Inherent Specificity and Mutational Sensitivity as Quantitative Metrics for RBP Binding

Soon Yi[1,2]*, Shashi S. Singh[1,2], Xuan Ye[3], Rohan Krishna[1,2], Eckhard Jankowsky[4], Joseph M. Luna[1,2]*

[1] Center for RNA Science and Therapeutics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA
[2] Department of Biochemistry, School of Medicine, Case Western Reserve University, Cleveland, OH, USA
[3] Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, USA
[4] Moderna, Inc., Cambridge, MA, USA
*Correspondence: sxy530@case.edu, jml371@case.edu

Abstract

Interactions between RNAs and RNA binding proteins (RBPs) regulate gene expression in eukaryotic cells. RNA-RBP affinities measured *in vitro* reveal diverse binding specificities, yet approaches to directly compare specificities across RBPs are lacking. Here, we introduce two quantitative metrics: *inherent specificity*, which measures how selectively an RBP distinguishes its strongest binding motif from all possible motifs, and *mutational sensitivity*, which assesses tolerance to single nucleotide variations within preferred motifs. Analyzing high-throughput sequencing datasets, we compared these metrics across 100 RBPs *in vitro* and 27 RBPs in cells, finding a strong correlation between *in vitro* and cellular measurements for RBPs that bind RNA independently of a local structural context. Through RNA binding domain swap CLIP experiments between low-specificity RBM25 and high-specificity HNRNPC, we demonstrated that sequence specificity can be measurably transferred between protein contexts. Using these insights, we developed mathematical models illustrating how inherent specificity and mutational sensitivity influence competitive RBP binding, observing that low-specificity RBPs can functionally enhance the specificity of high-specificity RBPs by occupying non-target sites. Together, our results provide a quantitative framework for modeling RNA-RBP interactions and designing RBPs with targeted specificity.

Keywords: RNA-Protein Interaction, RNA Binding Proteins, RBPs, CrossLinking and ImmunoPrecipitation (CLIP), RBP Specificity, HNRNPC, RBM25

Introduction

Gene expression in higher eukaryotic cells involves coordinated interactions between thousands of RNA-binding proteins (RBPs) and tens of thousands of RNAs.[1–4] These interactions create regulatory networks responsible for controlling all facets of post-transcriptional regulation of gene expression, including RNA splicing, export, translation, and decay.[5–8] Deregulations of RNA-RBP networks have been linked to a spectrum of human diseases, ranging from neuropathies and inflammations to various cancers.[9–11] The central biological roles of RNA-RBP networks for cellular functions emphasize the need to understand the principles that govern RNA-RBP interactions.

RBPs, like all nucleic acid binding proteins,[12] must discern biologically relevant RNA targets among many potential targets to exert their functional effects. How do RBPs differentiate between target and non-target RNA sites? This distinction often hinges on the nature of binding: RNA-RBP interactions are typically categorized as 'specific' when binding is concentrated on a few degenerate RNA sites around a primary motif, or 'non-specific' when RBPs associate with a wide array of RNA sites without an apparent preference for any specific motif.[13,14]

This binary framework has been useful in outlining basic parameters of RBP binding preferences at the motif level, exemplified by the use of tools such as logo plots and position weight matrices (PWMs) for visualizing and quantifying binding patterns.[15] However, evidence from high-resolution *in vitro* binding assays, such as RNACompete, RNA Bind-n-Seq (RBNS), and HITS-Eq,[4,16–18] has unveiled a wide range of RBP binding affinities. While this spectrum of binding affinities, defined as an affinity distribution, is unique to each RBP, currently there are no means to compare the affinity distributions across different RBPs. Our knowledge of RNA-RBP interactions within cells has further been revolutionized by CrossLinking and

ImmunoPrecipitation (CLIP) approaches that have been used to decipher transcriptome-wide binding patterns of hundreds of RBPs in a wide variety of biological systems,[5,19–23] yet the complexity of these datasets has made it challenging to directly compare the binding affinity distributions of these proteins within cells.

Here, we address this limitation by developing two straightforward metrics that quantify how RBPs discriminate between their RNA targets. The first metric, termed the *inherent specificity* (IS), quantifies the degree by which an RBP distinguishes the strongest binding RNA motif from the rest of its potential targets. The second, *mutational sensitivity* (MS), measures the RBP's tolerance to a single nucleotide change within its strongest binding motif.

Using these metrics, we analyzed binding behaviors across 100 RBPs and found that IS and MS values remain consistent regardless of motif length. Our analysis of *in vitro* versus cellular binding revealed that RBPs binding to unstructured RNA regions display consistent specificity metrics across both experimental contexts.

To test the biological relevance of these metrics, we performed RRM-swap experiments between two proteins: HNRNPC (high specificity) and RBM25 (low specificity). Using CLIP experiments on engineered cells expressing these domain-swapped proteins, we demonstrated that sequence specificity is partially RRM mediated, with the high-specificity HNRNPC RRM enhancing RBM25's apparent specificity for HNRNPC motifs while maintaining its original genomic targeting preferences. Our metrics successfully quantified these changes, confirming their utility for measuring manipulations of binding specificity.

Finally, we developed mathematical models of competitive binding among RBPs with different IS and MS values, revealing how these properties could influence target selection in complex cellular environments. Together, our findings provide a quantitative framework for understanding RNA-RBP binding specificity with applications for both basic RNA biology and the design of proteins with tailored binding properties.

Results

Inherent specificity and mutational sensitivity as distinct metrics to quantify RNA-RBP interactions

To understand how RBPs interact with their RNA targets, it is useful to consider an RNA-RBP affinity distribution as a histogram representation of interactions between an RBP and all possible *K*-mer motifs (**Fig. 1A**). From the affinity distribution, the most preferred binding motifs can be used to generate RNA-RBP models such as position weight matrices or sequence logos.[15] In this context, the binary classification of RNA-RBP interactions into specific and non-specific represents the two extreme cases in an affinity distribution, where a non-specific RBP interacts with all possible targets with equal affinity, while a specific RBP interacts with very few, or often a single RNA target (**Fig. 1A**).[24,25]

While the specific/non-specific categorization is useful, it does not account for the diverse binding affinities RBPs display across potential targets.[13] To illustrate this, we used RBNS data from the ENCODE consortium[24,25] and analyzed the binding characteristics of RNA binding motif protein 25 (RBM25) (**Fig. 1B**) and heterogenous ribonucleoprotein C (HNRNPC) (**Fig. 1C**) by enumerating the per position frequency of nucleotides for the top observed 5-mer motifs to generate PWMs and sequence logos. For the subsequent analyses, normalized RBNS enrichment (*R*) scores for RBPs were used as an approximation of RBP binding affinity to target RNAs (see methods).[26] The sequence logos of the two RBPs (**Fig. 1B** and **C**, top left) reveal distinct binding preferences, where RBM25 predominantly binds to G-rich motifs, while HNRNPC demonstrates a strong preference for U-rich motifs. However, while sequence logos suggest similar levels of binding specificity for both RBPs, their affinity distributions reveal different binding behaviors (**Fig. 1B** and **C**, bottom right). Whereas the affinity distribution of HNRNPC shows a unique preference towards a “UUUUU” motif, resulting in a skewed distribution, RBM25 displays a broader bell-shaped distribution, indicating a similar affinity

across a wider range of sequences. These distinct affinity distributions demonstrate that HNRNPC has substantially higher binding specificity than RBM25.

Position weight matrices (PWMs) assign a numerical score to each nucleotide at every position within a motif, offering a more quantitative perspective underlying sequence logos.[15] From PWMs, a predicted affinity distribution can be generated[27] and compared to the measured affinity distributions based on RBNS datasets (**Fig. 1B** and **C**, top right). This comparison reveals notable differences between the PWM and actual binding behaviors, with these discrepancies being more evident for HNRNPC than for RBM25. While sequence logos and PWMs provide valuable insights into RNA-RBP interactions, their complexity of interpretation underscores the need for simpler, more accessible metrics that reflect the subtleties of RNA-RBP interactions.

To address the limitations of PWMs and sequence logos, we propose two metrics for a more detailed quantitative assessment of RNA-RBP interactions. The first term, *inherent specificity (IS)*, quantifies how uniquely an RBP recognizes its strongest binding sequence motif from all possible motifs of a given length. Using the affinity of the median binding motif as the representative tendency of the affinity distribution, IS is numerically defined as the ratio of the affinity for the top binding motif versus the median binding motif. The IS values of the two RBPs, RBM25 and HNRNPC, delineate how specific each RBP binds to their preferred sequence motif (**Fig. 1B** and **C**, bottom right). Additionally, IS values enable a quantitative comparison of binding specificities across different RBPs, illustrating, for example, that HNRNPC is approximately 20 times more specific in its binding than RBM25.

The second term, *mutational sensitivity (MS)*, takes a different approach to capture the binding behavior of an RBP. Rather than contrasting the strongest binding motif with others, MS evaluates how an RBP differentiates between its top binding motif and closely related sequence variants (**Fig. 1B** and **C**, bottom left). MS is calculated by comparing the affinity of each single-nucleotide variant, derived from the top binding motif, against the affinity of the original top motif.

Like IS, MS values facilitate a direct quantitative comparison of how different RBPs respond to alterations in their preferred motifs, revealing, for instance, that HNRNPC is about four times as sensitive as RBM25 to variations in its top binding motif.

Inherent specificity and mutational sensitivity show strong inter- and intra-correlations across different *K*-mers

To understand how changes in *K*-mer length affect our proposed metrics, we calculated IS and MS values from 4 to 7-mers within RBNS datasets (**Table S3**). For this analysis, we focused on 26 RBPs that had a single RNA binding domain (RBD) purified for the RBNS assay[25] to exclude any confounding effects from multi-RBD interactions with RNA. Our analysis revealed a consistent increase in IS values across all 26 RBPs with increasing *K*-mer length (**Fig. 2A)**, averaging a 1.55-fold change between *K*-mers. Notably, this increase was more pronounced in RBPs with higher specificity. For example, significant fold-changes ($|Z \geq 2.6$) in IS values were observed in PRR3 (6-mer IS: 30.8, 7-mer IS: 96.6), RBFOX3 (4-mer IS: 17.2, 5-mer IS: 76.1), and RBM22 (5-mer IS: 29.9, 6-mer IS: 93.9).

Conversely, MS values remained stable (23 out of 26 RBPs) across different *K*-mer lengths (**Fig. S1**), with an average fold-change of 1.04, suggesting that mutational sensitivity is less affected by motif length than inherent specificity. However, certain RBPs did show notable changes ($|Z| \geq 2.6$) in MS values during specific K-1 to K transitions: DAZ3 experienced a decrease in MS from 5-mer to 6-mer, KHDRBS3 an increase from 6-mer to 7-mer, and SRSF11 an increase from 4-mer to 5-mer.

Despite the variations in IS and MS values with different *K*-mer lengths, there was a notable consistency in the ranking of inherent specificity and mutational sensitivity across *K*-mers. This was shown by both rank comparisons (**Fig. 2B**) and correlations within each metric for different *K*-mer combinations (**Fig. 2C**). These observations suggest that the relative dynamics of inherent specificity and mutational sensitivity remain stable for RBPs across

various *K*-mer lengths. Furthermore, despite IS and MS quantifying different aspects of RBP behavior, we observed strong correlations between these metrics across various *K*-mer lengths (**Fig. 2D**) suggesting that highly specific RBPs also tend to be more sensitive to mutations in their target motifs.

To examine how IS and MS values reflect the biological functions of RBPs, we next analyzed these metrics in RBPs with established collaborative roles. Serine-argine rich splicing factors (SRSFs) regulate pre-mRNA splicing by interacting with intronic and exonic splicing enhancers, while HNRNPs regulate splicing by interacting with splicing silencers.[28,29] Our analysis revealed that SRSFs exhibit lower IS and MS values, with an average IS of 2.4 and an average MS of 0.34, whereas HNRNPs have higher values with an average IS of 13.5 and an average MS of 0.52 (**Fig. S2**). These findings suggest that RBPs with similar functional roles may have similar inherent binding behaviors, providing redundancy in their RNA-RBP interaction capabilities alongside their functional redundancy.

Sequence-specific RBPs show cellular binding that is consistent with their inherent binding behaviors

Building from this, we next compared the inherent binding of RBPs *in vitro*, using IS and MS values, with their binding activities in cells. IS and MS values are obtained from *in vitro* experiments where a single RBP interacts with an equal mix of RNA substrates covering all possible *K*-mers.[17] Such a scenario, while tractable, represents a simplified model compared to the complex cellular environment, which involves numerous competing RBPs, variable protein expression levels, tissue-specific factors, and subcellular compartmentalization that creates an uneven distribution of RNA targets.[1–3]

Operating under the premise that CLIP datasets might capture RBP binding behavior in cells, we calculated motif enrichment scores for all 5-mers on eCLIP datasets for RBPs with corresponding RBNS data. We termed motif enrichment scores from eCLIP as 'apparent'

binding behaviors to distinguish them from 'inherent' metrics derived from RBNS data. With this approach, shown for the four RBPs EIF4G2, PCBP2, RBFOX, and HNRNPC, we compared the inherent binding behaviors with the eCLIP-derived apparent binding behaviors within cells (**Fig. 3A**).

EIF4G2, which has low IS and medium MS values (2.04 and 0.44, respectively), showed poor agreement between the inherent and the apparent binding measurements. Conversely, PCBP2 (IS: 21.6, MS: 0.49), which is approximately 10 times more specific than EIF4G2, displayed an improved correlation between *in vitro* and *in vivo* motif enrichments. For the RBPs RBFOX2 (IS: 31.5, MS: 0.90) and HNRNPC (IS: 50.8, MS: 0.84), which exhibit high specificity and mutational sensitivity, the strongest binding motifs identified *in vitro* ("GCAUG" for RBFOX2 and "UUUUU" for HNRNPC) remain dominant in cells. However, the correlation for other motifs between the inherent and apparent binding patterns is lower, indicating variability in binding preferences beyond the primary motif. These observations indicate that while RBPs with moderate specificity and sensitivity show consistent binding patterns, those at the extremes may exhibit more complex behaviors within cells compared to *in vitro*.

We next compared specificity and sensitivity values calculated from eCLIP datasets with their RBNS-derived counterparts (**Fig. 3B**). Similar to inherent specificities and sensitivities, their apparent counterparts correlated with each other (**Fig. 3B**), confirming that our metrics capture fundamental properties of RNA-RBP binding even when applied to cellular binding data. Intriguingly, we observed that several RBPs exhibit a shift in their specificity when comparing *in vitro* and cellular measurements. For instance, RBFOX2's specificity decreased close to 5-fold, from an IS of 31.5 to an apparent specificity (AS) of 6.63. Conversely, EIF4G2's specificity increased approximately 20-fold, from an IS of 2.04 to an AS of 40.3.

To understand the discrepancy between inherent and apparent specificity and sensitivity measurements, we separated the RBPs based on their preferences for RNA secondary structure context around the target motif.[25] If an RBP displayed increased enrichment in either a

stem loop, stem bulge, or stem of an RNA hairpin, we designated this RBP to have a high preference for a structural context around a target motif. Otherwise, the RBP was designated to have a low structural motif preference. We found the largest discrepancies between inherent and apparent specificity in RBPs that recognize RNA motifs in a local structural context (**Fig. 3C**). In contrast, we observed a positive correlation between inherent and apparent metrics for RBPs with a low structural context. MS values were more resistant to RBP preference for structural context as RBPs with a high preference for structural context also showed a positive correlation. Overall, these results suggest that RNA secondary structure significantly influences the differences between in vitro and in vivo binding behaviors. Our metrics provide a quantitative framework for understanding these context-dependent differences.

Inherent specificity encoded in RRMs is transferable between protein contexts

Our comparison of *in vitro* and cellular measurements revealed that RBPs with low structural context preferences, such as HNRNPC, display a strong correlation between inherent binding behaviors measured *in vitro* and apparent binding patterns observed in cells. However, *in vitro* RBNS experiments characterize partial proteins consisting primarily of RNA binding domains[24,25] while CLIP captures interactions of complete proteins in their native cellular environment. To isolate a single RRM domain's specific contribution to binding specificity within cells, we designed an RRM-swap experiment between two contrasting single-RRM containing RBPs: HNRNPC (IS: 50), which shows high specificity for U-rich motifs, and RBM25 (IS: 2.5), which exhibits low specificity for G-rich motifs. Despite sharing structural similarity, these RRMs share only 18% sequence homology (**Fig. S3**).

For this experiment, we engineered HEK293 cells to express V5-epitope tagged wild-type and RRM-swapped versions of HNRNPC and RBM25 under doxycycline induction (**Fig. 4A** and **Fig. S4**). While HNRNPC has been extensively studied using CLIP approaches,[21,30,31] RBM25 has been proposed to bind G-quadruplex structures[32] with low direct sequence

specificity (**Fig. 1B**) . We appended N-terminal V5 tags to WT and RRM-swapped constructs to enable consistent immunoprecipitation, performed UV CLIP experiments on all four cell types (**Fig. S4**), and analyzed the resulting libraries using the CLIPittyClip pipeline, comparing them with reanalyzed wild-type HNRNPC datasets[24,30] as baseline controls.

Comparison of CLIP datasets revealed consistent peak annotation distributions for HNRNPC across different experimental contexts, with strong intron enrichment (**Fig. 4B**). Remarkably, the RRM swap had minimal impact on genomic region preferences - both wild-type and mutant HNRNPC showed similar intronic enrichment, while both wild-type and mutant RBM25 maintained strong ncRNA preferences, particularly in rRNAs and tRNAs. This suggests that RNA biotype targeting for these proteins is largely independent of their RRMs.

The apparent sequence motif preferences of the RBPs were affected differentially between the two proteins (**Fig. 4C**). The mutant HNRNPC maintained U-rich motif binding similar to wild-type, showing little change in motif preference despite the RRM swap. In contrast, the mutant RBM25 demonstrated a lowered preference for G-rich motif compared to the wild-type RBM25, a change aligning with the HNRNPC RRM. While the two sets of PWM suggested that the RBPs were specific towards their respective preferential motifs, motif affinity distributions showed stark differences between HNRNPC and RBM25 (**Fig. 4D-E**). While both the wild-type and the mutant HNRNPC maintained nearly identical high apparent specificity for "UUUUU" motifs (wild-type: 23.5; mutant: 22.8), the mutant RBM25 showed a significant 2-fold increase in apparent affinity for "UUUUU" compared to wild-type (wild-type: 1.3; mutant: 2.6). The mutant RBM25's specificity for "UUUUU" increased while its specificity for other U-rich motifs generally decreased (mean-fold change: 0.98, **Table S4**), mimicking HNRNPC's characteristic preference for "UUUUU" over other poly-U motifs.

Apparent MS values also showed corresponding changes between wild-type and mutant RBPs. While HNRNPC maintained similar apparent MS values for "UUUUU" motifs (wild-type: 0.63; mutant: 0.61), mutant RBM25 exhibited an approximately 3-fold increase in

apparent MS for "UUUUU" motifs (wild-type: 0.11; mutant: 0.32) (**Fig. S5**). These results demonstrate that sequence specificity encoded in an RRM can be transferred to a different protein context, while suggesting that RNA biotype preferences are influenced by other protein features.

Finally, we examined the relative normalized read abundance at shared binding sites between wild-type and mutant proteins (**Fig. 4F-G**). For HNRNPC, most shared peaks showed reduced read per million counts (RPM) in the mutant compared to the wild-type, particularly at sites with higher wild-type binding. In contrast, mutant RBM25 containing the high-specificity RRM from HNRNPC exhibited increased read counts at most shared binding sites compared to wild-type RBM25. This asymmetric pattern demonstrates that the high-specificity HNRNPC RRM enhances the binding signal when transferred to RBM25, while the low-specificity RBM25 RRM reduces the binding signal when introduced into HNRNPC.

Taken together, these RRM-swap experiments reveal a directional transferability of binding properties, where high inherent specificity can be transferred to augment the sequence recognition of a typically low-specificity RBP, while the reverse transfer has more modest effects. These findings demonstrate that the binding behaviors quantified by our IS and MS metrics represent modular properties partially encoded within RRMs that can be experimentally manipulated.

Simulations of RNA binding for four model RBPs with unique inherent binding modes reveal the complexity of RBP decision-making

Our RRM-swap experiments showed that binding specificities can be partially transferred between proteins by exchanging their RRMs. This raised an important question: how do different combinations of IS and MS values influence competitive RNA binding when multiple RBPs target the same sequences? To explore this question, we developed a mathematical model examining how four hypothetical RBPs with varying IS and MS values would interact with

a single target RNA (**Fig. 5A**). The hypothetical RBPs were designed so that they all bind with the strongest affinity to uridine-rich motif and the weakest affinity to guanine-rich motif, but with distinct binding landscapes, as illustrated by their unique IS values (**Fig. 5B**) and MS values (**Fig. 5C**).

For the simulation, two important ground rules were set. First, all four RBPs were introduced simultaneously, ensuring equal chances of interaction with the RNA. Second, we considered each RNA binding site and its interactions with the RBPs as independent events. This allowed us to calculate the equilibrium concentration of RNA-RBP complexes for each binding site on the RNA (**Fig. 5D**). The equilibrium concentrations were then used to calculate the binding probability of each RBP on a given target site (see Methods).

Initially, we examined two different scenario concentrations of RBPs: one with all RBPs at equal concentrations (**Fig. S6A**), and another where RBPs with higher specificity were present at half the concentration (**Fig. S6B**). For both scenarios, a range of RNA concentrations was tested to examine how each RBP bound along the RNA. For the "UUUUU" motif, at equal RBP concentrations (**Fig. S6C, left**), high-specificity RBPs (blue) showed a higher probability of binding compared to low-specificity RBPs (red) across all RNA concentrations. Conversely, with varying RBP concentrations (**Fig. S6D, left**), low-specificity RBPs could better compete against high-specificity RBPs, particularly at higher RNA concentrations that offer more binding sites. Notably, in both scenarios, among RBPs with similar specificity, those with lower mutational sensitivity consistently outperformed those with higher mutational sensitivity in binding to their target.

The binding simulation also suggests how low-specificity RBPs augment the binding of high-specificity RBPs by occupying sites less likely to be targeted by their high-specificity counterparts. For example, the "GGGGG" motif represents the weakest binding target for all model RBPs. In this context, low-specificity RBPs occupied this binding site in all tested scenarios indicating their role in blocking access to suboptimal sites and thereby directing high-

specificity RBPs to their preferred motifs (**Fig. S6C and S6D, right**). Conversely, the occupation of the preferred motifs by high specificity RBPs led to an apparent avoidance of the "UUUUU" motif by low-specificity RBPs (**Fig. 5E)**.

Our simulation with model RBPs predicts that competitive binding between RBPs is informed by distinct binding behaviors that are captured by IS and MS calculations. To extend this analysis beyond *in silico*, we examined the competitive binding between U2AF2, PTBP1, and HNRNPC RBPs, all of which bind to poly-pyrimidine (poly-uridine) rich motifs. We first calculated IS values for these proteins using RNACompete datasets[33] and observed that U2AF2 and HNRNPC (IS: 27.0) are approximately four times more specific than PTBP1 (IS: 7.5).

Using these characteristics, we then simulated U2AF2 binding to the intronic region of the *PTBP2* transcript in both the absence and presence of PTBP1 or HNRNPC and compared our simulation results to *in vitro* iCLIP experimental data.[34] We hypothesized that given the differences in the IS values of these RBPs, non-specific PTBP1 binding would enhance U2AF2 binding to its target sites, while equally specific HNRNPC should not have an effect on U2AF2 target recognition. In the absence of PTBP1 (**Fig. 6A**), our simulation recapitulated the U2AF2 binding distribution across *PTBP2* target RNA, mirroring iCLIP results. When U2AF2 binding was simulated in the presence of PTBP1, we observed sustained U2AF2 binding to its principle binding sites on *PTBP2* target RNA (**Fig. 6B**). On the contrary, in the presence of HNRNPC (IS: 27.0), our simulation indicated that U2AF2 retained its distributed binding pattern across its target RNA (**Fig. 6C**). Overall, our RNA-RBP binding simulation results recapitulate experimental observations and demonstrate that the inherent characteristics of RBPs can help predict the complex binding dynamics of RBPs with their targets.

Discussion

Here, we developed two metrics, inherent specificity and mutational sensitivity, to quantitatively characterize RNA-RBP interactions. IS assesses how an RBP can identify its strongest binding motif among all potential motifs, while MS evaluates an RBP's ability to differentiate its primary binding motif from its close variants. Although both metrics focus on the strongest binding motif, they are independent of each other in contrasting the primary motif against two distinct sets of motifs. By translating complex RBP binding behaviors into quantifiable values, these metrics provide straightforward means to compare the binding behaviors of different RBPs.

Our analysis across various *K*-mer lengths showed consistent specificities and sensitivities of RBPs to their target motifs. Notably, even with significant increases in metric values at extended *K* lengths, there remained a strong correlation within each metric across different *K* lengths. Furthermore, we observed a pronounced inter-metric correlation, indicating that RBPs with high specificity for their top motif are adept at discriminating it from its close variants. Of the 26 RBPs with a single RBD purified, 6 (PRR3, RBFOX3, RBM22, DAZ3, KHDRBS3, and SRSF11) exhibited significant IS or MS changes based on motif length. Intriguingly, 5 of these 6 RBPs have previously been reported to require a structural context around their target motifs for binding,[25] hinting that our metrics may capture aspects of the structural dynamics in RNA-RBP interactions.

These metrics also highlight the connection between the inherent binding modes measured *in vitro* and the apparent binding behaviors observed in cells through eCLIP assays. In our reanalysis, we observed a low correlation between *in vitro* and cellular measurements for RBPs. This discrepancy reflects inherent differences between experimental contexts. Cellular environments feature variable RNA concentrations, complex RBP competition and cooperation networks, and additional biological constraints absent from controlled *in vitro* assays. Furthermore, RBNS experiments typically examine isolated RNA-binding domains plus flanking

residues rather than full-length proteins, potentially missing contributions from other protein regions. These experimental differences likely account for the limited correlation observed between *in vitro* and cellular measurements. However, within cellular contexts, we noted a strong correlation between apparent specificity and mutational sensitivity, suggesting that the relationship between specificity and sensitivity is maintained both *in vitro* and *in vivo*. Notably, the correlation between *in vitro* and *in vivo* datasets was consistent for RBPs with low structural context preferences within each metric. Similarly, our comparative analysis of SRSFs and HNRNPs—splicing factors with antagonistic roles—revealed a clear separation between these groups based solely on their *in vitro* IS and MS values. These findings reveal that RBPs with similar cellular functions tend to share characteristic IS and MS profiles, suggesting these metrics reflect biologically meaningful properties.

We tested the utility of our metrics in the biological context through CLIP experiments on RBPs with swapped RRMs. We observed that mutant RBM25 containing HNRNPC's RRM exhibits increased binding to "UUUUU" motifs, mimicking the inherent binding behavior of HNRNPC. However, we did not observe a corresponding change in mutant HNRNPC containing RBM25's RRM. This asymmetric effect raises important insights about RRM function in their native biological contexts.

Several factors likely contribute to this asymmetry. First, HNRNPC may contain structural elements outside its RRM that are optimized for recognizing U-rich motifs, effectively compensating for the introduced low-specificity RRM. For example, a mutation outside of the RRM in RBFOX2 has been shown to confer distinct specificities for sequence motif selection.[26] Second, high specificity could be a more readily transferable property than low specificity–it may be easier to make a promiscuous binder more specific than the reverse. Lastly, HNRNPC functions in multimeric complexes,[35,36] potentially allowing mutant hnRNPC to maintain specificity through interactions with wild-type HNRNPC or other complex partners despite the RRM swap.

The conclusions from RBNS assays, which use isolated RNA binding domains plus flanking residues, aren't contradictory to these findings. Rather, it suggests that for many RBPs, the isolated RBD is sufficient for sequence specificity under controlled *in vitro* conditions. Our swap experiments reveal that in cellular contexts, the full protein architecture significantly influences observed binding behavior. The minimal changes in binding peak annotations between wild-type and mutant proteins further support the role of non-RRM segments in dictating genomic target preferences.

These results support a modular but context-dependent view of RBP specificity: the RRM provides primary sequence recognition capability, but its effectiveness depends on the broader protein architecture in which it's embedded. This framework helps explain why transferring a high-specificity RRM can enhance the specificity of a typically promiscuous binder while transferring a low-specificity RRM has a limited impact on an already specific binder. These insights demonstrate how RBP binding *in vivo* can be modulated through targeted domain swapping, guided by quantitative metrics of inherent specificity and mutational sensitivity.

Our RRM-swap experiments revealed how inherent specificity can be transferred between proteins, but left open questions about how RBPs with different specificity profiles interact in competitive environments. To explore this systematically, we created a mathematical model simulating RNA-RBP interactions with four hypothetical uridine-rich motif binding RBPs, each characterized by distinct IS and MS values. This model showed how high-specificity RBPs preferentially bind to their target motifs, while low-specificity RBPs are often outcompeted at these sites. We further observed how this competition can make low-specificity RBPs appear to avoid certain target motifs when examined in isolation. Our simulations, corroborated by *in vitro* iCLIP data,[34] demonstrate how quantifying inherent specificity and mutational sensitivity provides insights into competitive binding dynamics in complex environments. How these predictions scale with additional cell biological constraints such as RNA modifications,

subcellular localization, and RBP post-translational status will be important avenues for further study.

These findings open intriguing possibilities for understanding RNA-RBP networks. For instance, the RNA-binding protein C5, previously deemed nonspecific, was later found to discriminate between various binding site variants in specific cellular contexts.[18] There is also evidence that RBPs can alter their apparent specificity through mechanisms like post-translational modifications that affect RNA binding or protein localization.[37,38] Emerging technologies for analyzing subcellular RNA[39,40] and RBP-target interactions[41,42] will be crucial for understanding how specificity varies across cellular compartments. Additionally, our metrics may inform protein engineering efforts aimed at designing RBPs with customized binding properties.[43–47] Overall, these quantitative metrics provide a framework for systematically analyzing RNA-RBP interactions, providing an intuitive lens through which we can view and understand the regulatory logic between RNAs and their binding proteins.

Data and code availability

All original code for analysis and data visualization is freely available on GitHub at https://github.com/S00NYI/Specificity_BITs. All the sequencing data generated by this study have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number GSE291358.

Acknowledgments

We are grateful to William Merrick, Maria Hatzoglou, and Boaz Tirosh for their helpful discussion and feedback. We further thank Frank Tedeschi and Thomas J. Sweet for feedback and a critical reading of this manuscript. This work was supported in part by the NIH under the award T32 GM007250 (S.Y.), funding from the Case Comprehensive Cancer Center P30 CA043703 (J.M.L), the American Cancer Society IRG-16-186-21 (J.M.L), and NIGMS R35GM154771 (J.M.L), as well as startup funds from the Department of Biochemistry at the Case Western Reserve University School of Medicine.

Author Contributions

Conceptualization, S.Y., X.Y., E.J.; Investigation, S.Y., S.S.S., R.K., J.M.L; Writing – Original Draft, S.Y.; Writing – Review & Editing, S.Y., J.M.L., E.J.; Funding Acquisition, J.M.L.; Supervision, E.J and J.M.L.

Declaration of Interests

E.J. is an employee of and shareholder in Moderna, Inc. The other authors declare no competing interests.

Materials and Methods

Collecting and Processing *In Vitro* RNA-RBP Affinity Measurements:
RBNS datasets were collected from the ENCODE portal[24] by searching "RNA Bind-n-Seq" in "Experiment search." RBNS affinity measurements for each RBP are performed at several different RBP concentrations and analyzed for different *K*-mer lengths. For each *K*-mer enrichment data, the RBP concentration that generated the highest R-value was selected as the representative sample and used throughout the analysis. The RBNS dataset that used only a single purified RNA binding domain as well as RBPs' structural context preferences were manually curated based on previous literature.[25] A list of the RBNS accession numbers is provided in Table S1. The collected RBNS *R* scores were then feature-scaled to a scale of 1-to-Euler's constant (*e*), followed by natural log transformation to fit the data into a scale of 0-to-1.[26] All normalized datasets used are provided in Table S2.

Position Weight Matrix and Sequence Logo Generation:
From RBNS datasets, the top 20 motifs were used as input for the *R* program *consensusMatrix* under the *Biostrings* package.[48] The output nucleotide frequency table was then used as an input for the R program *PWM* under the *Biostrings package* to generate the PWM for each tested RBP. To generate a sequence logo, the generated PWM was used as an input for *R* program *seqLogo.*[49] with an 'RNA' option to generate the sequence logo. To calculate the predicted binding scores based on the PWM, the sum of the weights for a nucleotide at each position was calculated for all possible 5-mers. The calculated binding scores were then normalized as described before.

Collecting and Processing eCLIP dataset:
eCLIP datasets were collected from the ENCODE portal[24] by searching "eCLIP" in "Experiment search." For each RBP, merged bed narrowPeak files were downloaded. To calculate the motif enrichment, the peak coordinates were first extended to 25 nucleotides upstream from the 5'-end of the peaks. Then, the number of motif appearances across all extended peaks was counted for all possible *K*-mers. To control for the background motif distributions around the peaks, genomic coordinates random distances away from the peaks (up to 500 nucleotides) in both upstream and downstream directions were generated and the same motif counting procedure was performed for the 'background' sets. The background set generation was repeated 100 times and average background counts for each *K*-mer were calculated. Finally, the background counts were subtracted from the peak motif counts to generate the motif enrichment near peaks. If eCLIP data were available for K562 and HepG2 cells, average motif enrichment was calculated. The eCLIP enrichments were normalized the same as the *in vitro* datasets. See Table S1 for a list of eCLIP accession numbers used.

Inherent Specificity and Mutational Sensitivity
Inherent specificity measures the degree by which an RBP differentiates the strongest binding *K*-mer motif from the rest of the *K*-mers. Inherent specificity is calculated by taking the ratio of the top affinity over the median affinity of the affinity distribution. Numerically, inherent specificity is calculated as:

$$IS = \frac{affinity_{top}}{affinity_{median}}$$

Mutational sensitivity measures the degree of changes in binding affinity for an RBP when a single nucleotide is changed in its strongest binding *K*-mer motif. Mutational sensitivity is calculated by taking the average changes in affinity between the strongest binding motif and the motif with one nucleotide changed from the strongest binding motif. Numerically, mutational sensitivity is calculated as:

$$MS = 1 - \frac{affinity_{mutated}}{affinity_{top}}$$

For the apparent specificity and apparent mutational sensitivity calculations, eCLIP motif enrichment was used instead of RBNS affinity measurements.

Statistical Analysis

All statistical analyses were performed in R using the stats package. For correlation calculation between different datasets, Pearson and Spearman correlations were used. For fold-change Z-scores calculation, fold-change values were transformed to Z-scores using means and standard deviations calculated for each dataset.

Plasmid DNA Construction

To generate stable cell lines expressing the wild-type and the mutant hnRNPC and RBM25, we cloned each RBP construct into enhanced Piggybac (ePB) doxycycline-inducible expression vector.[50] The wild-type and the mutant hnRNPC construct was prepared as a custom gBlock from Integrated DNA Technologies. The wild-type RBM25 construct was prepared as cDNA reverse transcribed using total RNA extracted from HEK293 cells. For the mutant RBM25 construct, the DNA segments encoding for RBM25 RRM was replaced with hnRNPC RRM through overlap PCR. Finally, V5 epitope tag[51] was appended to the N-terminus of all constructs. Synthesized constructs were cloned between BamHI and NotI sites in ePB vectors. All primers used to generate these constructs are listed in Table S6.

Cell Culture and Stable Cell Line Generation

HEK293 cells[52] were maintained at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium (DMEM, Fisher Scientific, cat. #11995065) supplemented with 0.1 mM nonessential amino acids (NEAA, Fisher Scientific, cat. #11140076) and 5% hyclone fetal bovine serum (FBS, ATLAS Biologicals, cat. #F-0500-D, Lot # F05D22D1). All cell lines tested negative for mycoplasma contamination.

Stable cell lines were generated by transfecting 1μg of ePB plasmid encoding each RBP construct with 1μg of pTransposase plasmid[50] using Lipofectamine 2000 (Invitrogen, cat. #11668027). Cells underwent selection with 1μg/ml Puromycin (Sigma, cat. #P8833-25MG) to eliminate non-transfected cells. Doxycycline (Sigma, cat. #D9891-1G) was used at 1μg/ml to induce RBP expression.

Expression of the RBPs and their localizations were confirmed through western blot analysis and immunofluorescence imaging. For western blot, HEK293 cell lysates were prepared in 1x

PXL lysis buffer supplemented with complete protease inhibitors (Roche, cat. #11873580001). All lysates were prepared from equal cell numbers per condition. Protein concentrations were determined by Bradford assay (Biorad), and 20µg total protein per sample was run on NuPAGE gels (Life Technologies) and transferred to fluorescence-compatible nitrocellulose membranes (Millipore). Membranes were blocked in Odyssey PBS-based buffer (LI-COR) for 1 hr, then primary antibodies were added for an overnight incubation at 4°C. Antibodies used for western blotting were: HNRNPC (ThermoFisher Scientific, cat. #PA5-22280; 1:1000), RBM25 (ThermoFisher Scientific, cat. #PA-90511, 1:1000), rabbit V5-tag (Cell Signaling Technology, cat. #13202S, 1:1000), mouse V5-tag (Cell Signaling Technology, cat. #80076S, 1:1000), and Beta-Actin (Sigma, cat. #A5441, 1:5000). After 3 washes in 1X PBS/0.05% Tween-20, membranes were incubated fluorescent secondary antibodies (LI-COR, 1:25,000) for 1 hr at room temperature. Membranes were washed 3 times in 1X PBS/0.05% Tween-20, rinsed in 1X PBS, and visualized on the Odyssey Imaging system (LI-COR).

For immunofluorescence, cells were plated either on glass coverslips or black-walled clear bottom 96-well plates (Corning, Cat# 3904) that were coated with poly-L-lysine. Upon harvest, cells were fixed with 4% paraformaldehyde (PFA) for 10 minutes. PFA was then removed, and cells were stored at 4°C in PBS containing 1% FBS until processing. Cells were washed in PBS containing 0.1% Tween-20 (PBST), permeabilized with PBS containing 0.1% Triton X-100 for 10 min at room temperature, and blocked for 1h at room temperature with a blocking solution of 5% BSA in PBST. Cells were stained with primary antibodies overnight at 4°C. Antibodies used for immunofluorescence were: HNRNPC (ThermoFisher Scientific, cat. #PA5-22280; 1:1000), RBM25 (ThermoFisher Scientific, cat. #PA-90511, 1:1000), rabbit V5-tag (Cell Signaling Technology, cat. #13202S, 1:1000), and mouse V5-tag (Cell Signaling Technology, cat. #80076S, 1:1000). After primary antibody incubation, cells were washed and stained with goat anti-rabbit IRDye 680RD antibody (LICOR, cat. #925-68971, 1:1000) and goat anti-mouse IRDye 800CW antibody (LICOR, cat. #925-32210, 1:1000). Nuclei were counterstained with DAPI (ThermoFisher Scientific cat. #D1306, RRID:AB_2629482) at 1 µg/ml, for 5 minutes prior to imaging. Fluorescent images were obtained on a Keyence BZ-X710 microscope or on a Cytation 7 (Agilent).

RBP CLIP

CLIP from crosslinked HEK293 cell pellets using V5 antibody was performed generally following previous work[42,53,54] with modifications. Detailed experimental procedures are outlined below.

Tissue culture and crosslinking

Cells were plated onto 10 or 15 cm dishes, and expression was induced for 2 days with 1µg/ml of doxycycline. The culture media was replaced with ice-cold PBS and cells were immediately irradiated under 254nm UV light on ice, once for 400 mJ/cm^2 and once again for 200 mJ/cm^2 using a Stratalinker 2400 (Agilent Genomics). These energies were based on our previous work.[42,53,54] Cells were scraped into tubes and pellets were stored at -80°C for until downstream preparation.

Bead preparation
For V5 antibody pull downs, protein G Dynabeads (Invitrogen, cat. #10004D) were washed 3x with and resuspended in antibody binding (AB) buffer (AB: PBS, 0.02% Tween-20). Beads were incubated with 5µg of V5 antibody (Addgene, cat. #218107-rAB) per 25µL beads for 30 min at room temperature or overnight at 4°C. Prior to IP, beads were washed in 1x PXL lysis buffer (1x PXL: 1x PBS tissue culture grade without magnesium or calcium, 0.1% SDS, 0.5% Sodium-deoxycholate, 0.5% NP-40, with protease inhibitors).

V5 IP and on-bead enzymatic steps
Lysates from crosslinked cells were prepared by adding 1ml of 1x PXL lysis buffer + quenchers + protease inhibitors (Roche, cat. #11873580001) and triturating to disrupt cells. Lysates were treated with 10µL DNase (RQ1, Promega) and placed on ice for 10 minutes. Lysates were then treated with RNase I (Thermofisher, cat. #EN0602), first diluted to the indicated concentration by volume (1:100 for over digest or 'OD', or 1:1,000 for Low RNase) in 1x PXL and then added at 10µl per mL of lysate. RNase I concentrations were empirically determined for optimal isolation of RNA-RBP complexes with sizes suitable for cloning (Moore et al. 2014). Lysates underwent thermomixing for 5 minutes at 37°C at 1100rpm before being spun at 4°C on max speed of a table-top microcentrifuge for 10 minutes. All subsequent steps were done on ice or at 4°C unless otherwise indicated. Supernatants, along with any lipid layer, were harvested and mixed with PXL-equilibrated V5 antibody-bound beads for immunoprecipitation. Samples were nutated with beads at 4°C for 2 hours. Beads were washed sequentially twice each with ice-cold 1x PXL, 5x PXL (same as 1x but using 5x PBS), and 1x PNK buffer (50mM Tris-HCl, pH 7.5, 10mM MgCl2, 0.5% NP-40).

To prepare RNA 3' ends for linker ligation, IPs were treated with alkaline phosphatase. Beads were resuspended in 40µL containing 1x dephosphorylation buffer, 3U of CIAP (Roche), RNasin inhibitor (Promega), and thermomixed for 20 minutes at 37°C, shaking at 1100rpm for 15 s every 2 min. Samples were washed as above sequentially in ice-cold 1x PNK, 1x PNK plus 20mM EGTA, and twice with 1x PNK.

Radiolabeled linkers were prepared with a poly-nucleotide kinase (PNK) reaction consisting of 100pmol of a 3' inverted ddT blocked L32 RNA linker, 0.5µL of 32P-gamma-ATP (Revvity, cat. #NEG035C005MC), 0.5µl 1x T4 PNK buffer, 0.25µL T4 PNK and 0.3µl RNasin inhibitor in a total volume of 5µl per sample. Radiolabeled linker was usually prepared for 10 ligation reactions in a volume of 50µl. Linkers were incubated for 20 min at 37°C, after which 2µl of 10mM ATP was added, and the reaction was incubated at 37°C for an additional 5 minutes. Linker was purified by passing through a G-25 column (GE Healthcare, cat. # 27-5325-01) and stored at -20°C until use.

Linker ligation at 3' ends was set up per sample after washes following alkaline phosphatase treatment by preparing a T4 RNA ligase 1 (NEB, cat. #M0204S) reaction in 40µL following the manufacturers' instructions with 100pmol of radiolabeled L32 RNA linker. Samples were incubated overnight at 16°C, shaking at 1100rpm for 15 s every 4 min. The next day, beads were washed twice each with 1x PXL and 5x PXL, before being equilibrated in 1x PXL.

RNA-RBP complex isolation and autoradiography
All beads were subsequently washed twice each with 1x PXL, 5x PXL, and 1x PNK. Protein was eluted off the beads by incubating with 30µl of 1X LDS loading buffer (Invitrogen) without reducing agent for 10 min at 70°C, shaking at 1100rpm. Supernatants were run on Novex NuPAGE 8% Bis-Tris gels (Invitrogen, cat. #WG1001BOX) in SDS-MOPS buffer at 4°C. Radiolabeled protein RNA complexes were transferred to BA85 nitrocellulose (Cytiva Amersham, cat. #45-004-007). After transfer, the membrane was rinsed with RNase-free PBS, and exposed to Biomax MR film (Kodak) at -80°C typically from 3 hr to up to 3 days. Alternatively, membranes were placed onto phosphor screens (GE cat. # BAS-IP MS 2025), and imaged with a Typhoon Scanner (GE Amersham).

Nitrocellulose membranes were aligned with the exposed film and regions of the membrane from low RNase IP lanes were excised corresponding to signal intensities for RNA-RBP complexes between 42-72kDa for hnRNPC and 160-180kDa for RBM25. The 10-30kDa above the RBP size region (32kDa for hnRNPC, 150kDa for RBM25) was chosen to confine the CLIP-seq analysis to RNA target fragments of 30-90nt nucleotides (assuming 3nt/kDa) to avoid confounding variables from higher molecular weight complexes. RNA was liberated from membrane fragments using 200µL of a 4mg/ml proteinase K solution (Roche, cat. #3115828001) diluted in PK buffer (100mM Tris-HCl, pH 7.5, 50mM NaCl, 1mM EDTA, 0.2% SDS) and incubated for 60 min at 50°C, shaking at 1100rpm for 15 s every 2 min. RNA fragments underwent acid phenol:chloroform extraction using and were precipitated overnight at -80°C. RNA was pelleted by spinning at max speed (> 13,000rpm) in a table top centrifuge at 4°C, and washed twice with 75% ethanol. Following the drying of the RNA at the bench, the pellet was dissolved in 8µL RNase-free water.

RBP footprint library generation and sequencing
CLIP footprints were reverse-transcribed using the Br-dU incorporation and bead-capture strategy described previously.[55,56] Indexed reverse transcription (RT) primers were used (listed in Table S6), allowing multiplexing of 22 samples per Nextseq 500 run. cDNA was circularized with CircLigase (Epicentre), pooled and then amplified with PCR primers with Illumina sequencing adapters as described previously.[55,56] Amplification was tracked with SYBR green (Life Technologies) on the Quantstudio real-time PCR machine (Thermo), and reactions were stopped once signal reached 100k relative fluorescence units (r.f.u.). Products were purified with Ampure XP beads (Beckman) and quantified via Qubit assay (ThermoFisher) and/or Tapestation system (Agilent). Multiplexed samples were run on the Illumina NExtSeq Mid-Output with 75 base pair single-end reads.

CLIP data analysis
CLIP data analysis was performed using the CLIPittyCLIP pipeline as described before.[42] CLIP motif analysis was performed as outlined in the previous section.

RRM Structure Prediction and Comparison

Amino acid sequences of hnRNPC (UniProt ID: P07910) and RBM25 (UniProt ID: P49756) were acquired through UniProt database.[57] The full length amino acid sequence was used as input for AlphaFold protein structure prediction program[58] and UCSF ChimeraX software was used to visualize the structures.[59] Amino acid sequence alignment was performed through MUSCLE alignment software.[60]

Mathematical Simulation of RNA-RBP Interactions

The four hypothetical RBPs with different 5-mer IS and MS combinations were created by curating the normalized affinity distributions. Four different affinity distributions for the 1,024 5-mer motif affinities were first created so that they follow a normal distribution. For simplicity, all four RBPs were set to bind 'UUUUU' with the highest affinity. Then, for RBPs with high IS, 5-mer motifs excluding the top binding motif were shifted to a lower affinity regime. Then, for RBPs with low MS, the 5-mer motifs with single nucleotide mutations from the top binding motif were placed near the top binding motif. For the RBPs with high MS, the same set of 5-mer motifs was placed at the lowest end of the affinity distribution. Finally, the affinities for motifs that are neither the top binding motif nor motifs with single nucleotide changes from the top binding motifs were shuffled and curated to give each RBP a unique binding affinity landscape. The equilibrium dissociation constant (K_D) towards the top motif for the RBPs was all set to 0.02 and K_D values for other motifs were scaled accordingly using the affinity distribution. See Table S5 for the normalized affinity distributions, K_D distributions, and the respective IS and MS values of the four hypothetical RBPs.

For a given RNA sequence of length l that interacts with two or more RBPs, there exists $l - (k-1)$ possible K-mer binding sites. For each binding site on the RNA, a kinetic reaction scheme can be constructed where P_i is an ith out of N RBP interacting with the binding site R on the given RNA:

$$P_i + R \leftrightarrow P_iR$$

Then, at equilibrium:

$$K_{D,i} = \frac{[P_{i,eq}][R_{eq}]}{[P_iR_{eq}]} \quad (1)$$

$$[P_{i,eq}] = [P_{i,0}] - [P_iR_{eq}] \quad (2)$$

$$[R_{eq}] = [R_0] - \sum_{i=1}^{N} [P_iR_{eq}] \quad (3)$$

where $K_{D,i}$ is the dissociation constant for P_i and $[P_{0,i}]$ and $[R_0]$ are the initial concentrations of the RBPs and RNA, respectively. Combining equations 1 and 2, and solving for the equilibrium concentration of the P_iR complex yields:

$$[P_iR_{eq}] = \frac{[P_{i,0}][R_{eq}]}{K_{D,i} + [R_{eq}]} \quad (4)$$

Combining equations 3 and 4 yields an equation in terms of initial conditions and the equilibrium concentration of the binding site:

$$[R_{eq}] = [R_0] - \sum_{i=1}^{N} \frac{[P_{i,0}][R_{eq}]}{K_{D,i} + [R_{eq}]} \quad 5$$

For the system of N=4, solving equation 5 in terms of $[R_{eq}]$ yields a 5th-degree polynomial, which can be solved with the boundary condition:

$$0 \leq [R_{eq}] \leq [R_0]$$

From the $[R_{eq}]$, the equilibrium concentration of the P_iR complex can be determined with equation 4.

Dividing the equilibrium P_iR complex concentration $[P_iR]$ by the initial concentration of the RNA $[R_0]$ yields a probability of the RNA binding site occupied by the protein, p_{site}. Using p_{site} for all possible binding sites on a given RNA, per position probability, p_{pos}, can be calculated by taking the average of the overlapping p_{site} for each position (up to k overlap). Then, for any given K-mer motif within the RNA sequence, average probability can be calculated by taking the average of the k p_{pos} that span the motif.

For RBP enrichments across the target RNA, the proportion of each RBP distributed across the RNA was calculated by normalizing the equilibrium P_iR complex concentration $[P_iR]$ by the sum of the equilibrium concentration for each RBP. Then, the RBP fold-enrichment was calculated by comparing the proportion to the uniform distribution of RBPs across the RNA. For simulation comparison to *in vitro* iCLIP, fold-enrichment was calculated by comparing the proportion of U2AF2 to the proportion of cofactor.

For simulation purposes, the initial conditions are set so that all $[P_{i,0}]$ are in excess of the target RNA $[R_0]$. For each binding site, effective initial RNA concentration needs to be determined as a given binding site may exist more than once in the given RNA sequence. Thus, the effective initial concentration of a binding site $[R_{0,e}]$ is calculated as n x $[R_0]$, where n is the number of the binding site motif appearances in the given RNA sequence.

To set up the polynomial and find the root that satisfies the boundary condition, *Python* packages *SymPy* and *NumPy* were used.

Figures:

Figure 1: Inherent specificity (IS) and mutational sensitivity (MS) quantify RBP's Binding behaviors in a distinct manner
Figure 2: IS and MS calculation across different *K*-mers show strong intra- and inter-metric correlation
Figure 3: Comparison of IS and MS calculated using *in vitro* affinity measurements (RBNS) and cellular RBP binding (eCLIP)
Figure 4: RRM-swap between HNRNPC and RBM25 reveals IS can be transferred between different cellular contexts
Figure 5: Mathematical modeling of binding behaviors of four hypothetical RBPs with different IS and MS
Figure 6: Comparisons of *in vitro* iCLIP and model simulations of U2AF2-RNA interactions

Figure 1

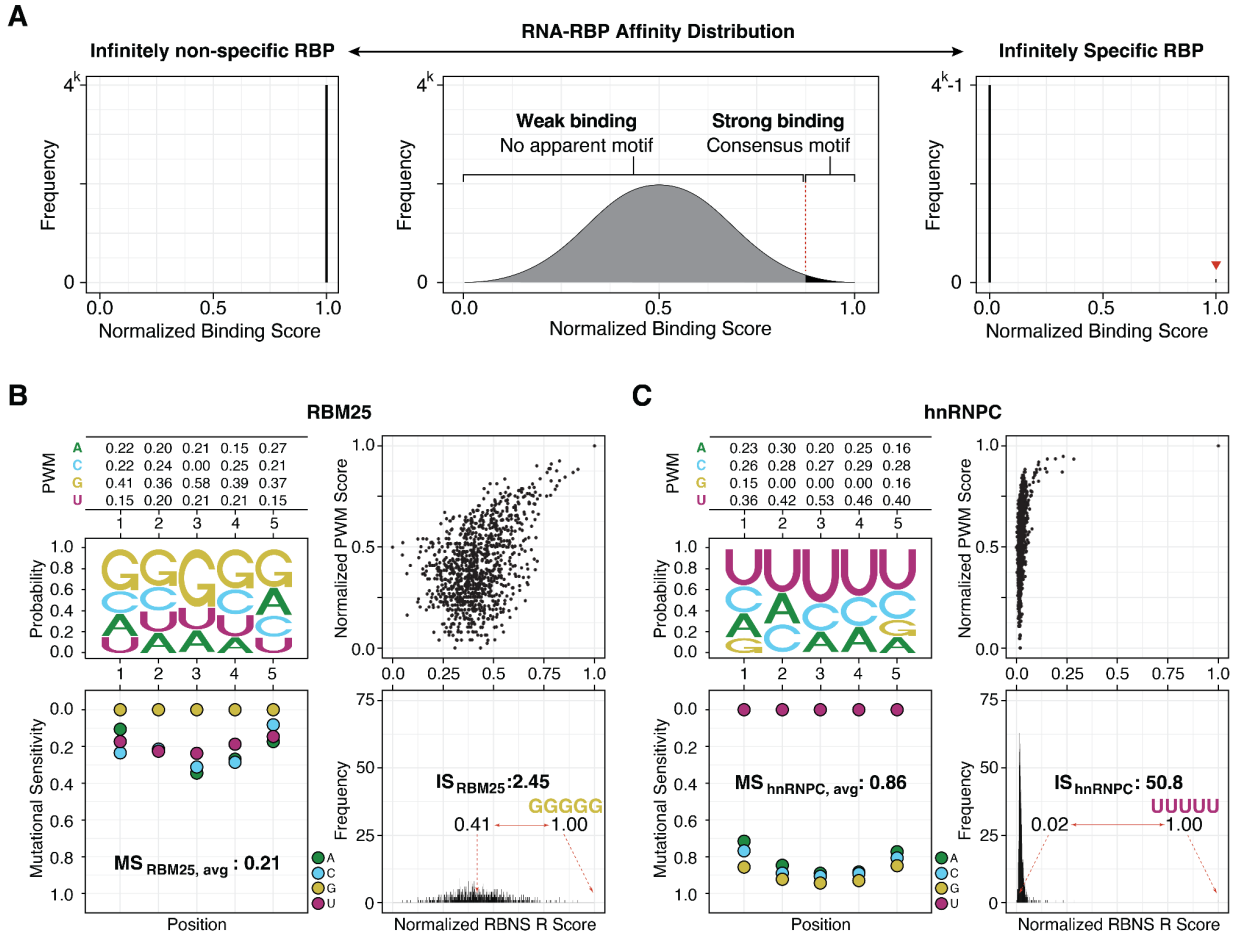

Figure 1: Inherent specificity (IS) and mutational sensitivity (MS) quantify RBP's Binding behaviors in a distinct manner. (A) Affinity distributions of RBPs with different binding behaviors. An infinitely non-specific RBP binds to all RNA sequence motifs with equal affinity (left), whereas an infinitely specific RBP binds only to a single RNA sequence motif (right). The affinity distributions of cellular RBPs reside between these two extremes (middle), where a consensus target motif can be generated from RNA sequence motifs in the high-affinity regime. **(B)** Graphical characterizations of the 5-mer RNA binding behavior of RBM25. The position weight matrix (PWM), derived from RBNS 5-mer affinity distribution, and the sequence logo, a graphical representation of the PWM, illustrate the traditional approach to characterizing the binding behavior of RBM25 (top left). A comparison of measured affinity distributions (from RBNS; x-axis) and reconstituted affinity distributions (from the PWM; y-axis) is shown (top right). Mutational sensitivity (bottom left) describes the average deviation in binding affinities to 5-mer sequence motifs with a single nucleotide change from the top binding motif (GGGGG). Inherent specificity (bottom right) measures the affinity ratio between the top and the median binding motifs. **(C)** The same as **(B)** but for HNRNPC.

Figure 2

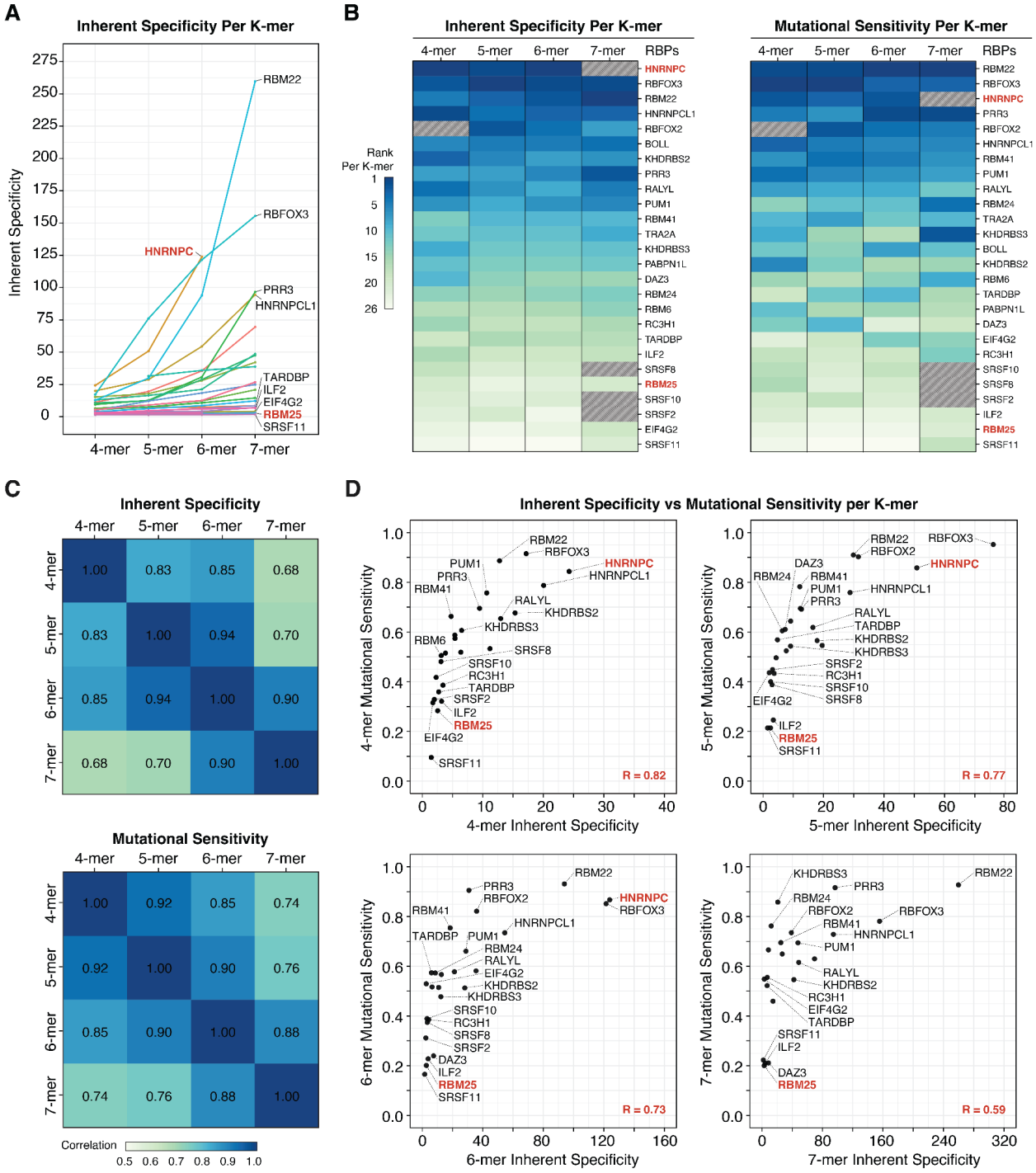

Figure 2: IS and MS calculation across different *K*-mers show strong intra- and inter-metric correlation. From the collected RBNS dataset, RBPs with only a single RNA binding domain purified were included in the analysis. **(A)** Change in the IS of RBPs across different *K*-mers. For the change in the MS across different *K*-mers, see figure S1. **(B)** Rank comparisons of IS and MS across different *K*-mers. RBPs without *K*-mer data available are marked in shaded

gray. **(C)** Pairwise correlation across different *K*-mers between IS (top) and MS (bottom) rank for RBPs examined in **(B)**. Pearson correlation coefficient values are shown. **(D)** Comparison between IS (x-axis) and MS (y-axis) across different *K*-mers. Pearson correlation coefficient values are shown at the bottom right in red. The two representative RBPs, RBM25 and HNRNPC, are marked in red.

Figure 3

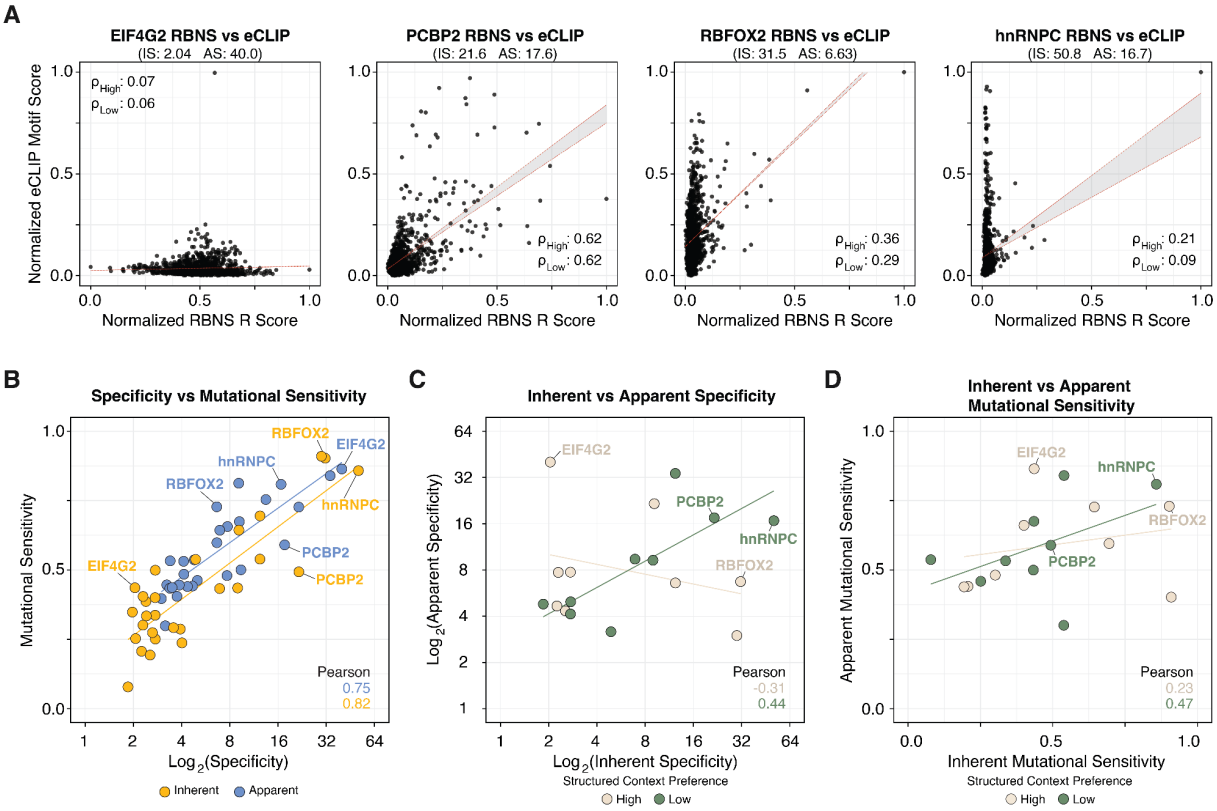

Figure 3: Comparison of IS and MS calculated using *in vitro* affinity measurements (RBNS) and cellular RBP binding (eCLIP). (A) Comparisons between 5-mer RNA-RBP motif enrichments measured by RBNS and eCLIP. Range of possible correlation values bounded by ρ_{High} and ρ_{Low} are shown, where ρ_{High} denotes the Pearson correlation coefficient calculated using the whole 5-mer comparison (1,024 data points) and ρ_{Low} denotes the coefficient calculated without the top 10 5-mer motifs for each RBP (1,014 data points). **(B)** Comparison between specificity and mutational sensitivity for RBPs with both RBNS and eCLIP datasets. Inherent metrics (yellow) are derived from RBNS measurements, while apparent metrics (blue) are derived from eCLIP measurements. The Pearson and Spearman correlation coefficients are shown in respective colors. RBPs shown in **(A)** are marked. **(C)** Comparison between inherent and apparent specificity for RBPs with known motif structural context preferences. The Pearson correlation coefficient for each group is shown in the respective color. **(D)** Same as **(C)** but for inherent and apparent mutational sensitivity.

Figure 4

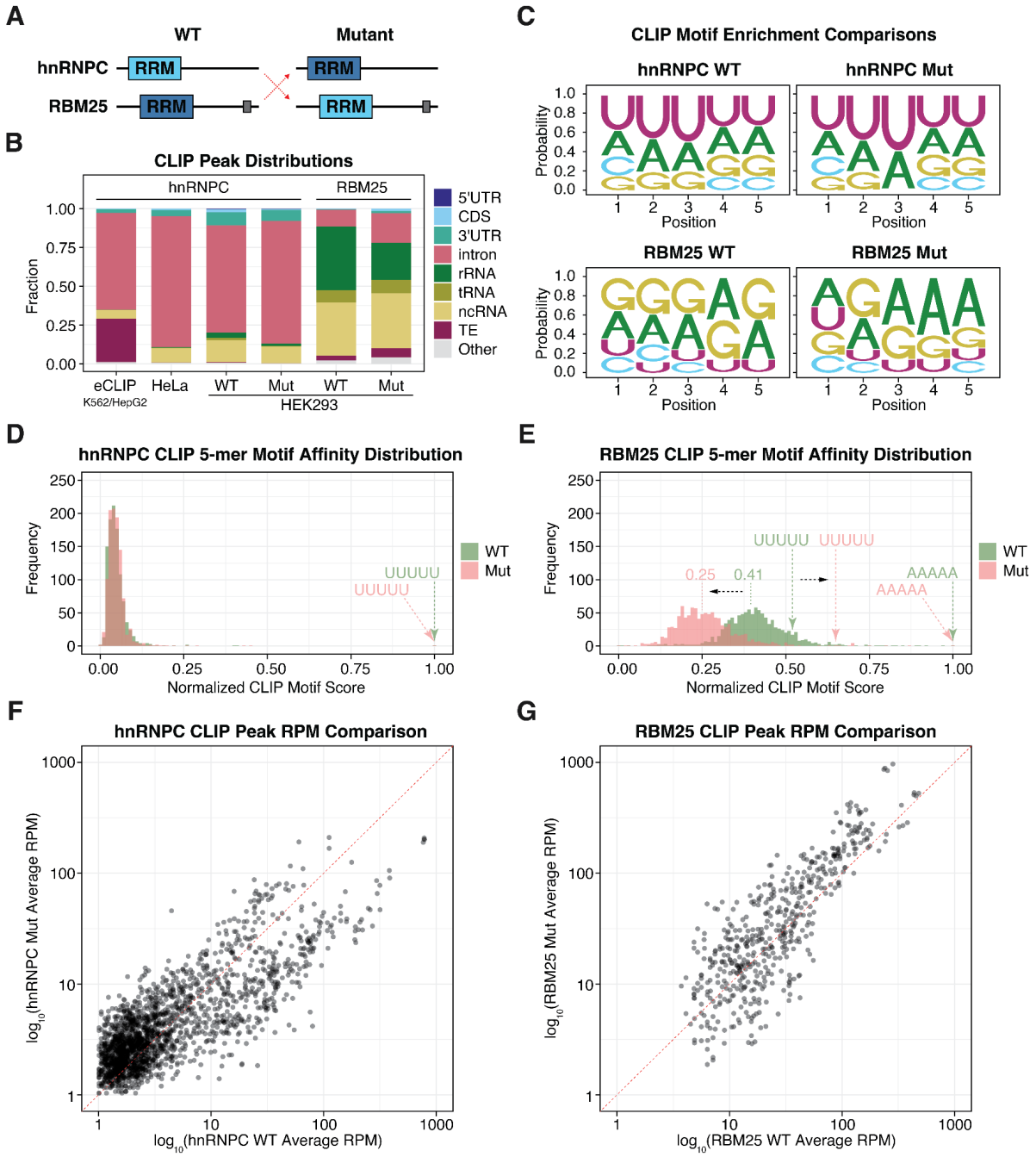

Figure 4: **RRM-swap between HNRNPC and RBM25 reveals IS can be transferred between different cellular contexts. (A)** Schematic representation of RRM swap between HNRNPC and RBM25. **(B)** CLIP peak annotation distributions in each analyzed dataset. For eCLIP analysis, peaks generated from K562 and HepG2 cell lines were combined. **(C)** PWM-based motif enrichment in CLIP peaks. The top 20 motifs were selected as representative sequence motifs and utilized for PWM generation. **(D)** 5-mer motif affinity distribution based on HNRNPC CLIP datasets. The position of the top motif "UUUUU" is noted. **(E)** same as **(D)** but for RBM25

CLIP datasets. The median normalized motif scores, as well as the positions of the “UUUUU” motif is shown. **(F)** Comparison of RPMs of CLIP peaks shared between the wild-type HNRNPC and the mutant HNRNPC. **(G)** Same as **(F)** but for RBM25.

Figure 5

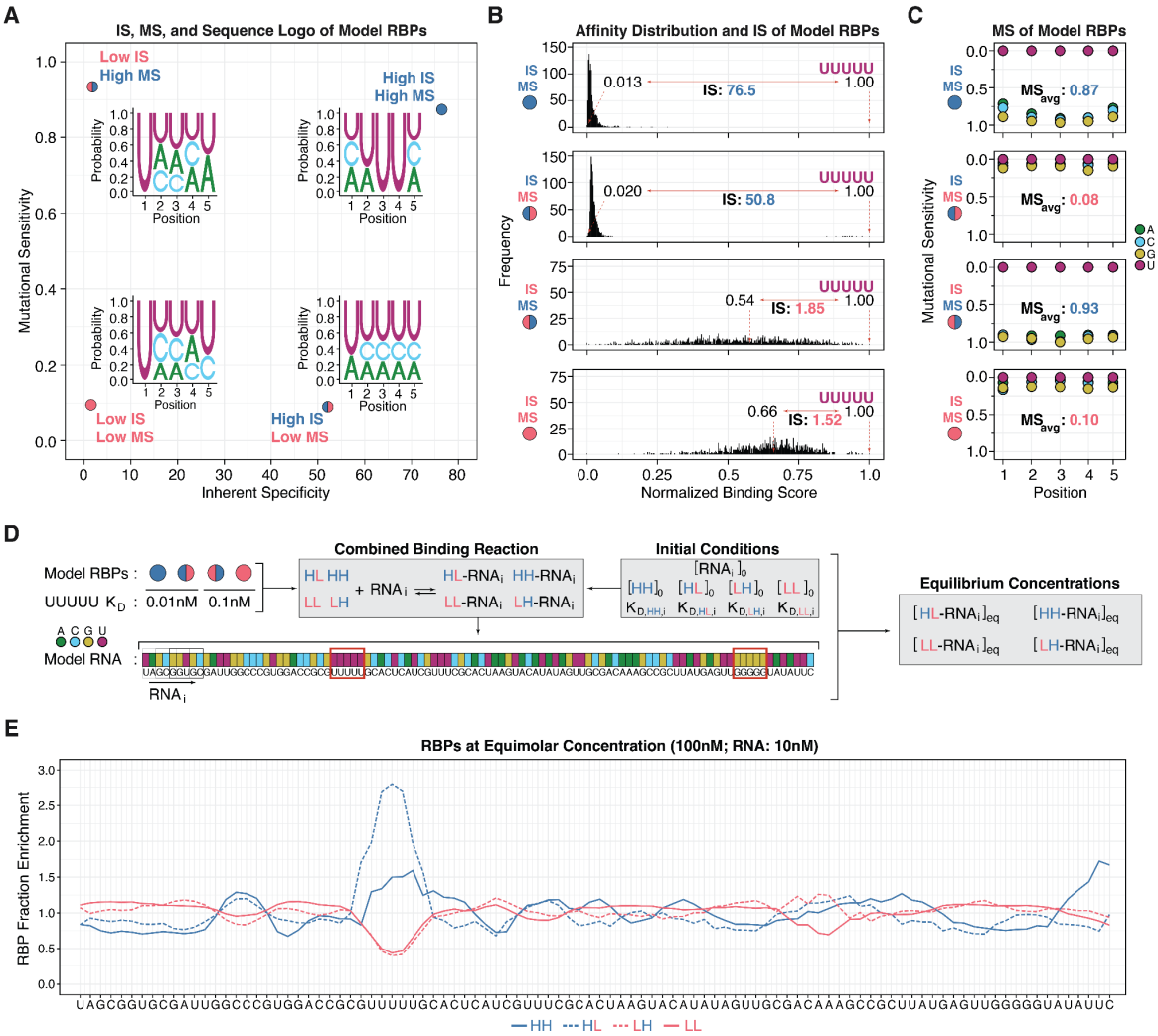

Figure 5: Mathematical modeling of binding behaviors of four hypothetical RBPs with different IS and MS. (A) Scatterplot showing the four hypothetical RBPs categorized by their respective IS and MS values. A sequence logo for each RBP is also shown adjacent to their respective position on the scatterplot. **(B)** Affinity distributions of the four RBPs, highlighting the median and top affinity values, as well as the IS value, for each RBP. **(C)** The MS plots and values for the four RBPs. **(D)** Graphical summary of the mathematical binding model. Dissociation constants (K_D) of the model RBPs toward the top binding motif "UUUUU" are shown. The K_D values for the other 5-mer motifs are calculated using the respective affinity distribution. The substrate for the binding reaction, model RNA, is also shown. For each moving window of 5-mer (RNA_i), K_D values toward RNA_i and initial conditions are used to calculate equilibrium concentrations of each RNA-RBP complex. Equilibrium concentrations across the model RNA are then utilized to calculate per-position binding probabilities for each RBP. (HH: High IS High MS, HL: High IS Low MS, LH: Low IS High MS, LL: Low IS Low MS). Two example binding sites, "UUUUU" and "GGGGG" are marked in red boxes. **(E)** Simulated RBP distribution across a model target RNA for the 4 model RBPs described in **(D)**.

Figure 6

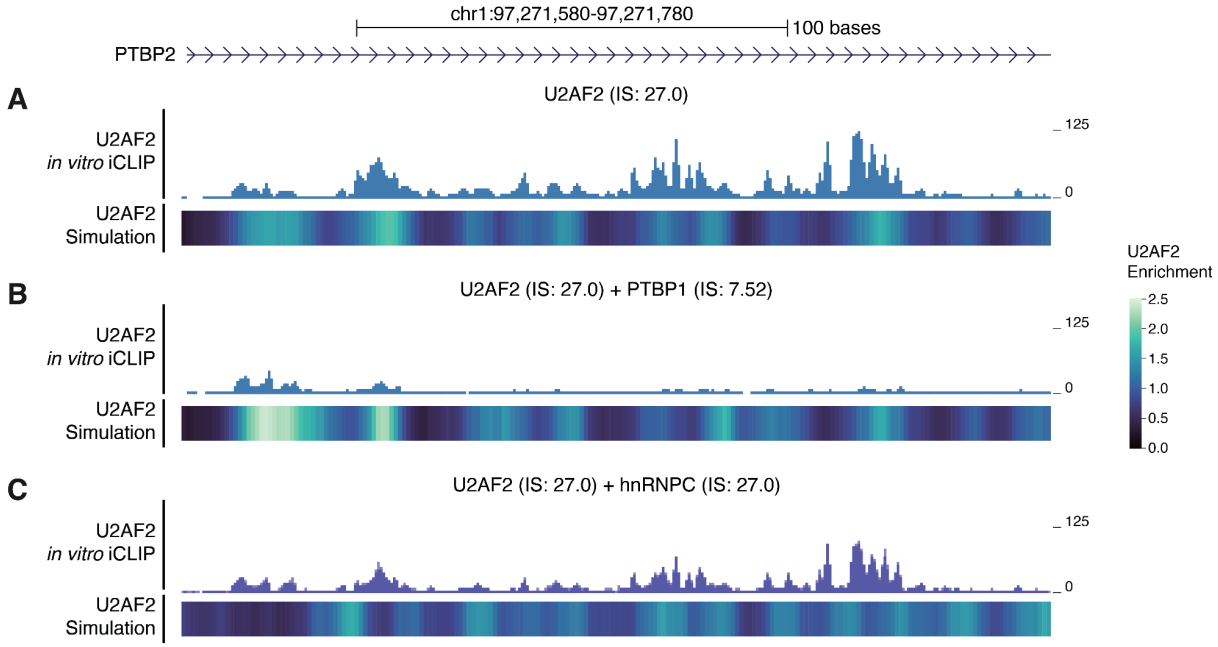

Figure 6: Comparisons of *in vitro* iCLIP and IS-informed model simulations of U2AF2-RNA interactions. The binding of U2AF2 to the intronic region of PTBP2 transcript was simulated and compared to *in vitro* U2AF2 iCLIP data [34]. The simulation was done in the absence **(A)** and presence of cofactor **(B)** PTBP1 or **(C)** HNRNPC. U2AF2 enrichment is calculated as U2AF2 distribution over uniform distribution of U2AF2 **(A)** or cofactor distribution **(B** and **C)**.

References

1. Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

2. Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393–1406.

3. Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327–341.

4. Licatalosi, D.D., Ye, X., and Jankowsky, E. (2020). Approaches for measuring the dynamics of RNA-protein interactions. Wiley Interdiscip Rev RNA *11*, e1565.

5. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719.

6. Mitchell, S.F., and Parker, R. (2014). Principles and properties of eukaryotic mRNPs. Mol. Cell *54*, 547–558.

7. Singh, G., Pratt, G., Yeo, G.W., and Moore, M.J. (2015). The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. Annu. Rev. Biochem. *84*, 325–354.

8. Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. *11*, 75–87.

9. Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M.W. (2020). RNA-binding proteins in human genetic disease. Nat. Rev. Genet. *22*, 185–198.

10. Darnell, R.B. (2010). RNA regulation in neurologic disease and cancer. Cancer Res. Treat. *42*, 125–129.

11. Turner, M., and Díaz-Muñoz, M.D. (2018). RNA-binding proteins control gene expression and cell fate in the immune system. Nat. Immunol. *19*, 120–129.

12. von Hippel, P.H., and Berg, O.G. (1986). On the specificity of DNA-protein interactions. Proc. Natl. Acad. Sci. U. S. A. *83*, 1608–1612.

13. Jankowsky, E., and Harris, M.E. (2015). Specificity and nonspecificity in RNA–protein interactions. Nat. Rev. Mol. Cell Biol. *16*, 533–544.

14. Gupta, A., and Gribskov, M. (2011). The role of RNA sequence and structure in RNA--protein interactions. J. Mol. Biol. *409*, 574–587.

15. Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

16. Ray, D., Kazan, H., Chan, E.T., Castillo, L.P., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. *27*, 667–670.

17. Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–900.

18. Guenther, U.-P., Yandek, L.E., Niland, C.N., Campbell, F.E., Anderson, D., Anderson, V.E., Harris, M.E., and Jankowsky, E. (2013). Hidden specificity in an apparently nonspecific RNA-binding protein. Nature *502*, 385–388.

19. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

20. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

21. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.

22. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.

23. M. Hafner, Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.-C., Munschauer, M., et al. (2010), Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).

24. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889.

25. 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., et al. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell *70*, 854–867.e9.

26. Ye, X., Yang, W., Yi, S., Zhao, Y., Varani, G., Jankowsky, E., and Yang, F. (2023). Two distinct binding modes provide the RNA-binding protein RbFox with extraordinary sequence specificity. Nat. Commun. *14*, 701.

27. Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16–23.

28. Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. *10*, 741–754.

29. Wang, Y., Liu, J., Huang, B.O., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., et al. (2015). Mechanism of alternative splicing and its regulation. Biomed Rep *3*, 152–158.

30. Yugami, M., Okano, H., Nakanishi, A., and Yano, M. (2020). Analysis of the nucleocytoplasmic shuttling RNA-binding protein HNRNPU using optimized HITS-CLIP method. PLoS One *15*, e0231450.

31. Ducoli, L., Zarnegar, B.J., Porter, D.F., Meyers, R.M., Miao, W., Riley, N.M., Srinivasan, S., Jackrazi, L.V., Yang, Y.-Y., Li, Z., et al. (2024). irCLIP-RNP and Re-CLIP reveal patterns of dynamic protein associations on RNA. bioRxiv. https://doi.org/10.1101/2024.09.27.615518.

32. Le Sénéchal, R., Keruzoré, M., Quillévéré, A., Loaëc, N., Dinh, V.-T., Reznichenko, O., Guixens-Gallardo, P., Corcos, L., Teulade-Fichou, M.-P., Granzhan, A., et al. (2023). Alternative splicing of BCL-x is controlled by RBM25 binding to a G-quadruplex in BCL-x pre-mRNA. Nucleic Acids Res *51*, 11239–11257.

33. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

34. Sutandy, F.X.R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.-S., Fallmann, J., Maticzka, D., Backofen, R., Stadler, P.F., et al. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome Res. *28*, 699–713.

35. Beyer, A.L., Christensen, M.E., Walker, B.W., and LeStourgeon, W.M. (1977). Identification and characterization of the packaging proteins of core 40S hnRNP particles. Cell *11*, 127–138.

36. Barnett, S.F., Friedman, D.L., and LeStourgeon, W.M. (1989). The C proteins of HeLa 40S nuclear ribonucleoprotein particles exist as anisotropic tetramers of (C1)3 C2. Mol Cell Biol *9*, 492–498.

37. Sternburg, E.L., Gruijs da Silva, L.A., and Dormann, D. (2022). Post-translational modifications on RNA-binding proteins: accelerators, brakes, or passengers in neurodegeneration? Trends Biochem. Sci. *47*, 6–22.

38. Cohen, T.J., Hwang, A.W., Restrepo, C.R., Yuan, C.-X., Trojanowski, J.Q., and Lee, V.M.Y. (2015). An acetylation switch controls TDP-43 function and aggregation propensity. Nat. Commun. *6*, 5845.

39. Fazal, F.M., Han, S., Parker, K.R., Kaewsapsak, P., Xu, J., Boettiger, A.N., Chang, H.Y., and Ting, A.Y. (2019). Atlas of Subcellular RNA Localization Revealed by APEX-Seq. Cell *178*, 473–490.e26.

40. Padrón, A., Iwasaki, S., and Ingolia, N.T. (2019). Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules. Mol. Cell *75*, 875–887.e5.

41. Tsue, A.F., Kania, E.E., Lei, D.Q., Fields, R., McGann, C.D., Marciniak, D.M., Hershberg, E.A., Deng, X., Kihiu, M., Ong, S.-E., et al. (2024). Multiomic characterization of RNA microenvironments by oligonucleotide-mediated proximity-interactome mapping. Nat Methods *21*, 2058–2071.

42. Yi, S., Singh, S.S., Rozen-Gagnon, K., and Luna, J.M. (2023). Mapping RNA-Protein Interactions with Subcellular Resolution Using Colocalization CLIP. bioRxiv, 2023.10.26.563984. https://doi.org/10.1101/2023.10.26.563984.

43. Schmok, J.C., Jain, M., Street, L.A., Tankka, A.T., Schafer, D., Her, H.-L., Elmsaouri, S., Gosztyla, M.L., Boyle, E.A., Jagannatha, P., et al. (2024). Large-scale evaluation of the ability of RNA-binding proteins to activate exon inclusion. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-02014-0.

44. Campbell, Z.T., Valley, C.T., and Wickens, M. (2014). A protein-RNA specificity code enables targeted activation of an endogenous human transcript. Nat. Struct. Mol. Biol. *21*, 732–738.

45. Wang, Y., Wang, Z., and Tanaka Hall, T.M. (2013). Engineered proteins with Pumilio/fem-3 mRNA binding factor scaffold to manipulate RNA metabolism. FEBS J. *280*, 3755–3767.

46. Chen, Y., and Varani, G. (2013). Engineering RNA-binding proteins for biology. FEBS J. *280*, 3734–3754.

47. Choudhury, R., Tsai, Y.S., Dominguez, D., Wang, Y., and Wang, Z. (2012). Engineering RNA endonucleases with customized sequence specificities. Nat. Commun. *3*, 1147.

48. Biostrings Bioconductor. http://bioconductor.org/packages/Biostrings/.

49. seqLogo Bioconductor. http://bioconductor.org/packages/seqLogo/.

50. Lacoste, A., Berenshteyn, F., and Brivanlou, A.H. (2009). An efficient and reversible transposable system for gene delivery and lineage-specific differentiation in human embryonic stem cells. Cell Stem Cell *5*, 332–342.

51. Hanke, T., Young, D.F., Doyle, C., Jones, I., and Randall, R.E. (1995). Attachment of an oligopeptide epitope to the C-terminus of recombinant SIV gp160 facilitates the construction of SMAA complexes while preserving CD4 binding. J Virol Methods *53*, 149–156.

52. Graham, F.L., Smiley, J., Russell, W.C., and Nairn, R. (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J Gen Virol *36*, 59–74.

53. Luna, J.M., Scheel, T.K.H., Danino, T., Shaw, K.S., Mele, A., Fak, J.J., Nishiuchi, E., Takacs, C.N., Catanese, M.T., de Jong, Y.P., et al. (2015). Hepatitis C virus RNA functionally sequesters miR-122. Cell *160*, 1099–1110.

54. Luna, J.M., Barajas, J.M., Teng, K.-Y., Sun, H.-L., Moore, M.J., Rice, C.M., Darnell, R.B., and Ghoshal, K. (2017). Argonaute CLIP Defines a Deregulated miR-122-Bound Transcriptome that Correlates with Patient Survival in Human Liver Cancer. Mol Cell *67*, 400–410.e7.

55. Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q., et al. (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep *6*, 1139–1152.

56. Moore, M.J., Blachere, N.E., Fak, J.J., Park, C.Y., Sawicka, K., Parveen, S., Zucker-Scharff, I., Moltedo, B., Rudensky, A.Y., and Darnell, R.B. (2018). ZFP36 RNA-binding proteins restrain T cell activation and anti-viral immunity. Elife *7*. https://doi.org/10.7554/eLife.33057.

57. UniProt Consortium (2025). UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res *53*, D609–D617.

58. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

59. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Sci *32*, e4792.

60. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792–1797.