μl and purified again on a silica column. DNA (1 μg) was mixed with 1 μl of 1 M 96× primer pool (96x1× amplicon reverse primers for total of 1 M oligonucleotides) and denatured in 6.25 μl total volume (with H2O) at 65 °C for 2 min, then chilled to 4 °C. Reverse transcription reaction conditions were as follows: 20 mM Tris–HCl pH 7.5, 7.5 mM KCl, 10 mM MgCl2, 5 mM DTT, 500 nM TGIRT (InGex, catalog no. TGIRT50), 1 U μl−1 Superscript-I; 9 μl total reaction volume. The reverse transcription (RT) reaction was preincubated at 25 °C for 30 min, then initiated with addition of 12.5 μl of 2×5 NTP extended trypsinization at 37 °C for 10 min. The reaction was heated at 95 °C for 2 min. Then 1 μl of 2.5 M HCl and 1 μl of 500 mM Tris–HCl pH 7.5 was added to neutralize. SPRINT select beads (29 μl; Beckman Coulter, catalog no. B23318) were used for purification of cDNA; the elution volume was 6 μl. Pooled reactions (4 × 96) for a total of 384 targets were performed for each sample. To each set of 96 pooled cDNA, multiplex PCR was performed with 5 μl of cDNA, 0.2 mM dNTP, 2.5 μM forward primer pool (96x1× 2 μM each primer for a total of 2 M oligonucleotides), 2 μM reverse primer pool, 1× SYBR Green I (ThermoFisher Scientific, catalog no. S7563), 0.02 U μl−1 Q5 HotStart DNA polymerase (NEB, catalog no. M0493S), 1× Q5 HotStart Reaction Buffer, 1× Q5 HotStart High GC Enhancer, in a total reaction volume of 30 μl. Cycling conditions were: 94°C for 30 s initial denaturation, followed by 15–25 cycles (terminated before plateau) of 98°C for 10s, 56°C for 40s, 76°C for 10s; PCR was run for 15–25 cycles. Each 96-pool multiplex PCR reaction was then used in a master mix of second PCR with 96 individual primer pair reactions: 0.2 mM dNTP, 500 nM forward primer, 500 nM reverse primer, 1× SYBR Green I, Q5 HotStart DNA polymerase, 1× Q5 HotStart Reaction Buffer, 1× Q5 HotStart High GC Enhancer, in a total reaction volume of 6 μl. Cycling conditions were: 94°C for 30 s initial denaturation, followed by 20 cycles of 98°C for 10 s, 62°C for 10 s, 76°C for 10 s. The 384 individual reactions were pooled and purified on silica columns (NEB, catalog no. T10305). The amplicon pool was end-prepared, Illumina adapter sequences were ligated, adapter-ligated DNA was size selected with SPRINT select beads (370 bp and three-cycle barcoding PCR were performed on Illumina HiSeq 4000 at Novogene. The 0.97×2 method from the TMM-normalized mutation count matrix and [rMxEM] for per-window accessibility pattern differences, we calculated the Anderson–Darling statistic in sliding 11-nucleotide windows between per-nucleotide rMxEM values of the window versus the whole amplicon. False discovery rate versus the whole amplicon.
Two-dimensional accessibility data analysis and structure models. Briefly, the normalized covariation matrix was clustered using multidimensional scaling, \( K = 2 \). Cluster average accessibility z-scores were used to constrain partition function calculation in Vienna RNA (v.2.4.14)\(^9\). A total of 250 structures were sampled for each cluster and used as input suboptimals to REEFFIT (v.0.6.3)\(^92\). For visualization of the landscape, we used pairwise distance metrics, structure clustering and medoid assignment produced by REEFFIT, and sum of weights for structures belonging to each of the three clusters represented by a medoid structure are presented in the main figure. Bootstrapping was used to estimate population fraction errors. See Supplementary Notes for the full description.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

Raw sequencing data (related to Figs. 4, 5 and 6) are deposited to GEO with accession code GSE155656. Processed reactivity data have been deposited in the RNA Mapping Database (RMDB) with accession codes CSDE1_DMS_0000 and CSDE1_DMS_0001. Sources for publicly available data are described in the Methods.

**Code availability**

All software used to analyze the study data are listed in the Methods and in the Nature Research Reporting Summary and are publicly available. All codes used to analyze icM2 data are available through a Github repository: github.com/barnalab/icm2p.

**References**

85. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
86. Motenko, H., Neuhauser, S. B., O’Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325–330 (2015).
87. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
88. Yoon, A. et al. Impaired control of IRES-mediated translation in X-linked dyskeratosis congenita. *Science* **312**, 902–906 (2006).
89. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
90. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
91. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
92. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Hyperconserved 5'UTRs in vertebrate genomes. **a**, Left: heatmap of the positions of LOD ≥ 500 PhastCons elements in each hSUTR. Middle: heatmap of the relative positions (calculated in 100 bins across the hSUTRs) of the elements. Right: plot of average element overlap across the 100 bins to illustrate the positional preference. **b**, Histogram of the length of hSUTRs. Average length is 674nt. **c**, Histogram of the number of nucleotides overlap between LOD ≥ 500 PhastCons elements and hSUTRs. Average overlap is 389nt. **d**, Distributions of cross-tissue transcriptome-proteome correlations for all genes, genes with hSUTRs, or genes with variance-matched non-conserved 5'UTRs. Indicated p-values are from two-sided Wilcoxon rank sum tests for cross-tissue correlation values between hSUTR genes and all genes or between hSUTR genes and variance-matched non-conserved controls. **e**, Distributions of the number of annotated alternative 5'UTRs for all genes, genes with hSUTRs, or genes with size-matched non-conserved 5'UTRs. Indicated p-values are from two-sided Wilcoxon rank sum tests for the number of alternative 5'UTRs between hSUTR genes and all genes or between hSUTR genes and size-matched non-conserved controls. **f**, Scatter plot illustrating the lack of significant term enrichments for a size-matched set of non-conserved 5'UTRs. X-axis and y-axis plots expected and the observed number of genes for each term. Blue dashed line indicates the minimum observed/expected ratio cutoff of 3. Green line indicates expected and observed counts where Fisher's test p-value (p_f) is estimated to have FDR = 0.05. Neighbor-weighted test p-value (p_{fw}) ≤ 0.05 is further used as an additional cutoff. The final set of enriched terms passing filter is colored by p_f and sized by p_{fw}. 
Extended Data Fig. 2 | Non-canonical translation activation by hyperconserved 5’UTRs across cell types. **a**, Density plots of non-canonical translation initiation activities from h5UTRs by bicistronic reporter assay. X-axis is the luciferase reporter activity ratios. Jittered dots mark individual reporter ratios for each h5UTR in each cell type. **b**, Summarized plot of ribosome load (sum of % mRNA times the ribosome number for each fraction) differential ratio between NSCs and ESCs calculated from polysome profiles for each gene shown in Extended Data Fig. 2c–l. Red indicates significant increase in NSCs and black indicates significant decrease (two-sided t-test p ≤ 0.05, n = 3, marked by asterisk). **c–l**, Endogenous polysome profiles of NSCs versus ESCs for genes with h5UTRs that show high non-canonical translation reporter activities in NSCs compared to ESCs. Distribution of mRNAs across sucrose gradient fractions are plotted. Y-axis plots the mean percent mRNA. Error bars indicate standard error. Asterisk indicates two-sided t-test p ≤ 0.05 for each fraction between the two cell types. n = 3 for each cell type. Indicated p-value (p_f) is calculated by Fisher’s method across all fractions. Note that Extended Data Fig. 2c shows the profile of 18S rRNA, which indicates lower global translation in NSCs compared to ESCs.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Non-canonical activation by hyperconserved 5’UTRs substantially contributes to translation. a, Scatter plot of luciferase activity versus RNA level ratios (mean from n = 3) observed for the bicistronic reporters of 90 h5UTRs measured in 10T1/2 cells. Dashed line marks the 10% FDR used in Fig. 3a. Spearman correlation indicated on top left. b, The effect of various truncations of the h5UTRs on non-canonical initiation and total translation efficiency (also see Fig. 3d). Left: positions of truncations. Dashed lines indicate truncations. Purple horizontal lines indicate uORFs; yellow and red lines indicate in-frame and out-of-frame uAUGs, respectively. Middle: non-canonical initiation efficiency. Right: total translation efficiency. X-axis indicates the mean of luciferase reporter ratios relative to the wild-type. Error bars indicate standard error. Dashed line marks the wild-type 5’UTR activity. Asterisk indicates two-sided t-test p ≤ 0.05 for each truncation versus the full-length. The numbers to the left of the bars indicate n and p-values.

c, Comparison of translational activities between the full-length long, non-conserved 5’UTRs versus the only first 300nt truncation. 11 different pairs are tested. X-axis indicates the mean log2 luciferase reporter ratios of each truncation relative to its full-length wild-type. Error bars indicate standard error. Bars colored in red indicate significantly reduced translation in the shorter, truncated 300nt fragment; black indicates significant increase (two-sided t-test, paired n = 3, p ≤ 0.05, marked by asterisk). The numbers to the left of the bars indicate p-values. d, Violin plot of full-length/truncated reporter activity ratios (log2) from hyperconserved and non-conserved 5’UTRs. p indicates two-sided Wilcoxon rank sum test p-value. Box hinges: 25% quantile, median, 75% quantile, respectively from left to right. Whiskers: lower or upper hinge ±1.5*IQR. e, Scatter plot of change in translation efficiency between full-length and truncated h5UTRs shown in Fig. 3e versus change in uAUG density (change in number of AUGs / change in length between each pair of full-length and truncated h5UTRs). r indicates pearson’s correlation coefficient and p indicates two-tailed p-value.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Cellular remodeling of hyperconserved 5'UTR RNA structures. a, Stacked bar plots showing proportions of significant (FDR ≤ 0.05) or not significant windows that overlap uAUG in black versus that do not overlap uAUG in red. OR indicates odds ratio for overlaps uAUG / does not overlap uAUG, and p indicates Fisher’s test p-value (one-sided, H$_0$ = odds ratio>0). b, Stacked bar plots showing proportions of significant (FDR ≤ 0.05) or not significant windows that overlap uORF in black versus that do not overlap uORF in red. OR indicates odds ratio for overlaps uORF / does not overlap uORF, and p indicates Fisher’s test p-value (one-sided, H$_0$ = odds ratio>0). c, Zoomed-in view of differential accessibilities along h5UTRs with one or more significantly different windows under ATP depletion. Top plot shows -log10 p-value for each window. Highlighted boxes mark significantly different windows, above the dashed line indicating 5% FDR. Middle plot shows differential accessibility on the y-axis, where greater than zero indicates increased accessibility upon ATP depletion and less than zero indicates decreased accessibility. Bottom plot shows differential accessibility for in vitro refolded RNA. Error bars in each plot show standard error, n = 3. The three profiled regions shown on the left side exhibit discordant profiles between accessibility changes observed in cells following ATP depletion and accessibility changes observed for in cell versus in vitro refolded RNA. The other three on the right side exhibit concordant profiles.
Extended Data Fig. 5 | icM² reveals structured elements in the hyperconserved Csd1 5’UTR. a, Boxplot of average PhastCons scores in significant windows of ATP-dependent remodeling versus all windows shown in Fig. 4c. p indicates two-sided Wilcoxon rank sum test p-value. b, Same as Extended Data Fig. 5a, but showing the distribution of average PhyloP scores.
Extended Data Fig. 6 | In-vitro M² analysis of Csde1 5'UTR. a, Heatmap of in-vitro M² accessibility matrix for Csde1 5'UTR from position 190 to 386. For each row, the chemical mapping profile of a single-nucleotide variant of the RNA is plotted across the columns, where the colors indicate z-scaled accessibility change values from the wild-type RNA. 1D data from each mutant are vertically stacked to display a 2D matrix. White boxes mark perturbation signals that support the model shown in Extended Data Fig. 6b; color bars at the bottom indicate the nucleotide positions of the stems that match the same color in the model. b, The model for the in-vitro structure of Csde1 5'UTR from position 190 to 386. Also see Extended Data Fig. 6a.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

☐ n/a  Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of all covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

☐ Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

ThermoAlign 1.0.0; PrimerPooler 1.41; cutadapt 1.18; BBMerge 38.22; bowtie2 2.3.4.3; samtools 1.9; UMI-tools 0.5.4; shapemapper 2.1.5

Data processing using these softwares is detailed in Online Methods section, and the pipeline scripts are available on github.com/barnalab/icom2p

Data analysis

ViennaRNA 2.4.14; REEFIT 0.6.3; R 3.6.2 with packages tidyverse 1.3.0, limma 3.42.2, edgeR 3.28.1, Biostrings 2.54.0, viridis 0.5.1, data.table 1.12.8, KSamples 1.2.9, ggrepel 0.8.2, impute 1.60.0, mixtools 1.2.0, topGO 2.38.1, ggrridges 0.5.2, GOSemSim 2.12.1, zoo 1.8-7, igraph 1.2.5, cowplot 1.0.0

Analysis methods and parameters are described in Online Methods. The analysis codes used are available in full at github.com/barnalab/icom2p

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data (related to Figs. 4, 5 and 6) are deposited to GEO with accession code GSE155656. The following lists the version and sources of publicly available data used in this study. RefSeq: release 84, https://ftp.ncbi.nlm.nih.gov/refseq/; 60-way PhastCons: UCSC mm10, http://hgdownload.soe.ucsc.edu/
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [ ] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | Samples analyzed were chosen based on the expected potential effect sizes and based on the variability typically seen for each types of experiments performed as previously observed in the literature or practically experienced by the field. |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | Low coverage amplicons are excluded from analysis. While the quantitative criteria for exclusion were not pre-established prior to the study, dropouts in amplicon sequencing are both expected and obviously distinguished by analyzing its coverage and per-amplicon replicate correlations. Final quality control criteria for the amplicon sequencing data are empirically determined and detailed in the Online Methods section. |
| Replication | Each replicate comprises an independent cell culture, sample collection and quantitative analysis per genotype or condition. The exact number of replications vary for different experiments but are always clearly indicated in each figure panel or legend. |
| Randomization | Randomization is not relevant because genotypes or conditions were constructed and there was no subjective allocation of samples to experimental groups. |
| Blinding | Blinding was not considered in this study as the experiments are not based on subjective measurements. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a | n/a |
| [x] Involved in the study | [x] Involved in the study |
| [x] Antibodies | [x] ChiP-seq |
| [x] Eukaryotic cell lines | [x] Flow cytometry |
| [x] Palaeontology and archaeology | [x] MRI-based neuroimaging |
| [x] Animals and other organisms | |
| [x] Human research participants | |
| [x] Clinical data | |
| [x] Dual use research of concern | |

Eukaryotic cell lines

Policy information about cell lines

| Cell line source(s) | F14Tg2a.4; C3H/10T1/2, Clone 8 (ATCC CCL-226); NIH/3T3 (ATCC CRL-1658) |
| Authentication | Cell lines are not authenticated. |
| Mycoplasma contamination | Cell lines were not tested for mycoplasma. |
| Commonly misidentified lines (See ICTAC register) | No commonly misidentified cell lines were used in this study. |
## Animals and other organisms

Policy information about [studies involving animals](#), **ARRIVE guidelines** recommended for reporting animal research.

| Category               | Information                                                      |
|------------------------|------------------------------------------------------------------|
| Laboratory animals     | Species: *Mus musculus*; Strain: C57Bl/6J; Sex: female; Age: 2-6 months |
| Wild animals           | This study did not involve wild animals.                          |
| Field-collected samples| This study did not involve samples collected from the field.     |
| Ethics oversight       | All animal work was reviewed and approved by the Stanford Administrative Panel on Laboratory Animal Care (APLAC). |

Note that full information on the approval of the study protocol must also be provided in the manuscript.