nature portfolio

Peer Review File

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

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In the manuscript entitled “Understanding species-specific and conserved RNA-protein interactions in vivo and in vitro”, the authors evaluated evolutionary conservation and properties of RBP-RNA interaction sites. They examined and compared the in vivo and in vitro binding of the neuronal RNA-binding protein Unkempt (UNK) mostly focusing on human and mouse. While they found conserved transcript binding for around 45% between species, the binding within transcripts were less conserved. To understand the underlying mechanism of species-specific binding, they mainly utilized in vitro RNA-bind-n-seq (RBNS) data. They propose that contextual sequence and structural features are important contributors to binding-site turnover. The authors further found that there is correlation between evolutionary distance, individual binding site conservation and UNK binding strength. They ultimately propose three insightful models to explain differences in species-specific. The “moderate binding” and “complex binding” models invoke a combination of multiple RNA binding domains, motifs, and secondary structure. However, these the direct impact of the multiple RNA binding domains is not explicitly tested. Despite these concerns, this study represents one of the most thorough examinations of species specific RBP-RNA interactions (excluding miRNA binding sites).

Major:
1. Line 203 and figure 2: The authors conclude “Thus, nsRBNS captures binding features derived from in vivo CLIP”. While this statement is supported for 3’ UTR binding sites, it is not for the CDS, which the authors acknowledge in their discussion.

2. In this same section the authors suggest the discrepancy for CDS binding could be due to differences in sequence composition between CDS and 3’ UTR. However, it is difficult determine this from the plots in Fig S2A as there is no direct comparison e.g. scatter plot of UTR vs CDS frequencies. Moreover, the contribution of UAG, UAA, and UUU are independently evaluated (separate CDF plots). It would be beneficial for the authors to evaluate the relative contribution of each to the RBNS enrichment. This could be done using a linear model or partial correlation or similar methods.

3. In figure 2E and all other uses of ribosome profiling data, is the comparison (x-axis) only changes in ribo-seq data and not translational efficiency i.e. normalizing for changes in RNA levels? It should be TE or it should be demonstrated there is no UNK-dependent differential expression.

4. Line 333-334 and figure 6: The authors suggest that the secondary RBD of UNK engages with U/A rich downstream sequences of the core motif, but this was not explicitly tested. RBNS data of the UNK with deletions or mutations of the secondary RBDs would be one way to provide support for this model.

Minor:
1. It would be beneficial to change the result titles to match the main conclusion in each section.

2. Line 68: Clarify if the 95% identity for the whole TF or the DNA-binding domain

3. It was unclear if the results in Figure 1A/Line 120 only consider 1-1 orthologues

4. Line 125: It would be helpful to stick with clearly defined terminology i.e. conserved instead of homologous in text vs figure.

5. Are the results/conclusion of Figure 2D, S2D, S2E different for CDS vs UTR?

6. Line 209: Where is the 60% of binding sites mirrored the in vivo trend?

7. Figure S2B,C: in the figure, please indicate which species you're referring to.

8. Figure 2B, C: x-axis – enrichment of what? Inset Y-axis – enrichment of what? Presumably RNBS

9. Figure 3C + line 256: What's the difference between "perfectly conserved" vs "conserved" – how identical is conserved %-wise?

10. Figure 3C,D: Please clarify the definition of "CDS-all" and "CDS-motif conserved". Does that mean in C there is no motif in human? And in D there is motif?

11. It is clear that the perfectly conserved oligos are enriched the most. Line 261-262: "when only regions with UAG motifs in both human and mouse were considered". Does that mean that C does not have UAG in both? According to the figure that could be true. So that would mean that there binding sites without the core UAG motif were considered?

12. Figure 5D/Line 376: How much of the relationship between evolutionary distance and binding is explained by the difference in sequence identity?

Reviewer #2 (Remarks to the Author):

In this manuscript, Harris et al interrogate how post-transcriptional regulation evolves across species. To this end, the authors focus on an RNA-binding protein, Unkempt (UNK), and its RNA interactions in mouse and human. To determine what are the UNK binding sites in mRNA of these species, they use previous available CLIP data and perform RNA bind-n-seq in vitro.
The first immediate result is the realization that while UNK may bind the same transcript in both mouse and human, the position of the binding site is not conserved. This finding aligns with the principles of evolvability, that states that the core mechanisms of regulation are conserved, not the way to conduct them.
Next, the massive parallel library used in Bind-n-Seq allows the authors to confirm that a central UAG facilitates binding and that low secondary structure around this motif is important for the effect of the UAG motif in binding. The degree of conservation of this UAG

between human and mouse orthologus positions determined the strength of binding, followed up with the degree of identity in the surrounding ~120 nucleotides.
Among all the interspecies variation, the sites with more degree of conservation were still the more functional, as determined by the ability to repress the translation of the target. Interestingly, the authors determine that UNK binds 51% of the same transcripts in different cell types.
Next, in an elegant approach that takes advantage of the high-throughput capabilities of Bind-n-Seq, the authors focus on sites that were bound in human but not in mice, and exchange segments of the sites among species and determine the gain or loss of binding. The authors conclude that the most important region contributing to binding is the central region with the UAG motif and the positions downstream of it.
Finally, the authors expand their analysis to all available vertebrate sites of UNK and conclude that the strength of binding is correlated with sequence identity conservation and evolutionary distance. Still the most functional sites are the ones with deeper conservation in vertebrates.

Overall, this manuscript presents a deep functional analysis at how post-translational regulatory elements evolve across species. The high-throughput approach of the authors and their systematic analysis allows them to reach solid conclusions that are of interest to the board audience of Nature Communications. However, the paper cannot be accepted in the current format until the authors address some of the following comments.

Minor comments:

1.- In Figure 1A, when the authors analyze CLIP data to determine if a site is bound by UNK in human and/or mouse, it is not clear if the authors impose the rule that the gene analyzed must be expressed in the input of human and mouse samples. If a gene is not expressed in one of the samples, the absence of binding by CLIP is not informative. A similar situation occurs in figure S3C where the authors examine the conservation of binding between cell types. The authors should clarify the analysis and make sure that only analyze genes that are expressed in both CLIP datasets.

2.- The Methods sections does not have a section detailing the statistical analysis used thought the manuscript. It is also unclear why some cumulative plots have associated p values and others not. The statistical analysis and display should be standardized thought the manuscript and figures.

3.- It is not clear the number of oligos and the corresponding permutations represented in each Bind-n-Seq library. For Figure 1, the authors should specify the total number of sequences analyzed. For Figure 4, the authors should specify the total of single and double chimeras.

4.- The authors analyze Bind-n-Seq as the “frequency of an oligo in the protein bound sample divided by the frequency in the input”. It is not clear if “frequency” refers to number of reads or total number of normalized reads (normalized by the size of the library). The authors should clarify this point of the analysis.

5.- To understand to what degree we expect or not differences in the binding of UNK across species, it would be important to provide a protein sequence alignment, and a structure showing the binding domain where the divergent sites are highlighted.

6.- In Figure 5B, the red discontinuous lines are not defined in the figure legend and are not easy to interpret. Also, it is not intuitive that % identity refers to target RNA sequence and that %Similarity refers to UNK aminoacid sequences.

Notes: Our responses to reviewers are in blue and anything regarding text changes within manuscript are indented with changes denoted in red.

All sequencing data has been uploaded to GEO under GSE262560. The data is private pending manuscript acceptance; however, it can be reviewed with the following token: gdozociilbyxlin

Comments from Reviewer 1

Comment 1: *Line 203 and figure 2: The authors conclude "Thus, nsRBNS captures binding features derived from in vivo CLIP". While this statement is supported for 3' UTR binding sites, it is not for the CDS, which the authors acknowledge in their discussion… In this same section the authors suggest the discrepancy for CDS binding could be due to differences in sequence composition between CDS and 3' UTR. However, it is difficult determine this from the plots in Fig S2A as there is no direct comparison e.g. scatter plot of UTR vs CDS frequencies.*

The reviewers raise an important question regarding modeling in vivo binding in vitro. We too were surprised to see discrepancies between in vivo and in vitro preferences for transcript regions. UNK iCLIP clearly shows preferential binding to CDS over 3'UTR (at least in number of binding sites detected in each region) (Murn et al., 2015). However, it is a challenge to determine from iCLIP if the sites bound within UTRs are bound more strongly (e.g. greater affinity). Overall, we propose that the high A- and U- content of 3'UTRs is a driver of this feature. In vivo, the interaction between UNK and ribosomes (Murn et al., 2016) may promote CDS interactions, something that was not modeled in this study in vitro.

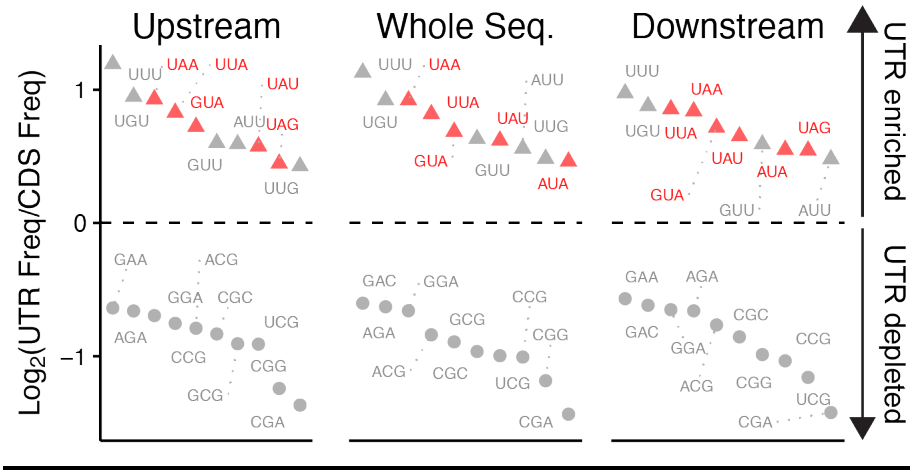
To more clearly convey the differences in composition we revised Fig S2A to reflect 3'UTR enriched or depleted 3mers as a $\log_2$ fold change over CDS frequency. We evaluated these for the whole sequence as well as separately for upstream and downstream of the central UAG. As is shown 3mers in red are those enriched in UNK RBNS experiments on randomized pools (performed in Dominguez et al., 2018).

Citations:

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., van Nostrand, E. L., & Pratt, G. A. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Molecular Cell, 70(5), 854–867.

Murn, J., Zarnack, K., Yang, Y. J., Durak, O., Murphy, E. A., Cheloufi, S., Gonzalez, D. M., Teplova, M., Curk, T., & Zuber, J. (2015). Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. Genes & Development, 29(5), 501–512.

Murn, J., Teplova, M., Zarnack, K., Shi, Y., & Patel, D. J. (2016). Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. Nature Structural and Molecular Biology, 23(1), 16–23. https://doi.org/10.1038/nsmb.3140



Legend: Scatter plot of the $\log_2$ *k*mer frequency fold change (UTR/CDS) of the top and bottom ten 3mers of all (left) motif-upstream (center) whole sequence and (right) motif-downstream sequences colored by UNK bound *k*mer as identified via RBNS.

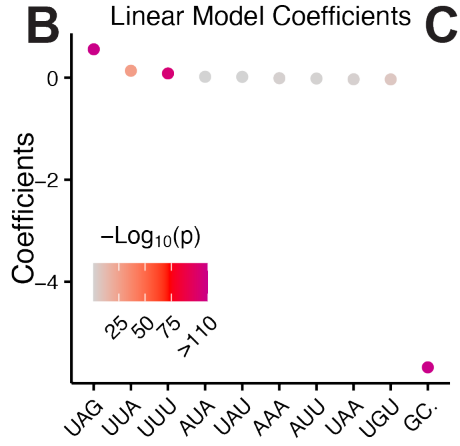
Comment 2: *Moreover, the contribution of UAG, UAA, and UUU are independently evaluated (separate CDF plots). It would be beneficial for the authors to evaluate the relative contribution of each to the RBNS enrichment. This could be done using a linear model or partial correlation or similar methods.*

Thank you for this suggestion. While some of our linear modeling efforts were limited based on the number of sequences utilized in nsRBNS, we were able to fit a linear model that has an adjusted R^2 of 0.58. We selected UNK bound kmers from available RBNS data (Dominguez et al., 2018) as well as some prevalent kmers from RBNS on the individual domains. Below in Comment 4 we discuss in more detail which kmers are enriched and their relative positions to the central UAG. We have updated the text to include this model:

Citations:

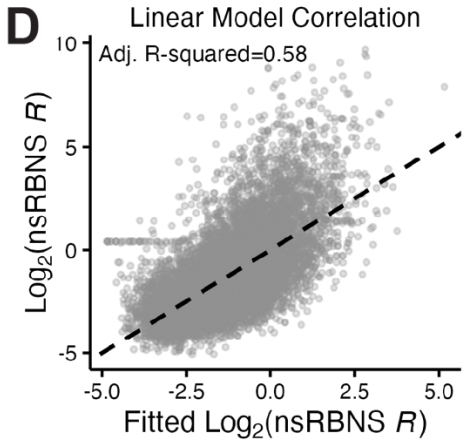
Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., van Nostrand, E. L., & Pratt, G. A. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Molecular Cell, 70(5), 854–867.

We modeled all natural (i.e. non-mutated, non-chimeric) sequences using a linear model (**Fig. S4B-D**) and unsurprisingly 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.



C

*k*mer	Coefficients	-Log_{10}(p)
UAG	0.555	>110
UUA	0.1345	27.3892
UUU	0.0818	103.3791
AUA	0.0165	0.7367
UAU	0.0147	0.6793
AAA	-0.0113	2.0037
AUU	-0.0178	1.2112
UAA	-0.0322	2.1536
UGU	-0.0347	6.695
GC-content	-5.6743	>110



Legend: B-D) Linear modeling of all natural (non-mutated, non-chimeric) sequences. B) Plot of linear model coefficients for top UNK motifs as defined by RBNS (Dominguez et al., 2018), colored by -log_{10} p. C) Table of linear model coefficients and -log_{10} p for top UNK motifs. D) Correlation of fitted log2 nsRBNS enrichment via linear model versus observed log2 nsRBNS enrichment.

Comment 3: *In figure 2E and all other uses of ribosome profiling data, is the comparison (x-axis) only changes in ribo-seq data and not translational efficiency i.e. normalizing for changes in RNA levels? It should be TE or it should be demonstrated there is no UNK-dependent differential expression.*

The data presented shows changes in ribo-seq that are not normalized for changes in expression. In conferring with our collaborator and co-author Jernej Murn, we believe it best to show changes in unnormalized RiboSeq data, rather than translational efficiency. The main reason is that in addition to its translational repression activity, UNK has recently been demonstrated to also destabilize RNA, thus further affecting the protein output (Shah et al., 2024). To address the question raised, below are three plots that demonstrate these differences. While translational efficiency (aka Norm. RiboSeq Fold Change, Log_2 or TE) still demonstrates that conserved oligos are more translationally repressed, the effect is muted, likely due to the RNA destabilization present in the RNAseq data. We have updated the text to include this information as follows:

"UNK regulates neuronal morphology, is a negative regulator of translation, *mildly destabilizes RNA targets,* and associates with polysomes (Murn *et al*., 2015; Murn *et al*., 2016; Shah *et al*., 2024)."

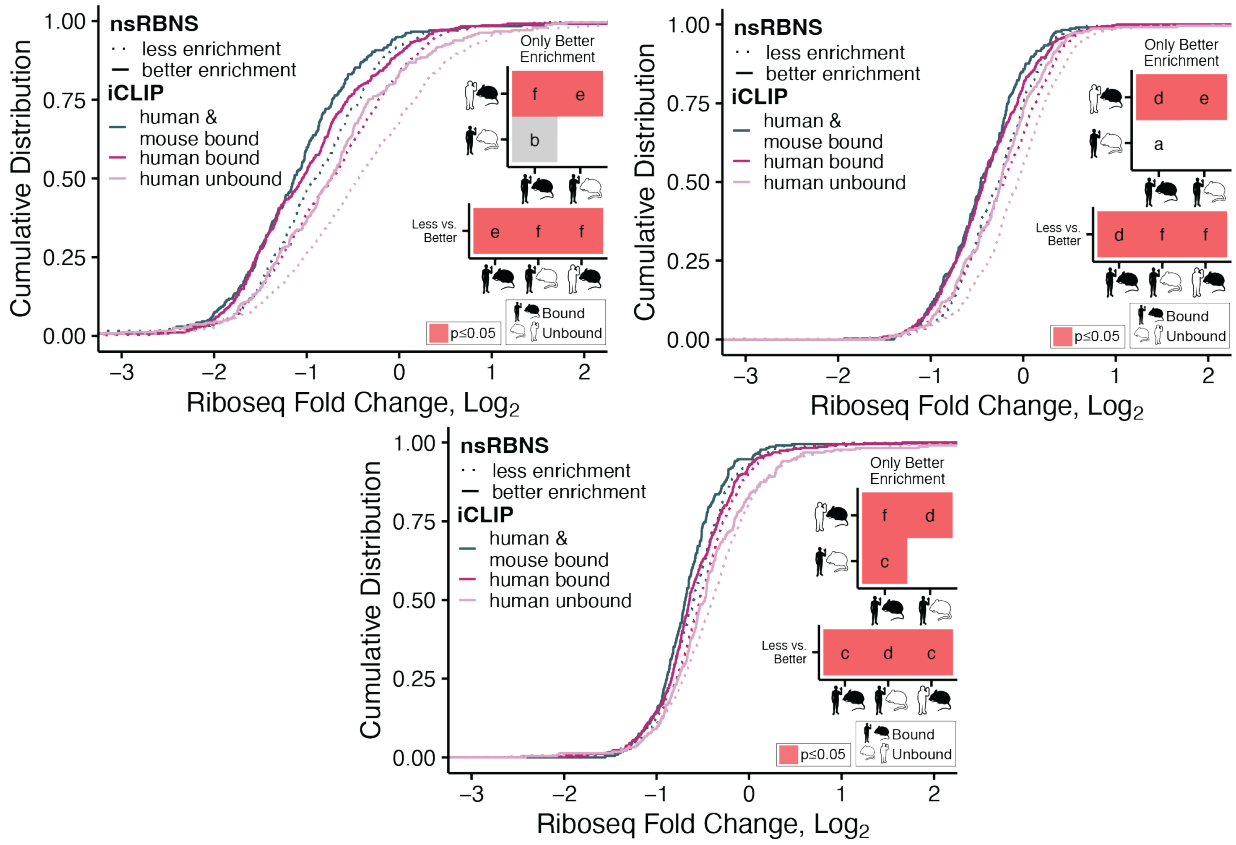
"UNK is a translational repressor (Murn *et al*., 2015) *and mildly destabilizes its target RNAs (Shah et al., 2024)*, thus UNK-regulated RNAs are predicted to have decreased translation as previously shown (Murn *et al*., 2015)."

Citations:

Murn, J., Zarnack, K., Yang, Y. J., Durak, O., Murphy, E. A., Cheloufi, S., Gonzalez, D. M., Teplova, M., Curk, T., & Zuber, J. (2015). Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. Genes & Development, 29(5), 501–512.

Murn, J., Teplova, M., Zarnack, K., Shi, Y., & Patel, D. J. (2016). Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. Nature Structural and Molecular Biology, 23(1), 16–23. https://doi.org/10.1038/nsmb.3140

Shah, K., He, S., Turner, D. J., Corbo, J., Rebbani, K., Dominguez, D., Bateman, J. M., Cheloufi, S., Igreja, C., Valkov, E., & Murn, J. (2024). Regulation by the RNA-binding protein Unkempt at its effector interface. Nature Communications, 15(1), 3159. https://doi.org/10.1038/s41467-024-47449-4



Comment 4: *Line 333-334 and figure 6: The authors suggest that the secondary RBD of UNK engages with U/A rich downstream sequences of the core motif, but this was not explicitly*

tested. RBNS data of the UNK with deletions or mutations of the secondary RBDs would be one way to provide support for this model.

We have approached this with a combination of RBNS with the individual binding domains (ZnF1-3 or ZnF4-6) as well as fluorescence polarization. As can be shown from the randomized RNA pool RBNS, the preference for ZnF4-6 is primarily UAG, while the preference for ZnF1-3 is more UA-rich, confirming the difference in specificity.

We also tested binding directly against specific oligos using fluorescence polarization. Of note is the fact that ZnF4-6 has much stronger binding affinity based on fluorescence polarization assays compared to ZnF1-3. This likely explains why the UAG motif is so critical for overall UNK binding. As expected, the full-length protein binds much better than the individual domains. Finally, removal of the AU-rich sequence downstream of the UAG displays drastically reduced binding by full length UNK. In fact, this binding is very similar to the binding displayed by ZnF4-6 (which lacks the ability to bind AU-rich sequences).

Finally, we used our existing nsRBNS data to identify positionally (relative to the central UAG) enriched 3mers in human bound vs mouse not bound sequences. As is shown below these sequences tend to be AU-rich. The following text and figures have been added to the manuscript:

Citations:

Achsel, T., & Bagni, C. (2016). Cooperativity in RNA–protein interactions: the complex is more than the sum of its partners. Current Opinion in Neurobiology, 39, 146–151.

Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Molecular Cell, 78(1), 9–29.

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., van Nostrand, E. L., & Pratt, G. A. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Molecular Cell, 70(5), 854–867.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., & Burge, C. B. (2014). RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. Molecular Cell, 54(5), 887–900. https://doi.org/10.1016/j.molcel.2014.04.016

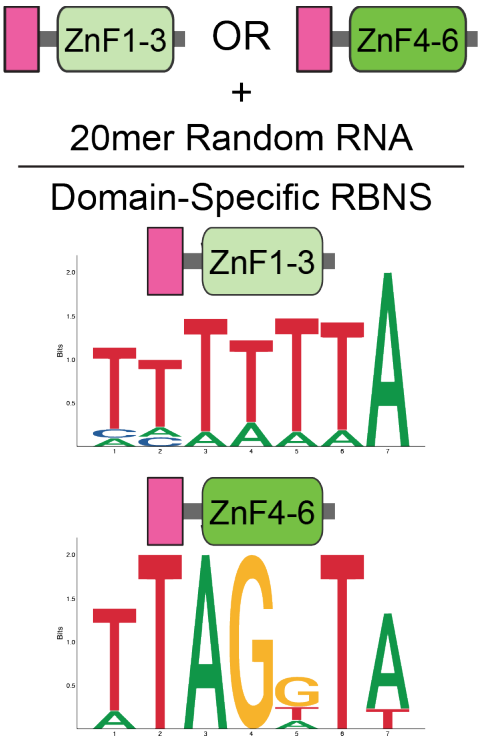
Murn, J., Teplova, M., Zarnack, K., Shi, Y., & Patel, D. J. (2016). Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. Nature Structural and Molecular Biology, 23(1), 16–23. https://doi.org/10.1038/nsmb.3140

> "Indeed, when we tested the individual domains (ZnF1-3 or ZnF4-6) via random RBNS as previously described (Dominguez *et al*., 2018; Lambert *et al*., 2014), we observed strong UAG binding with ZnF4-6 (the primary domains) and U/A rich motifs with ZnF1-3 (Fig. S1B). These data support previous crystal structures showing UAG binding with ZnF4-6 and U/A binding via ZnF1-3 (Murn *et al*., 2016)."

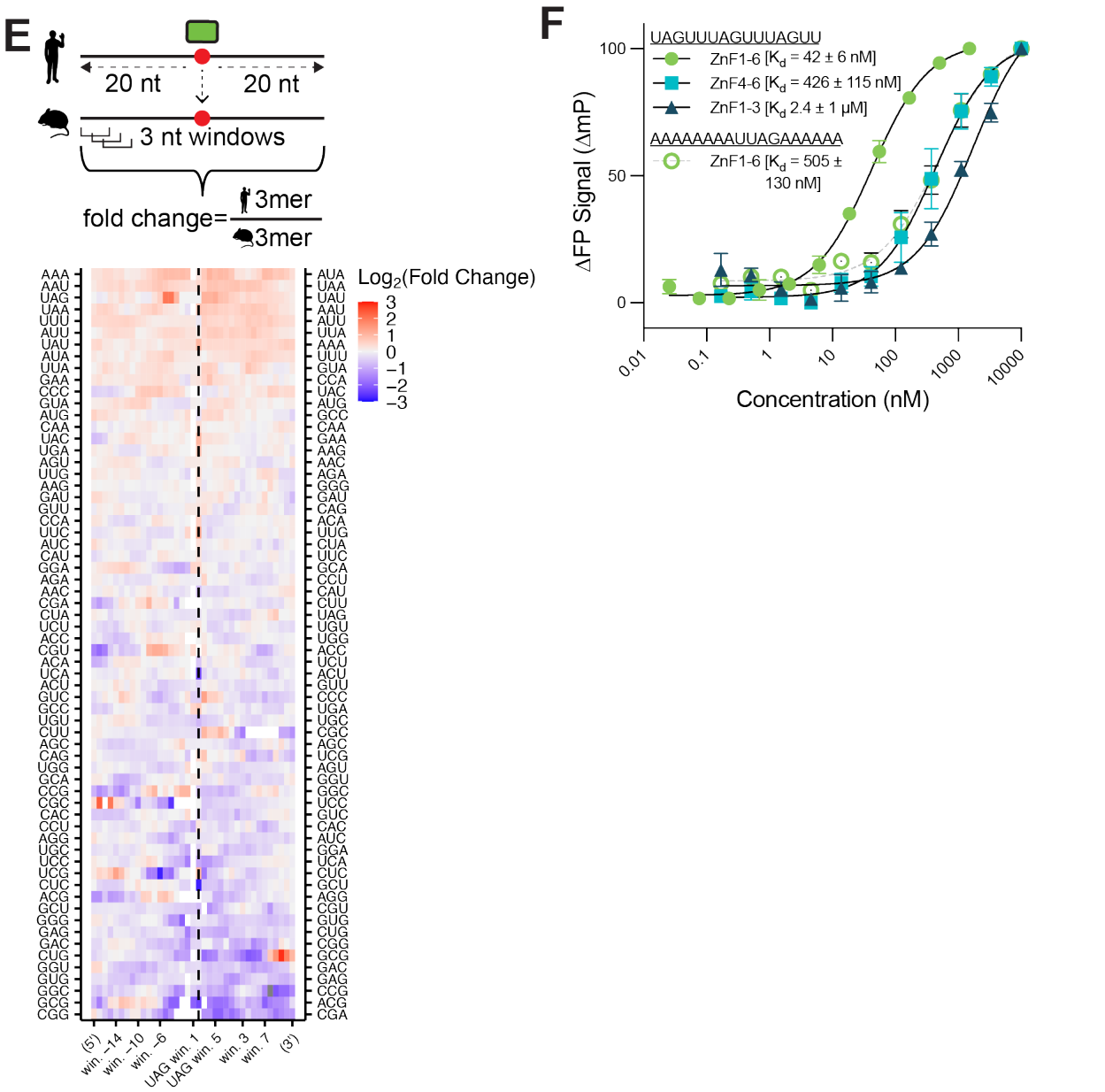
> "To examine these inter-species sequence differences on a global scale more specifically, we analyzed the 3mer enrichment across human and mouse where the human oligo was bound better, despite maintenance of UAG content. Looking across all possible 3mers upstream and downstream of the central UAG, we observe that human bound sequences are more enriched in A and U-rich motifs centrally than their unbound mouse counterparts (Fig. S3E). 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 (Murn *et al.,* 2016)."

"To test this, we turned to fluorescence polarization (FP) to understand to what extent the dual-RBD infrastructure 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 Kd 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, the Kd decreases another 10-fold at ~40 nM (Fig. S3F). When comparing this to the binding patterns of ZnF1-6 with an RNA oligo with only a UAG motif, the Kd increases to match that of ZnF4-6 with the full motif at ~500 nM (Fig. S3F). These data highlight the importance of multiple domains for selecting its targets and is supported by previous work on cooperativity and avidity for other RBPs (reviewed by Achsel and Bagni 2016; Corley *et al.,* 2020)."



Legend: Design and motif logos of RBNS with ZnF1-3 and ZnF4-6.



Legend: E) Heat map of 3mer human over mouse enrichment upstream and downstream of central UAG in orthologs bound better in humans. F) Delta fluorescence polarization binding curves for UNK ZnF1-6 (green circle), ZnF1-3 (blue triangle), and ZnF4-6 (teal square) incubated with a tri-UAG-containing RNA oligo graphed with delta fluorescence polarization binding curves for UNK ZnF1-6 (hollow green circle) incubated with a mono-UAG-containing RNA oligo. Each curve was normalized to its minimum and maximum fluorescence polarization signal to produce delta fluorescence polarization values.

Comment 5: *It would be beneficial to change the result titles to match the main conclusion in each section.*

Thank you for the suggestion. The results titles have been changed as follows:

- Conserved and Species-Specific *in vivo* Binding Patterns → UNK RNA-Binding Patterns Vary Across Species
- Understanding the UNK-RNA Interactome *in vitro* at Massive Scale → nsRBNS Measures Natural Sequence Binding Differences *in vitro* at Massive Scale
- Recapitulation of *in vivo* Binding Patterns and Regulation → *In vivo* Binding Patterns and Regulation can be Recapitulated *in vitro*
- Binding Strength and in vivo Regulation → *In vitro* Binding Patterns Correlate with *in vivo* Regulation
- Species-Specific Binding Site Patterns → *In vivo* Binding Differences can be Recapitulated at the Binding Site
- Binding Site Patterns Across Cell Types Within Species → Intra-Species Binding Patterns are Dependent on Cellular Factors
- Species-Specific Regional Impacts on Binding → Sequence Contextual Changes Impact Species-Specific Binding
- Evolutionary Conservation of Binding → Sequence Differences across 100 Vertebrates affect UNK-RNA Interactions *in vitro*

Comment 6: *Line 68: Clarify if the 95% identity for the whole TF or the DNA-binding domain*

We have edited the text for clarification:

> "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."

Comment 7: *It was unclear if the results in Figure 1A/Line 120 only consider 1-1 orthologues*

We have clarified the text as follows:

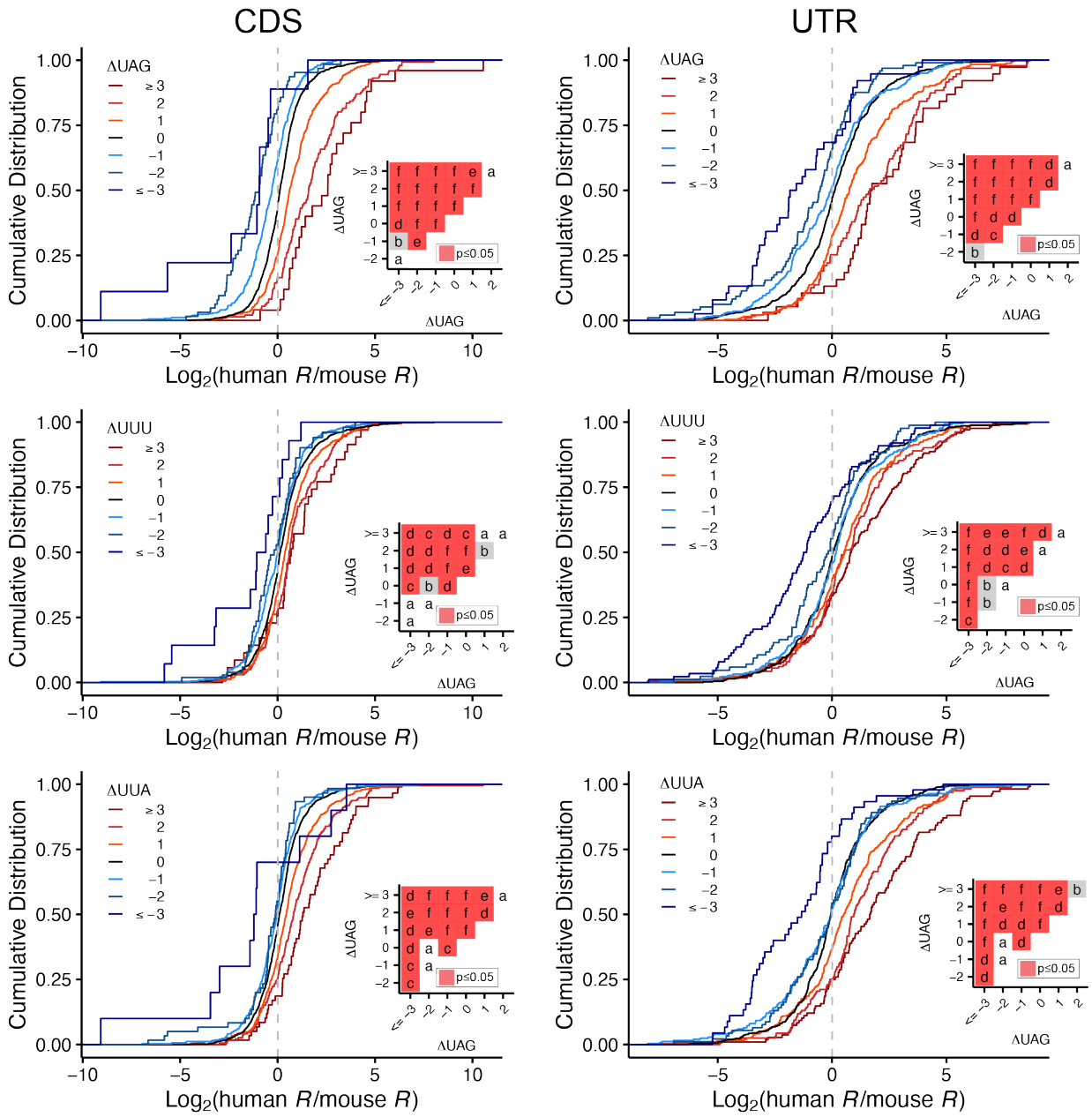
> "Comparing *one-to-one orthologous* binding sites across species at the transcript level, we observe that ~45% of transcripts are bound in both species."

Comment 8: *Line 125: It would be helpful to stick with clearly defined terminology i.e. conserved instead of homologous in text vs figure.*

In the given section, "conserved" refers to transcript-level binding whereas "homologous" is in reference to the specific region within the transcript.

Comment 9: *Are the results/conclusion of Figure 2D, S2D, S2E different for CDS vs UTR?*

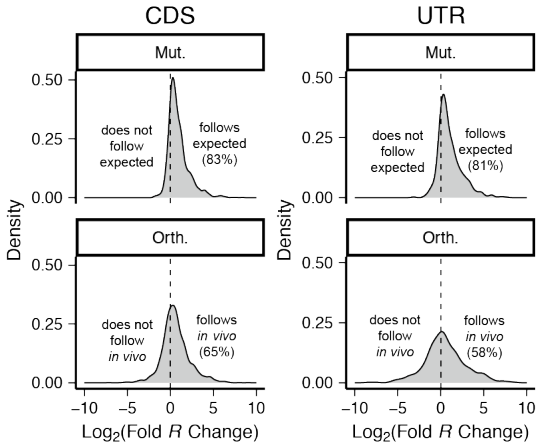
While the data shifts slightly depending on transcript region, the results are similar where motif changes across species contribute to species-specific binding patterns. These patterns can be seen below:



Comment 10: *Line 209: Where is the 60% of binding sites mirrored the in vivo trend?*

When comparing the log_2 fold change of in vivo bound vs. motif mutants or unbound orthologs, we can see this trend emerge. This is shown in the boxplots inset in Fig. 2 B,C. We've added histograms to Supp. Fig 2 to further emphasize these trends. From the histograms (below), 65% of CDS bound oligos and 58% of UTR bound oligos have a higher enrichment over their unbound orthologs. We have edited the text, included the histograms, and edited the wording to clarify this point:

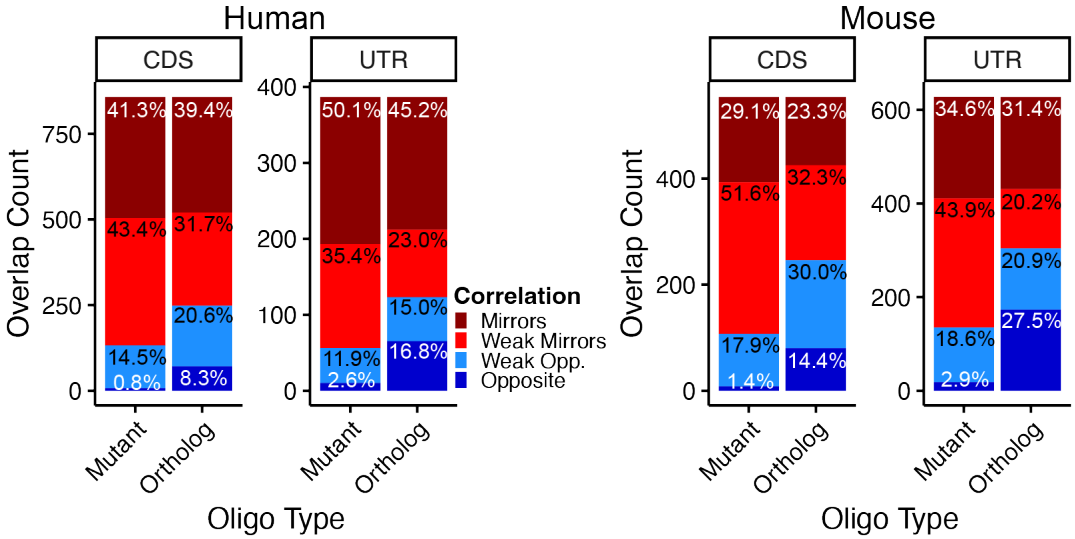
> "We measured how often a species-specific site was better bound than its non-bound ortholog and found that ~60% *(65% for CDS and 58% for UTR)* of binding sites mirrored the *in vivo* trend."



Legend: Density plot of in vivo binding versus *in vitro* binding patterns for "motif mutant" and "orthologous" oligos versus *in vivo* bound oligos for B) CDS and C) UTR oligos.

Comment 11: *Figure S2B,C: in the figure, please indicate which species you're referring to.*

The figure has been edited for clarity.

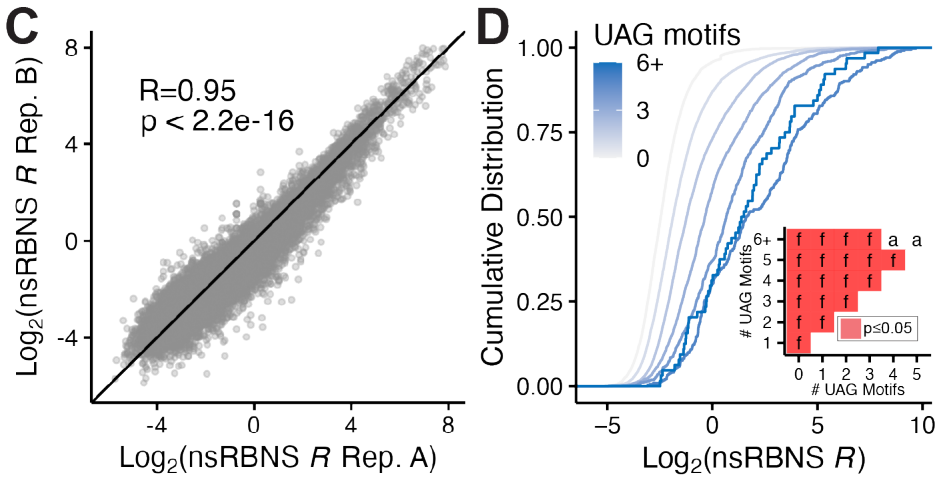


Comment 12: *Figure 2B, C: x-axis – enrichment of what? Inset Y-axis – enrichment of what? Presumably RNBS*

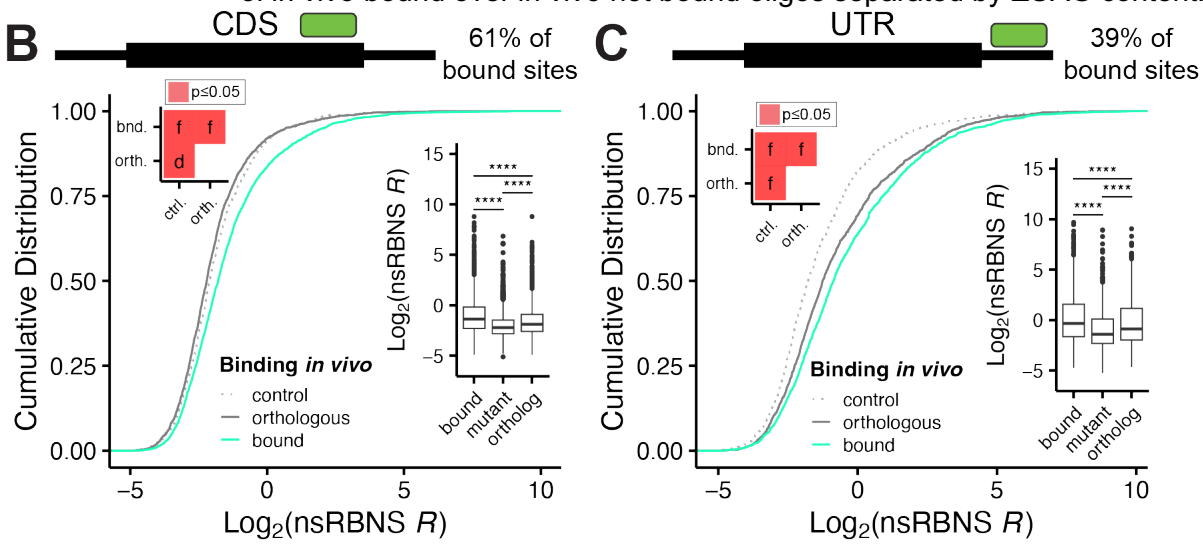
Thank you for pointing this out, we agree that this will make the data easier to follow. We have adjusted axes where space allows to denote nsRBNS R and updated figure legends as shown below. As shown below we have done this for all figures not just figure 2.

- Fig 1
  - "D) Cumulative distribution function of $\log_2$ *nsRBNS* enrichment of all oligos separated by UAG motif content."

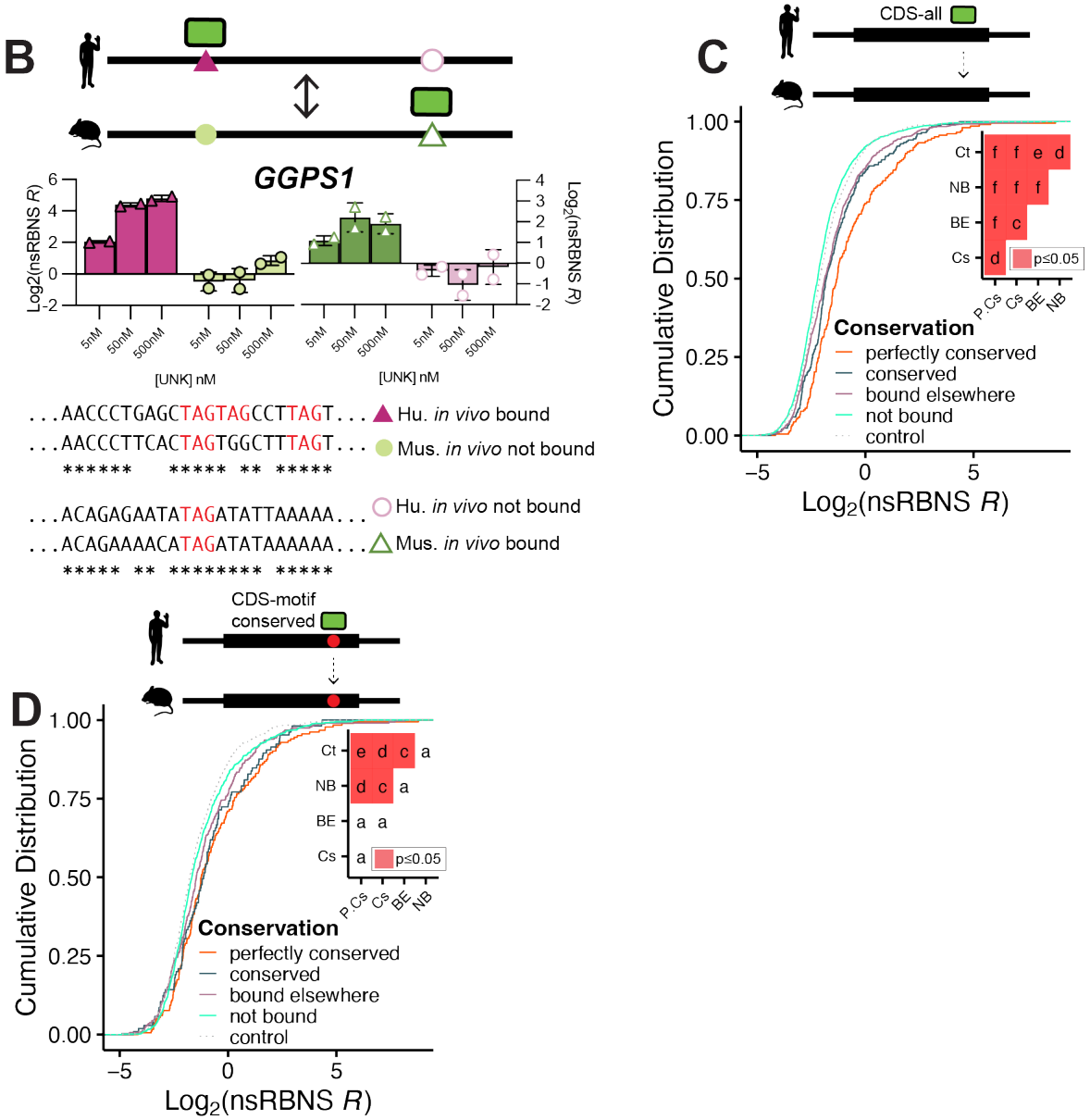
- "E) Scatter plot of $\log_2$ *nsRBNS* enrichment of wildtype (Y-axis) versus motif mutant (X-axis) oligos."



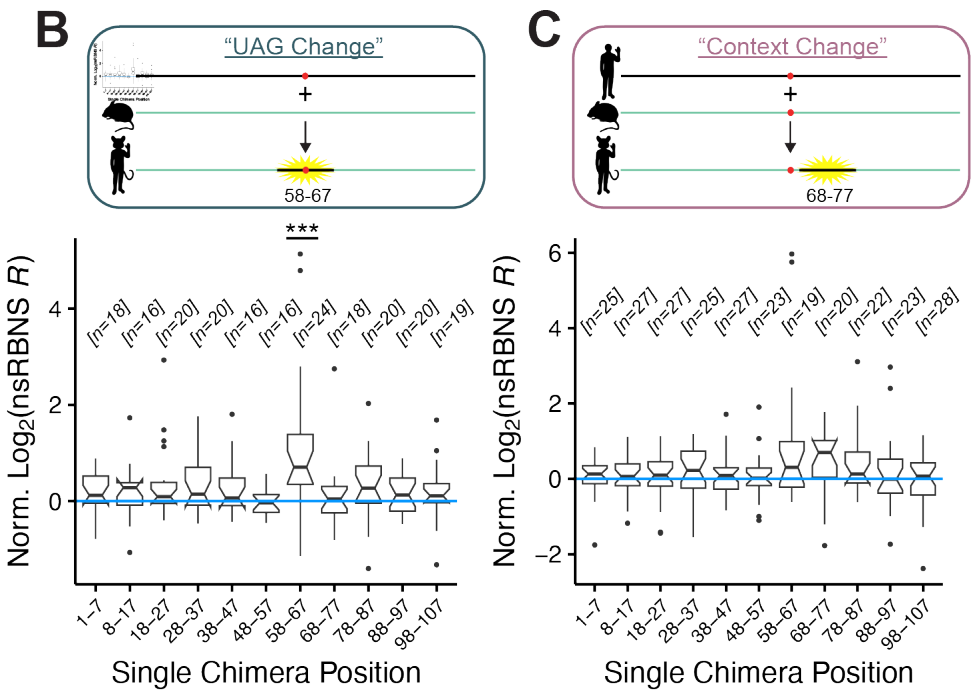
- Fig 2
  - "B-C) Cumulative distribution function of $\log_2$ *nsRBNS* enrichment of all iCLIP hits: control (light grey; dotted), orthologous (dark grey), and bound (teal) of B) CDS and C) UTR oligos."
  - "D) Cumulative distribution function of $\log_2$ fold *nsRBNS* enrichment change of *in vivo* bound over *in vivo* not bound oligos separated by ΔUAG content."

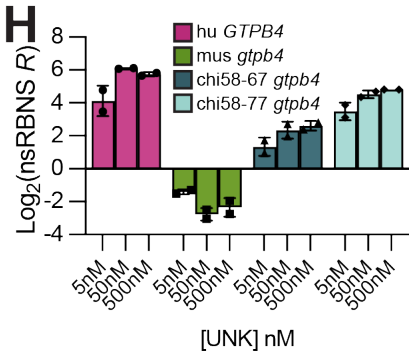


- Fig 3
  - "B) Conservation and binding of *GGPS1* orthologous pairs. (left) $\log_2$ *nsRBNS* enrichment values from nsRBNS for human bound…"
  - "C-D) Cumulative distribution function of $\log_2$ *nsRBNS* enrichment of control…"

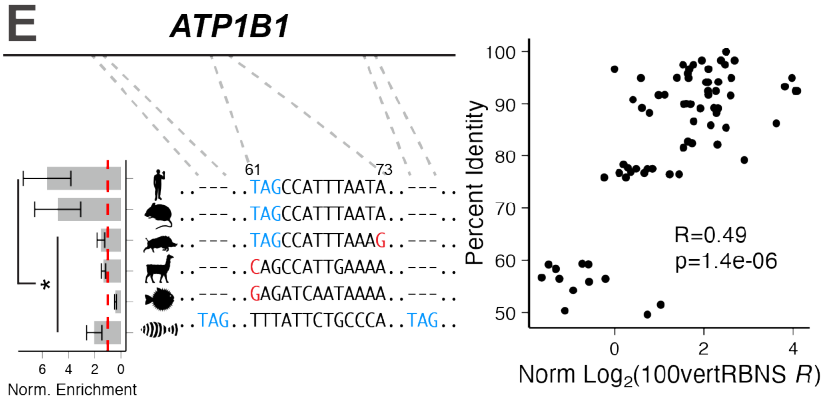


- Fig 4
  - "B) Design and box and whisker plot of normalized $\log_2$ *nsRBNS* enrichment…"
  - "C) Design and box and whisker plot of normalized $\log_2$ *nsRBNS* enrichment…"
  - "D) Heat map of median normalized $\log_2$ *nsRBNS* enrichment…"
  - "F) Heat map of median normalized $\log_2$ *nsRBNS* enrichment…"
  - "H) $\log_2$ *nsRBNS* enrichment values ~~from nsRBNS~~…"

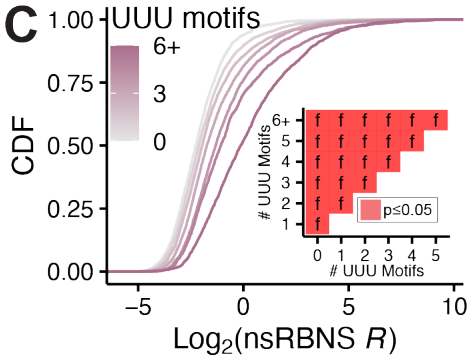


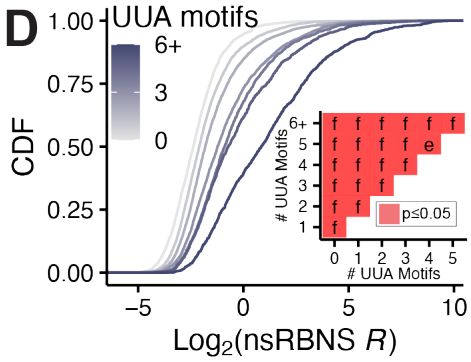


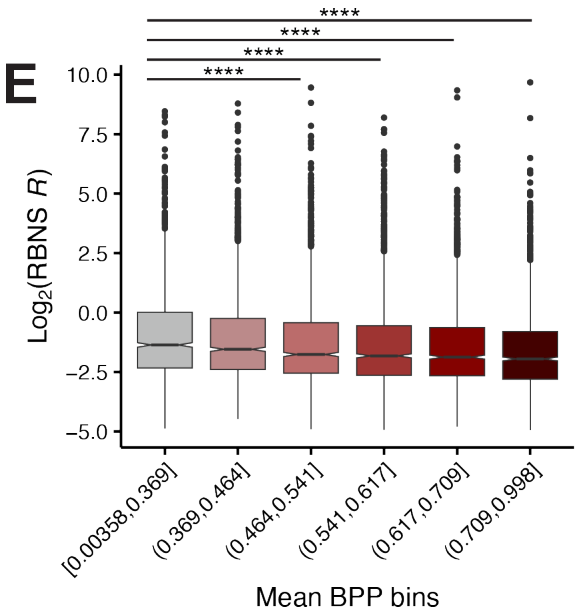
- Fig 5
  - "B) Delta $\log_2$ *100vertRBNS* enrichment, percent RNA sequence identity, percent UNK similarity…"
  - "C) Mean percent RNA sequence identity (Y-axis) versus mean delta $\log_2$ *100vertRBNS* enrichment (X-axis) for each aligned oligo."
  - "D) Evolutionary distance in millions of years (Y-axis) versus mean delta $\log_2$ *100vertRBNS* enrichment (X-axis) for each aligned oligo."
  - "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. (right) Percent RNA sequence identity (Y-axis) versus normalized delta $\log_2$ *100vertRBNS* enrichment (X-axis)."
  - "F) Scatter plot of $\log_2$ normalized ~~UNK binding~~ *100vertRBNS* enrichment by evolutionary distance."



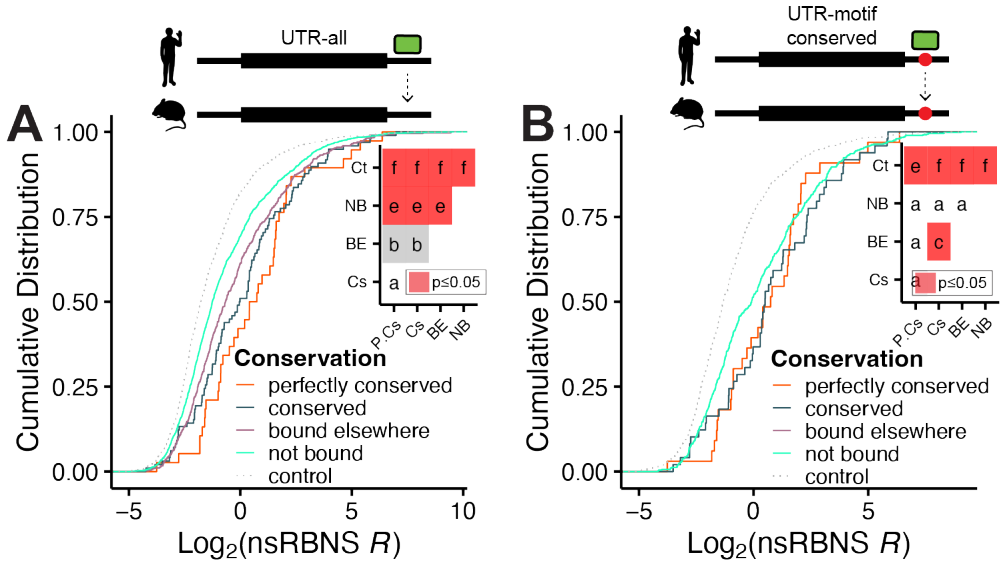
- Fig S1
  - "B-C) Cumulative distribution function of $\log_2$ *nsRBNS* enrichment of all oligos separated by B) UUU and C) UUA motif content."
  - "E) Box and whisker plot of $\log_2$ *nsRBNS* enrichment of all oligos separated by quantile-binned mean base pair probability (BPP) of the central region (54-64)."

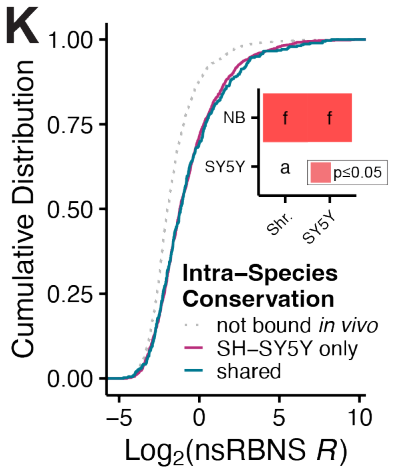




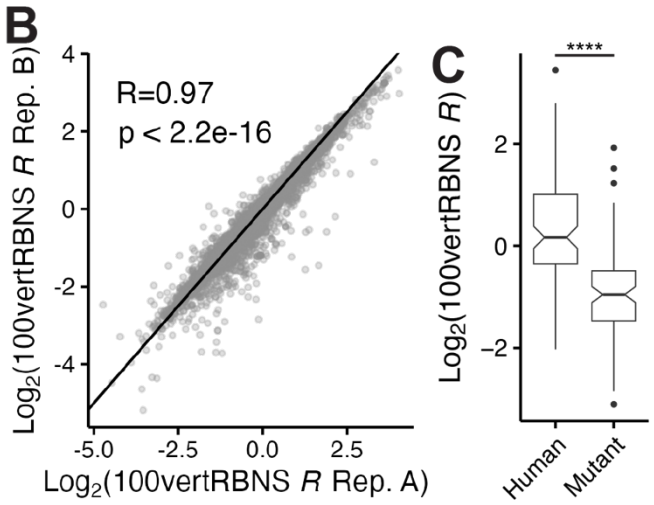


- Fig S2
  - "D-E) Cumulative distribution function of $\log_2$ fold *nsRBNS* enrichment change of *in vivo* bound over *in vivo* not bound oligos separated by D) ΔUUU and E) ΔUUA content."
  - F-G) Cumulative distribution function of $\log_2$ fold *nsRBNS* enrichment change of *in vivo* bound over *in vivo* not bound oligos separated by percent conservation for F) all and G) *k*mer loss cross-species comparisons."
- Fig S3
  - "A-B) Cumulative distribution function of $\log_2$ *nsRBNS* enrichment of control…"
  - "K) Cumulative distribution function of $\log_2$ *nsRBNS* enrichment of human not bound *in vivo*…"





- Fig S5
  - "C) Box and whisker plot of $\log_2$ *100vertRBNS* enrichment for human and total motif mutants."
  - "E) Normalized $\log_2$ *100vertRBNS* enrichment of *ATP1B1*."
  - "F) …(right) Percent RNA sequence identity (Y-axis) versus normalized delta $\log_2$ *100vertRBNS* enrichment (X-axis)."
  - "H) Normalized $\log_2$ *100vertRBNS* enrichment of *NFATC3*."
  - "I) …(right) Percent RNA sequence identity (Y-axis) versus normalized delta $\log_2$ *100vertRBNS* enrichment (X-axis)."
  - "J) (left) Multiple sequence alignment for *PPP2R5C*. (right) Normalized $\log_2$ *100vertRBNS* enrichment of *PPP2R5C*."



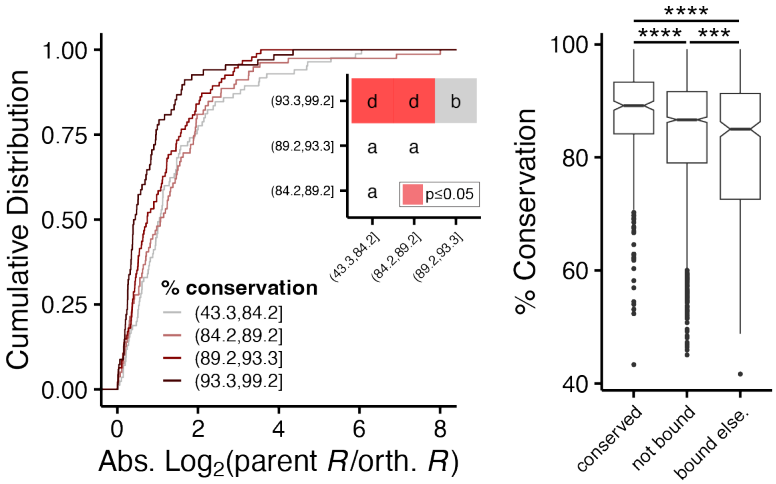
Comment 13: *Figure 3C + line 256: What's the difference between "perfectly conserved" vs "conserved" – how identical is conserved %-wise?*

This is a great question. "Perfectly conserved" refers to sequence conservation (100% identical in the aligned region between human and mouse) whereas "conserved" refers to binding conservation of homologous human-mouse regions. To clarify, we've updated our terminology to

"binding conserved" in the manuscript. On average, "binding conserved" oligos have ~82% sequence homology within the aligned region.

As might be expected, higher sequence conserved oligo pairs have more similar nsRBNS enrichments than less conserved pairs. Additionally, "binding conserved" pairs are more conserved at the sequence level than the other two categories. "Bound elsewhere" had the lowest sequence conservation of the three (which tracks with the lowest degree of binding). Generally, the low level of conservation of "bound elsewhere" sites can be explained by being a culmination of two "not bound" categories (human compared to mouse and mouse compared to human). We've included the following panels in Fig. S3 to emphasize this point with the "binding conserved" oligo subset.

"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 (Fig. S3C). 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 where in vitro bound "binding conserved" oligo pairs are more conserved than "not bound" which are more conserved than "bound elsewhere" (Fig. S3D; p≤0.001, Wilcoxon test). Strikingly, "bound elsewhere" pairs had the lowest average conservation of the three categories, further highlighting the shifting nature of these sites."



Legend: C) Cumulative distribution function of log2 fold nsRBNS enrichment change (parent/ortholog) of "binding conserved" oligos pairs separated by percent conservation. Inset shows significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), b (p≤0.1), d (p≤0.01). D) Boxplot of percent conservation of binding conserved, not bound, and bound elsewhere oligo pairs where the parent was bound in human, but the aligned orthologous region was unbound. Significance was determined via KS tests and corrected for multiple comparisons via the BH procedure. Statistical marks are as follows: *** — p ≤ 0.001, **** — p ≤ 0.0001.

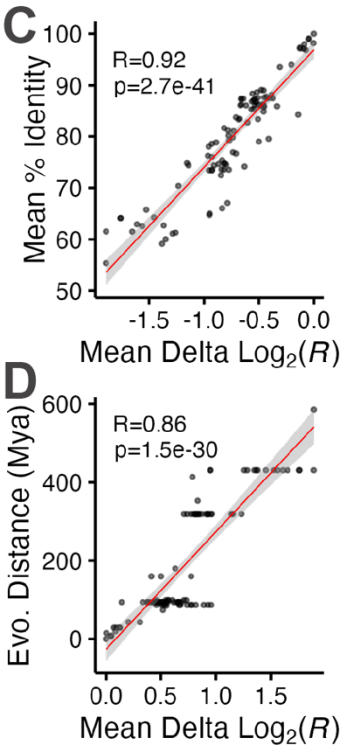
Comment 14: *Figure 3C,D: Please clarify the definition of "CDS-all" and "CDS-motif conserved". Does that mean in C there is no motif in human? And in D there is motif?... It is clear that the perfectly conserved oligos are enriched the most. Line 261-262: "when only*

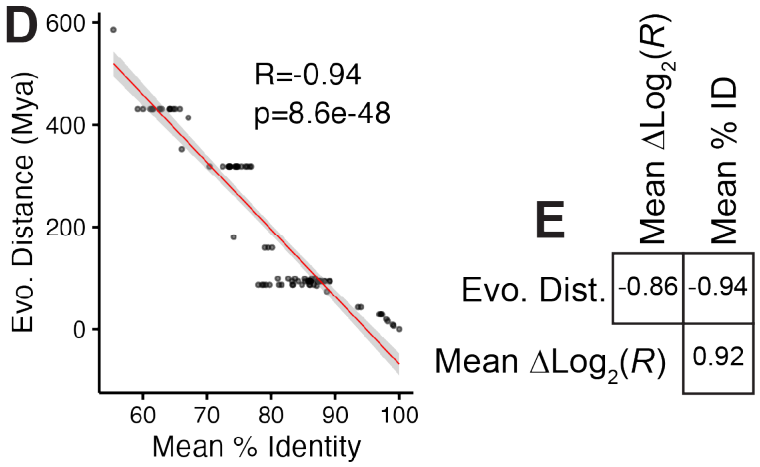
regions with UAG motifs in both human and mouse were considered". Does that mean that C does not have UAG in both? According to the figure that could be true. So that would mean that there binding sites without the core UAG motif were considered?

"CDS-all" in panel C refers to oligo pairs where a UAG motif was not required. Nearly every in vivo binding site had a core UAG motif, however, unbound orthologous regions did not always. "CDS-motif conserved" in panel D, on the other hand, refers only to oligo pairs where the UAG motif was present in both species, despite loss of binding in the orthologous species. Therefore, "CDS-motif conserved" is a subset of "CDS-all," not an opposite. "CDS-motif conserved" represents ~46% of all CDS occurrences.

Comment 15: *Figure 5D/Line 376: How much of the relationship between evolutionary distance and binding is explained by the difference in sequence identity?*

Good question and it may be somewhat redundant to show both as evolutionary distance and sequence identity are highly related (especially in CDS regions). We've added a correlation plot between mean percent identity and evolutionary distance to supp. fig 5 and performed a three-way correlation between mean percent identity, mean log_2 100vert enrichment, and evolutionary distance. These three are very high correlated (>0.85 across all comparisons). As the RNA sequence is rapidly evolving, binding patterns are also rapidly evolving despite little to no change in protein sequence. To the reviewer's point, the primary driver of binding differences is the identity between RNAs.





Legend: D) Evolutionary distance in millions of years (Y-axis) versus mean percent RNA sequence identity (X-axis) for each aligned oligo. Pearson's correlation coefficient included. E) Full correlation of evolutionary distance in millions of years, mean percent RNA sequence identity, and mean delta $\log_2$ 100vertRBNS enrichment.

Comments from Reviewer 2

Comment 1: *In Figure 1A, when the authors analyze CLIP data to determine if a site is bound by UNK in human and/or mouse, it is not clear if the authors impose the rule that the gene analyzed must be expressed in the input of human and mouse samples. If a gene is not expressed in one of the samples, the absence of binding by CLIP is not informative. A similar situation occurs in figure S3C where the authors examine the conservation of binding between cell types. The authors should clarify the analysis and make sure that only analyze genes that are expressed in both CLIP datasets.*

Apologies for not being clear on this. We have edited the text as follows:

"We used UNK iCLIP data in human and mouse neuronal cells and tissue, respectively(Murn et al. 2015), to identify species-specific and conserved UNK binding sites. *Only genes expressed at greater than 5 transcripts per million (TPM) in both cell lines were included.* Comparing *one-to-one orthologous* binding sites across species at the transcript level, we observe that ~45% of transcripts are bound in both species."

"To compare these binding preferences to intra-species changes, we examined available iCLIP data from HeLa cells overexpressing UNK from the same study. *Only genes with greater than 5 TPM expression in both cell lines and one-to-one orthologs across species were included.* Surprisingly, when looking at transcript-level conservation, we observed that approximately 51% of UNK transcripts were bound in both cell types, similar to that observed in human vs. mouse comparisons."

"Using these data, accounting for only genes with similar expression across cell types, 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."

- **Comment 2:** *The Methods sections does not have a section detailing the statistical analysis used thought the manuscript. It is also unclear why some cumulative plots have associated p values and others not. The statistical analysis and display should be standardized thought the manuscript and figures.*

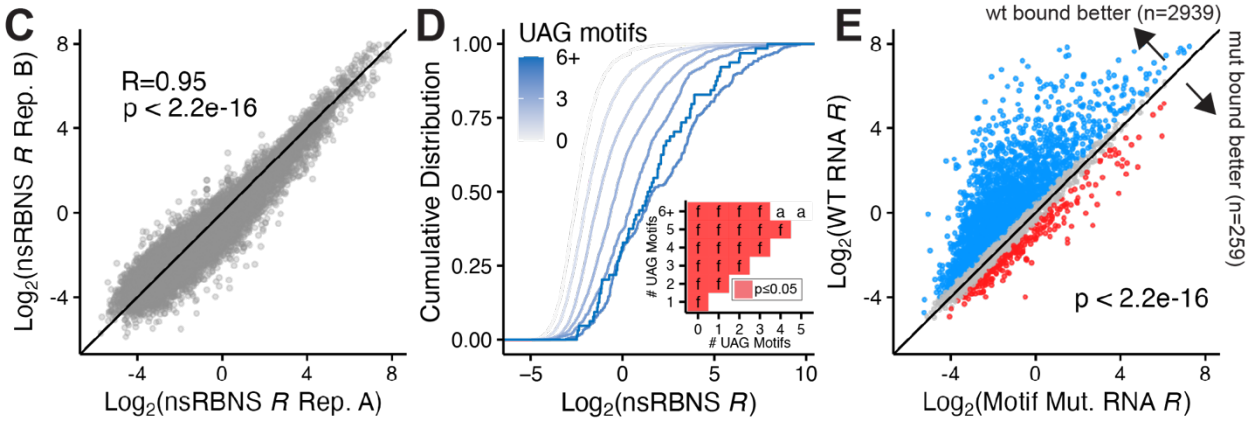
We have adjusted the figures, figure legends, and methods section as follows:

Citations:

Murn, J., Zarnack, K., Yang, Y. J., Durak, O., Murphy, E. A., Cheloufi, S., Gonzalez, D. M., Teplova, M., Curk, T., & Zuber, J. (2015). Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. Genes & Development, 29(5), 501–512.

Methods: "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 pvals shown are for the correlation. For wild type versus mutant group and chimerized comparisons, paired, one-sided Wilcoxon tests were with the expectation that chimerization would increase binding. 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, pvals were corrected via the BH procedure based on number of comparisons."

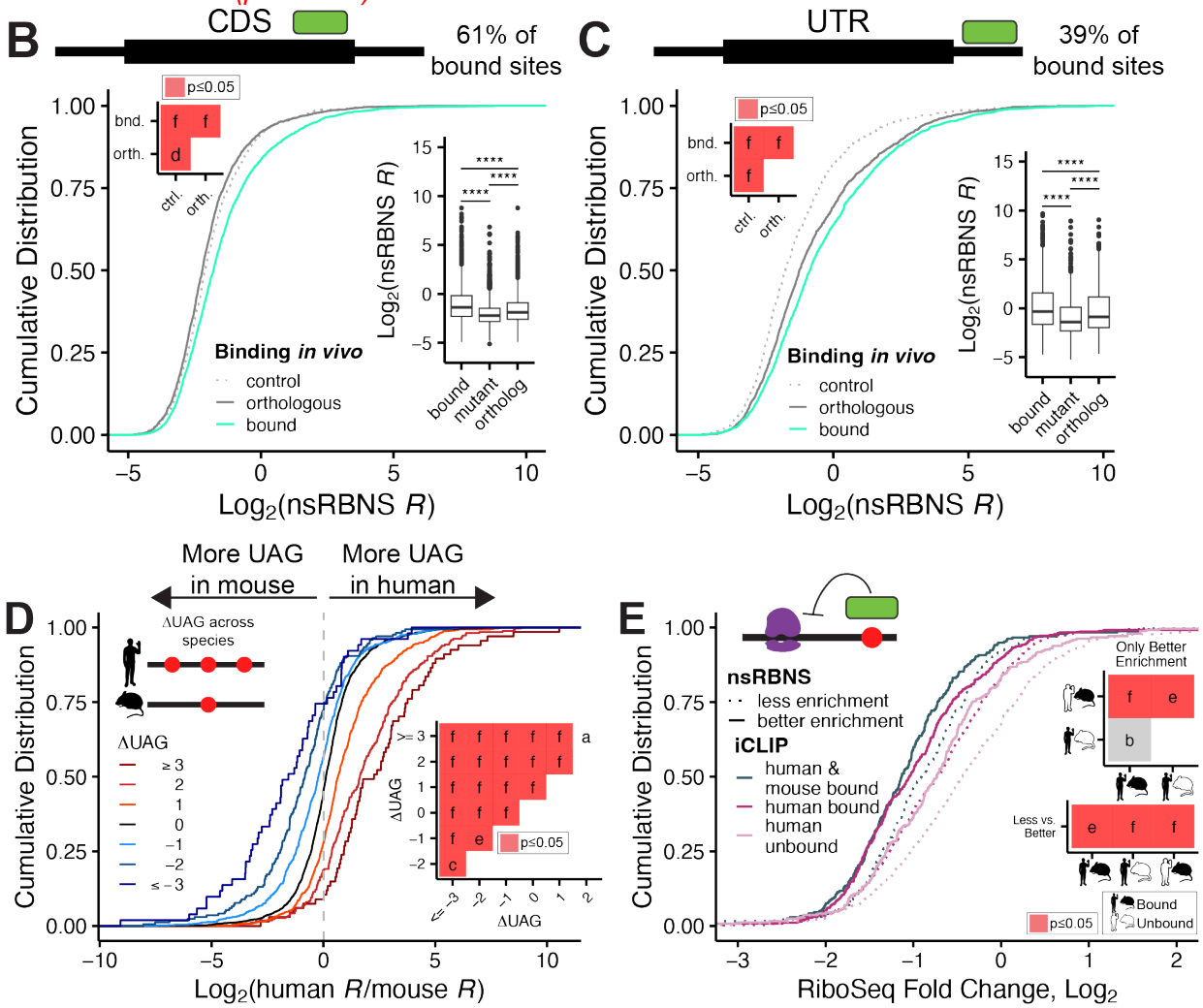
- Fig 1
  - D) Cumulative distribution function of $\log_2$ nsRBNS enrichment of all oligos separated by UAG motif content. *Inset shows significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), f (p≤0.0001).*
  - E) Scatter plot of $\log_2$ nsRBNS enrichment of wildtype (Y-axis) versus motif mutant (X-axis) oligos…*Significance determined via paired, one-sided Wilcoxon test.*



- Fig 2
  - B) CDS and C) UTR oligos. Significance of bound vs. orthologous was determined via KS test. Insets ~~show~~ boxplot ~~of~~ *shows in vitro* binding patterns for "bound," "motif mutant," and "orthologous" oligos. Significance was

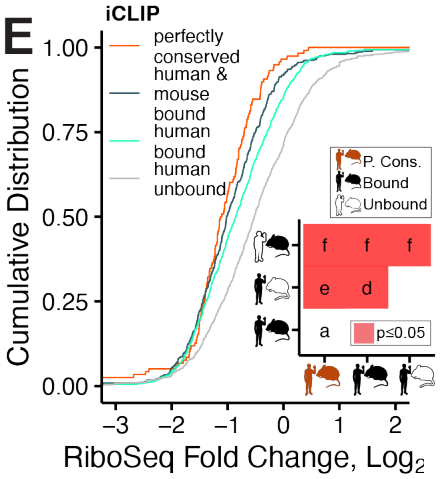
determined via two-sided Wilcoxon test. *Inset heatmap shows significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significance (p≤0.05). Values are as follows: d (p≤0.01), f (p≤0.0001).*

- 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 KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), c (p≤0.05), e (p≤0.001), f (p≤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 KS test and corrected for multiple comparisons via the BH procedure. Grey denotes nearing significance (p≤0.1). Red denotes significant (p≤0.05). Values are as follows: b (p≤0.1), e (p≤0.001), f (p≤0.0001).*

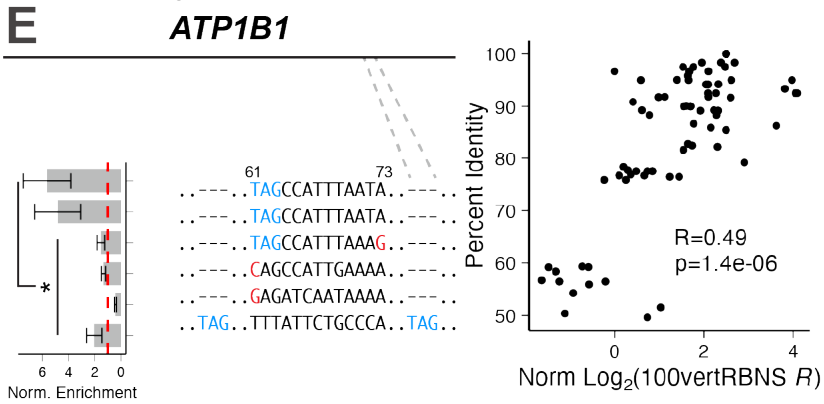


- Fig 3
  - E) Cumulative distribution function of RiboSeq fold change, $\log_2$ separated via iCLIP detection and sequence conservation. *Inset shows significance*

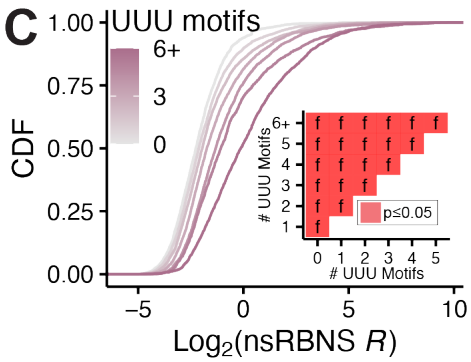
values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), d (p≤0.01), e (p≤0.001), f (p≤0.0001).

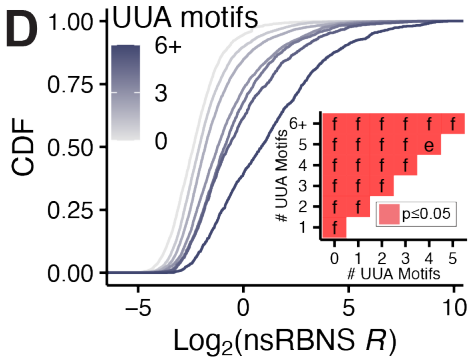


- Fig 5

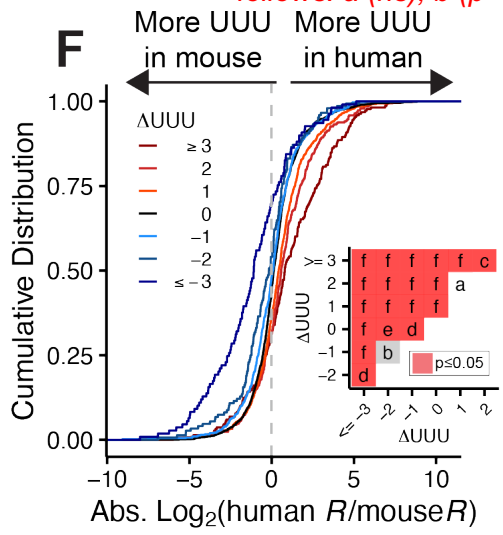


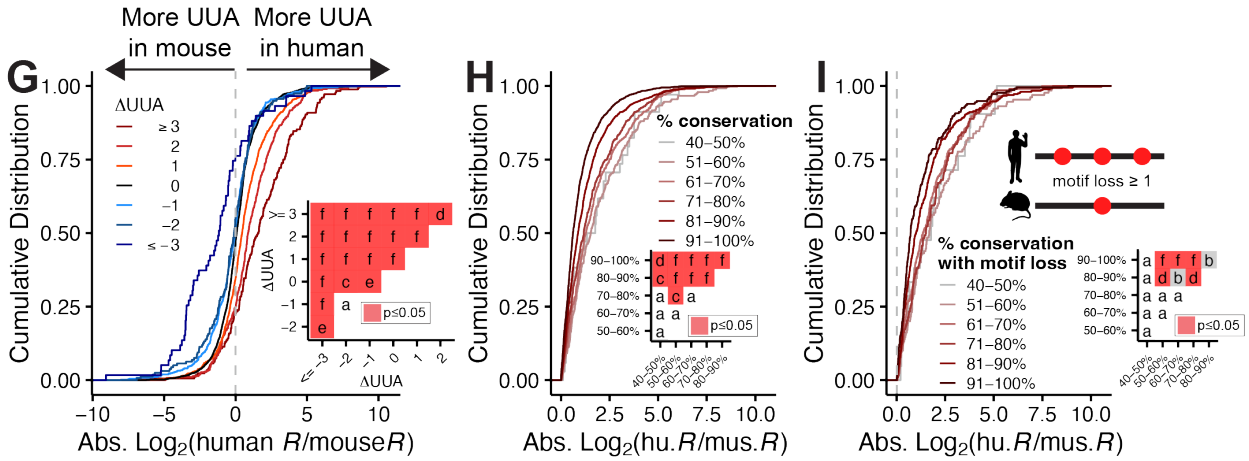
- Supp Fig 1
  - B) UUU and C) UUA motif content. *Insets show significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: e (p≤0.001), f (p≤0.0001).*



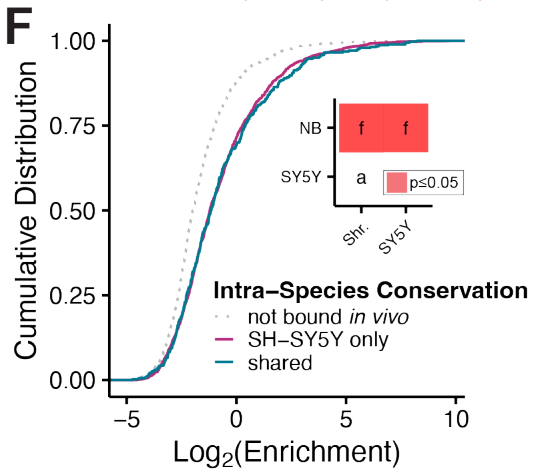


- Supp Fig 2
  - F) ΔUUU and G) ΔUUA content. *Insets show significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), b (p≤0.1), c (p≤0.05), d, (p≤0.01), e (p≤0.001), f (p≤0.0001).*
  - H) all and I) *k*mer loss cross-species comparisons. *Insets show significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), b (p≤0.1), c (p≤0.05), d, (p≤0.01), f (p≤0.0001).*

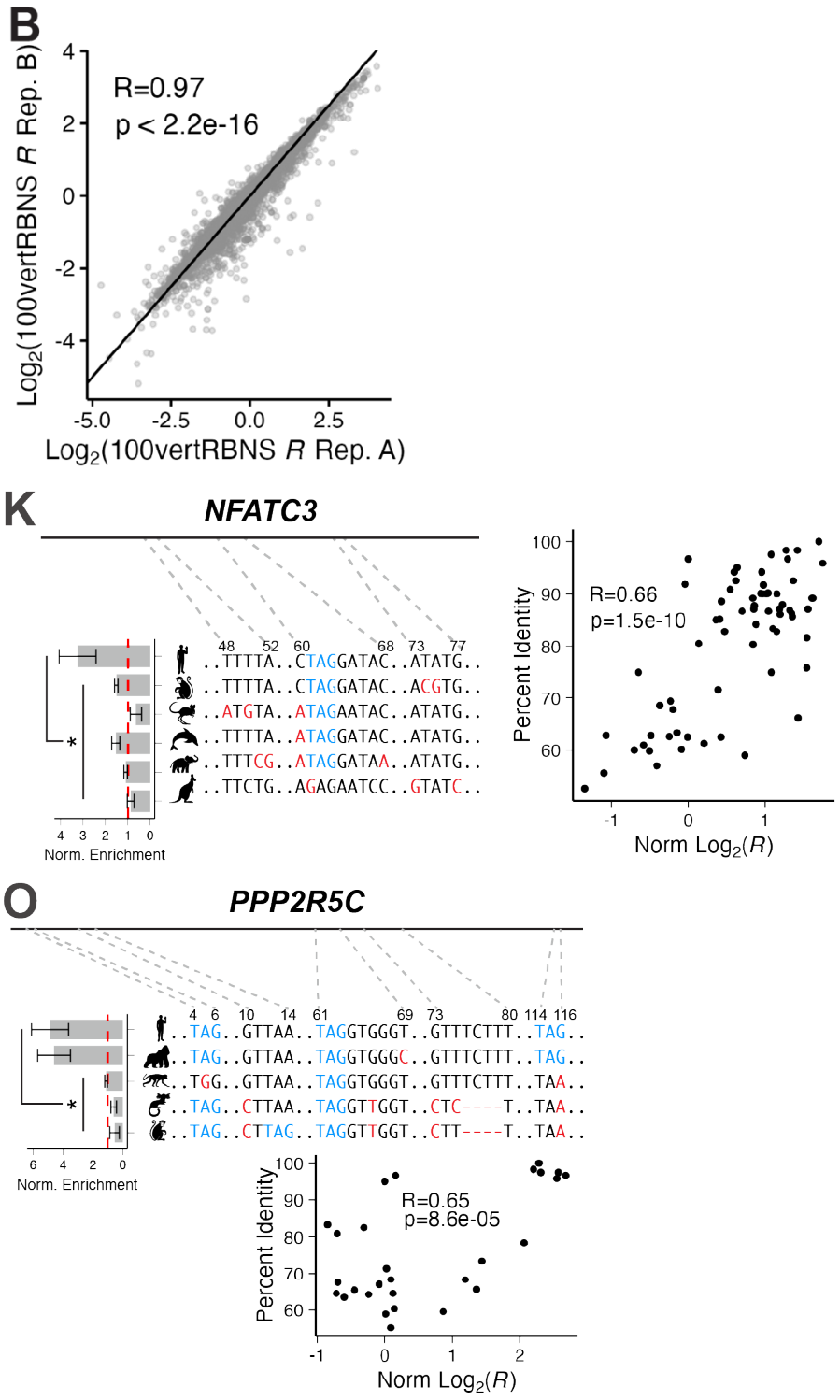




- Supp Fig 3
  - F) Cumulative distribution function of $\log_2$ nsRBNS enrichment of human not bound *in vivo* (light grey; dotted), SH-SY5Y-specific oligos (purple), and SH-SY5Y and HeLa shared oligos (blue) *Inset shows significance values for all comparisons via KS test and corrected for multiple comparisons via the BH procedure. Red denotes significant (p≤0.05). Values are as follows: a (ns), b (p≤0.1), d, (p≤0.01).*



- Supp Fig 5

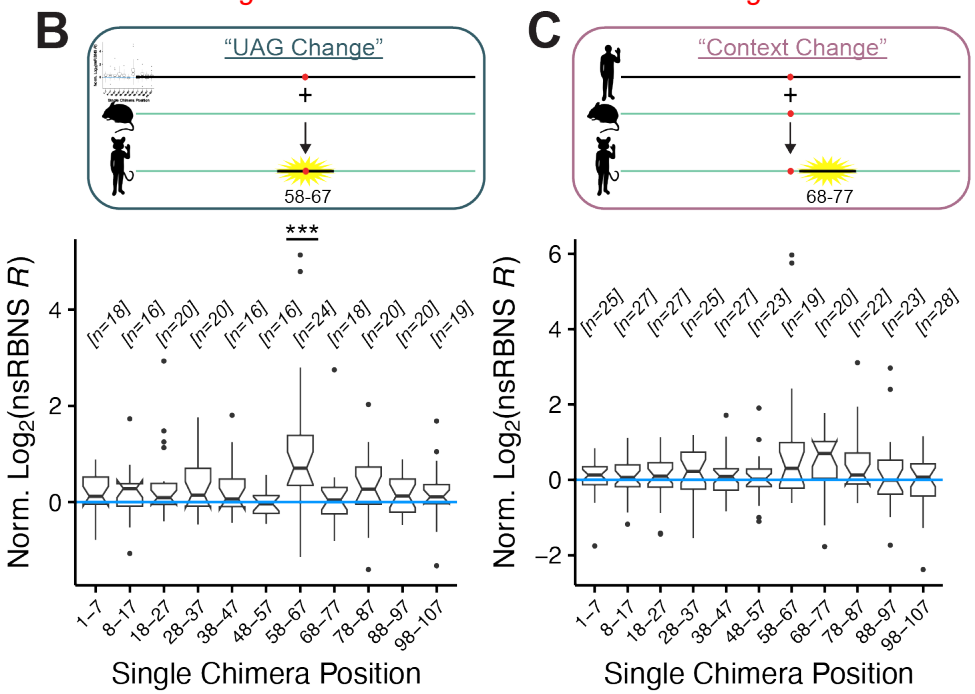


Comment 3: *It is not clear the number of oligos and the corresponding permutations represented in each Bind-n-Seq library. For Figure 1, the authors should specify the total number of sequences analyzed. For Figure 4, the authors should specify the total of single and double chimeras.*

The text and figures have been edited to include numbers as follows:

"We derived UNK binding sites from iCLIP data in one-to-one orthologous human and mouse genes and designed 12,287 natural RNA sequences, each 120 nucleotides long. Contained within this "pool" were UNK binding sites identified via iCLIP in human *(n=2,023)* and mouse *(n=2,346)* neuronal cells and tissue, respectively, as well as orthologous regions *(human: n=2,335; mouse: n=1,906)* whether or not they displayed evidence of binding in cells. Sequences were designed such that UAGs identified via iCLIP were located in the center of each oligo whenever possible. Non-bound control regions *(n=2,474)* were also selected to have similar UAG content. Additionally, 11,967 mutated oligos were also included and are discussed below."

"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*."



Comment 4: *The authors analyze Bind-n-Seq as the "frequency of an oligo in the protein bound sample divided by the frequency in the input". It is not clear if "frequency" refers to number of reads or total number of normalized reads (normalized by the size of the library). The authors should clarify this point of the analysis.*

We agree with this suggestion and have modified the text to include the appropriate description:

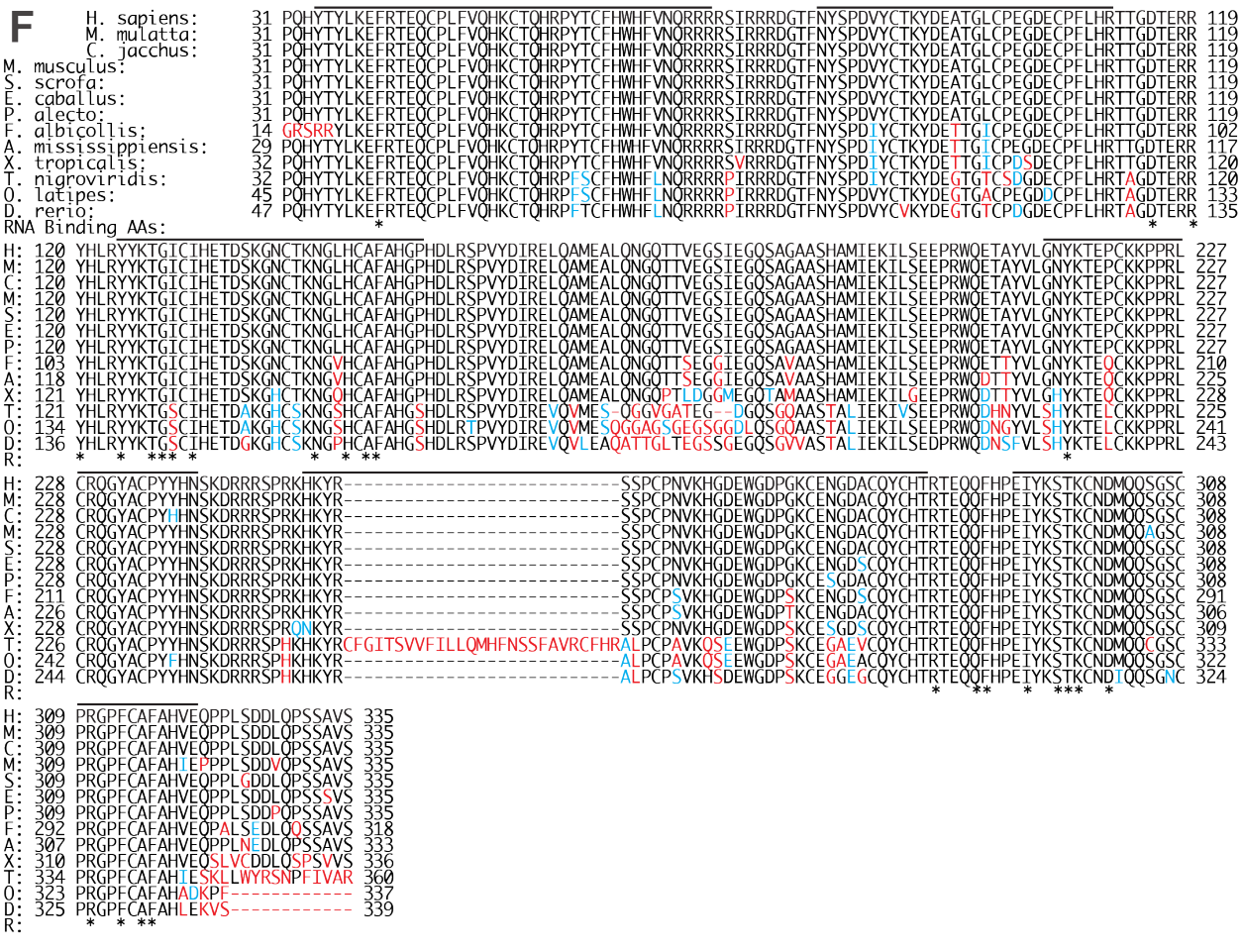
"RNA sequencing was used to quantify the abundance of each RNA bound to UNK as well as the abundance of each RNA in the input RNA pool. These experiments yield binding enrichments (*R* values) for each oligo which are defined as the frequency *(normalized count for library size)* of a given oligo bound to UNK vs the frequency of that oligo in the input RNA."

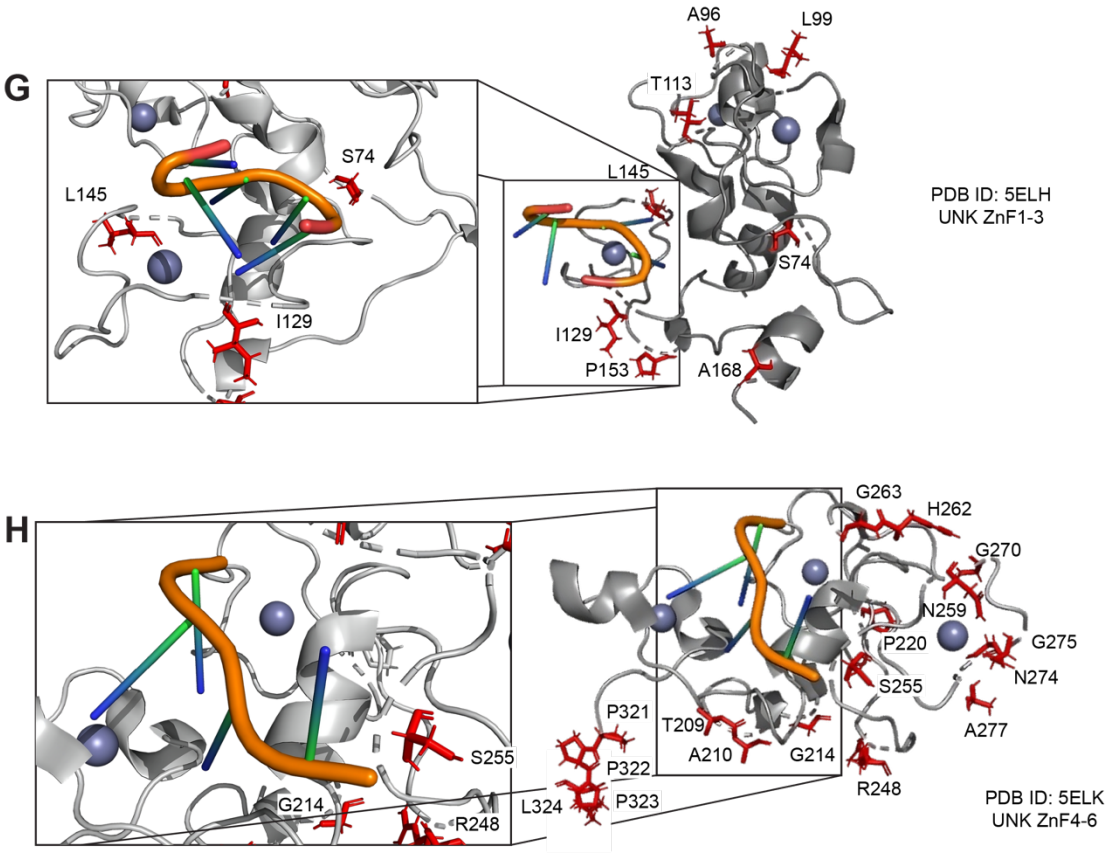
Comment 5: *To understand to what degree we expect or not differences in the binding of UNK across species, it would be important to provide a protein sequence alignment, and a structure showing the binding domain where the divergent sites are highlighted.*

Thank you for the suggestion. Due to the complexity of the display for the alignment of 100 protein sequences, we have elected to select representative species within the tree and used their sequences for alignment, including only the predicted RNA-binding domains based on alignment. Additionally, we have included two structures from Murn et al., 2016 and highlighted changed residues across species in red. These have been added to supp. figure 5.

Citations:

Murn, J., Teplova, M., Zarnack, K., Shi, Y., & Patel, D. J. (2016). Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. Nature Structural and Molecular Biology, 23(1), 16–23. https://doi.org/10.1038/nsmb.3140

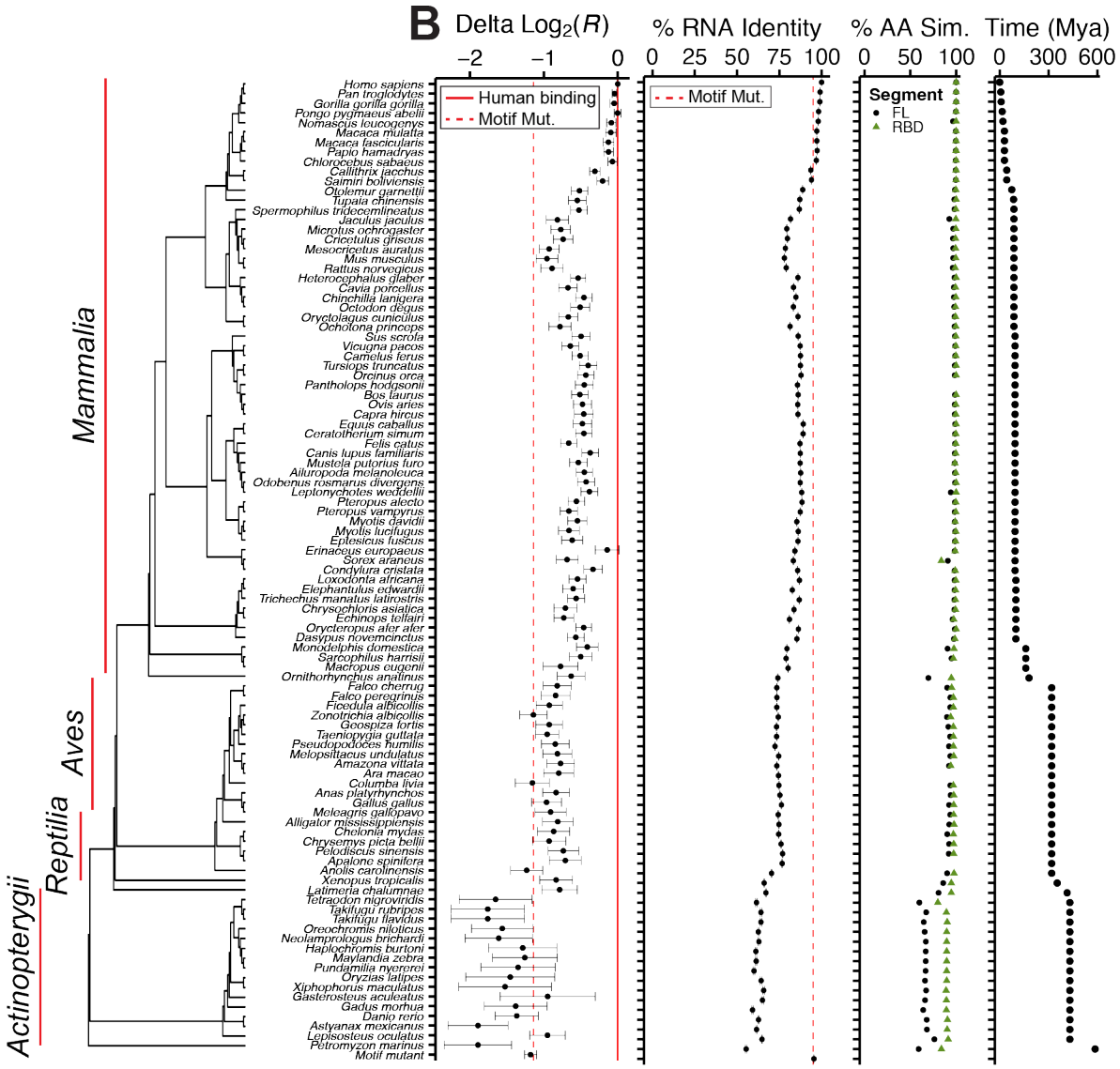




Legend: F) Multiple sequence alignment of the RNA-binding domains of UNK across select vertebrates. Individual ZnF domains are highlighted via black bar. Blue denotes similar amino acids while red denotes non-similar sequence divergence as predicted by BLAST. Asterisks denote direct RNA-contacting residues as predicted based on Murn et al., 2016 G) PDB of UNK ZnF1-3 from Murn et al., 2016 with less stringently conserved residues highlighted in red. Note: only I129 is predicted to have direct RNA contacts based on the crystal structure. H) PDB of UNK ZnF4-6 from Murn et al., 2016 with less stringently conserved residues highlighted in red. Note: no evolving residues are predicted to have direct RNA contacts based on the crystal structure.

Comment 6: *In Figure 5B, the red discontinuous lines are not defined in the figure legend and are not easy to interpret. Also, it is not intuitive that % identity refers to target RNA sequence and that %Similarity refers to UNK amino acid sequences.*

Thank you for pointing this out. We have adjusted the figure legend and figure as follows:



Legend: 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).

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors addressed all of our previous comments in the rebuttal and revisions to the manuscript.

Reviewer #2 (Remarks to the Author):

In this revised version of the manuscript, the authors have addressed all my comments and clarification requests. Specifically, the authors have:
- clarified that only use CLIP data for genes that are expressed to compare site conservation between human and mouse.
- added and standardized statistical analysis across main figures and supplementary figures.
- clarified the number of oligos and their configuration in the Bind-n-Seq library.
- added both protein alignments and structures to showcase the differences and similarities of UNK across species.

With these clarifications and additional information, the manuscript now is ready for publication in Nature Communications.