

Understanding species-specific and conserved RNA-protein interactions *in vivo* and *in vitro*

Sarah E. Harris^{1,2}, Maria S. Alexis^{3,4}, Gilbert Giri^{2,5}, Francisco F. Cavazos², Jernej Murn^{6,7}, Maria M. Aleman², Christopher B. Burge³, and Daniel Dominguez^{1,2,5,8,†}

¹Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC

²Department of Pharmacology, University of North Carolina, Chapel Hill, NC

³Department of Biology, Massachusetts Institute of Technology, Cambridge, MA

⁴Current address: Remix Therapeutics, Cambridge, MA

⁵Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC

⁶Department of Biochemistry, University of California, Riverside, CA

⁷Center for RNA Biology and Medicine, Riverside, CA

⁸RNA Discovery Center, University of North Carolina, Chapel Hill, NC

†corresponding author: Daniel Dominguez (didoming@email.unc.edu)

ABSTRACT

While evolution is often considered from a DNA- and protein-centric view, RNA-based regulation can also impact gene expression and protein sequences. Here we examined interspecies differences in RNA-protein interactions using the conserved neuronal RNA binding protein, Unkempt (UNK) as model. We find that roughly half of mRNAs bound in human are also bound in mouse. Unexpectedly, even when transcript-level binding was conserved across species differential motif usage was prevalent. To understand the biochemical basis of UNK-RNA interactions, we reconstituted the human and mouse UNK-RNA interactomes using a high-throughput biochemical assay. We uncover detailed features driving binding, show that *in vivo* patterns are captured *in vitro*, find that highly conserved sites are the strongest bound, and associate binding strength with downstream regulation. Furthermore, subtle sequence differences surrounding motifs are key determinants of species-specific binding. We highlight the complex features driving protein-RNA interactions and how these evolve to confer species-specific regulation.

Introduction

Species divergence and adaptation rely on a delicate balance of robustness — the ability to withstand mutations without serious deleterious effects on fitness — and evolvability — the susceptibility to developing a novel phenotype^{1,2}. Driving this balance are changes in gene expression programs and coding sequences²⁻⁵. Understanding how changes in *trans* (nucleic acid-binding proteins) and *cis* (nucleic acid sequences) impact gene regulation across species remains an important challenge. While changes in *cis* and *trans* across species are both important, *cis*-regulatory elements change more rapidly than *trans* factor amino acids or binding preferences^{6,7}. For example, transcription factors (TFs) and RNA-binding proteins (RBPs) are highly conserved over long evolutionary distances while regions harboring *cis*-regulatory elements that these proteins bind can vary drastically over the same distances^{8,9}.

Most previous studies have taken a DNA- and protein-centric view (reviewed by Villar *et al.*¹⁰ and Mitsis *et al.*¹¹); however, RNA regulation influences both expression levels as well as protein coding sequences, resulting in potential widespread effects¹²⁻¹⁴. RBPs constitute a large class of pan-essential regulatory factors^{15,16} that drive RNA regulation, contributing significantly to transcription, splicing, and trans-

lation to influence the expression and identity of proteins produced¹⁷⁻²⁰. These processes are dictated by the strength of the interaction between RBPs and their RNA targets²¹⁻²⁵. In the simplest model of RNA regulation, RBPs bind short sequence motifs (3-8 nucleotides) within RNA to influence its regulation²⁶. However, these interactions are complex as change, loss, or gain of a single nucleotide within or surrounding a motif can greatly impact binding²⁷⁻²⁹.

RBPs themselves have a striking level of amino acid conservation with many RNA-binding domains (RBDs) remaining nearly identical, even over hundreds of millions of years³⁰. Generally, RBPs tend to be more conserved than their DNA-binding counterparts, transcription factors⁹. Paradoxically, RNA processing events regulated by RBPs, such as alternative splicing and translation, have been found to be more species-specific and to evolve more rapidly than gene expression programs (*i.e.*, tissue-specific expression across species)³¹⁻³⁴.

How often are regulatory elements that control gene expression and RNA processing conserved across species? If binding has changed, what are the mechanisms underlying that change? Previous studies on TF binding to regulatory elements have addressed this in a number of species^{6,7,35-38} (and reviewed by Villar *et al.*¹⁰). Multiple studies — in-

cluding one employing chromatin immunoprecipitation and sequencing (ChIP-Seq) across five vertebrates — have found that although TFs are highly conserved, *cis*-regulatory elements evolve rapidly and primarily dictate TF binding profiles⁶. More specifically, TF binding profiles (*i.e.*, bound genes) demonstrate less than 40% conservation between human and mouse, even though the individual TFs studied are nearly identical (>95% identity) at the amino acid level and have identical or near-identical binding preferences⁷. When a human chromosome is placed in a mouse context, TF binding predominantly follows the human binding patterns rather than that of the mouse³⁵, indicating that binding pattern changes are primarily *cis*-directed. Of course, these forms of interactome evolution are partially dependent on evolutionary time and the individual TFs being assessed³⁹.

Few similar studies of species-specific RNA-protein interactions have been conducted. But some emerging themes parallel the similarities to that of TF-DNA interactions. For example, previous work has examined the conservation of the Pumilio and FBF (Puf) superfamily of proteins and their interactomes^{40–43}. Puf3 exhibits highly similar RNA binding specificities across fungal species⁴²; however, Puf3 targets change significantly between *S. cerevisiae* and *N. crassa*⁴³. More strikingly, targets bound by Puf3 in one species are bound by a different RBP — Puf4/5 — in another⁴³, highlighting the complex nature of RNA binding site evolution and the interplay between *cis* and *trans*. Within species, single nucleotide polymorphisms (SNPs) have been shown to impact RBP-RNA interactions. A comprehensive analysis of RBP-RNA interaction studies in two cell types identified over a thousand cases of allele-specific RBP-RNA interactions, some of which were validated biochemically and had functional impact on RNA regulation⁴⁴. The complex paths in which RNA regulation evolves have been previously reviewed⁴⁵, but much work is still needed to understand how the underlying driving forces, namely RBP-RNA interactions, drive changes in regulation.

To understand species-specific RNA binding, we used available individual-nucleotide resolution crosslinking and immunoprecipitation (iCLIP) data²⁷ from a neuronal RBP, unkempt (UNK), in human and mouse. UNK regulates neuronal morphology, is a negative regulator of translation, and associates with polysomes^{27,46,47}. We identified species-specific and shared UNK binding sites and found that 45% of UNK transcript binding was conserved across species. Importantly, while binding of transcripts was conserved, the individual motifs that were bound were far less conserved, often switching between species. We reconstituted the *in vivo* UNK-RNA interactomes of human and mouse *in vitro* to understand the driving forces underlying species-specific binding and regulation. We found that while motif turnover is an important mediator of species-specific binding, contextual sequence and structural features in which motifs are embedded are of comparable importance and contribute to binding site turnover. We extended our studies across 100 vertebrates to understand how

sequence changes over longer time scales affect binding and found striking correlations between evolutionary distances, individual binding site conservation, and strength of UNK binding. This work deepens our understanding of *cis*-regulatory evolution and highlights the complex nature of evolving RNA binding sites.

Results

Conserved and Species-Specific *in vivo* Binding Patterns

We undertook an RNA-centric view and sought to determine how RNA binding sites change or are conserved across species. We focused on the conserved neuronal RBP, unkempt (UNK) for the following reasons: i) UNK has a well-defined RNA-binding motif supported by structural studies⁴⁶; ii) UNK is 95% conserved between human and mouse with only one amino acid difference within the RNA-binding zinc finger domains^{48–50}; iii) Murn and coworkers demonstrated that even the sea sponge (*Amphimedon queenslandica*) UNK paralog functionally rescues knockdown of UNK in human cell lines⁴⁶ even though these species only share 53% similarity at the protein level and 80% similarity within the RBDs^{48–50}. Thus, this level of functional conservation provides an opportunity to study changes in UNK binding sites across species primarily driven by changes in RNA sequence rather than in the protein's binding properties.

We used UNK iCLIP data in human and mouse neuronal cells and tissue, respectively²⁷, to identify species-specific and conserved UNK binding sites (**Fig. S1A**). Comparing binding sites across species at the transcript level, we observe that 45% of transcripts are bound in both species (**Fig. 1A**-Venn diagram; $p=6e-94$, hypergeometric test). As iCLIP allows for individual nucleotide level determination of binding sites, we further investigated where on each transcript UNK was bound. UNK binding sites require a UAG core motif, which has been identified both *in vitro* and *in vivo*^{26,27}. In instances where transcript level binding was conserved between human and mouse, roughly half of binding was observed at aligned (homologous) motifs across species. In cases where binding sites within transcripts changed across species, motif loss only accounted for a minority of these changes. That is, in many cases both human and mouse preserved a UAG motif in the same location, yet binding was often identified elsewhere on the transcript (**Fig. 1A**-top pie chart). Likewise, when comparing motif differences in transcripts only bound in human or mouse, motifs were preserved across species in over 70% of orthologous regions (**Fig. 1A**-bottom pie charts) yet binding was differential. Thus, while UNK protein is highly conserved, engagement of UNK with specific UAG motifs often varies across species.

Understanding the UNK-RNA Interactome *in vitro* at Massive Scale

While iCLIP is a powerful technique that allows for the derivation of nucleotide-level binding sites, several experimental

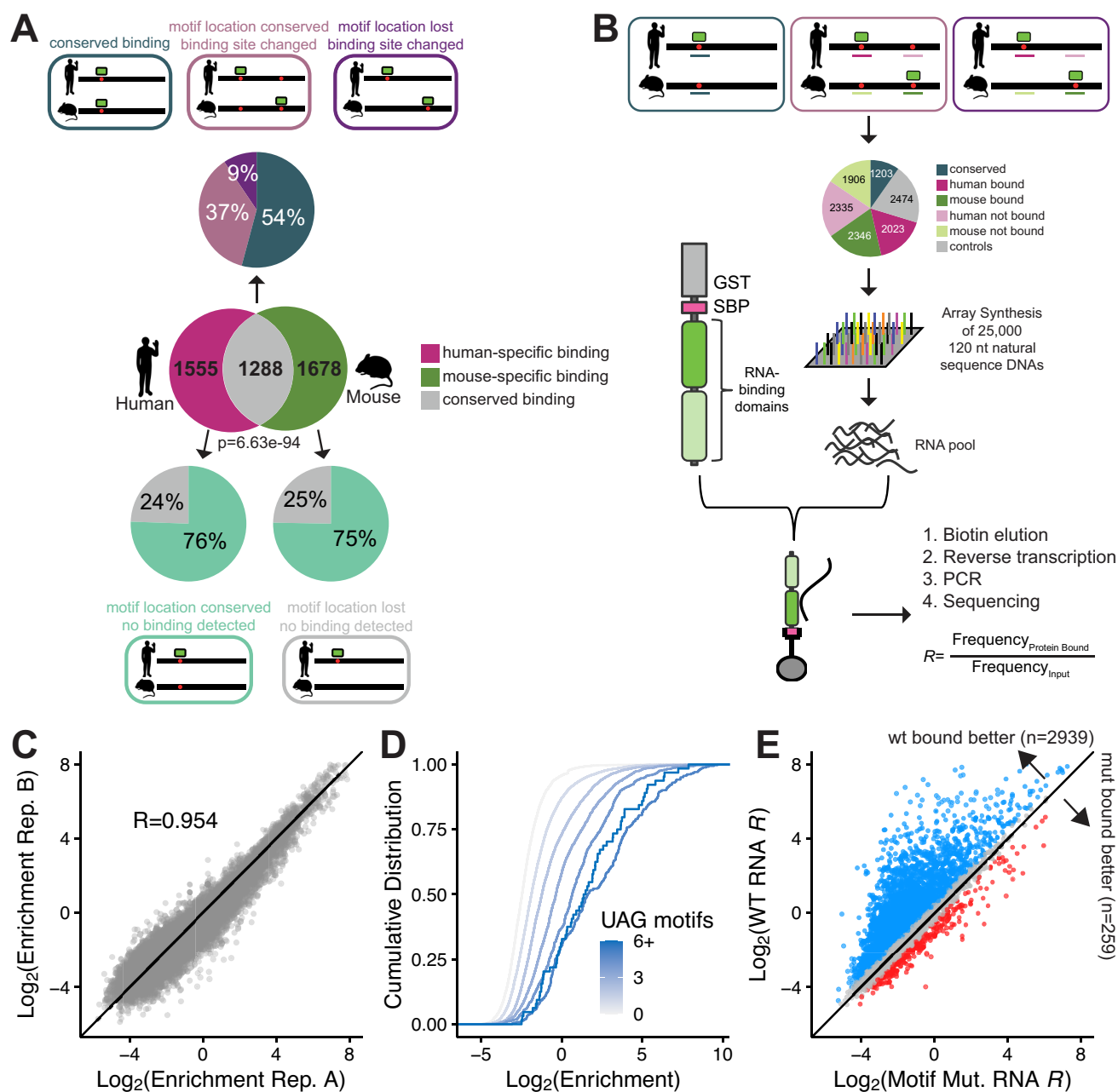


Figure 1. Design and validation of natural sequence RNA bind-n-seq (nsRBNS). A) (Venn diagram) Transcript level conservation of iCLIP UNK hits between human neuronal cells (SH-SY5Y) and mouse brain tissue. Significance determined via hypergeometric test. (Pie charts) Motif level conservation of iCLIP UNK hits between human neuronal cells (SH-SY5Y) and mouse brain tissue. B) Design of natural sequence oligo pool and layout of nsRBNS. C) Correlation plot of two experimental UNK nsRBNS replicates. Pearson's correlation coefficient included. D) Cumulative distribution function of \log_2 enrichment of all oligos separated by UAG motif content. E) Scatter plot of \log_2 enrichment of wildtype (Y-axis) versus motif mutant (X-axis) oligos. \log_2 change in enrichment (wt-mut) was calculated for each sequence pair: > 0.5 defined as bound better in wt (blue), < -0.5 defined as bound better in mut (red), 0 ± 0.5 defined as similar binding (grey).

factors, including RNA crosslinking efficiencies and biases across RBPs, cell types, and tissues, complicate its interpretation. Another important consideration for understanding binding site conservation across species via iCLIP is that false

negative rates of CLIP experiments are largely unknown⁵¹. Finally, the strength of RBP-RNA interactions found in CLIP-based experiments have a limited dynamic range. That is, binding affinity and occupancy are not easily determined, and

binding is often interpreted as binary when a continuum of occupancy levels likely occur *in vivo*. To mitigate CLIP biases and understand binding differences across species due to the intrinsic properties of the RNA-protein interactions, we sought to reconstitute the human and mouse UNK-RNA interactomes *in vitro*.

We derived UNK binding sites from iCLIP data in one-to-one orthologous human and mouse genes (see **Methods**) and designed 12,287 natural RNA sequences, each 120 nucleotides long. Contained within this “pool” were UNK binding sites identified via iCLIP in human and mouse neuronal cells and tissue, respectively, as well as orthologous regions whether or not they displayed evidence of binding in cells (**Fig. 1B**). Sequences were designed such that UAGs identified via iCLIP were located in the center of each oligo whenever possible (**Methods**). Non-bound control regions were also selected and matched for UAG content (**Fig. 1B**). Additionally, 11,967 mutated oligos were also included and are discussed below.

An array of these natural sequence DNA oligos was synthesized and underwent *in vitro* transcription to generate an RNA pool. To determine how UNK protein binds these 25,000 sequences, we performed natural sequence RNA Bind-n-Seq^{26,52} (nsRBNS), a quantitative large-scale *in vitro* binding assay (**Fig. 1B**). Briefly, the RNA pool of natural sequences was incubated with recombinant protein, protein-RNA complexes were immobilized on magnetic beads, washed, and RNA was isolated. RNA sequencing was used to quantify the abundance of each RNA bound to UNK as well as the abundance of each RNA in the input RNA pool. These experiments yield binding enrichments (*R* values) for each oligo which are defined as the frequency of a given oligo bound to UNK vs the frequency of that oligo in the input RNA (**Methods**). Greater *R* values indicate a higher degree of binding. Previous work has demonstrated that nsRBNS correlates well with *in vivo* binding and regulation^{53,54}. This approach enabled us to test the binding of nearly 25 thousand sequences in tandem, with a wide range of *in vivo* binding properties.

UNK nsRBNS experiments were performed in duplicate and at different protein concentrations with robust cross-replicate correlation (**Fig. 1C**; $R=0.954$, Pearson’s correlation). We first asked whether nsRBNS is capable of capturing binding differences based on previously derived UNK motifs. UNK is known to bind a primary core UAG motif with secondary U/A-rich motifs^{26,27}. Presence of more than one UNK motif within an RNA has been shown to enhance binding²⁶, driven by engagement with the tandem zinc fingers of UNK⁴⁶. Within our pool, we observe that binding enrichment increased with increasing UAG count (**Fig. 1D**), consistent with previous studies²⁶. Similar but slightly more modest increases in binding occurred with increasing counts of UUU and UUA (**Fig. S1B,C**).

To demonstrate that the core UAG motif is important for binding, we included central motif mutants. For these sequences, if there was a central UAG motif present within the 120 nt region, it was mutated to CCG to assess whether

binding is reduced (**Methods**). As expected, and as reported previously²⁷, mutating the central UAG motif is enough to drastically diminish binding (**Fig. 1E**). This observation was further validated via an *in vitro* qPCR-based binding assay for one gene, *GART*. We observed that mutation of the central UAG motif to a CCG drastically diminished binding (**Fig. S1D**; $p=0.004$, one-sided Wilcoxon test). Finally, given that UNK is known to bind single-stranded RNA²⁶, we computed the base-pairing probability of the central 10 nt region harboring binding sites using a thermodynamic RNA folding algorithm⁵⁵. As the mean base-pairing probability of this central region decreased (e.g., more structure occluding the region) enrichment values decreased (**Fig. S1E**). These data confirm nsRBNS as a replicable *in vitro* assay, capable of measuring binding differences based on sequence features for 25,000 sequences in parallel.

Recapitulation of *in vivo* Binding Patterns and Regulation

We next tested whether *in vivo* binding patterns could be recapitulated *in vitro*. For our *in vitro* analysis, we defined three classes of binding sites: “control” where no evidence of binding was detected via iCLIP in either species; “bound,” where binding was detected via iCLIP; and “orthologous not bound,” where sites were bound in one species and not bound in the other (**Fig. 2A**, diagram). As UNK has been shown to bind primarily within the coding sequences (CDS) and secondarily within 3’ untranslated regions (UTRs)²⁷, we assessed binding patterns individually for these regions. Within CDS binding sites, we found that “orthologous not bound” oligos had similar enrichments as “control” oligos whereas “bound” oligos were significantly more enriched (**Fig. 2B**, $p<2.2e-16$, KS test). In UTRs, “bound” oligos were again the most enriched, but in this case “orthologous not bound” sites in UTRs had greater enrichments than controls (**Fig. 2C**). In fact, UTR sites overall had better enrichments than CDS, perhaps due to UTRs being generally enriched for U- and A- rich 3mers that are bound by UNK (**Fig. S2A**). Thus, nsRBNS captures binding features derived from *in vivo* iCLIP.

nsRBNS enrichment values span several orders of magnitude, driven by differences in affinity and avidity (**Fig. 1D**). Although occurring in a far more complex environment, binding in cells likely also occurs on a spectrum driven by affinity, though more difficult to capture experimentally. To compare *in vivo* to *in vitro* patterns, we asked what proportion of species-specific binding observed *in vivo* could be captured *in vitro*. We measured how often a species-specific site was better bound than its non-bound ortholog and found that 60% of binding sites mirrored the *in vivo* trend (**Fig. 2B,C**-inset & **Fig. S2B,C**). However, the degree to which species-specific binding was recapitulated *in vitro* ranged from no difference to greater than 100-fold difference between the bound and unbound orthologous site. To better understand these patterns, we turned to *in vivo*-bound sites where we also mutated the UAG motif (see above **Fig. 1E**). We reasoned that because

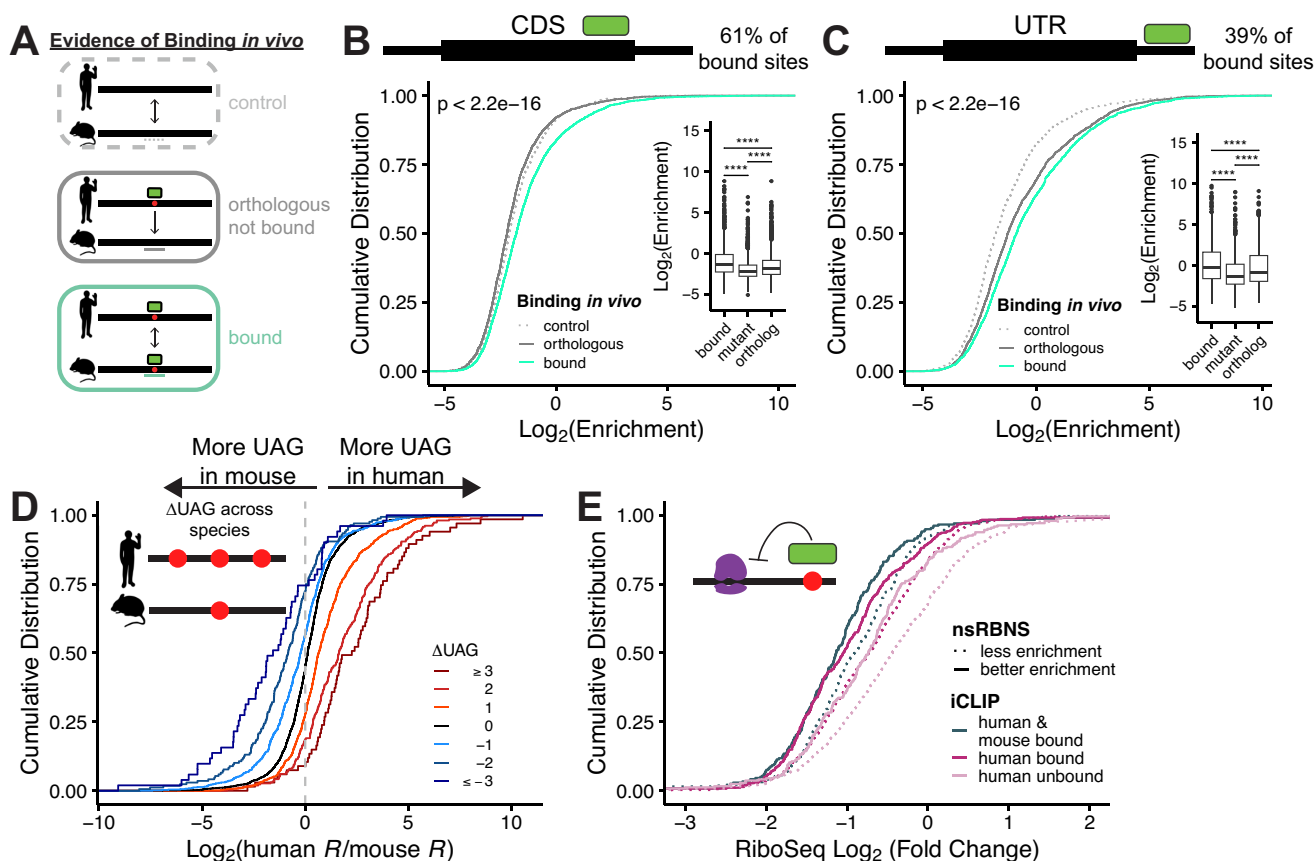


Figure 2. Analysis of species-specific binding patterns. A) Schematic of “control,” “orthologous,” and “bound” oligo classes used for species-specific transcript-level binding analysis. B-C) Cumulative distribution function of all iCLIP hits: control (light grey; dotted), orthologous (dark grey), and bound (teal) of B) CDS and C) UTR oligos. Significance of bound vs. orthologous was determined via KS test. Insets show boxplot of *in vitro* binding patterns for “bound,” “motif mutant,” and “orthologous” oligos. Significance was determined via two-sided Wilcoxon test. D) Cumulative distribution function of log₂ fold enrichment change of *in vivo* bound over *in vivo* not bound oligos separated by ΔUAG content. E) Cumulative distribution function of RiboSeq log₂ fold change separated via iCLIP detection. nsRBNS enrichment cutoffs defined as “less enrichment” <1 and “better enrichment” >1.

UAG drives binding, these mutants would be representative of minimal binding. Indeed, in 80% of cases UAG mutation diminished binding (Fig. 2B,C-inset & Fig. S2B,C). In aggregate, we found that “orthologous not bound” sites had an intermediate enrichment, that is, not as weakly bound as UAG mutants but significantly less bound *in vitro* than the “bound” category. Consistent with these findings and what is known about UNK-RNA interactions, the difference in UAG content between human and mouse orthologous sites had a large impact on differential binding (Fig. 2D), with gain of UAG enhancing binding and loss decreasing binding. The same was true of the known secondary motifs UUU (Fig. S2D) and UUA (Fig. S2E). Additionally, the greater the difference in percent identity between the 120 nt human and mouse binding sites, the greater the absolute difference in enrichments across species (Fig. S2F,G). These data highlight that *in vivo* binding patterns can be recapitulated *in vitro*. However, we note that some *in vivo* differences are not captured *in vitro*,

likely reflecting a combination of the cellular environment and limitations of *in vivo* (CLIP) and *in vitro* (nsRBNS) assays.

Binding Strength and *in vivo* Regulation

To determine whether these *in vitro* binding patterns also correspond with *in vivo* regulatory patterns, we examined ribosome profiling data upon UNK induction⁴⁷. UNK is a translational repressor²⁷, thus UNK-regulated RNAs are predicted to have decreased translation as previously shown²⁷. Genes with peaks identified via iCLIP²⁷ in both human and mouse were more translationally repressed than genes with peaks only identified in human or genes lacking UNK peaks (Fig. 2E). These data highlight that conserved targets display stronger regulation, consistent with what has been observed for microRNAs and splicing factors^{56–59}. We next asked whether the strength of binding from *in vitro* (nsRBNS) also had an impact on regulation. Greater *in vitro* binding enrichments were associated with increased translational suppression for

transcripts that were common to mouse and human or to human alone (**Fig. 2E**). These data support a direct relationship between interactions measured *in vitro* and binding and regulation assessed *in vivo*.

Species-Specific Binding Site Patterns

As noted above, iCLIP analysis revealed that while binding can be conserved at the transcript level, specific binding locations often change (**Fig. 3A**). For example, within the *GGPS1* transcript, we observe two species-specific binding sites (**Fig. 3B**). *In vivo*, this transcript is bound in both human and mouse, but different binding sites are used (located 122 nt apart in the alignment) even though the UAG motifs are conserved (**Fig. 3B**, bottom). In our *in vitro* nsRBNS assay, we see enrichments that mirror these *in vivo* patterns (**Fig. 3B**, top) prompting us to use this biochemical data to examine binding site preferences.

To determine how binding location changes across species in an *in vitro* context, we included four additional classes of oligos within our pool: “conserved,” where both the motif and binding location were maintained across species; “bound elsewhere,” where transcript binding was conserved, yet there was still differential motif usage across species (even when a motif was preserved across species); “not bound,” where the motif was maintained yet binding was not detected in the orthologous species (**Fig. 3A**); and “perfectly conserved,” the subset of “conserved” oligos with identical sequences between human and mouse. In aggregate, the degree of conserved binding *in vivo* correlated with *in vitro* enrichments. Least enriched were the “not bound” category followed by “bound elsewhere,” then “conserved,” and most enriched were “perfectly conserved” sites (**Fig. 3C,D & Fig. S3A,B**). Surprisingly, *in vitro* binding followed *in vivo* binding even when only regions with UAG motifs in both human and mouse were considered (**Fig. 3D & Fig. S3B**). Similar trends were observed in CDS and UTR regions (CDS in **Fig. 3C,D**; UTR in **Fig. S3A,B**). These data demonstrate that factors beyond the core motif impact RBP-RNA interactions, as our data shows that UNK can switch UAG motif usage between species. The fact that these preferences can be captured *in vitro* indicates that *cis* sequence changes surrounding the motifs are an important driver of binding.

Of note, binding sites perfectly conserved (100% identity) between human and mouse were among the strongest bound. In fact, of these regions we found that fewer than 3% were bound in only one species *in vivo*, indicating that a high degree of conservation within larger sequence regions is associated with conserved binding. To associate this degree of conservation with *in vivo* regulation, we again turned to ribosome profiling after UNK induction and found that transcripts with perfectly conserved binding sites were more translationally suppressed than other bound transcripts (**Fig. 3E**).

We hypothesized that these high-affinity binding sites may be more accessible (*i.e.*, have reduced levels of RNA secondary structure). To this end, we aligned sequences by their

central UAG, performed *in silico* folding⁵⁵, and compared each of the above categories. Indeed, perfectly conserved binding sites were the most accessible (with lower base pair probabilities (BPP)) at and downstream of the motif (**Fig. 3F**). Consistent with the preferences of these and many other RBPs for single-stranded RNA²⁶, accessibility appears to drive evolutionary changes in RNA binding. Simply put, conservation of context is a critically important mediator of conserved RNA-protein interactions.

Binding Site Patterns Across Cell Types Within Species

To compare these binding preferences to intra-species changes, we examined available iCLIP data from HeLa cells overexpressing UNK from the same study²⁷. Surprisingly, when looking at transcript-level conservation, we observed that approximately 51% of UNK transcripts were bound in both cell types (**Fig. S3C**; $p=2.3e-202$, hypergeometric test), similar to that observed in human vs. mouse comparisons. Looking further at the binding site level, only 41% occurred at the same motif (**Fig. S3D**), again similar to the cross-species comparisons. However, when we turned to *in vitro* nsRBNS, sites that were bound in both cell types versus only bound in one had no biochemical difference in binding as enrichments were largely similar (**Fig. S3E**). These data suggest that differing cellular environments (likely including presence of different complements of RBPs) can influence binding locations to a substantial degree.

To compare these patterns more generally to a larger group of RBPs across cell types within humans, we assessed binding of 14 RBPs (with well-defined motifs) from available enhanced CLIP (eCLIP) data in HepG2 and K562 cells^{60,61} (**Fig. S3F**). Although eCLIP differs from iCLIP^{61,62} (reviewed by Hafner *et al.*⁶³), we reasoned that both types of experiments should yield similar information. Using these data, we found that RBP binding sites — although variable from RBP to RBP — are well-conserved at the transcript level across cell types with 64% conservation on average for exonic binding and 53% conservation on average for non-exonic binding (*e.g.*, introns) between HepG2 and K562 cells. At the binding site level, approximately 54% of exonic peaks and 41% of non-exonic peaks are bound at the same motif across cell types (**Fig. S3G**). As expected, peaks with well-defined motifs displayed a greater degree of overlap between cell lines (**Fig. S3H,I**). These observations are similar to what we observed for UNK between SH-SY5Y versus HeLa cells with iCLIP (**Fig. S3C,D**). Thus, although limited to a small cohort of RBPs, these data suggest that whereas inter-species differences can be largely influenced by *cis* changes that can be captured biochemically (as discussed above), intra-species differences may be dictated by changing cellular environment across tissues (*i.e.*, RNA/RBP expression levels, levels of other RBPs, etc.).

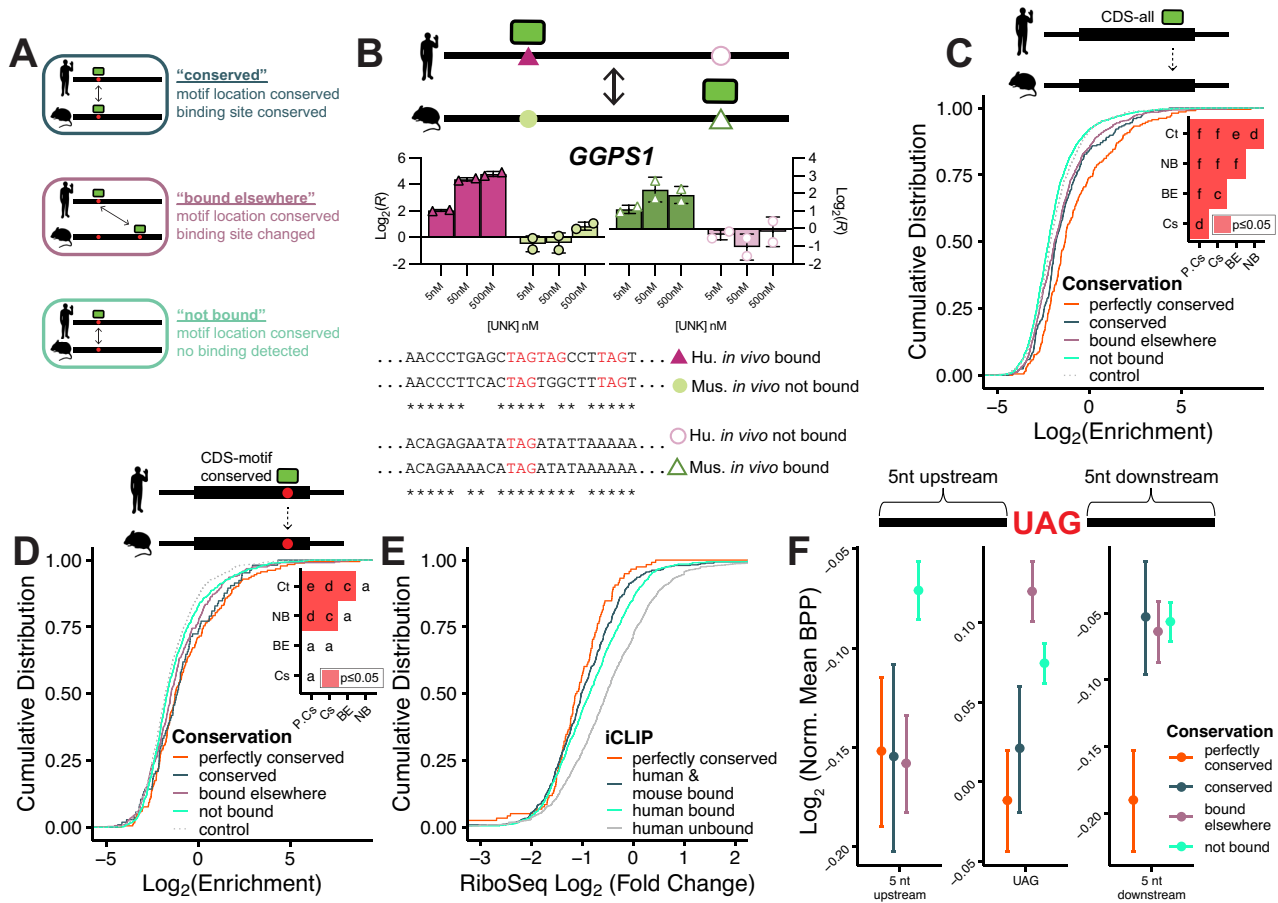


Figure 3. Analysis of species-specific syntenic motif level binding patterns. A) Definition of “conserved,” “bound elsewhere,” and “not bound” oligo classes used for species-specific transcript regional binding analysis. B) Conservation and binding of *GGPS1* orthologous pairs. (left) Log₂ enrichment values from nsRBNS for human bound (purple triangle), mouse not bound (light green circle), mouse bound (green open triangle), and human not bound (light purple open circle). (right) Alignment of human bound (purple triangle) to mouse not bound (light green circle) and mouse bound (green open triangle) to human not bound (purple open circle). Note: full oligos were used for alignment but only central region is shown. C-D) Cumulative distribution function of log₂ enrichment of control (light grey; dotted), not bound (teal), bound elsewhere (purple), conserved (blue), and perfectly conserved (orange) C) all CDS and D) motif conserved CDS oligos. Insets show significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p<=0.05). Values are as follows: a (ns), c (p<=0.05), d (p<=0.01), e (p<=0.001), f (p<=0.0001). E) Cumulative distribution function of RiboSeq log₂ fold change separated via iCLIP detection and sequence conservation. F) Log₂ fold change of mean base pair probability of the central region of “perfectly conserved,” “conserved,” “bound elsewhere,” and “not bound,” oligos normalized to UAG-containing CDS controls (see **Methods**). Error bars show standard error of the mean.

Species-Specific Regional Impacts on Binding

When binding is species-specific, is it possible to identify the sequences that drive binding in one species but not the other? In the simplest scenario this would be a region harboring a UAG motif that is found in only one species. To test whether introduction of sequences from an *in vivo*-bound species to the orthologous region that displayed no *in vivo* binding could restore binding, we designed chimeric mutants. Starting with the unbound mouse sequence, we substituted 10 nucleotide segments of the bound sequence into the unbound sequence to test which parts, if any, of the human sequence could confer

binding (Fig. 4A).

Within these chimeric oligos we included two classes: “UAG Change,” where the central UAG was present in the bound sequence but not in the unbound mouse sequence; and “Context Change,” where the UAG was conserved in both. As expected, in a “UAG Change” example, substitution of the central 10 bases which include the UAG motif (58-67) significantly enhanced binding (Fig. 4B). Supporting the importance of contextual features, other positions not harboring the central UAG could also confer enhanced binding but no single chimerized position contributed as significantly as position

58-67 which harbored the UAG (Fig. 4B).

Of particular interest were the “not bound” cases where a motif was conserved across species yet binding was lost (Fig. 3A). In “Context Change” chimeras, we noted a boost in binding upon changing of positions 58-67 (that harbor the central UAG) despite the motif being present in both species, suggesting contextual differences. Importantly, we also found swapping the segment just downstream (68-77) appeared to enhance binding (Fig. 4C), though statistical significance was not reached after correcting for multiple tests in this small cohort of binding sites tested (n=22). Enhancing chimeric sequences — mostly downstream of the core motif — tend to be U/A rich (Fig. S4A), likely leading to increased avidity and further engagement of UNK’s secondary RBD⁴⁶.

RNA binding is complicated and multi-factorial. Therefore, we hypothesized that chimerization of 10 nucleotides might not be sufficient to recover binding across species and that longer-range effects could be at play. Thus, we also included double chimeras where every possible combination of single chimeras was tested for binding (Fig. 4A). As we expected, double chimerization of both “UAG Change” (Fig. 4D) and “Context Change” (Fig. 4F) cases improved binding in many cases where single substitutions did not. Interestingly, almost all combinations with position 58-67 for “UAG Change” chimeras significantly enhanced binding (Fig. S4B-D). Double chimerization enhanced binding for 70% of “UAG Change” chimeras whereas single chimerization only achieved 21% restoration (Fig. 4E). Likewise, for “Context Change” double chimerization enhanced binding up to 52% from 24% with single chimeras (Fig. 4G). For approximately 50% of double chimeras, double chimerization not only restores binding but also led to binding at the level of the bound human sequence (Fig. S4E,F).

In one example, UNK bound human *GTPB4* approximately 500-fold better than mouse *Gtpb4*. Single chimerization of the central UAG-containing region (58-67) restored binding 30-fold (Fig. 4H), while double chimerization with positions 68-77 improved binding an additional 3.5-fold (Fig. 4H). We validated these binding patterns for *GTPB4* via fluorescence polarization (FP) with a 6-FAM-tagged RNA and fit a Kd for human *GTPB4* at 262.6 nM, mouse *Gtpb4* at >5 μ M, and chi58-*Gtpb4* at 304.8 nM (Fig. 4I). Thus, evolution of UNK-RNA binding involves substantial contributions of both primary motif level and contextual changes.

Evolutionary Conservation of Binding

To expand our phylogenetic scope beyond human/mouse, we investigated binding patterns across 100 vertebrates. Selecting the top 250 bound human sequences from our original nsRBNS assay, we sought to identify orthologous regions from 100 vertebrates^{64,65}, keeping only those where 25 or more species were aligned (Fig. 5A & S5A; see Methods). Within this set of sequences, we also included total motif mutants for each human oligo where every UAG was mutated to a CCG to define cutoffs for null binding (Methods). We

performed nsRBNS as described above (Fig. S5B). As an initial analysis, we measured the decrease in binding between wild-type regions and total motif mutant regions in human. As expected, we found that wild-type human sequences are better bound across the assay than mutant counterparts (Fig. S5C; $p=1.1e-15$, two-sided, paired Wilcoxon test).

To understand how binding diverges along the evolutionary timeline, we set the difference between each wild-type region and its UAG-mutant as the dynamic range of max binding to no binding (Methods). As we progress to more distant species from human, we observed that binding enrichment decreased (Fig. 5B). To understand the driving force behind loss or maintenance of binding, we computed sequence identity between human and all other vertebrates tested for every binding site (Fig. 5B, center). Some species and families along the tree have RNA sequence identity more similar to human which is mirrored by an increase in binding enrichment (Fig. 5B, center). These changes are reflected with a remarkably high degree of correlation between mean percent identity to human and binding enrichments (Fig. 5C; $R=0.919$, Pearson’s correlation). Similarly, evolutionary distances are also well-correlated with binding (Fig. 5D; $R=0.861$, Pearson’s correlation). A large drop-off in binding with increased variance was observed in fish (class *Actinopterygii*) (Fig. 5B, left), suggesting a loss of human UNK compatibility and species-specific RNA-protein interactions. Not surprisingly, however, while the RNA sequences evolve rapidly, with percent identity dropping quickly, UNK protein is highly conserved across species and does not reach 60% similarity (compared to human) until pufferfish (*Tetraodon nigroviridis*; 431 million years divergence), while the RBDs never drop below 70% similarity in vertebrates^{48,49} (Fig. 5B, center right).

In one example, *ATP1B1*, the central UAG motif is well conserved through pufferfish; however, binding begins to drop off around Cape golden mole (*Chrysochloris asiatica*). Similar to the trend for all binding sites tested, the percent identity of *ATP1B1* orthologs to human positively correlated with UNK binding (Fig. 5E & Fig. S5D,E). While this binding dropoff is not mirrored by any apparent changes in UAG content, a subtle shift in downstream sequence may be responsible for this binding difference. This can be observed through wild boar (*Sus scrofa*) *Atp1b1* which still maintains a central UAG motif but has a decrease in A/U content just downstream. Interestingly, alpaca (*Vicugna pacos*) *Atp1b1* binds with similar enrichment to wild boar *Atp1b1* even though it has completely lost its central UAG motif, perhaps indicating that the downstream changes in A/U content in wild boar are as important as loss of the UAG. At further evolutionary distances, zebrafish (*Danio rerio*) has completely lost the UAG that confers binding in human but has gained a UAG motif upstream and downstream which is mirrored through an increase in binding enrichment. Similarly complex binding trajectories are observed for *NFATC3* orthologs (Fig. S5F-H). While many primate sequences were perfectly identical, we observed with *PPP2R5C* that even across short evolutionary

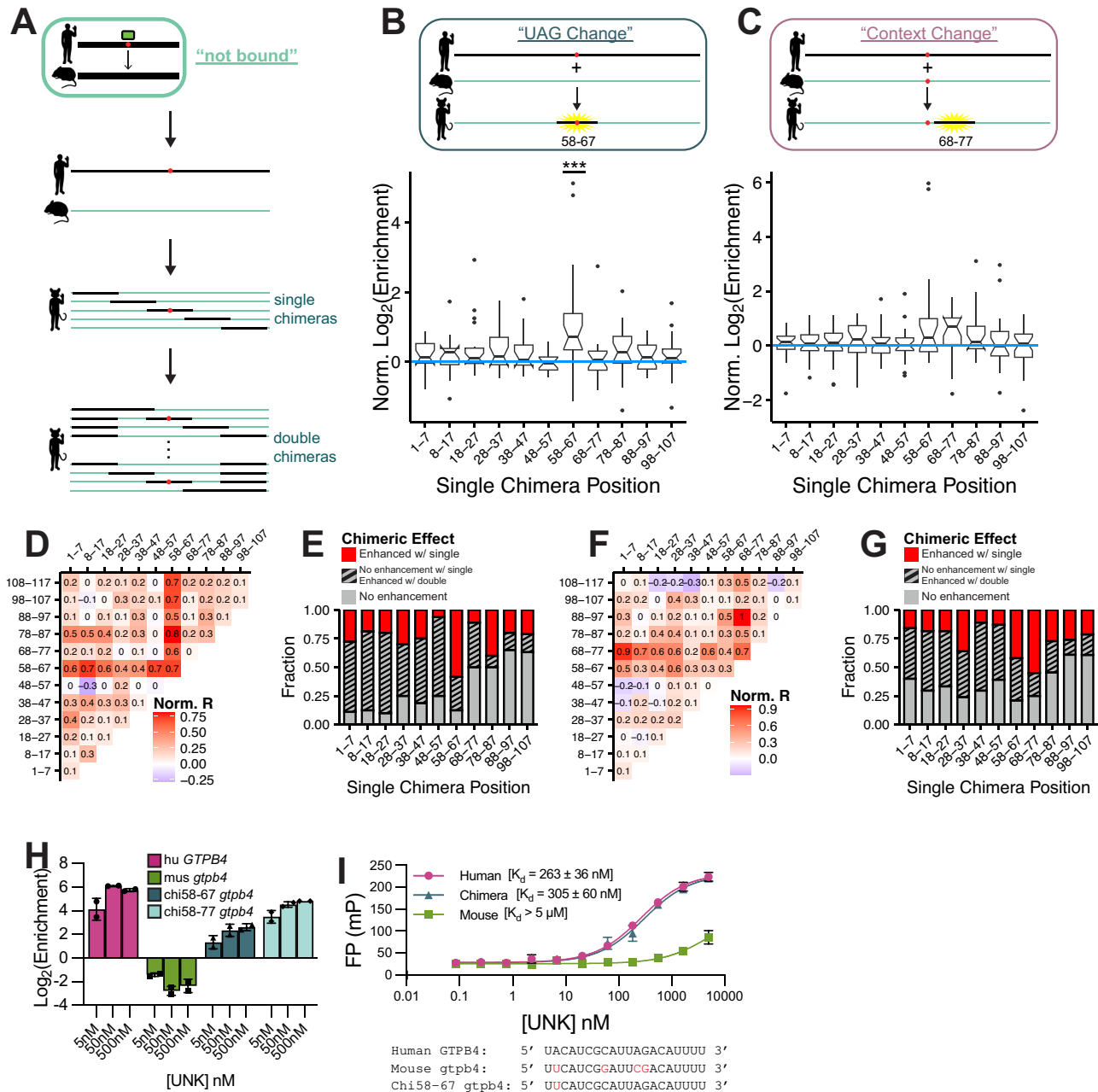


Figure 4. Analysis of regional impacts on binding. A) Design of single and double chimera oligos. B) Design and box and whisker plot of normalized \log_2 enrichment (chimera/wt) for "UAG Change" single chimeras. Significance was determined via paired, one-sided Wilcoxon test and corrected for multiple comparisons via the BH procedure. Chimerization at positions 58-67 was found to be significant ($p < 0.001$). C) Design and box and whisker plot of normalized \log_2 enrichment (chimera/wt) for "Context Change" single chimeras. Significance was determined via paired, one-sample Wilcoxon test and corrected for multiple comparisons via the BH procedure. Following multiple comparison correction, no positions were determined to be significant. D) Heat map of median normalized \log_2 enrichment (chimera/wt) for "UAG Change" single and double chimeras. E) Fraction of "UAG Change" chimeras enhanced with binding after single (red) or double (grey, striped) chimerization. F) Heat map of median normalized \log_2 enrichment (chimera/wt) for "Context Change" single and double chimeras. G) Fraction of "Context Change" chimeras enhanced with binding after single (red) or double (grey, striped) chimerization. H) \log_2 enrichment values from nsRBNS for human, mouse, single chimera, and double chimera *GTPB4* at 5, 50, and 500 nM UNK. I) Fluorescence polarization binding curves for human *GTPB4*, mouse *gtpb4*, and chi58-67 *gtpb4* RNA oligos incubated with UNK.

distances, large motif changes can occur, leading to drastic changes in binding (Fig. S5I,J).

To examine how binding changes across 100 vertebrates correlates with RNA regulation, we once again turned to ribosome profiling data⁴⁷. As discussed above, we found correlations between UNK binding and evolutionary distance and sequence conservation (Fig. 5F). Interestingly, binding sites within mRNAs that UNK translationally suppressed displayed a higher degree of UNK binding conservation across vertebrates (Fig. 5F), spanning large evolutionary distances. This effect may be explained by the fact that translational targets were modestly more conserved than those that are not translational targets (Fig. S5K).

These data support a model wherein both subtle (often difficult to discern) and large changes greatly influence RNA-protein interactions and therefore RNA regulation (Fig. 6). Often adjacent sequences and RNA structure change — sometimes driven by single substitutions — resulted in loss of binding *in vitro*. These subtle sequence differences can seemingly have impacts akin to total motif loss, implicating these differences in loss of *in vivo* binding. Future work will expand these studies beyond UNK, however, given the prototypical nature of UNK-RNA interactions and the high degree of conservation at the protein level, we expect that other RBPs behave similarly.

Limitations of This Study

Our approach relies on assessing how well *in vivo* data is reflected using high-throughput *in vitro* approaches. This strategy is well-suited for understanding direct protein-RNA interactions influenced by changes in RNA sequence and local RNA structure; however, it does not include the complex cellular environment including factors that can impact binding *in vivo*: **i)** Individual mRNA and RBP concentrations can vary drastically in cells, across cell types, and across species. The concentration of the RNA, the RBP, other components of mRNPs, and ribosomal interactions (as UNK is primarily involved in translational control²⁷) can all be anticipated to affect UNK-RNA interactions and accessibility (see **Supp. Note 1**). **ii)** Our work has been guided by *in vivo* binding data, which — while incredibly informative — have limitations (discussed above) which ultimately prompted us to reconstitute the UNK-RNA interactome *in vitro*. Because these iCLIP experiments partially guided our experimental designs, technical biases present in iCLIP may have impacted our choice of sequences to study (see **Supp. Note 1**). **iii)** Due to difficulties in purifying full-length UNK protein, we have performed our experiments on the RBDs alone, which exhibit tight and specific binding to known UNK motifs (as shown above and previously²⁶); additional components of UNK, such as its disordered domain, may impact binding in unexpected ways⁴⁷. **iv)** For practical reasons we have used human UNK protein for these studies. For smaller evolutionary distances such as human versus mouse, this is unlikely to impact binding. However, for larger distances where UNK protein RBDs may

be more different, one might expect to see co-evolution of the *cis*-elements with changes in RNA-binding properties of UNK (see **Supp. Note 1**). **v)** RBNS is in nature a zero-sum experiment in that all RNA molecules compete with each other for protein binding. Thus, in pools that contain mostly high affinity RNA targets, some targets will appear enriched and others depleted despite all being high-affinity binders. This effect is most evident in our assays using binding sites derived from 100 vertebrates (Fig. S5B; see **Supp. Note 2**). **vi)** The nature of oligo pool design removed sequences with poor alignments. Therefore, the percent identity analysis shown is reflective of only those binding sites that had a minimal level of alignment (**Methods**), while those not having sufficient alignments were excluded. The above should be considered when interpreting enrichments presented in this study.

Discussion

How differences in RNA sequence impact RNA-protein interactions and downstream regulation remains poorly understood. Until recently, studies focused primarily on TFs and their binding sites; however, recent work has begun to incorporate studies on RNA^{31,32,40–45,66}. Here, we examine species-specific and conserved RNA binding using the neuronal RBP, UNK, as a model. We find that roughly 45% of UNK binding sites have been maintained between human and mouse, and far fewer maintain these binding sites down to the motif level (Fig. 1). Using a high-throughput *in vitro* assay to test binding, we found that species-specific binding patterns and regulation can be explained to a substantial effect by biochemically measurable RNA-protein interactions (Fig. 2 & Fig. 3). Although more limited in scope, binding changes across cell types within species appear to be driven cellular context (e.g., the *trans* environment) (Fig. S3). Evolutionarily, while RBPs are highly conserved, binding differences occur across species and are driven primarily by *cis* RNA evolution (Fig. 5B). These differences can emerge through evolution of primary motifs; however, substitutions within secondary binding sites can also lead to drastic binding differences across species (Fig. 5E & Fig. S5D-J).

RBP binding appears to be more complex than one might expect. While primary motifs can serve as “on/off” switches, the full mechanism is often more elaborate. For UNK specifically, the primary motif reported previously (UAGNNUUU) is only bound by UNK in 23% of all occurrences within the CDS, highlighting that other factors influence binding²⁷. As has been reported before^{53,54}, secondary motifs and local structure have evolved to modulate binding and regulation of individual transcripts. Given that many RBPs bind similar motifs, evolution of high-affinity binding sites presents an interesting balancing act. While enhanced accessibility and context may enhance individual RNA-RBP interactions, it may make these regions more accessible to other RBPs and thus enable RBP-RBP competition for binding.

While it's difficult to look at the effects of single nucleotide variants (SNVs) on a global scale, the framework

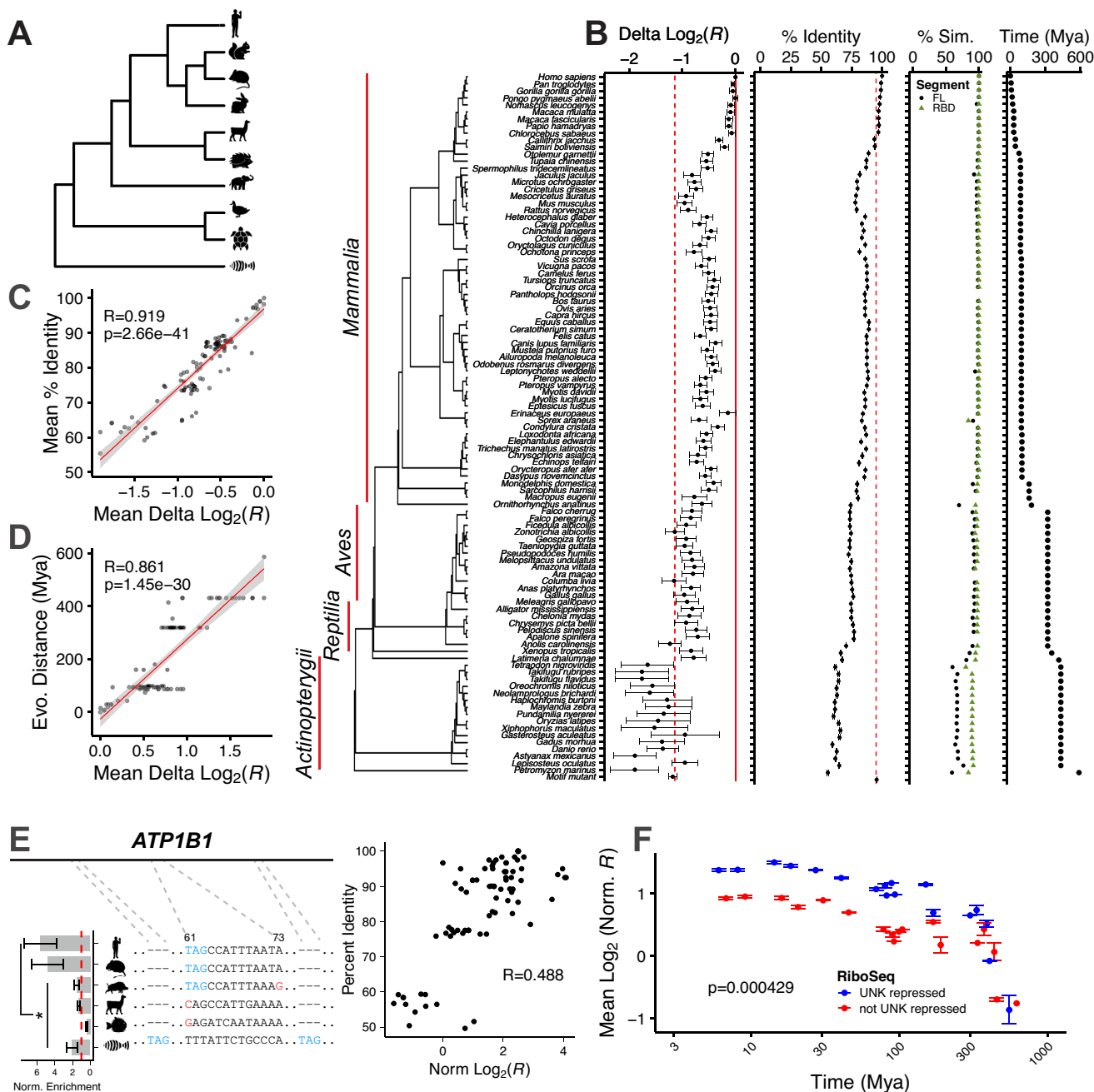


Figure 5. Evolutionary Conservation of Binding. A) Simplified alignment of 100 vertebrate natural sequence alignment. B) Delta log_2 enrichment, percent RNA sequence identity, percent UNK similarity (full length-grey and RBDs-green), and evolutionary distance in millions of years against 100 vertebrates for the aligned sequences from the top human bound oligos. Error bars show standard error of the mean (SEM). C) Mean percent RNA sequence identity (Y-axis) versus mean delta log_2 enrichment (X-axis) for each aligned oligo. Pearson's correlation coefficient included. D) Evolutionary distance in millions of years (Y-axis) versus mean delta log_2 enrichment (X-axis) for each aligned oligo. Pearson's correlation coefficient included. E) (left) Multiple sequence alignment for *ATP1B1* for *Homo sapiens*, *Mus musculus*, *Sus scrofa*, *Vicugna pacos*, *Tetradon nigroviridis*, and *Danio rerio* with normalized enrichment by species. (right) Percent RNA sequence identity (Y-axis) versus normalized delta log_2 enrichment (X-axis). Pearson's correlation coefficient included. F) Scatter plot of log_2 normalized UNK binding enrichment by evolutionary distance. X-axis plotted on log_{10} scale. Error bars show SEM. Data were separated by regulation as determined via RiboSeq where blue reflects UNK repression of translation [higher than average log_2 fold change (>0.9)] and red reflects lack of UNK repression [less than average log_2 fold change (<-0.9)]. Significance was determined via KS test.

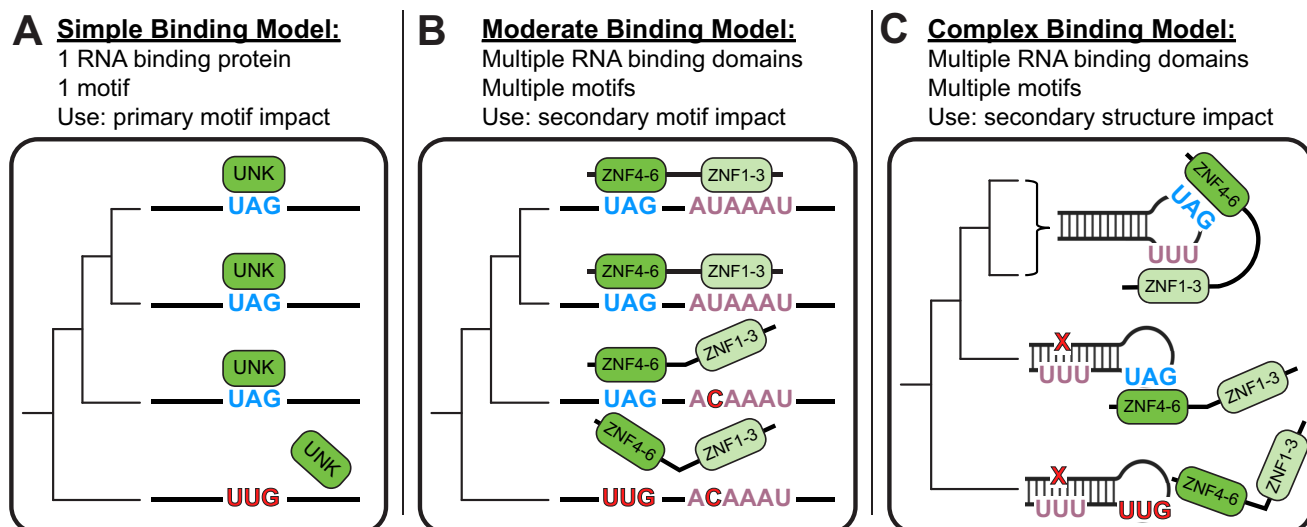


Figure 6. Models of RNA Binding. A) Simple binding model: considers only primary motifs. B) Moderate binding model: considers primary and secondary motifs. C) Complex binding model: considers primary and secondary motifs as well as RNA secondary structure.

presented here should be broadly applicable for such studies. Indeed, previous work has shown that SNVs themselves can impact direct RBP-RNA interactions. When examining a “simple binding model,” SNVs within primary motifs can be understood to totally abrogate binding (Fig. 6A). When we look at a more “moderate binding model” and begin to consider secondary motif contributions and increased valency, we can see how UNK’s secondary motif may contribute to differential binding patterns across evolution (Fig. 6B). Finally, when examining the most “complex binding model,” which considers secondary structure as well, global structural rearrangements that affect motif access are the most striking examples (Fig. 6C). All three of these models — taking into account primary motifs, secondary motifs, and RNA structure — likely apply in different situations. Understanding selection for or against these complex features that impact binding will help explain how regulation is conserved or species-specific.

These binding differences have been observed throughout several previous studies, where RBP binding appears to be dynamic and cell-type specific^{21,67–69}. Our findings support that genomic evolution of regulatory sites most frequently occurs in *cis*, rather than *trans*^{6,7}, at least on shorter time frames (as shown with TFs³⁹). However *cis* evolution can also result in *trans* evolution as regulatory elements like TFs and RBPs also have self-regulatory features⁷⁰, and at longer distances *trans* variation may play a larger role. Additionally, the cellular environment is also changing. Thus, genomic evolution relies on a delicate balance of binding site mutations, protein conservation and expression, and cellular context.

UNK is a translational inhibitor that binds primarily in the CDS but can also bind in the 3’ UTR of its target mRNAs²⁷. *In vitro*, UNK bound UTR sequences more tightly than CDS (Fig. 2 & Fig. 3), driven by the presence of more UNK motifs in UTRs (Fig. S2A). However, in cells, UNK binds primarily to CDS²⁷; this preference may be driven by increased local concentration of UNK near CDS, resulting from its association with ribosomes⁴⁶. UNK is not unique in exhibiting gene region preference⁷¹ and the mechanisms driving these preferences are not well understood. One recent study has demonstrated that UNK can associated with CCR4-NOT on some targets⁴⁷, thus additional cellular factors like these may impact target selection.

The work presented here parallels studies performed on a handful of transcription factors^{6,7,38,72–74}. For example, Schmidt and coworkers⁶ examined the binding profiles of two TFs in human, mouse, dog, opossum, and chicken and found that binding tends to be species-specific even when the proteins are highly conserved. Additionally, they observed that these species-specific binding preferences are largely due to *cis* sequence element changes across species⁶. For a more direct comparison, Odom *et al.*⁷ examined the binding profiles of four highly conserved TFs between human and mouse and found that while the TFs themselves are not readily changing, their DNA-interactome changes readily (up to 60%) often due to changes in motif content. Our work also highlights the complexity of translating defined regulatory elements from one species to another as the precise location of these sites may frequently change even when motifs appear to be conserved.

Methods

Expression and Purification of Recombinant UNK

Plasmid pGEX-GST-SBP-UNK (30-357)²⁶ was transformed into Rosetta *E. coli* competent cells (Novagen). Cultures were grown to an OD of 0.8 in LB media, adjusted to 16°C, and induced with 0.5 mM IPTG (Thermo Scientific) for 24 hours. Cells were collected via centrifugation at 4,000 \times g for 15 minutes and lysed in lysis buffer (200 mM NaCl, 5 mM DTT, 50 mM HEPES, 3 mM MgCl₂, 2 mM PMSF, 1 PierceTM protease inhibitor mini tablet/2 L; Thermo Scientific). The lysate was sonicated then incubated with 500 units/1 L culture Benzonase Nuclease (Sigma-Aldrich) for 30 minutes then with 5 units/1 L RQ1 RNase-free DNase (Promega) for 10 minutes at room temperature. NaCl concentration was adjusted to 1 M and the lysate was clarified by centrifugation at 17,800 \times g for 30 minutes. 0.05% polyethyleneimine (PEI) was added to precipitate excess nucleotides and was centrifuged at 17,800 \times g for 10 minutes.

Supernatant was passed over a 0.45-micron filter. Recombinant protein was purified via GST-trap FF column (GE). The column was washed in low salt buffer (300 mM NaCl, 50 mM HEPES), ATP buffer (300 mM NaCl, 50 mM HEPES, 5 mM ATP, 500 mM MgCl₂), and high salt buffer (1 M NaCl, 50 mM HEPES). For 100vertRBNS only, 1:50 PreScission Protease (Cytiva) was loaded on column in cleavage buffer (20 mM HEPES, 100 mM NaCl, 5 mM DTT, 10% glycerol, 0.01% triton X-100) to cleave the GST-tag prior to SEC. Protein was incubated at 4°C overnight on-column. Cleaved SBP-tagged protein was eluted in cleavage buffer. For nsRBNS, proteins were eluted off column in glutathione buffer (50 mM Tris pH 8.0, 20 mM reduced glutathione, final pH = 7).

All proteins were concentrated via centrifugation to 500 μ L before further purification via size exclusion chromatography [Superdex 200 Increase 10/300 GL (Cytiva)] in size exclusion buffer (20 mM HEPES, 1 M NaCl, 10 mM DTT, 0.01% triton X-100). 0.5-mL fractions were collected. Purity was assessed via SDS-PAGE (4-12% gradient) and Coomassie staining. Fractions corresponding to SBP-UNK (42.3 kDa) were pooled and dialyzed into 20 mM HEPES, 100 mM NaCl, and 5% glycerol. Concentration was determined via Pierce 660 nm assay (Thermo Scientific).

iCLIP Data Analysis and Oligo Design

UNK iCLIP-seq data was obtained from E-MTAB-2279²⁷. Mouse coordinates were converted from mm9 to mm10 and human coordinates were converted from hg19 to hg38 using liftOver⁷⁵ (version 1.24.0) in RStudio⁷⁶ (version 2023.03.0) with R platform⁷⁷ (version 4.2.2). Peaks were selected such that only the maximum scoring peak within 20 nucleotides of other peaks would be recorded. This was done in a rolling fashion such that if several peaks were back-to-back, each within 20 nucleotides of each other, only the maximum scoring peak would be recorded. Peaks were mapped back to their respective genes/transcripts using RStudio package ‘AnnotationHub⁷⁸’ (version 3.8.0) with ‘BSGenome.Hsapiens.NCBI.GRCh38⁷⁹’ and ‘BSGenome.Mmusculus.UCSC.mm10⁸⁰’. Only peaks mapping to exons were included for subsequent analysis.

For overlap analysis, peaks were expanded to 101 nucleotides and sequences were obtained using ‘getSeq’ using BSgenomes mentioned above. TAG-containing regions were identified. Previously reported RNA-seq data for SH-SY5Y cells (Murn *et al.*:²⁷ E-MTAB-2277⁸¹) mouse brain tissue (ENCODE^{60,82}: ENCSR000BZJ), and HeLa cells (ENCODE^{60,82}: ENCSR552EGO) was mapped to the mm10 or hg38 genomes using STAR⁸³ (version 2.7.10b) with default parameters. RSEM was used to calculate gene level expression values. Tximport was used to read RSEM output and transcripts per million (TPM) was used for comparison. Peaks were filtered to genes with \geq 5 TPM in the respective cell lines. Gene level intersecting peaks (SH-SY5Y) were converted from hg38 to mm10 using the liftOver utility from UCSC⁸⁴ then intersected with mouse iCLIP peaks using BEDtools⁸⁵ (version 2.31.0). RStudio package ‘VennDiagram⁸⁶’ was used to produce Venn diagrams.

For final oligo pool, sequences were expanded to 170 nucleotides and coding frame was determined where applicable. Using previously reported RBNS data for UNK²⁶, sequences were recentered around the highest ranking *k*mer closest to the center. Final sequences were trimmed down to 120 nucleotides (**Fig. S1A**). A subset of oligos were selected from the bound and unbound human and mouse orthologs where the central UAG motifs were mutated to a CCG. Additionally, single and double chimeras were designed such that 10 (single) or 20 (double) nucleotides of the bound ortholog were placed into the unbound sequence. Due to the differences between human and mouse, the chimerization was not always perfect, meaning that placing 10 nt of human into the same exact syntenic mouse region may not be 100% accurate. This should be considered when interpreting the chimera data.

Natural Sequence RNA Bind-n-Seq (nsRBNS)

Target RNAs were identified from iCLIP experiments performed in human and mouse neuronal cells²⁷ (see above). An array of 24,254 natural sequence oligos was synthesized by Twist Biosciences and transcribed to RNA with T7 polymerase. RBNS was performed as previously described^{26,52}. Briefly, MyOne Streptavidin T1 Dynabeads were washed in binding buffer (25 mM Tris pH 7.5, 150 mM KCl, 3 mM MgCl₂, 0.01% tween, 500 μ g/mL BSA, 1 mM DTT) and incubated with 0, 5, 50, or 500 nM recombinant GST-SBP-UNK. After 30-minute incubation, 1 μ M RNA was added to the reaction. After 1 hour, UNK-RNA complexes were isolated and unbound RNA was washed away. Complexes were eluted in 4 mM biotin. The eluted RNA was

reverse transcribed with Superscript III (Invitrogen) with RBNS RT primer (IDT; see below), amplified by PCR with Phusion DNA polymerase (NEB) with RBNS index primers and RBNS reverse primer I (see below), and sequenced on an Illumina HiSeq 300 instrument.

Primer name	Sequence
RBNS RT primer	GCCTTGGCACCCGAGAATTCCA
RBNS index primer	CAAGCAGAAGACGGCATACGAGAT[N ₆ barcode]GTGACTGGAGTTCCTTGGCACCC GAGAATTCCA
RBNS reverse primer	AATGATACGGCGACCACCGAGATCTACACGTTTCAGAGTTCTACAGTCCGA

nsRBNS Mapping and Enrichment Analysis

Reads were trimmed using fastx_toolkit⁸⁷ (version 0.0.14) as needed. Mapping was performed with STAR⁸³ (version 2.7.10b). STAR mapping parameters were set to `-outFilterMultimapNMax 1` and `-outFilterMismatchNmax 1` to generate counts files. Fasta file for reference was trimmed for adapters using seqtk⁸⁸ (version 2.3.0) as needed. SAMtools⁸⁹ (version 1.16) was used for processing alignment files as needed. Enrichment was calculated as frequency of an oligo in the protein bound sample divided by the frequency in the input.

nsRBNS Data Analysis

Data was compiled and analyzed in RStudio⁷⁶ (version 2023.03.0) with R platform⁷⁷ (version 4.2.2). R packages ggplot2⁹⁰ (version 3.4.1), 'ggpattern'⁹¹ (version 1.0.1), 'ggpubr'⁹² (version 0.6.0), and 'ggrepel'⁹³ (version 0.9.3) were used to make publication quality figures. Other RStudio packages — including 'dplyr'⁹⁴ (version 1.1.0), 'Hmisc'⁹⁵ (version 4.8.0), 'msa'⁹⁶ (version 1.30.1), 'org.Hs.eg.db'⁸⁶ (version 3.17.0), 'reshape2'⁹⁷ (version 1.4.4), 'rstatix'⁹⁸ (Kassambara 2023b) (version 0.7.2), and 'stringr'⁹⁹ (version 1.4.4) — were used for data analysis as needed. GraphPad Prism (version 10) was also used to make publication quality figures as needed. Data tables used for all analyses with sequences, relevant iCLIP information, enrichment values, relevant sequence information, and relevant oligo information can be found in **Supplemental Table 1**.

Ribosome Profiling Data Analysis

Ribosome profiling data was obtained from Shah *et al*⁴⁷. Genes bound in both human and mouse, human only, or mouse only (human not bound) were identified via iCLIP as described above. For genes with multiple peaks, nsRBNS enrichment values were summed. Only human nsRBNS enrichments and Ribo-Seq log₂ fold changes were used as ribosome profiling after UNK over-expression in mouse is not currently available.

Mean Base Pair Probability Analysis

DNA sequences for all hg38 genes were obtained from Ensembl¹⁰⁰. Genes not bound in human neuronal cells as identified by iCLIP²⁷ were selected for subsequent analysis. For genes with multiple isoform sequences, only one was kept: This was done randomly with RStudio function "sample." 120 nucleotide sequences were selected and centered around the downstream TAG motif just upstream of the stop codon to match the binding pattern of UNK which increases near the stop codon but does not bind the stop codon itself. Individual base pair probabilities were calculated with Vienna RNAfold⁵⁵ -partfunc to calculate the partition function and base pairing probability matrix for both the CDS controls as well as "perfectly conserved," "conserved," "bound elsewhere," and "not bound" sequences aligned perfectly at the central UAG. Mean base pair probability (bpp) was calculated for each category positionally and divided by CDS controls positionally to normalize. Mean bpp was further averaged across the central motif (UAG), five nucleotides upstream, and five nucleotides downstream.

100 Vertebrate DNA Pool Assembly

The top 250 human *in vivo* and *in vitro* bound sequences were selected from the nsRBNS experiment. UCSC's BLAT¹⁰¹ was used to determine the chromosome as well as the start and stop position for each sequence. The start and stop positions were expanded out 65 nucleotides each such that the region was 250 nucleotides in total to account for insertions and deletions across species. Multiple alignment blocks were selected using maf_parse in PHAST module¹⁰² (version 1.5) against UCSC's MAFs for 100 vertebrates. Sequences with less than 25 alignments were filtered out, and the remaining sequences were collapsed down with gaps removed. Sequences were aligned centrally in RStudio with package 'msa'^{103,104} (version 1.30.1) and trimmed back to 120 nucleotides. RStudio packages 'stringr'⁹⁹ (version 1.5.0) and 'data.table'¹⁰⁵ (version 1.14.8) were also used for string manipulation and data table formation as needed. Finally, total motif mutants were included where all TAG motifs in human oligos were mutated to CCG to define a null cut-off for binding. From the starting 112 human oligos, we found 5,641 orthologous sequences.

100 Vertebrate RNA Bind-n-Seq

RBNS was performed similarly as described above and previously^{26,52}. An array of 5,753 oligos corresponding to 112 human sequences and their 100 vertebrate alignments was ordered from Twist Biosciences. *In vitro* transcription was performed with T7 RiboMAX Express Large Scale RNA Production System (Promega) according to manufacturer protocols. RNA was purified via denaturing gel electrophoresis, eluted via RNA crush-n-soak into H₂O, and concentrated with phenol chloroform extraction.

Binding reactions were assembled as follows: MyOne Streptavidin T1 Dynabeads (Thermo Fisher) were first washed in new RBNS binding buffer (25 mM Tris, pH 7.5, 150 mM KCl, 3 mM MgCl₂, 0.01% triton-X 100, 500 µg/mL BSA, 20 units/mL SUPERase-In (Thermo Fisher)) then were incubated with 100 nM recombinant SBP-UNK for 30 minutes at 4°C. Following incubation, 1 µM RNA was added and incubated for 1 hour at 4°C. Unbound RNA was then washed away in new RBNS wash buffer (25 mM Tris pH 7.5, 150 mM KCl, 0.01% triton X-100, 20 units/mL SUPERase-In) three times. Bound RNA was eluted in 0.1% SDS and 0.3 mg/mL proteinase K (Thermo Fisher) at 60°C for 30 minutes. Elution was performed twice and the elutions were pooled. Following elution, reverse transcription was performed with Superscript IV (Invitrogen) with RBNS RT primer (see above) and amplified as described above. Sequencing was performed on an Illumina NextSeq 500.

100 Vertebrate nsRBNS Data Analysis

Data was analyzed similarly to nsRBNS data as discussed above with a few additions. RStudio package ‘ape’¹⁰⁶ (version 5.7) was used to assemble 100 vertebrate phylogenetic tree according to available data from UCSC^{64,107}. Additionally, RStudio package ‘msa’^{103,104} (version 1.30.1) was used for sequence alignment and percent identity analysis. RStudio packages ‘ggmsa’¹⁰⁸ (version 1.3.4), ‘ggprism’¹⁰⁹ (version 1.0.4), and ‘scales’¹¹⁰ (version 1.2.1) were used to make publication quality alignment figures. Data tables used for all analyses with sequences, relevant species information, enrichment values, relevant sequence information, and relevant oligo information can be found in **Supplemental Table 2**. For protein percent similarity analysis, human protein sequences were pulled from UniProt⁵⁰ and BLAST⁴⁸ was used for all species alignments. The RBDs were annotated as amino acids 31 to 335 based on previous work by Murn *et al*⁴⁶. All enrichments were normalized to their respective total motif mutant enrichment. Where indicated, delta enrichments were used for analysis and represent the divergence from the normalized human enrichment ($\log_2[\text{norm species}/\text{norm human}]$).

eCLIP Peak Overlap Analysis

RNA-seq data for HepG2 and K562 normal cell lines were downloaded from ENCODE^{60,82} and mapped to hg38 with STAR⁸³ (version 2.7.10b) using default parameters. Expression values were quantified using RSEM¹¹¹ (version 1.3.1) and differential analysis was conducted using RStudio package ‘DESeq2’¹¹¹ (version 1.40.1). Genes with no significant differences in expression between K562 and HepG2 cell lines were filtered using mean expression >10 and absolute log₂ fold change (L2FC) <=1.

14 RBPs which have eCLIP peaks⁶¹ in HepG2 and K562 and an RBNS motif²⁶ were selected. The peaks were downloaded from ENCODE (see table below) and filtered for enrichment of log₂ fold change > 1 and p-value of <0.001. Each peak was extended by 50 base pairs upstream to account for experimental limitations of eCLIP⁶¹. Replicates for each cell line were combined and filtered for the ones that fell within the genes that did not have significantly differential expression between HepG2 and K562. BEDtools⁸⁵ (version 2.31.0) was used to collapse the overlapping peaks that were within 20 base pairs into a single peak spanning the region. Sequences under the combined peaks were taken from RStudio package ‘BSgenome.Hsapiens.NCBI.GRCh38’⁷⁹ (version 1.3.1). Peaks were grouped into two groups based on presence or absence of RBNS motif within the peak²⁶.

Peaks were further grouped based on whether they overlapped exons or not. Exon annotations for hg38 (v109) genome were obtained from RStudio package ‘Annotation Hub’⁷⁸ (version 3.8.0) and a peak was assigned as an exonic peak if it had at least 20 base pair overlap with an exon. Thus, peaks for an RBP in each cell line were grouped into four following groups: peaks within an exon and had an RBNS motif, peaks within an exon and did not have an RBNS motif, peaks not within an exon and had an RBNS motif, and peaks not within an exon and did not have an RBNS motif. For each of these groups, we calculated the number of overlapping peaks between K562 and HepG2. Since the proportion of overlapping peaks between two cell lines is relative to total peaks identified in each cell line, we took the maximum after calculating the proportion using both cell lines respectively. Additionally, similar analysis was conducted for overlaps between parent genes of the peaks as well (**Fig. S3F**).

RNA-Seq	Cell Line	Rep
ENCFF267RKD,ENCFF455VYN	K562	1
ENCFF606ZTR,ENCFF444KCV	K562	2
ENCFF713MNU,ENCFF478DZZ	HepG2	1
ENCFF936SLY,ENCFF446UEC	HepG2	2

eCLIP	RBP Target	Cell Line	eCLIP	RBP Target	Cell Line
ENCFF073PCD	HNRNPL	HepG2	ENCFF077OSY	HNRNPL	K562
ENCFF082QGS	RBFOX2	HepG2	ENCFF103PRM	HNRNPL	HepG2
ENCFF105GZJ	TRA2A	HepG2	ENCFF121XCN	PCBP1	K562
ENCFF127WMZ	KHSRP	K562	ENCFF145YYK	IGF2BP1	HepG2
ENCFF150ZOO	RBM22	K562	ENCFF185IDD	IGF2BP1	K562
ENCFF230QOU	HNRNPC	HepG2	ENCFF241AOZ	HNRNPK	K562
ENCFF253ZSN	RBM22	HepG2	ENCFF288MWL	HNRNPK	HepG2
ENCFF327JJE	SRSF9	HepG2	ENCFF348TPU	TIA1	K562
ENCFF374XQF	KHSRP	K562	ENCFF378HWC	TIA1	HepG2
ENCFF383ZAQ	TAF15	HepG2	ENCFF390PJW	RBFOX2	HepG2
ENCFF401YRZ	HNRNPL	K562	ENCFF402AIE	TRA2A	K562
ENCFF409DPS	TAF15	K562	ENCFF421FJD	TARDBP	HepG2
ENCFF432ASF	SRSF9	HepG2	ENCFF443KJS	FUS	K562
ENCFF502OYV	PCBP1	HepG2	ENCFF526OQL	HNRNPC	K562
ENCFF534YQS	HNRNPC	HepG2	ENCFF545NBF	RBM22	HepG2
ENCFF565ILV	FUS	K562	ENCFF606RXB	TARDBP	K562
ENCFF611AHG	KHSRP	HepG2	ENCFF613UUR	SRSF9	K562
ENCFF618ZPP	TRA2A	K562	ENCFF626XAA	TARDBP	HepG2
ENCFF664RLU	HNRNPC	K562	ENCFF669TNM	TARDBP	K562
ENCFF674TKN	TIA1	K562	ENCFF685MZA	FUS	HepG2
ENCFF705SDK	IGF2BP1	HepG2	ENCFF754XAQ	HNRNPK	HepG2
ENCFF766DUS	PCBP1	K562	ENCFF779OIO	RBFOX2	K562
ENCFF824IDO	HNRNPK	K562	ENCFF853FGC	IGF2BP1	K562
ENCFF856EHA	TIA1	HepG2	ENCFF860QZG	SRSF9	K562
ENCFF899HGF	RBM22	K562	ENCFF899ZEH	TAF15	HepG2
ENCFF910WLP	RBFOX2	K562	ENCFF914VUW	PCBP1	HepG2
ENCFF951IBI	FUS	HepG2	ENCFF966KQG	TRA2A	HepG2
ENCFF988MWD	KHSRP	HepG2	ENCFF996BXS	TAF15	K562

Fluorescence Polarization (FP)

RNA oligos were synthesized by Integrated DNA Technologies (IDT) with a 3' 6-FAM label (see below) and incubated at 5 nM with serially diluted recombinant SBP-UNK (10.9, 50.8, 152, 457 pM, 1.37, 4.12, 12.3 37.0, 111, 333, or 1000 nM) in FP binding buffer (20 mM HEPES, 5 mM DTT, 137.5 mM NaCl, 0.01% triton X-100, 10 ng/μL BSA, 2 units/mL SUPERase•In™; Thermo Scientific) for 15 min at 4°C. Plates were centrifuged at 1000 *x g* for 1 min and fluorescence polarization was measured with a PHERAstar plate reader (BMG Labtech) at 25°C. Data were fit to a single site binding model and a K_D was determined.

Oligo name	Sequence
Human <i>GTPB4</i> FP RNA	UACAUCGCAUUAGACAUUUU /6-FAM/
Mouse <i>gtpb4</i> FP RNA	UUCAUCGGAUUCGACAUUUU /6-FAM/
Chi58-67 <i>gtpb4</i> FP RNA	UUCAUCGCAUUAGACAUUUU /6-FAM/

In vitro Transcription of RNA for qPCR Binding Assay

DNA fragments for wild-type and mutant GART were ordered from IDT (see below) and PCR amplified with Phusion DNA polymerase (New England Biolabs) with FWD_adapter and REV_adapter primers from IDT (see below), resulting in full-length DNA oligos for *in vitro* transcription. DNA was purified via agarose gel extraction, and transcribed with a T7 RiboMAX Express Large Scale RNA Production System (Promega). RNA was purified with RNA crush-n-soak and concentrated with phenol chloroform isolation.

Oligo name	Sequence
wt_GART_fwd	TCCGACGATCGTTATCTCCAACAAAGCCGCAGTAGCTGGGTTAGATAAAGCGGAAAGAGC
wt_GART_mid	TTAGATAAAGCGGAAAGAGCTGGTATTCCCAGTAGAGTAATTAATCATAAACTGTATAAA
wt_GART_mid-comp	TTTATACAGTTTATGATTAATTACTCTAGTGGAATACCAGCTCTTTCCGCTTTATCTAA
wt_GART_rev	TTTATACAGTTTATGATTAATTACTCTAGTGGAATACCAGCTCTTTCCGCTTTATCTAA
mut_GART_fwd	TCCGACGATCGTTATCTCCAACAAAGCCGCAGTAGCTGGGTTAGATAAAGCGGAAGAGC
mut_GART_mid	TTAGATAAAGCGGAAAGAGCTGGTATTCCCACCCGAGTAATTAATCATAAACTGTATAAA
mut_GART_mid_comp	TTTATACAGTTTATGATTAATTACTCGGGTGGGAATACCAGCTCTTTCCGCTTTATCTAA
mut_GART_rev	TCCTCTGTTGGTCAATTGCACTGTCAAATTCTACACGATTTTATACAGTTTATGATCTAA
FWD_adapter	TAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATC
REV_adapter	TGATGCTCAATCCTCTGTTG
WT GART RNA	GAGUUCUACAGUCCGACGAUCGUUAUCUCCAACAAAGCCGCAGUAGCUGGGU UAGAUAAAAGCGGAAAGAGCUGGUUAUCCCCACUAGAGUAAUUAUCAUAAACU GUAUAAAAAUCGUGUAGAAUUUGACAGUGCAAUUGACCAACAGAGGGAUUGAG CAUCA
Mutant GART RNA	GAGUUCUACAGUCCGACGAUCGUUAUCUCCAACAAAGCCGCAGUAGCUGGGU UAGAUAAAAGCGGAAAGAGCUGGUUAUCCCCACCCGAGUAAUUAUCAUAAACU GUAUAAAAAUCGUGUAGAAUUUGACAGUGCAAUUGACCAACAGAGGGAUUGAG CAUCA

qPCR Binding Assay

We performed our qPCR-based binding assay as previously described¹¹² with a few modifications. Dynabeads MyOne Streptavidin T1 magnetic beads (Thermo Fisher) were washed in blocking buffer (25 mM Tris pH 8.0, 150 mM KCl, 3 mM MgCl₂, 1 mg/mL BSA, 2 units/1 μL SUPERase-In (Invitrogen), and 1 mg/mL yeast tRNA (Fisher Scientific)), and then in qPCR binding buffer (25 mM Tris pH 8.0, 150 mM KCl, 3 mM MgCl₂, 1 mg/mL BSA, 2 units/1 mL SUPERase-In, and 50 nM random sequence RNA). Beads were incubated with two concentrations of SBP-UNK (167 and 1500 nM) at 25°C for 10 minutes. Bead-protein complexes were separated on the magnet and resuspended in 0.1 nM RNA. Beads, protein, and RNA were incubated at 25°C for 30 minutes. Unbound RNA was removed, and bound RNA was eluted in 4 mM biotin and 25 mM Tris, pH 8.0 at 37°C for 30 minutes. Reverse transcription was performed on unbound and bound RNA with iScript Reverse Transcription Supermix (Bio-Rad) following manufacturer's protocols with qPCR_REV primer (see below). An RNA calibration curve was assembled at RT with the following amounts of RNA: 45.7 fM, 0.137, 0.412, 1.23, 3.70, 11.1, 33.3, and 100 pM. RT reactions were diluted 2-fold and qPCR was performed in duplicate with SsoAdvanced SYBR Green Supermix (Bio-Rad) according to manufacturer's protocols with qPCR_FWD and qPCR_REV primers from IDT (see below). The threshold cycle (C_t) was determined using Bio-Rad's CFX Maestro software, and fraction bound was determined against the RNA calibration curve.

Primer name	Sequence
qPCR_FWD	GAGTTCTACAGTCCGACGATC
qPCR_REV	TGATGCTCAATCCTCTGTTG

References

1. Wagner, A. Robustness and evolvability: a paradox resolved. *Proc. Royal Soc. B: Biol. Sci.* **275**, 91–100 (2008).
2. Masel, J. & Trotter, M. V. Robustness and evolvability. *Trends Genet.* **26**, 406–414 (2010).
3. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their

- biological differences. *Science* **188**, 107–116 (1975).
- Britten, R. J. & Davidson, E. H. Gene Regulation for Higher Cells: A Theory: New facts regarding the organization of the genome provide clues to the nature of gene regulation. *Science* **165**, 349–357 (1969).
 - Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Q. review biology* **46**, 111–138 (1971).
 - Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
 - Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. genetics* **39**, 730–732 (2007).
 - Wang, Z.-Y. *et al.* Transcriptome and translome co-evolution in mammals. *Nature* **588**, 642–647 (2020).
 - Gerstberger, S., Hafner, M., Ascano, M. & Tuschl, T. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. In *Systems biology of RNA binding proteins*, 1–55 (Springer, 2014).
 - Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat. Rev. Genet.* **15**, 221–233 (2014).
 - Mitsis, T. *et al.* Transcription factors and evolution: An integral part of gene expression. *World Acad. Sci. J.* **2**, 3–8 (2020).
 - Nussbacher, J. K., Tabet, R., Yeo, G. W. & Lagier-Tourenne, C. Disruption of RNA metabolism in neurological diseases and emerging therapeutic interventions. *Neuron* **102**, 294–320 (2019).
 - Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).
 - Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. reviews Mol. cell biology* **18**, 437–451 (2017).
 - Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
 - Brinegar, A. E. & Cooper, T. A. Roles for RNA-binding proteins in development and disease. *Brain research* **1647**, 1–8 (2016).
 - Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters* **582**, 1977–1986 (2008).
 - Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat. reviews Mol. cell biology* **3**, 195–205 (2002).
 - Orphanides, G. & Reinberg, D. A unified theory of gene expression. *Cell* **108**, 439–451 (2002).
 - Lasko, P. Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. *Sci. STKE* **2003**, re6—re6 (2003).
 - Achsel, T. & Bagni, C. Cooperativity in RNA–protein interactions: the complex is more than the sum of its partners. *Curr. opinion neurobiology* **39**, 146–151 (2016).
 - Kitov, P. I. & Bundle, D. R. On the nature of the multivalency effect: a thermodynamic model. *J. Am. Chem. Soc.* **125**, 16271–16284 (2003).
 - Stefan, M. I. & Le Novère, N. Cooperative binding. *PLoS Comput. Biol* **9**, e1003106 (2013).
 - Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol. cell* **78**, 9–29 (2020).
 - Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. reviews Mol. cell biology* **8**, 479–490 (2007).
 - Dominguez, D. *et al.* Sequence, structure, and context preferences of human RNA binding proteins. *Mol. cell* **70**, 854–867 (2018).
 - Murn, J. *et al.* Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. *Genes development* **29**, 501–512 (2015).
 - Robert, F. & Pelletier, J. Exploring the impact of single-nucleotide polymorphisms on translation. *Front. genetics* **9**, 507 (2018).
 - Shatoff, E. & Bundschuh, R. Single nucleotide polymorphisms affect RNA-protein interactions at a distance through modulation of RNA secondary structures. *PLoS computational biology* **16**, e1007852 (2020).
 - Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845, [10.1038/nrg3813](https://doi.org/10.1038/nrg3813) (2014).
 - Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593–1599 (2012).
 - Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
 - Barr, K. A., Rhodes, K. L. & Gilad, Y. The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. *Genome Biol.* **24**, 1–21 (2023).
 - Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome biology* **16**, 1–11 (2015).
 - Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438 (2008).

36. Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324**, 659–662 (2009).
37. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85–88 (2004).
38. Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. & McCallion, A. S. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276–279 (2006).
39. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. biology evolution* **19**, 1114–1121 (2002).
40. Hogan, G. J., Brown, P. O. & Herschlag, D. Evolutionary conservation and diversification of Puf RNA binding proteins and their mRNA targets. *PLoS biology* **13**, e1002307 (2015).
41. Najdová, V., Stairs, C. W., Vinopalová, M., Voleman, L. & Doležal, P. The evolution of the Puf superfamily of proteins across the tree of eukaryotes. *BMC biology* **18**, 1–18 (2020).
42. Wang, M., Ogé, L., Perez-Garcia, M.-D., Hamama, L. & Sakr, S. The PUF protein family: overview on PUF RNA targets, biological functions, and post transcriptional regulation. *Int. journal molecular sciences* **19**, 410 (2018).
43. Wilinski, D. *et al.* Recurrent rewiring and emergence of RNA regulatory networks. *Proc. Natl. Acad. Sci.* **114**, E2816–E2825 (2017).
44. Yang, E.-W. *et al.* Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nat. communications* **10**, 1338 (2019).
45. Mantica, F. & Irimia, M. The 3D-Evo Space: Evolution of Gene Expression and Alternative Splicing Regulation. *Annu. Rev. Genet.* **56**, 315–337 (2022).
46. Murn, J., Teplova, M., Zarnack, K., Shi, Y. & Patel, D. J. Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. *Nat. Struct. Mol. Biol.* **23**, 16–23, [10.1038/nsmb.3140](https://doi.org/10.1038/nsmb.3140) (2016).
47. Shah, K. *et al.* A paradigm for regulation at the effector interface with RNA-binding proteins. *bioRxiv* 2009–2023 (2023).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. molecular biology* **215**, 403–410 (1990).
49. Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **35**, D5–D12 (2007).
50. UniProt. UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
51. Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M. & Ule, J. Data science issues in studying protein–RNA interactions with CLIP technologies. *Annu. Rev. Biomed. Data Sci.* **1**, 235–261 (2018).
52. Lambert, N. *et al.* RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol. Cell* **54**, 887–900, [10.1016/j.molcel.2014.04.016](https://doi.org/10.1016/j.molcel.2014.04.016) (2014).
53. Taliaferro, J. M. *et al.* RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. *Mol. cell* **64**, 294–306 (2016).
54. Begg, B. E., Jens, M., Wang, P. Y., Minor, C. M. & Burge, C. B. Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. *Nat. structural molecular biology* **27**, 901–912 (2020).
55. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for molecular biology* **6**, 1–14 (2011).
56. Yang, J.-S. *et al.* Widespread regulatory activity of vertebrate microRNA* species. *Rna* **17**, 312–326 (2011).
57. Ha, M., Pang, M., Agarwal, V. & Chen, Z. J. Interspecies regulation of microRNAs and their targets. *Biochimica et Biophys. Acta (BBA)-Gene Regul. Mech.* **1779**, 735–742 (2008).
58. Mazin, P. V. *et al.* Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques. *Rna* **24**, 585–596 (2018).
59. Lareau, L. F. & Brenner, S. E. Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol. biology evolution* **32**, 1072–1079 (2015).
60. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
61. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. methods* **13**, 508–514 (2016).
62. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. structural molecular biology* **17**, 909–915 (2010).
63. Hafner, M. *et al.* CLIP and complementary methods. *Nat. Rev. Methods Primers* **1**, 20 (2021).
64. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* **12**, 996–1006 (2002).
65. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921, [10.1038/35057062](https://doi.org/10.1038/35057062) (2001).

66. Torres-Méndez, A. *et al.* Parallel evolution of a splicing program controlling neuronal excitability in flies and mammals. *Sci. advances* **8**, eabk0445 (2022).
67. Beckmann, B. M., Castello, A. & Medenbach, J. The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. *Pflügers Arch. J. Physiol.* **468**, 1029–1040 (2016).
68. Iadevaia, V. & Gerber, A. P. Combinatorial control of mRNA fates by RNA-binding proteins and non-coding RNAs. *Biomolecules* **5**, 2207–2222 (2015).
69. Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. A combinatorial code for CPE-mediated translational control. *Cell* **132**, 434–448 (2008).
70. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
71. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719, [10.1038/s41586-020-2077-3](https://doi.org/10.1038/s41586-020-2077-3) (2020).
72. Paris, M. *et al.* Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. *PLoS genetics* **9**, e1003748 (2013).
73. Bradley, R. K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* **8**, [10.1371/journal.pbio.1000343](https://doi.org/10.1371/journal.pbio.1000343) (2010).
74. Wong, E. S. *et al.* Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nat. Commun.* **8**, 1092 (2017).
75. Bioconductor Package Maintainer. *liftOver: Changing genomic coordinate systems with rtracklayer::liftOver*, [10.18129/B9.bioc.liftOver](https://doi.org/10.18129/B9.bioc.liftOver) (2023).
76. RStudio, T. RStudio: integrated development for R. *Rstudio Team, PBC, Boston, MA URL <http://www.rstudio.com>* (2020).
77. R Core Team, R. R: A language and environment for statistical computing. (2013).
78. Morgan, M. & Shepherd, L. *AnnotationHub: Client to access AnnotationHub resources*, [10.18129/B9.bioc.AnnotationHub](https://doi.org/10.18129/B9.bioc.AnnotationHub) (2023).
79. Team, T. B. D. *BSgenome.Hsapiens.NCBI.GRCh38: Full genome sequences for Homo sapiens (GRCh38)* (2014).
80. Team, T. B. D. *BSgenome.Mmusculus.UCSC.mm10: Full genome sequences for Mus musculus (UCSC version mm10, based on GRCm38.p6)* (2021).
81. Sarkans, U. *et al.* The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic acids research* **46**, D1266–D1270 (2018).
82. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* **48**, D882–D889 (2020).
83. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
84. Lee, B. T. *et al.* The UCSC genome browser database: 2022 update. *Nucleic acids research* **50**, D1115–D1122 (2022).
85. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
86. Chen, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots* (2022).
87. Liu, X. *et al.* FastProNGS: fast preprocessing of next-generation sequencing reads. *BMC bioinformatics* **20**, 1–6 (2019).
88. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS one* **11**, e0163962 (2016).
89. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
90. Villanueva, R. A. M. & Chen, Z. J. ggplot2: elegant graphics for data analysis (2019).
91. FC, M., Davis, T. L. & authors, G. *ggpattern: 'ggplot2' Pattern Geoms* (2022).
92. Kassambara, A. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (2023).
93. Slowikowski, K. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'* (2023).
94. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *dplyr: A Grammar of Data Manipulation* (2023).
95. Harrell Jr, F. E. *Hmisc: Harrell Miscellaneous* (2023).
96. Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C. & Hochreiter, S. *msa: an R package for multiple sequence alignment*. *Bioinformatics* **31**, 3997–3999, [10.1093/bioinformatics/btv494](https://doi.org/10.1093/bioinformatics/btv494) (2015).
97. Wickham, H. Reshaping data with the reshape package. *J. statistical software* **21**, 1–20 (2007).
98. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots* (2023).
99. Wickham, H. & Wickham, M. H. Package 'stringr'. Website: <http://stringr.tidyverse.org>, <https://github.com/tidyverse/stringr> (2019).
100. Cunningham, F. *et al.* Ensembl 2022. *Nucleic acids research* **50**, D988–D995 (2022).
101. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656–664 (2002).

102. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings bioinformatics* **12**, 41–51 (2011).
103. May, A. C. W. Percent sequence identity: the need to be explicit. *Structure* **12**, 737–738 (2004).
104. Raghava, G. P. S. & Barton, G. J. Quantification of the variation in percentage identity for protein sequence alignments. *BMC bioinformatics* **7**, 1–4 (2006).
105. Dowle, M. & Srinivasan, A. *data.table: Extension of 'data.frame'* (2023).
106. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
107. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
108. Zhou, L. *et al.* ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Briefings Bioinforma.* **23**, bbac222 (2022).
109. Dawson, C. *ggprism: A 'ggplot2' Extension Inspired by 'GraphPad Prism'* (2022).
110. Wickham, H. & Seidel, D. *scales: Scale Functions for Visualization* (2022).
111. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 1–16 (2011).
112. Hinkle, E. R. *et al.* Alternative splicing regulation of membrane trafficking genes during myogenesis. *RNA* **28**, 523–540 (2022).

Acknowledgements

We would like to thank Michael Love, Peter Sudmant, Matthew Taliaferro, and Eric Van Nostrand for helpful and insightful feedback on this manuscript. This work was in part supported by NIH T32 GM008570 (S.E.H.) and R35GM142864 (D.D.) as well as startup funds from UNC Chapel Hill to (D.D.).

Author Contributions

Conceptualization, S.E.H., M.S.A., C.B., and D.D.; Methodology, S.E.H., M.S.A., M.M.A., and D.D.; Software, S.E.H., M.S.A., G.G., F.F.C., and D.D.; Validation, S.E.H., M.S.A., G.G., M.M.A., and D.D.; Formal Analysis, S.E.H., M.S.A., F.F.C., and D.D.; Investigation, S.E.H., M.S.A., M.M.A., and D.D.; Resources, C.B. and D.D.; Data Curation, S.E.H., M.S.A., G.G., J.M., and D.D.; Writing – Original Draft, S.E.H. and D.D.; Writing – Review & Editing, S.E.H., M.S.A., G.G., J.M., M.M.A., C.B.B., and D.D.; Visualization, S.E.H., G.G., and D.D.; Supervision, C.B.B. and D.D.; Funding Acquisition, D.D.

Conflict of Interests

The authors declare no conflicts.