nature communications

Article

https://doi.org/10.1038/s41467-024-52231-7

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

Received: 20 December 2023

Accepted: 28 August 2024

Published online: 27 September 2024

Check for updates

Sarah E. Harris[1,2], Maria S. Alexis[3,8], Gilbert Giri[2,4], Francisco F. Cavazos Jr[2], Yue Hu[2], Jernej Murn[5,6], Maria M. Aleman[2], Christopher B. Burge[3] & Daniel Dominguez[1,2,4,7]

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 examine 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 reconstitute 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.

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 sequences[2–5]. 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 effects[12–14]. RBPs constitute a large class of pan-essential regulatory factors[15,16] that drive RNA regulation, contributing significantly to transcription, splicing, and translation to influence the expression and identity of proteins produced[17–20]. These processes are dictated by the strength of the interaction between RBPs and their RNA targets[21–25]. In the simplest model of RNA regulation, RBPs bind short sequence motifs (3–8 nucleotides) within RNA to

[1]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, USA. [2]Department of Pharmacology, University of North Carolina, Chapel Hill, NC, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC, USA. [5]Department of Biochemistry, University of California, Riverside, CA, USA. [6]Center for RNA Biology and Medicine, Riverside, CA, USA. [7]RNA Discovery Center, University of North Carolina, Chapel Hill, NC, USA. [8]Present address: Remix Therapeutics, Cambridge, MA, USA. e-mail: didoming@email.unc.edu

influence its regulation[26]. However, these interactions are complex as change, loss, or gain of a single nucleotide within or surrounding a motif can greatly impact binding[27–29].

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 years[30]. Generally, RBPs tend to be more conserved than their DNA-binding counterparts, transcription factors[9]. 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—including 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 profiles[6]. 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% amino acid conservation for the full-length protein) at the amino acid level and have identical or near-identical binding preferences[7]. When a human chromosome is placed in a mouse context, TF binding predominantly follows the human binding patterns rather than that of the mouse[35], 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 assessed[39].

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 species[42]; 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 another[43], 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 impacts on RNA regulation[44]. The complex paths in which RNA regulation evolves have been previously reviewed[45], 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 use available individual-nucleotide resolution crosslinking and immunoprecipitation (iCLIP) data[27] from a neuronal RBP, unkempt (UNK), in human and mouse. UNK regulates neuronal morphology, is a negative regulator of translation, mildly destabilizes RNA targets, and associates with polysomes[27,46,47]. We identify species-specific and shared UNK binding sites and find that ~45% of UNK transcript binding was conserved across species. Importantly, while the binding of transcripts is conserved, the individual motifs that are bound are far less conserved, often switching between species. We reconstitute the in vivo UNK-RNA interactomes of human and mouse in vitro to understand the driving forces underlying species-specific binding and regulation. We find 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 extend our studies across 100 vertebrates to understand how sequence changes over longer time scales affect binding and find striking correlations between evolutionary distances, individual binding site conservation, and strength of UNK binding. This work deepens our understanding of *cis*-regulatory element evolution and highlights the complex nature of evolving RNA binding sites.

Results

UNK-RNA binding patterns vary across species

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 studies[46]; ii) UNK is 95% conserved between human and mouse with only one amino acid difference within the RNA-binding zinc finger domains (ZnFs)[48–50]; iii) Murn and coworkers demonstrated that even the sea sponge (*Amphimedon queenslandica*) UNK paralog functionally rescues knockdown of *UNK* in human cell lines[46] 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, respectively[27], to identify species-specific and conserved UNK binding sites (Supplementary Fig. 1A). Only genes expressed at greater than 5 transcripts per million (TPM) in both cell lines were included. Comparing one-to-one 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.

nsRBNS measures natural sequence binding differences in vitro at massive scale

While iCLIP is a powerful technique that allows for the derivation of nucleotide-level binding sites, several experimental 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 unknown[51]. 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

https://doi.org/10.1038/s41467-024-52231-7

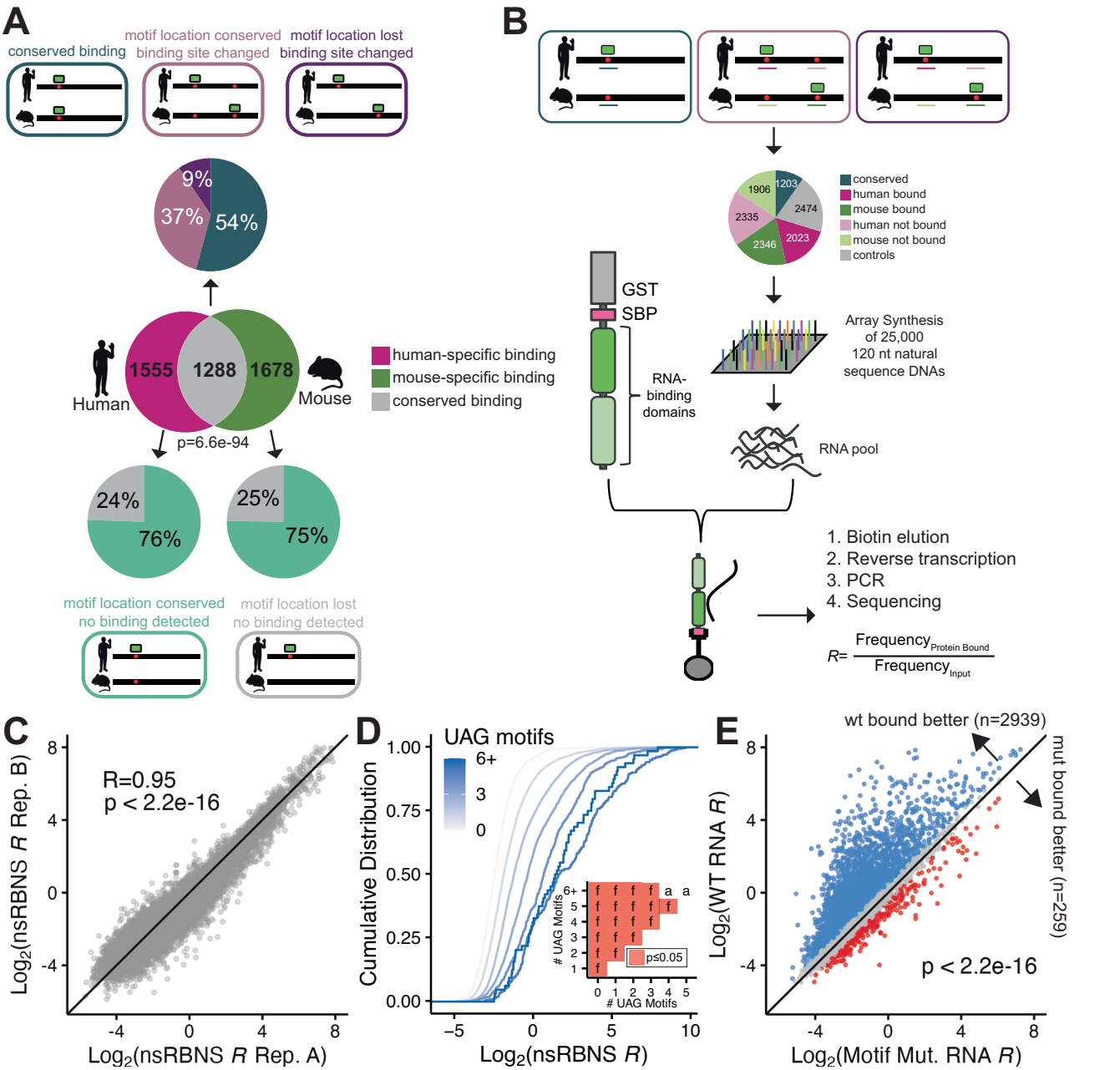

Fig. 1 | Design and validation of natural sequence RNA bind-n-seq (nsRBNS). **A** (Venn diagram) Transcript-level conservation of iCLIP UNK hits between human neuronal cells (SH-SY5Y) and mouse brain tissue. Significance determined via hypergeometric test. (Pie charts) Motif level conservation of iCLIP 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 and *p* val included. **D** Cumulative distribution function of $\log_2$ nsRBNS enrichment of all oligos separated by UAG motif content. Inset shows significance values for all comparisons via two-sided KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant ($p \leq 0.05$). Values are as follows: a (ns), f ($p \leq 0.0001$). **E** Scatter plot of $\log_2$ nsRBNS enrichment of wild-type (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). Significance determined via paired, one-sided Wilcoxon test.

12,287 natural RNA sequences, each 120 nucleotides long. Contained within this pool were UNK binding sites identified via iCLIP in human neuronal cells (n = 2023) and mouse brain tissue (n = 2346), as well as orthologous regions (human: n = 2335; mouse, n = 1906) 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 (n = 2474) 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 (normalized count for library size) 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 = 0.95, 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 binding[26], driven by engagement with the tandem ZnFs of UNK[46]. Indeed, when we tested the individual domains (ZnF1-3 or ZnF4-6) via random RBNS as previously described[26,52], we observed strong UAG binding with ZnF4-6 (the primary domains) and U/A rich motifs with ZnF1-3 (Supplementary Fig. 1B). These data support previous crystal structures showing UAG binding with ZnF4-6 and U/A binding via ZnF1-3[46]. Within our pool, we observe that binding enrichment increases with increasing UAG count (Fig. 1D), consistent with previous studies[26]. Similar but slightly more modest increases in binding occurred with increasing counts of UUU and UUA (Supplementary Fig. 1C, D).

To further 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 previously[27], 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 (Supplementary Fig. 1E; $p \le 0.01$, one-sided, paired Wilcoxon test). Finally, given that UNK is known to bind single-stranded RNA[26], we computed the base-pairing probability of the central 10 nt region harboring binding sites using a thermodynamic RNA folding algorithm[55]. As the mean base-pairing probability of this central region increased (e.g., more structure occluding the region) enrichment values decreased (Supplementary Fig. 1F; $p \le 0.0001$, KS test). 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.

In vivo binding patterns and regulation can be recapitulated in vitro

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, $p \le 0.01$, 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, $p \le 0.0001$, KS test). In fact, UTR sites overall had better enrichments than CDS, perhaps due to UTRs being generally more enriched over CDS for U- and A- rich 3mers that are bound by UNK (Supplementary Fig. 2A). 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% (65% for CDS and 58% for UTR) of binding sites mirrored the in vivo trend (Supplementary Fig. 2B–E). 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 UAG drives binding, these mutants would be representative of minimal binding. Indeed, in ~80% (83% for CDS and 81% for UTR) of cases UAG mutation diminished binding (Supplementary Fig. 2B, 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 (Fig. 2B, C-inset). 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 (Supplementary Fig. 2F) and UUA (Supplementary Fig. 2G). 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 (Supplementary Fig. 2H, I). 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.

In vitro binding patterns correlate with 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 induction[47]. UNK is a translational repressor[27], and mildly destabilizes its target RNAs[47], thus UNK-regulated RNAs are predicted to have decreased translation as previously shown[27]. Genes with peaks identified via iCLIP[27] 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.

In vivo binding differences can be recapitulated at the binding site

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: binding conserved, where both the motif and binding location were maintained across species; bound elsewhere, where

https://doi.org/10.1038/s41467-024-52231-7

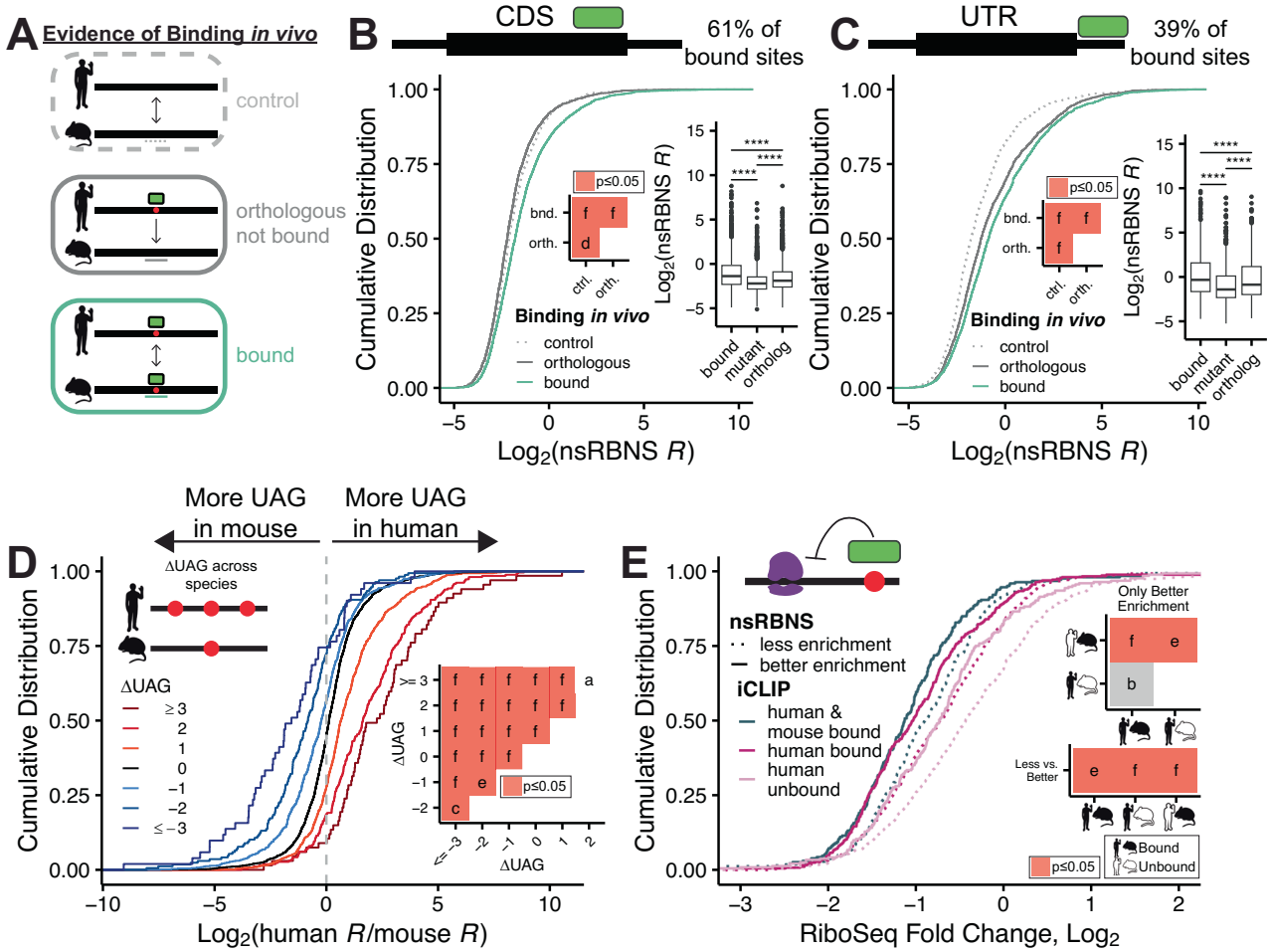

Fig. 2 | Analysis of species-specific binding patterns. A Schematic of "control," "orthologous," and "bound" oligo classes used for species-specific transcript-level binding analysis. **B**, **C** Cumulative distribution function of $\log_2$ nsRBNS enrichment of all iCLIP hits: control (light grey; dotted), orthologous (dark grey), and bound (teal) of (**B**) CDS and (**C**) UTR oligos. Inset boxplots show in vitro binding patterns for "bound," "motif mutant," and "orthologous" oligos. Significance of inset boxplots was determined via two-sided paired Wilcoxon test ($n = 1373$ sequences in (**B**) and 987 sequences in (**C**)). Significance marks are as follows: ****($p \leq 0.0001$). Centre line denotes median (50th percentile) with bounds of box representing 25th to 75th percentiles and the whiskers denoting 5th to 95th percentiles. Outliers are denoted as individual points. Inset heatmaps show significance values for all comparisons via two-sided KS test and corrected for multiple comparisons via the BH procedure for the cumulative distribution curves. Red denotes significant ($p \leq 0.05$). Values are as follows: d ($p \leq 0.01$), f ($p \leq 0.0001$). **D** Cumulative distribution function of $\log_2$ fold nsRBNS enrichment change of in vivo bound over in vivo not bound oligos separated by ΔUAG content. Inset shows significance values for all comparisons via two-sided KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant ($p \leq 0.05$). Values are as follows: a (ns), c ($p \leq 0.05$), e ($p \leq 0.001$), f ($p \leq 0.0001$). **E** Cumulative distribution function of RiboSeq fold change, $\log_2$ separated via iCLIP detection. nsRBNS enrichment cutoffs defined as "less enrichment" <1 and "better enrichment" >1. Insets show significance values for all comparisons via two-sided KS test and corrected for multiple comparisons via the BH procedure. Grey denotes nearing significance ($p \leq 0.1$). Red denotes significant ($p \leq 0.05$). Values are as follows: b ($p \leq 0.1$), e ($p \leq 0.001$), f ($p \leq 0.0001$).

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 binding 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 binding conserved, and most enriched were perfectly conserved sites (Fig. 3C, D and Supplementary Fig. 3A, B). Surprisingly, in vitro binding followed in vivo binding even when only regions with UAG motifs conserved across human and mouse were considered (Fig. 3D and Supplementary Fig. 3B). Similar trends were observed in CDS and UTR regions (CDS in Fig. 3C, D; UTR in Supplementary Fig. 3A, 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.

Broadly, when examining sequence conservation effects on in vitro enrichment differences, we observe that more sequence conserved oligo pairs have more similar nsRBNS enrichments than less sequence conserved oligo pairs, highlighting the robustness of sequence evolution to binding sites (Supplementary Fig. 3C). Interestingly, when we compare sequence conservation for all three categories of oligo pairs—binding conserved, bound elsewhere, and not bound—we observe a categorial breakdown in percent sequence identity. The binding conserved oligo pairs are most conserved at the sequence level followed by the not bound category and the bound elsewhere group being the least conserved at the sequence level (Supplementary Fig. 3D; $p \leq 0.001$, Wilcoxon test). It should be noted that both the not bound and bound elsewhere categories involve a species-specific in vivo binding event, that is, binding is observed in one species but not the other.

https://doi.org/10.1038/s41467-024-52231-7

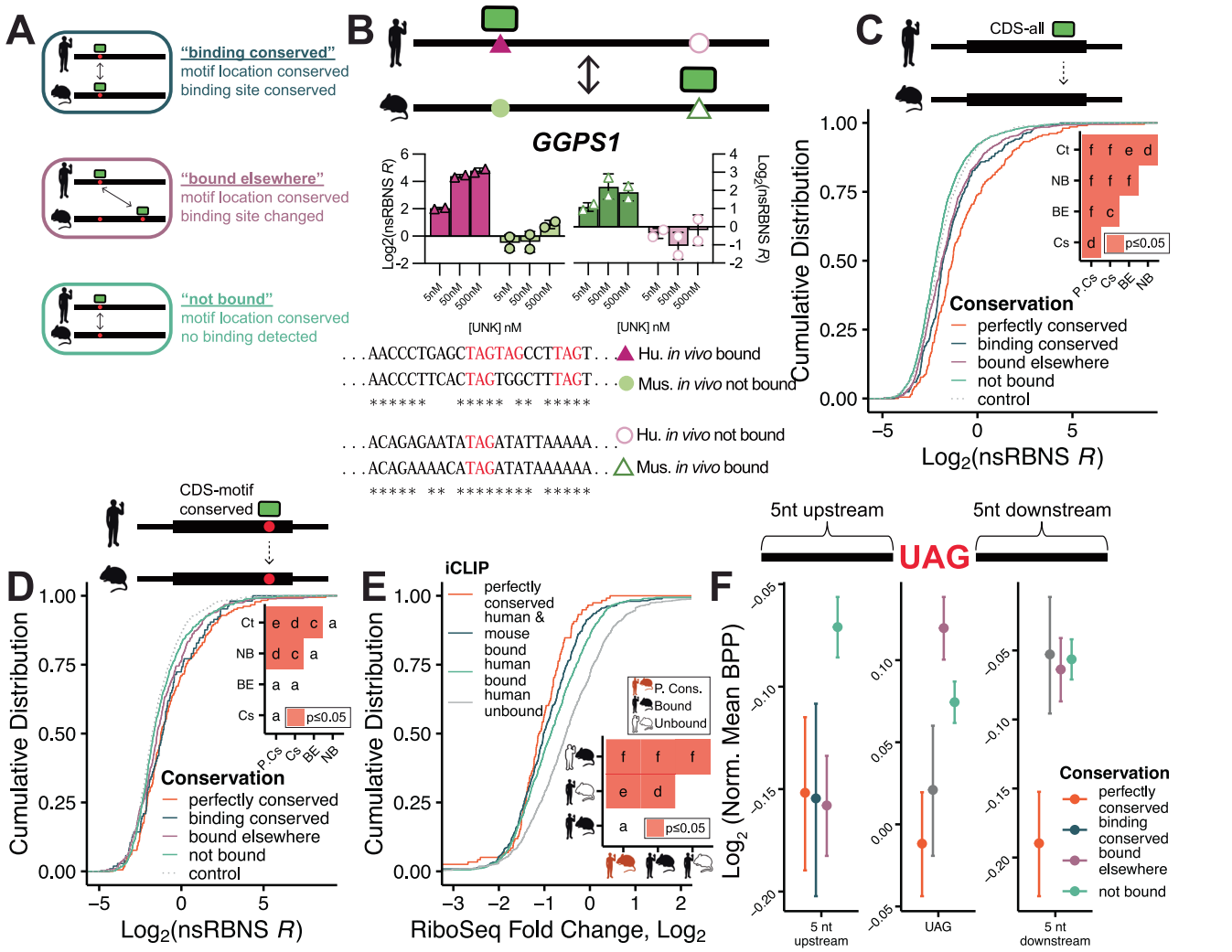

Fig. 3 | Analysis of species-specific syntenic motif level binding patterns. **A** Definition of binding conserved, bound elsewhere, and not bound oligo classes used for species-specific transcript regional binding analysis. **B** Conservation and binding of *GGPS1* orthologous pairs. (left) Log_2 nsRBNS 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) ($n = 2$). (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 the central region is shown. **C**, **D** Cumulative distribution function of log_2 nsRBNS enrichment of control (light grey; dotted), not bound (teal), bound elsewhere (purple), binding conserved (blue), and perfectly conserved (orange) **C** all CDS and **D** motif conserved CDS oligos. Insets show significance values for all comparisons via two-sided KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant ($p \le 0.05$). Values are as follows: a (ns), c ($p \le 0.05$), d ($p \le 0.01$), e ($p \le 0.001$), f ($p \le 0.0001$). **E** Cumulative distribution function of RiboSeq fold change, log_2 separated via iCLIP detection and sequence conservation. Inset shows significance values for all comparisons via two-sided KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant ($p \le 0.05$). Values are as follows: a (ns), d ($p \le 0.01$), e ($p \le 0.001$), f ($p \le 0.0001$). **F** Log_2 fold change of mean base pair probability of the central region of perfectly conserved (n = 221 sequences), binding conserved ($n = 155$ sequences), bound elsewhere ($n = 574$ sequences), and not bound ($n = 1395$ sequences) oligos normalized to UAG-containing CDS controls (see "Methods"). Error bars show standard error of the mean.

To examine these inter-species sequence differences on a global scale more specifically, we analyzed the 3mer enrichment across human and mouse where the human oligo was bound better, despite maintenance of a UAG. Looking across all possible 3mers upstream and downstream of the central UAG, we observe that human bound sequences are more enriched in A and U-rich motifs centrally than their unbound mouse counterparts (Supplementary Fig. 3E). We hypothesized that these contextual sequence differences may drive UNK binding due to the dual-RBD architecture of UNK where ZnF4-6 mediates primary UAG association while ZnF1-3 binds secondarily to U/A rich motifs[46].

To test this, we turned to fluorescence polarization (FP) to understand to what extent the dual-RBD architecture aids in UNK binding patterns. When comparing the binding preferences of ZnF1-3, ZnF4-6, and ZnF1-6 to a UAG-containing oligo with downstream U-rich content, we observe that ZnF1-3 binds weakly with a K_d of ~2 μM while ZnF4-6 binds more than 5-fold better at ~420 nM. However, when ZnF1-3 and ZnF4-6 bind in combination, binding is enhanced 10-fold to a K_d ~ 40 nM (Supplementary Fig. 3F). When comparing this to the binding patterns of ZnF1-6 with an RNA oligo with only a UAG motif, the K_d increases similar to that of ZnF4-6 (Supplementary Fig. 3F). These data highlight the importance of UNK's multiple domains for selecting targets and is supported by previous work on cooperativity and avidity for other RBPs (reviewed by Achsel and Bagni[21] and Corley et al.[24]).

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 folding[55], 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 UNK and many other RBPs for single-stranded RNA[26], accessibility appears to drive evolutionary changes in RNA binding. Simply put, conservation of context is a critically important mediator of conserved RNA-protein interactions.

Intra-species binding patterns are dependent on cellular factors

To compare these binding preferences to intra-species changes, we examined available iCLIP data from HeLa cells overexpressing UNK from the same study[27]. Only genes with greater than 5 TPM expression in both cell lines and one-to-one orthologs across species were included. When looking at transcript-level conservation, we observed that approximately 51% of UNK transcripts were bound in both cell types (Supplementary Fig. 3G; $p = 2.3e\text{-}202$, 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 (Supplementary Fig. 3H), 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 (Supplementary Fig. 3I). 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] (Supplementary Fig. 3J). 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, accounting for only genes with similar expression across samples, 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 (Supplementary Fig. 3K). As expected, peaks with well-defined motifs displayed a greater degree of overlap between cell lines (Supplementary Fig. 3L, M). These observations are similar to what we observed for UNK between SH-SY5Y versus HeLa cells with iCLIP (Fig S3G, H). 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.).

Sequence contextual changes impact species-specific 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. On average, 18 chimeras for UAG Change and 24 chimeras for Context Change were considered per position. 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; $p \leq 0.001$; one-sided, paired Wilcoxon test). 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. Enhancing chimeric sequences—mostly downstream of the core motif—tend to be U/A rich (Supplementary Fig. 4A), likely leading to increased avidity and further engagement of UNK's secondary RBD[46]. We used all natural (i.e. non-mutated, non-chimeric) sequences to create a linear model (Supplementary Fig. 4B–D) and found that UAG has the strongest positive correlation with enrichment, with a coefficient of 0.55. Additionally, U/A rich 3mers also had positive and significant contributions, highlighting the importances of downstream motifs in binding. Further, GC had a strong negative correlation with enrichment, highlighting the importance of structure (or lack thereof) to binding.

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 (Supplementary Fig. 4E–G). 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 (Supplementary Fig. 4H,I; $p \leq 0.01$, one-sided, paired Wilcoxon test).

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 FP with a 6-FAM-tagged RNA and fit a K_d for human *GTPB4* at 262.6 nM, mouse *Gtbp4* at >5 μM, 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.

Sequence differences across 100 vertebrates affect UNK-RNA interactions in vitro

To expand our phylogenetic scope beyond human/mouse, we investigated binding patterns across 100 vertebrates. Selecting the top 250

https://doi.org/10.1038/s41467-024-52231-7

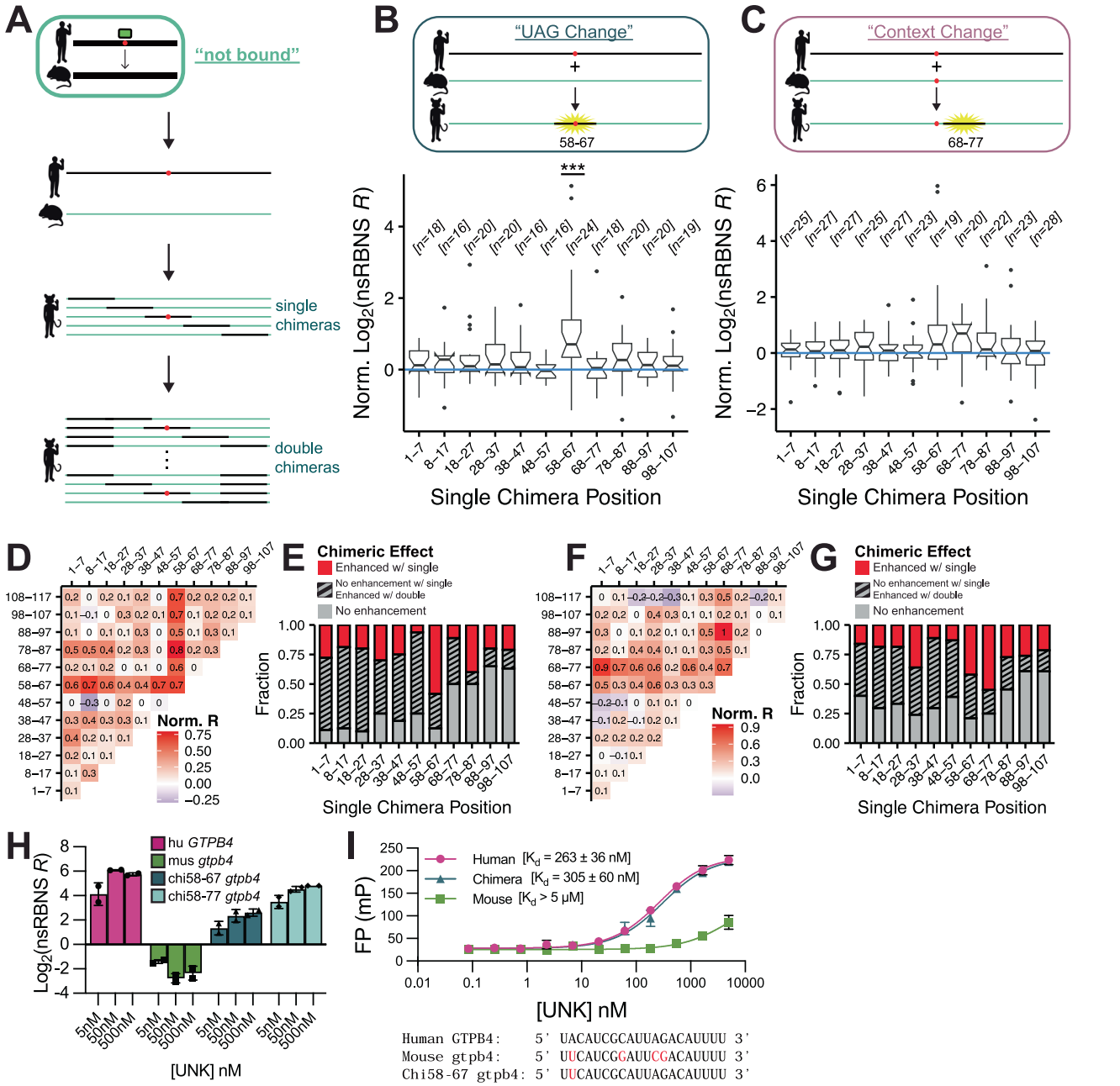

Fig. 4 | Analysis of regional impacts on binding. A Design of single and double chimera oligos. **B** Design and box and whisker plot of normalized $\log_2$ nsRBNS enrichment (chimera/wt) for UAG Change single chimeras. Significance was determined via paired, one-sided Wilcoxon test and corrected for multiple comparisons via the BH procedure. Chimerization at positions 58–67 was found to be significant ($p = 0.0005$). Centre line denotes median (50th percentile) with bounds of box representing 25th to 75th percentiles and the whiskers denoting 5th to 95th percentiles. Outliers are denoted as individual points. **C** Design and box and whisker plot of normalized $\log_2$ nsRBNS enrichment (chimera/wt) for Context Change single chimeras. Significance was determined via paired, one-sided Wilcoxon test and corrected for multiple comparisons via the BH procedure. Following multiple comparison correction, no positions were determined to be significant. Centre line denotes median (50th percentile) with bounds of box representing 25th to 75th percentiles and the whiskers denoting 5th to 95th percentiles. Outliers are denoted as individual points. **D** Heat map of median normalized $\log_2$ nsRBNS 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 $\log_2$ nsRBNS 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** $\log_2$ nsRBNS enrichment values ($n = 2$) for human, mouse, single chimera, and double chimera *GTPB4* at 5, 50, and 500 nM UNK. Data are presented as mean values ± SD. **I** Fluorescence polarization binding curves ($n = 3$) for human *GTPB4* (purple circle), mouse *gtpb4* (green square), and chi58-67 *gtpb4* (blue triangle) RNA oligos incubated with UNK. Data are presented as mean values ± SD.

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 and 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 and found robust and reproducible binding (Supplementary Fig. 5B; R = 0.97, Pearson's correlation). 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

https://doi.org/10.1038/s41467-024-52231-7

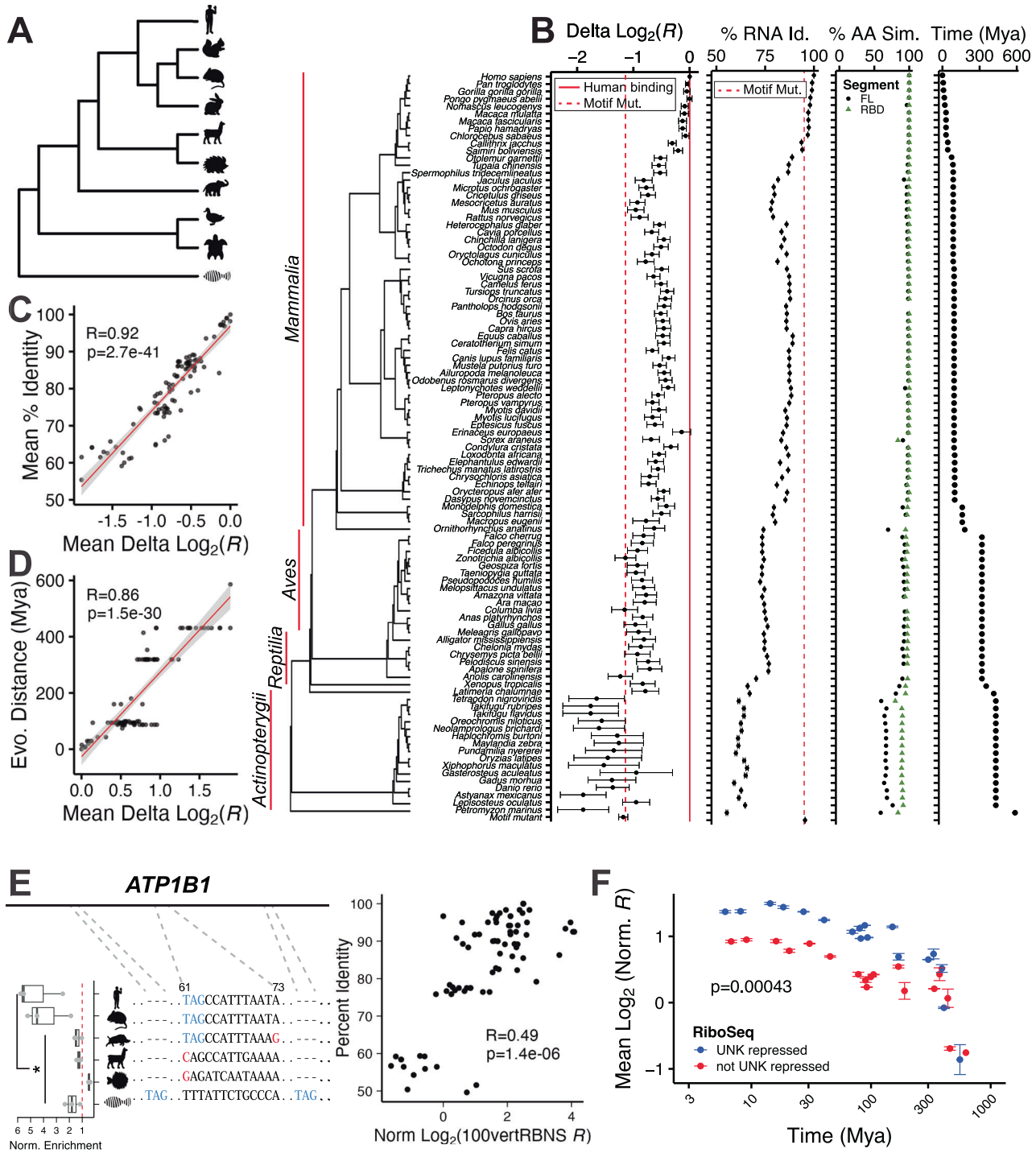

Fig. 5 | Evolutionary conservation of binding. A Simplified tree schematic of vertebrates used for natural sequence RBNS (not all species shown). **B** Delta $\log_2$ 100vertRBNS enrichment, percent RNA sequence identity, percent UNK similarity (full length-grey and RBDs-green), and evolutionary distance in millions of years against 100 vertebrates for the aligned sequences from the top human bound oligos. Red dotted line shows average for total motif mutant. Red solid line shows average for human binding. Error bars show standard error of the mean (SEM). **C** Mean percent RNA sequence identity (*Y* axis) versus mean delta $\log_2$ 100vertRBNS enrichment (*X* axis) for each aligned oligo. Pearson's correlation coefficient and *p*val included. Line is presented as mean fit ± SEM. **D** Evolutionary distance in millions of years (*Y* axis) versus mean delta $\log_2$ 100vertRBNS enrichment (*X* axis) for each aligned oligo. Pearson's correlation coefficient and *p*val included. Line is presented as mean fit ± SEM. **E** (left) Multiple sequence alignment for *ATP1B1* for *Homo sapiens, Mus musculus, Sus scrofa, Vicugna pacos, Tetradon nigroviridis,* and *Danio rerio* with normalized 100vertRBNS enrichment by species ($n = 3$). Significance determined via one-sided, paired Wilcoxon tests. Significance marks are as follows: * ($p \leq 0.05$). Centre line denotes median (50th percentile) with bounds of box representing 25th to 75th percentiles and the whiskers denoting 5th to 95th percentiles. All data included as individual points. (right) Percent RNA sequence identity (*Y* axis) versus normalized delta $\log_2$ 100vertRBNS enrichment (*X* axis). Pearson's correlation coefficient and *p*val included. **F** Scatter plot of $\log_2$ normalized 100vertRBNS 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 a two-sided KS test.

https://doi.org/10.1038/s41467-024-52231-7

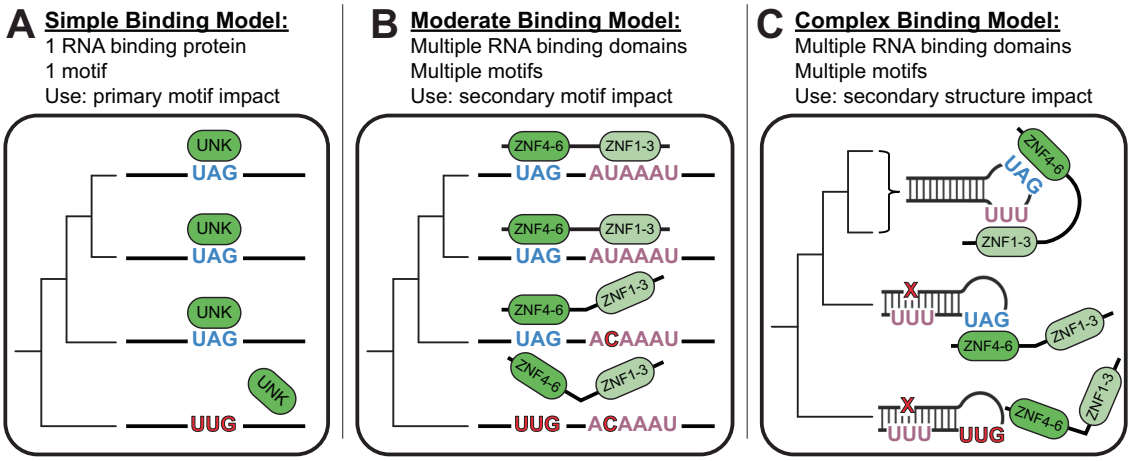

Fig. 6 | Models of RNA binding. A Simple binding model: considers only primary motifs. **B** Moderate binding model: considers primary and secondary motifs. **C** Complex binding model: considers primary and secondary motifs as well as RNA secondary structure.

wild-type human sequences are better bound across the assay than mutant counterparts (Supplementary Fig. 5C; $p \leq 0.0001$, one-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 decreases (Fig. 5B). To understand the driving force behind loss or maintenance of binding, we computed RNA 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 high degree of correlation between mean percent identity to human and binding enrichments (Fig. 5C; $R = 0.92$, Pearson's correlation). Similarly, evolutionary distances are also well-correlated with binding (Fig. 5D; $R = 0.86$, Pearson's correlation). As expected evolutionary distance and RNA percent identity are also very highly correlated (Fig. 5SD; $R = -0.94$), suggesting a strong relationship between the three (Supplementary Fig. 5E). A large drop-off in binding with increased variance was observed in fish (*class Actinopterygii*) (Fig. 5B, left), suggesting a loss of human UNK and RNA target 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). To examine UNK sequence conservation more closely, we aligned UNK ZnF1-6 amino acid sequences across 100 vertebrates (Supplementary Fig. 5F). For all but one RNA-contacting residue[46], the amino acid sequence is highly conserved, with only slight divergence to similar amino acids as annotated by BLAST[48] (Supplementary Fig. 5G, H).

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 and Supplementary Fig. 5I, J). 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 (Supplementary Fig. 5K–M). While many primate sequences were perfectly identical, we observed with *PPP2R5C* that even across short evolutionary distances, large motif changes can occur, leading to drastic changes in binding (Supplementary Fig. 5N, O).

To examine how binding changes across 100 vertebrates correlates with RNA regulation, we once again turned to ribosome profiling data[47]. 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; $p \leq 0.001$, KS test), spanning large evolutionary distances. This effect can be explained by the fact that translational targets were modestly more conserved than those that are not translational targets (Supplementary Fig. 5P).

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 control[27]) 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 Supplementary 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 previously[26]); additional components of UNK, such as its disordered domain, may impact binding in unexpected ways[47]. (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 Supplementary 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 (Supplementary Fig. 5B; see Supplementary Note 2). (vi) The nature of oligo pool design removed sequences with poor alignments. Therefore, the percent identity analyses 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. 1A). Using a high-throughput in vitro assay to test binding, we found that species-specific binding patterns and regulation can be partially explained by biochemically measurable RNA-protein interactions (Figs. 2 and 3). Although more limited in scope, binding changes across cell types within species appear to be driven cellular context (e.g., the *trans* environment) (Supplementary Fig. 3). 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 and Supplementary Fig. 5I–O).

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 binding[27]. 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 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, the most "complex binding model," considers RNA secondary structure as well, global structural rearrangements may affect motif access and unlike sequence structures may be dynamically regulated in specific context (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 TFs[39]). However *cis* evolution can also result in *trans* changes as regulatory elements like TFs and RBPs also have self-regulatory features[70]. This is especially true for RBPs genes that produce multiple mRNA isoforms via alternative splicing that result in proteins with different functions. Splicing itself is species-specific and under the control of *cis* regulatory elements[71,72]. 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 mRNAs[27]. In vitro, UNK bound UTR sequences more tightly than CDS (Figs. 2 and 3), driven by the presence of more UNK motifs in UTRs (Supplementary Fig. 2A). However, in cells, UNK binds primarily to CDS[27]; this preference may be driven by increased local concentration of UNK in the CDS, resulting from its association with ribosomes[46]. UNK is not unique in exhibiting gene region preference[73] and the mechanisms driving these preferences are not well understood. One recent study has demonstrated that UNK can associated with CCR4-NOT on some targets[47], thus additional cellular factors like these may impact target selection.

The work presented here parallels studies performed on a handful of transcription factors[6,7,38,74–76]. For example, Schmidt and coworkers[6] 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 species[6]. 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

Plasmids pGEX-GST-SBP-UNK-ZnF1-3 and pGEX-GST-SBP-UNK-ZnF4-6 were cloned from pGEX-GST-SBP-UNK (30–357)[26] with primers listed in Supplementary Table 1. All plasmids are available through Addgene. UNK_6xZNF_pGEX is associated with Dominguez et al.[26]. UNK_ZNF1-3_pGEX and UNK_ZNF4-6_pGEX have been deposited under this manuscript. Plasmids were 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 4000 × *g* for 15 minutes and lysed in lysis buffer (200 mM NaCl, 5 mM DTT, 50 mM HEPES, 3 mM $MgCl_2$, 2 mM PMSF, 1 Pierce™ 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 × *g* for 30 minutes. 0.05% polyethyleneimine (PEI) was added to precipitate excess nucleotides and was centrifuged at 17,800 × *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 $MgCl_2$), and high salt buffer (1 M NaCl, 50 mM HEPES). For SBP-UNK ZnF1-3, SBP-UNK ZnF4-6, and 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. 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). Purity for SBP-UNK ZnF1-3 (22.1 kDa) and SBP-UNK ZnF4-6 (20.0 kDa) was assessed following GST-cleavage via SDS-PAGE (4–12% gradient) and Coomassie staining.

GST-SBP-UNK (30–357) and SBP-UNK (30–357) were concentrated via centrifugation to 500 μL 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. All constructs were 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-2279[27]. Mouse coordinates were converted from mm9 to mm10 and human coordinates were converted from hg19 to hg38 using liftOver[77] (version 1.24.0) in RStudio[78] (version 2023.03.0) with R platform[79] (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 'AnnotationHub[80]' (version 3.8.0) with 'BSGenome.Hsapiens.NCBI.GRCh38[81]' and 'BSGenome.Mmusculus.UCSC.mm10[82]'. 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 RNAseq data for SH-SY5Y cells (Murn et al.[27]: E-MATB-2277[83]) mouse brain tissue (ENCODE[60,84]: ENCSR000BZJ), and HeLa cells (ENCODE[60,84]: ENCSR552EGO) was mapped to the mm10 or hg38 genomes using STAR[85] (version 2.7.10b) with default parameters. RSEM was used to calculate gene level expression values. Tximport was used to read RSEM output and TPM was used for comparison. Peaks were filtered to genes with ≥ 5 TPM in the respective cell lines. Gene level intersecting peaks (SH-SY5Y) were converted from hg38 to mm10 using the liftOver utility from UCSC[86] then intersected with mouse iCLIP peaks using BEDtools[87] (version 2.31.0). RStudio package 'VennDiagram[88]' was used to produce Venn diagrams.

For final oligo pool, sequences were expanded to 170 nucleotides. Using previously reported RBNS data for UNK[26], sequences were recentered around the highest ranking *kmer* closest to the center. Final sequences were trimmed down to 120 nucleotides (Supplementary Fig. 1A). 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 was not always perfect. 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 cells[27] (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 RBNS binding buffer I (25 mM Tris pH 7.5, 150 mM KCl, 3 mM $MgCl_2$, 0.01% tween, 500 μg/mL BSA, 1 mM DTT) and incubated with 0, 5, 50, or 500 nM recombinant GST-SBP-UNK. After 30-minute incubation, 1 μM RNA was added to the reaction. After 1 hour, UNK-RNA complexes were isolated and unbound RNA was washed away in RBNS wash buffer I (25 mM Tris pH 7.5, 150 mM KCl, 0.01% tween). Complexes were eluted in 4 mM biotin. The eluted RNA was reverse transcribed with Superscript III (Invitrogen) with RBNS RT primer (IDT; Supplementary Table 2), amplified by PCR with Phusion DNA polymerase (NEB) with RBNS index primers and RBNS reverse primer I (Supplementary Table 2), and sequenced on an Illumina HiSeq 2000 instrument.

nsRBNS mapping and enrichment analysis

Reads were trimmed using fastx_toolkit[89] (version 0.0.14) as needed. Mapping was performed with STAR[85] (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 seqtk[90] (version 2.3.0) as needed. SAMtools[91] (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 RStudio[78] (version 2023.03.0) with R platform[79] (version 4.2.2). R packages ggplot2[92] (version 3.4.1), 'ggpattern[93]' (version 1.0.1), 'ggpubr[94]' (version 0.6.0), and 'ggrepel[95]' (version 0.9.3) were used to make publication figures. Other RStudio packages—including 'cowplot[96]' (version 1.1.1), 'dplyr[97]' (version 1.1.0), 'flextable[98]' (version 0.9.6), 'grid[78]' (version 4.3.1), 'Hmisc[99]' (version 4.8.0), 'lsr[100]' (version 0.5.2), 'magick[101]' (version 2.8.3), 'msa[102]' (version 1.30.1), 'org.Hs.eg.db[103]' (version 3.17.0), 'reshape2[104]' (version 1.4.4), 'rstatix[105]' (version 0.7.2), and 'stringr[106]' (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 Supplementary Data 1.

Linear modeling

To predict the nsRBNS enrichment value of UNK binding, we used a linear model based on selected 3mer frequencies and the GC content. Due to previously reported RBNS data for UNK[26], only A/U-rich 3mers were considered in the model. The GC content was incorporated to account for structural information that may influence binding. Modeling was performed in RStudio[78] with the built-in 'lm' function to predict the log-scale *r* value, excluding the intercept. This approach allowed us to directly assess how the selected 3mers and GC content contribute to the enrichment of UNK binding against natural sequences (i.e. motif mutants and chimeras were excluded).

Random RBNS

RBNS was performed as previously described with slight modifications[26]. SBP-UNK ZnF1-3 or SBP-UNK ZnF4-6 were incubated with beads and 20mer random RNA in RBNS binding buffer II (25 mM Tris, pH 7.5, 150 mM KCl, 3 mM $MgCl_2$, 0.01% triton X-100, 500 μg/mL BSA, 20 units/mL SUPERase·In (Thermo Fisher)) then washed in RBNS wash buffer II (25 mM Tris pH 7.5, 150 mM KCl, 0.01% triton X-100, 20 units/mL SUPERase·In). Proteins were incubated at 250, 500, or 1000 nM. 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.

Enrichments were calculated as detailed in Dominguez et al.[26]. The top 20 6mers were used to generate logos. Each 6mer was aligned to the top enriched, allowing for one mismatch and/or one offset, or two mismatches. A Position Weight Matrix (PWM) was constructed with the aligned 6mers, where the enrichment values were used to add weight to each position. Logos were trimmed if the edges of the PWM had minimal aligned sequences. Final logos were plotted in RStudio[78] (version 2023.03.0) with R package 'ggseqlogo[107]' (version 0.2).

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 RiboSeq log_2 fold changes were used as RiboSeq data for mouse is not currently available.

Mean base pair probability analysis

DNA sequences for all hg38 genes were obtained from Ensembl[108]. Genes not bound in human neuronal cells as identified by iCLIP[27] were selected for subsequent analysis. For genes with multiple isoform sequences, only one was kept: This was done randomly with RStudio function "sample." 120 control 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 RNAfold[55] --partfunc to calculate the partition function and base pairing probability matrix for both the CDS controls as well as "perfectly conserved," "binding 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 BLAT[109] 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 module[110] (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'[111,112] (version 1.30.1) and trimmed back to 120 nucleotides. RStudio packages 'stringr'[106] (version 1.5.0) and 'data.table'[113] (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. This resulted in a total of 5753 oligos aligning to 112 human sequences.

100 vertebrate RNA bind-n-seq

RBNS was performed similarly as described above and previously[26,52]. An array of 5753 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 H_2O, and concentrated with phenol chloroform extraction.

Binding reactions were performed as detailed above (see *Random RBNS*) for individual domain constructs with slight modifications. 100 nM recombinant SBP-UNK was used. Sequencing was performed on an Illumina NextSeq 500.

100 vertebrate nsRBNS mapping and enrichment analysis

Mapping was performed as detailed above for nsRBNS. Inputs with less than 25 counts mapped were excluded from analysis. Enrichment analysis was performed as detailed for nsRBNS.

100 vertebrate nsRBNS data analysis

Data was analyzed similarly to nsRBNS data as discussed above with a few additions. RStudio package 'ape'[114] (version 5.7) was used to assemble 100 vertebrate phylogenic tree according to available data from UCSC[64,115]. Additionally, RStudio package 'msa'[111,112] (version 1.30.1) was used for sequence alignment and percent identity analysis. RStudio packages 'ggmsa'[116] (version 1.3.4), 'ggprism'[117]' (version 1.0.4), and 'scales[118]' (version 1.2.1) were used to make 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 Supplementary Data 2. For protein percent similarity analysis, human protein sequences were pulled from UniProt[50] and BLAST[48] 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[norm species/norm human]).

eCLIP peak overlap analysis

RNAseq data for HepG2 and K562 normal cell lines were downloaded from ENCODE[60,84] (Supplementary Table 3) and mapped to hg38 with STAR[85] (version 2.7.10b) using default parameters. Expression values were quantified using RSEM[119] (version 1.3.1) and differential analysis was conducted using RStudio package 'DESeq2'[120] (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 log_2 fold change (L2FC) <=1.

14 RBPs which have eCLIP peaks[61] in HepG2 and K562 and an RBNS motif[26] were selected. The peaks were downloaded from ENCODE (Supplementary Table 4) and filtered for enrichment of log_2 fold change > 1 and p-value of <0.001. Each peak was extended by 50 base pairs upstream to account for experimental limitations of eCLIP[61]. 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. BEDtools[87] (version 2.31.0) was used to collapse the overlapping peaks that were within 20 base pairs into a single peak spanning the region. Fold change was summed for collapsed peaks. Sequences under the combined peaks were taken from RStudio package 'BSgenome.Hsapiens.NCBI.GRCh38'[81] (version 1.3.1). Peaks were grouped into two groups based on presence or absence of RBNS motif within the peak[26].

Peaks were further grouped based on whether they overlapped exons or not. Exon annotations for hg38 (v109) genome were obtained from RStudio package 'AnnotationHub'[80] (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 (Supplementary Fig. 3F).

Fluorescence polarization (FP)

Synthetic UAG oligos. RNA oligos were synthesized by Integrated DNA Technologies (IDT) with a 3′ 6-FAM label (Supplementary Table 5) and incubated at 5 nM with serially diluted recombinant SBP-UNK ZnF1-3, SBP-UNK ZnF4-6, or SBP-UNK (30–357) (16.9, 50.8 pM, 1.5, 4.6, 13.7, 41.2, 123, 370 nM, 1.11 3.33, 10 μM) or for tri-UAG FP RNA with serially diluted recombinant SBP-UNK (30–357) (2.5, 7.6, 22.9, 68.6 pM, 6.17, 18.5, 55.6, 167, 500, or 1500 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 × *g* for 1 min and fluorescence polarization was measured with a PHERAstar plate reader (BMG Labtech) at 25 °C. FP values were normalized within sample to account for minimum and maximum FP, resulting in ΔFP. Data were fit to a single site binding model and a K_D was determined.

GTPB4 chimeras. FP was performed as above with slight modifications. SBP-UNK (30–357) was serial diluted as follows: 10.9, 50.8, 152, 457 pM, 1.37, 4.12, 12.3 37.0, 111, 333, or 1000 nM. FP values were considered, not ΔFP. Oligo sequences are available in Supplementary Table 6.

In vitro transcription of RNA for qPCR binding assay

DNA fragments for wild-type and mutant *GART* were ordered from IDT and PCR amplified with Phusion DNA polymerase (New England Biolabs), 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 (Supplementary Table 7).

qPCR binding assay

We performed our qPCR-based binding assay as previously described[121] 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 $MgCl_2$, 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 $MgCl_2$, 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 magnetic 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 (Supplementary Table 8). The threshold cycle (C_t) was determined using Bio-Rad's CFX Maestro software, and fraction bound was determined against the RNA calibration curve.

Statistical analyses

Individual statistical analyses are detailed in figure legends. For iCLIP gene overlaps, hypergeometric tests were used where the universe was defined as only one-to-one orthologous genes expressed in both cell lines at greater than 5 TPM. For correlation plots, Pearson's correlation was used and *p* vals shown are for the correlation. For wild type versus mutant group and chimerized comparisons, paired, one-sided Wilcoxon tests were used as previous data for motif mutants have demonstrated diminished binding[27] and recovery was expected for chimerization. For orthologous group comparisons, paired Wilcoxon tests were used. For orthologous and wild type versus mutant single transcript comparisons, one-sided Wilcoxon tests were used. For all other population comparisons, KS tests were used. Where multiple comparisons were done, *p* vals were corrected via the BH procedure based on number of comparisons.

Reporting summary

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

Data availability

The enrichments for RBNS, nsRBNS, and 100vertRBNS generated in this study have been deposited in Gene Expression Omnibus under accession code GSE262560. The raw fastqs are also available under accession code GSE262560. Previously published data used in this study have been accessed through the following sources: UNK iCLIP-seq data[27]: E-MTAB-2279 [https://www.ebi.ac.uk/biosamples/samples/SAMEA2341158], ENCODE K562 RNAseq[61]: ENCFF267RKD, ENCFF455VYN, ENCFF606ZTR, ENCFF444KCV, ENCODE HepG2 RNAseq[61]: ENCFF713MNU, ENCFF478DZZ, ENCFF936SLY, ENCFF446UEC, ENCODE K562 eCLIP[61]: ENCFF077OSY, ENCFF121XCN, ENCFF127WMZ, ENCFF150ZOO, ENCFF185IDD, ENCFF241AOZ, ENCFF348TPU, ENCFF374XQF, ENCFF401YRZ, ENCFF402AIE, ENCFF409DPS, ENCFF443KJS, ENCFF526OQL, ENCFF565ILV, ENCFF606RXB, ENCFF613UUR, ENCFF618ZPP, ENCFF664RLU, ENCFF669TNM, ENCFF674TKN, ENCFF766DUS, ENCFF779OIO, ENCFF824IDO, ENCFF853FGC, ENCFF860QZG, ENCFF899HGF, ENCFF910WLP, ENCFF996BXS, ENCODE HepG2 eCLIP[61]: ENCFF073PCD, ENCFF082QGS, ENCFF103PRM, ENCFF105GZJ, ENCFF145YYK, ENCFF230QOU, ENCFF253ZSN, ENCFF288MWL, ENCFF327JJE, ENCFF378HWC, ENCFF383ZAQ, ENCFF390PJW, ENCFF421FJD, ENCFF432ASF, ENCFF502OYV, ENCFF534YQS, ENCFF545NBF, ENCFF611AHG, ENCFF626XAA, ENCFF685MZA, ENCFF705SDK, ENCFF754XAQ, ENCFF856EHA, ENCFF899ZEH, ENCFF914VUW, ENCFF951IBI, ENCFF966KQG, ENCFF988MWD.

Code availability

All scripts for nsRBNS and 100vertRBNS table assembly, data processing, and figure generation are available at https://github.com/DominguezRNAGroup/Binding_Site_Evolution.

References

1. Wagner, A. Robustness and evolvability: a paradox resolved. *Proc. R. 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).
4. 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).
5. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev. Biol.* **46**, 111–138 (1971).
6. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
7. Odom, D. T. et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39**, 730–732 (2007).
8. Wang, Z.-Y. et al. Transcriptome and translatome co-evolution in mammals. *Nature* **588**, 642–647 (2020).
9. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
10. 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).
11. Mitsis, T. et al. Transcription factors and evolution: an integral part of gene expression. *World Acad. Sci. J.* **2**, 3–8 (2020).
12. 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).
13. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).
14. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
15. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
16. Brinegar, A. E. & Cooper, T. A. Roles for RNA-binding proteins in development and disease. *Brain Res.* **1647**, 1–8 (2016).
17. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
18. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.* **3**, 195–205 (2002).
19. Orphanides, G. & Reinberg, D. A unified theory of gene expression. *Cell* **108**, 439–451 (2002).
20. Lasko, P. Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. *Sci. STKE* **2003**, re6–re6 (2003).
21. Achsel, T. & Bagni, C. Cooperativity in RNA–protein interactions: the complex is more than the sum of its partners. *Curr. Opin. Neurobiol.* **39**, 146–151 (2016).
22. Kitov, P. I. & Bundle, D. R. On the nature of the multivalency effect: a thermodynamic model. *J. Am. Chem. Soc.* **125**, 16271–16284 (2003).
23. Stefan, M. I. & Le Novère, N. Cooperative binding. *PLoS Comput. Biol.* **9**, e1003106 (2013).
24. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell* **78**, 9–29 (2020).
25. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).
26. Dominguez, D. et al. Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* **70**, 854–867 (2018).
27. Murn, J. et al. Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. *Genes Dev.* **29**, 501–512 (2015).
28. Robert, F. & Pelletier, J. Exploring the impact of single-nucleotide polymorphisms on translation. *Front. Genet.* **9**, 507 (2018).
29. Shatoff, E. & Bundschuh, R. Single nucleotide polymorphisms affect RNA-protein interactions at a distance through modulation of RNA secondary structures. *PLoS Comput. Biol.* **16**, e1007852 (2020).
30. Gerstberger, S., Hafner, M., Ascano, M. & Tuschl, T. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv. Exp. Med. Biol.* **825**, 1–5 (2014).
31. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593–1599 (2012).
32. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
33. 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).
34. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16**, 1–11 (2015).
35. 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. Biol. Evol.* **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 Biol.* **13**, e1002307 (2015).
41. Najdrová, V., Stairs, C. W., Vinopalová, M., Voleman, L. & Doležal, P. The evolution of the Puf superfamily of proteins across the tree of eukaryotes. *BMC Biol.* **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. J. Mol. Sci.* **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. Commun.* **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 (2016).
47. Shah, K. et al. Regulation by the RNA-binding protein Unkempt at its effector interface. *Nat. Commun.* **15**, 3159 (2024).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
49. Wheeler, D. L. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **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 (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. Struct. Mol. Biol.* **27**, 901–912 (2020).
55. Lorenz, R. et al. ViennaRNA package 2.0. *Algorithms Mol. Biol.* **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. *Biochim. Biophys. Acta* **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. Biol. Evol.* **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. Struct. Mol. Biol.* **17**, 909–915 (2010).
63. Hafner, M. et al. CLIP and complementary methods. *Nat. Rev. Methods Prim.* **1**, 20 (2021).
64. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
65. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
66. Torres-Méndez, A. et al. Parallel evolution of a splicing program controlling neuronal excitability in flies and mammals. *Sci. Adv.* **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. Eur. 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. Gueroussov, S. et al. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. *Cell* **170**, 324–339 (2017).
72. Gueroussov, S. et al. An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349**, 868–873 (2015).
73. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
74. Paris, M. et al. Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS Genet.* **9**, e1003748 (2013).
75. Bradley, R. K. et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol.* **8**, e1000343 (2010).
76. Wong, E. S. et al. Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nat. Commun.* **8**, 1092 (2017).
77. Bioconductor Package Maintainer. liftOver: changing genomic coordinate systems with rtracklayer::liftOver. Available at: https://doi.org/10.18129/B9.bioc.liftOver (2023).
78. RStudio, T. RStudio: integrated development for R. *Rstudio Team, PBC*, Boston, MA http://www.rstudio.com (2020).
79. R Core Team, R. R: a language and environment for statistical computing. (2013).
80. Morgan, M. & Shepherd, L. AnnotationHub: Client to access AnnotationHub resources. Available at https://doi.org/10.18129/B9.bioc.AnnotationHub (2023).
81. Team, T. B. D. BSgenome.Hsapiens.NCBI.GRCh38: full genome sequences for Homo sapiens (GRCh38). Available at: https://www.bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.NCBI.GRCh38.html (2014).
82. Team, T. B. D. BSgenome.Mmusculus.UCSC.mm10: full genome sequences for Mus musculus (UCSC version mm10, based on GRCm38.p6). Available at: https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Mmusculus.UCSC.mm10.html (2021).
83. Sarkans, U. et al. The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* **46**, D1266–D1270 (2018).
84. Luo, Y. et al. New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
85. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
86. Lee, B. T. et al. The UCSC genome browser database: 2022 update. *Nucleic Acids Res.* **50**, D1115–D1122 (2022).
87. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
88. Chen, H. VennDiagram: generate high-resolution venn and euler plots. Available at: https://cran.r-project.org/package=VennDiagram (2022).
89. Liu, X. et al. FastProNGS: fast preprocessing of next-generation sequencing reads. *BMC Bioinform.* **20**, 1–6 (2019).
90. 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).
91. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
92. Villanueva, R. A. M. & Chen, Z. J. ggplot2: elegant graphics for data analysis. Available at: https://ggplot2-book.org/ (2019).
93. FC, M., Davis, T. L. & ggplot2 authors. ggpattern: 'ggplot2' pattern geoms. Available at https://cran.r-project.org/package=ggpattern (2022).
94. Kassambara, A. ggpubr: 'ggplot2' based publication ready plots. Available at: https://cran.r-project.org/package=ggpubr (2023).
95. Slowikowski, K. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. Available at: https://cran.r-project.org/package=ggrepel (2023).
96. Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2'. Available at: https://CRAN.R-project.org/package=cowplot (2020).

97. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: a grammar of data manipulation. Available at: https://cran.r-project.org/package=dplyr (2023).
98. Gohel, D. & Skintzos, P. flextable: functions for tabular reporting. Available at: https://CRAN.R-project.org/package=flextable (2024).
99. Harrell Jr, F. E. Hmisc: Harrell miscellaneous. Available at: https://cran.r-project.org/package=Hmisc (2023).
100. Navarro, D. Learning statistics with R: a tutorial for psychology students and other beginners. (Version 0.6). Available at: https://learningstatisticswithr.com (2015).
101. Ooms, J. *magick: advanced graphics and image-processing in R*. Available at https://CRAN.R-project.org/package=magick (2024).
102. Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999 (2015).
103. Carlson, M. org.Hs.eg.db: genome wide annotation for human. Available at: https://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/org.Hs.eg.db/html/00Index.html (2023).
104. Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20 (2007).
105. Kassambara, A. rstatix: pipe-friendly framework for basic statistical tests. Available at https://cran.r-project.org/package=rstatix (2023).
106. Wickham, H. & Wickham, M. H. Package 'stringr'. Website: http://stringr.tidyverse.org, https://github.com/tidyverse/stringr (2019).
107. Wagih, O. ggseqlogo: a 'ggplot2' extension for drawing publication-ready sequence logos. Available at: https://CRAN.R-project.org/package=ggseqlogo (2017).
108. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
109. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
110. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).
111. May, A. C. W. Percent sequence identity: the need to be explicit. *Structure* **12**, 737–738 (2004).
112. Raghava, G. P. S. & Barton, G. J. Quantification of the variation in percentage identity for protein sequence alignments. *BMC Bioinform.* **7**, 1–4 (2006).
113. Dowle, M. & Srinivasan, A. Data.table: extension of 'data.frame'. Available at: https://cran.r-project.org/package=data.table (2023).
114. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
115. Murphy, W. J. et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
116. Zhou, L. et al. ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Brief. Bioinform.* **23**, bbac222 (2022).
117. Dawson, C. ggprism: a 'ggplot2' extension inspired by 'GraphPad Prism'. Available at: https://cran.r-project.org/package=ggprism (2022).
118. Wickham, H. & Seidel, D. scales: scale functions for visualization. Available at: https://cran.r-project.org/package=scales (2022).
119. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 1–16 (2011).
120. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
121. 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.; validation, S.E.H., M.S.A., G.G., M.M.A., and D.D.; formal analysis, S.E.H., M.S.A., Y.H., 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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52231-7.

Correspondence and requests for materials should be addressed to Daniel Dominguez.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.