A comprehensive thermodynamic model for RNA binding by the *Saccharomyces cerevisiae* Pumilio protein PUF4

Christoph Sadée1,10, Lauren D. Hagler1,10, Winston R. Becker2, Inga Jarmoskaite1,3, Pavanapuresan P. Vaidyanathan1,8, Sarah K. Denny2,9, William J. Greenleaf3,4,5 & Daniel Herschlag1,6,7

Received: 1 November 2021
Accepted: 7 July 2022

Genomic methods have been valuable for identifying RNA-binding proteins (RBPs) and the genes, pathways, and processes they regulate. Nevertheless, standard motif descriptions cannot be used to predict all RNA targets or test quantitative models for cellular interactions and regulation. We present a complete thermodynamic model for RNA binding to the *S. cerevisiae* Pumilio protein PUF4 derived from direct binding data for 6180 RNAs measured using the RNA on a massively parallel array (RNA-MaP) platform. The PUF4 model is highly similar to that of the related RBPs, human PUM2 and PUM1, with one marked exception: a single favorable site of base flipping for PUF4, such that PUF4 preferentially binds to a non-contiguous series of residues. These results are foundational for developing and testing cellular models of RNA-RBP interactions and function, for engineering RBPs, for understanding the biophysical nature of RBP binding and the evolutionary landscape of RNAs and RBPs.

RNAs and their interactions are integral to the regulation of gene expression. With the realization that RNA binding proteins (RBPs) bind many related RNAs to coordinate their function1–4 and that at least one in 20 proteins in the genome bind RNA5–8, several important challenges have emerged: (i) to identify the RNA targets of RBPs; (ii) to determine the functional consequences of these interactions; (iii) to determine the biophysical basis of RBP binding and specificity; (iv) to trace the evolution of RBPs and their interactions; and (v) to develop rules for engineering RBP affinity and specificity.

Following the recognition in the early 2000s that individual RBPs bind large sets of related RNAs, a major goal has been to identify the RNA binding partners for the many cellular RBPs9–13. Subsequent studies have helped to elucidate the roles of RNA–protein interactions in cellular processes14–15. As genomic-scale methods have become more and more sensitive, RBP target lists have grown in length16–18. However, RNAs have a range of affinities, and RBPs exhibit robust non-specific binding. It is unclear where, or even if, one can draw a clean cutoff between binders and non-binders, and, importantly, it is unlikely that all RNAs that bind a particular RBP are equally affected by that RBP and by changes in its levels, modifications, and localization. Most simply, RNAs with higher and lower RBP occupancies are expected to be more and less functionally affected by those RBPs, respectively, all else being equal. Thus, predictive models of RNA and RBP function must start from predictions of occupancies.

A zeroth-order model is that occupancy is simply a reflection of RNA-RBP thermodynamics—i.e., the intrinsic affinities for each sequence and the accessibility of that sequence. Genomic-scale RNA-RBP cross-linking (CLIP) experiments can in principle be used to test this model in cells. However, there are limitations to the information one can extract from genomic-wide studies. The number of
sequencing reads from genome-wide eCLIP is quantitative, but it is subject to distortions at each of several experimental steps and also subject to statistical limitations from low sequence read depth. As a result, the number of reads at a particular RNA site may not faithfully reflect quantitative RBP occupancies in the cell. Testing this zeroth-order model, determining when it holds and where it breaks down, and uncovering what cellular factors and features are responsible for its breakdown requires starting with comprehensive binding models to predict relative occupancies. Such simple quantitative models are most critical when attempting to disentangle complex multi-variate systems such as molecular interactions and outcomes in cells.

Pumilio proteins have eight pseudo repeats each recognizing one residue (Fig. 1b). Intriguingly, human PUM1/2 and yeast PUF4 have highly similar RNA-contacting amino acids (Fig. 1c) and binding motifs and logos that are similar (Fig. 1a); yet the identified RNA targets are considerably different. These observations are fascinating from both evolutionary and biophysical perspectives. Evolutionarily, the similar binding specificities likely allowed the exchange of >100 RNA targets between S. cerevisiae PUF3 (an ortholog of PUM1/2) and PUF4 during fungal evolution to alter their regulation. Biophysically, determining the altered sequence specificity provides a foundation for deciphering the molecular origins for these changes.

We previously derived a complete thermodynamic model for human Pumilio proteins, PUM2 and PUM1 (referred to together as PUM1/2, Fig. 1a) binding to all RNAs. PUM1/2 have identical specificities and are described by the same mathematical model contained 59 terms, with a single constant affinity offset. The binding to an engineered variant of PUM1 with a single mutation within one of its Pumilio modules was quantitatively accounted for by varying just one of the 59 terms, one representing specificity within the mutated module. Here, we derive an analogous mathematical model for the yeast PUF4 targets that provides similar high accuracy, despite the greater complexity of its motif descriptions. The PUF4 thermodynamic model is based on direct high-throughput binding data for 6180 RNAs measured with the RNA on a massively parallel array (RNA-Map) platform. The resulting model, consisting of 56 terms, is highly similar to that for the related RBPs, PUM1/2, with one marked exception. The difference in RNA binding to PUF4 versus PUM1/2 arises from a single favorable flipping term for PUF4, such that PUF4 binds preferentially to a non-contiguous series of residues. All other flips and

![Fig. 1 | Comparison of PUF4, PUM2, and PUF3 motifs, structure, and sequence.](https://doi.org/10.1038/s41467-022-31968-z)

---

**Fig. 1** | Comparison of PUF4, PUM2, and PUF3 motifs, structure, and sequence.  
**a** Literature PUF4 sequence motifs compared to a subset of those for PUM2 and PUF3. The experimental method for each representation is shown with the corresponding ref. 9, 10, 26, 27, 31, 32, 36, 54, 55 (RIP; RNA precipitation; RIP-Chip: RNA-binding immunoprecipitation microarray profiling; SEQRS: high-throughput sequencing and sequence specificity landscapes; PAR-CLIP: photoactivatable ribonucleotide-enhanced cross-linking and immunoprecipitation; HITS-CLIP: cross-linking and immunoprecipitation coupled with high-throughput sequencing). A full list of PUM2 and PUF3 motifs is given in ref. 34. **b** Crystal structures of the RNA-binding domain of PUF4 bound to UGUAAUAUA (left, PDB: 3BX2) and human PUM2 bound UGUAUAUA (right, PDB: 3Q0Q). **c** PUM2, PUF3, and PUF4 repeat residues involved in base-specific interactions (see also Supplementary Fig. 1). For simplicity, the eight binding sites (R1-R8) are numbered in the 5’ to 3’ order of bound consensus RNA (5’S-GUAUA-AUA-3’), the reverse order of the protein primary sequence and opposite to literature convention. Red residues make specific contacts with RNA bases. Yellow boxed residues indicate differences between PUF4 and PUM2/PUF3. The green, orange, and red RNA bases correspond to the base favored at each position of PUM2/PUF3 based on the consensus motifs.
non-contiguous binding modes are unfavorable for PUF4 as they are for PUM1/2, and the other S2 terms in the PUF4 and PUM1/2 binding models are highly similar, indicating that binding to each of the individual Pumilio repeats is the same or nearly the same. These results are foundational for testing cellular models of RNA-PUF interactions and function, for engineering RBPs, and for understanding the biophysical nature of Pumilio binding and the evolutionary landscape of RNAs and Pumilio proteins.

Results

RNA-MaP to measure the binding affinities of PUF4 to 6180 RNAs

To derive a thermodynamic PUF4-binding model, we followed the approach we used previously with the related human Pumilio proteins, PUM1/2, designing a series of RNA sequences that vary relative to the previously-identified consensus sequence (Fig. 1a)³. In this case, the library was designed, rather than randomized, to ensure systematic variation relative to the consensus sequence, while also allowing exploration further into sequence space without sacrificing large amounts of the library to non-binders (Fig. 2a). Other library approaches allow more sequences to be explored but sacrifice thermodynamic rigor. Also, the models from RNA-MaP, while not generated based on all possible sequences, are quantitative and predictive and can be expanded or modified if confronted by new affinities that, by rigorous statistical treatment, are not predicted by the current model.

In our library, we included mutations (1–4 nt) to probe sequence specificity and insertions (1–5 nt) to assay potential non-contiguous binding sites. Additionally, we varied the immediate flanking residues (0–3 nt) to assess possible extended binding sites (Fig. 2a and Supplementary Fig. 2a). The library included the sequence variants used to develop the PUM1/2 binding models³⁴ and variants from the consensus of PUF5 (an S. cerevisiae Pumilio protein related to PUF4)³⁵ to provide a range of related but different sequence variants (see Source data for full library). The RNA sequence variants were embedded in four different scaffolds to control for RNA structures that could form from interactions with particular sequence variants or flanking sequences (Fig. 2a).

We used RNA-MaP to measure PUF4 equilibrium dissociation constants directly for the entire library (Fig. 2b and Source data). We began with a DNA library encoding our RNA variants; the DNA library was sequenced on an Illumina MiSeq flow cell, followed by in situ transcription on a custom-built imaging and fluidics setup³⁷,³⁸. RNA transcripts were immobilized by stalling E. coli RNA polymerase at the end of the DNA template, and RNA–protein association was measured by equilibrating the RNA with increasing concentrations of fluorescently-labeled PUF4 and by imaging each cluster (comprising ~1000 copies of that RNA variant)³⁹. The resulting binding curves were used to obtain the dissociation constant (K_D) and the corresponding free energy of binding, ΔG (−RTlnK_D), of the protein for each RNA variant, as described in the “Methods” section. Figure 2b shows representative binding curves for a tight and weak binding RNA, with the tight binder

Fig. 2 | Quantitative, high-throughput measurements of RNA binding to PUF4. 

(a) Designed RNA library to probe PUF4 binding specificity. A variable region (yellow bases) is embedded into each of four scaffolds (green bases) S1a, S1b, S2a, and S2b. Variations are introduced with reference to PUF4 and other Pumilio protein consensus sequences (PUM2 and PUF5), including one to four base mutations, one to five insertions at each consensus position, and flanking insertions (N = 0–3 nt). Additional sequence information is shown in Supplementary Fig. 2a. 

(b) Schematic representation of RNA-MaP procedure³⁷, including representative fluorescence images for Cy3B-PUF4 binding to a small region of a sequencing chip. 

(c) Representative binding curves from RNA-MaP of two RNA variants (UGUAUAUAUA in S1a scaffold and UGUAUCGCAC in S2b scaffold). Black circles indicate median fluorescence at each protein concentration for all clusters of the respective variants normalized to the background fluorescence of the RNA channel. The number of replicate clusters is denoted by n. Error bars represent 95% confidence intervals (CI) across the clusters determined by bootstrap analysis. The green lines indicate the fits to the binding model. The grey shaded area indicates the 95% CI of the fit (K_D, (consensus) = 0.88 nM, CI_{95%} = (7.86 nM; 1.03 nM); K_D (mutant) > 2.4 μM, corresponding to the upper limit for binding affinities that could be confidently distinguished from background). For high affinity binders an additional increase in fluorescence is observed at high protein concentrations that is accounted for by a non-specific binding term for PUF4 binding to the RNA-PUF4 complex (see ref. 34 and “Methods”). The x-axis is logarithmic.
corresponding to the PUF4 consensus sequence. As observed previously with PUM2, there is a second, non-specific binding component that represents binding of the protein to the RNA–protein complex, and these data were fit as previously done for PUM2 (ref. 34 and “Methods”).

We assessed the binding of 15,272 distinct RNAs. Of these, affinities were obtained for 6180 RNAs which had measurable affinities (≤2.3 μM) and ≥5 technical replicates to provide high affinity and robust statistics. Prior work demonstrated precise measurements within -0.28 kcal/mol with this number of replicates, with the accuracy increasing with more replicates and decreasing with weaker binding, and our results mirrored these prior findings (Supplementary Fig. 2c; ref. 34). Equilibrium binding was measured for two salt conditions (low and high; see “Methods”). Fully independent datasets measured on two chips with high salt (2 mM MgCl2, 100 mM KOAc) gave excellent agreement, with a root-mean-square-error (RMSE) of 0.3 kcal/mol, which corresponds to an average error in dissociation constants of less than two-fold (Supplementary Fig. 3). The low salt condition (0.75 mM KOAc) gave excellent agreement, with a root-mean-square-error (RMSE) of 0.1 kcal/mol, which corresponds to an average error in dissociation constants of less than two-fold (Supplementary Fig. 3).
MgCl₂, 30 mM KOAc) was used to develop and test the model as the wider range of affinities that could be measured provided more extensive data to develop and test the quantitative binding model. The binding data from RNA-MaP agreed well with reliable literature affinity measurements for PUF4 and PUM2 (Supplementary Fig. 5 and refs. 34, 41).

**PUF4 and PUM2 predictive binding models differ in one flipping term**

We previously developed a binding model for PUM2 containing three classes of free energy terms (Fig. 3a); (i) terms for binding at each recognition site \( \Delta \Delta G_k^f \), \( b = 1 \sim 9 \), representing the position of the bound residue to eight Pumilio repeats and an additional 3’ site that provides a small contribution to binding; \( X = A, C, U, \) or \( G \) is the base at that position; (ii) terms for binding non-contiguously, which we refer to as flipping terms \( \Delta \Delta G_k^{f \text{flip}} \); \( f = 1/2, 2/3, 3/4, 4/5, 5/6, \) or \( 6/7 \) corresponds to either a single residue \( Y = A, C, U, \) or \( G \) inserted between two positions or any two residues \( Y = \text{NN} \) flipped between these positions; and (iii) coupling terms to account for non-additive energetics when the value of a particular energetic term depends on the residue or residues present at one or more other positions \( \Delta \Delta G_{k+i}^f \). c is the position of the coupled bases and \( Z \) is the combination of residues that meet all of the conditions, see also ref. 34). For PUM2, we fitted 149 terms needed, and extended as

\[
\Delta \Delta G_k^{f \text{flip}} = \Delta \Delta G_k^f + \Delta \Delta G_{k+i}^f,
\]

which could be measured provided more than one position (Supplementary Table 1 and Table 1 below). The striking asymmetry of the error, \( \Delta \Delta G_k^{f \text{flip}} \text{ term differed substantially from the PUM2 model. Whereas residue flips were strongly disfavored at all positions for PUM2. Flipping was strongly favorable at position 6/7 for RNA bound to PUF4 but not at any of the other positions (Supplementary Table 1 and Table 1 below). We discuss this finding and its implications in more detail after further consideration of the data fitting.

Fitting the PUF4 binding data to the PUM2 model with all parameters varied improved the fit (Fig. 3d versus 3c), with the RMSE decreasing from 0.71 to 0.53 kcal/mol and with fewer outliers (554 versus 246 with z-score >3, respectively), and both models provided large improvements relative to the unaltered PUM2 model, which gave a RMSE of 1.43 kcal/mol with 2296 outliers with z-scores of >3 (Fig. 3b). While there was some variation in the best-fit values for the individual site PUM2 and PUF4 terms, these values strongly correlated (Fig. 4a, b, blue points, \( R^2 = 0.86 \) and see below); the only clear difference was the -0.9 kcal/mol discrimination against A at position 5 for PUF4 but not for PUM2 (Fig. 4a and Table 1; see also Fig. 5 below). The flipping terms were also similar except for the terms for the single residue 6/7 flip, which strongly deviated and were favorable for PUF4 but unfavorable for PUM2 as noted above (Fig. 4c, d, closed & open red points).

Further test of the PUF4 binding model

The agreement of the model of Fig. 3c with the PUF4 data might lead one to conclude that this model is correct. However, the complexity of

| Table 1 | Thermodynamic parameter values for the PUF4 additive non-consecutive model |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Term I**      | **ΔΔG_k^f (kcal/mol)** | **X = A**       | **C**           | **U**           | **G**           |
| Bound residue position \( b \) | **ΔΔG_k^f (kcal/mol)** | **X = A**       | **C**           | **U**           | **G**           |
| 1                | 3.69            | 2.85            | 0.00            | 3.97            |
| 2                | 2.50            | 4.41            | 4.39            | 0.00            |
| 3                | 2.67            | 3.01            | 0.00            | 2.52            |
| 4                | 0.00            | 2.29            | 1.75            | 1.51            |
| 5                | 0.89            | 0.03            | 0.00            | 1.10            |
| 6                | 0.00            | 0.93            | 0.85            | 1.41            |
| 7                | 2.48            | 1.82            | 0.00            | 2.07            |
| 8                | 0.00            | 1.47            | 1.33            | 1.33            |
| 9                | 0.60            | 0.37            | 0.00            | 0.24            |

| **Term II**      | **ΔΔG_k^f (kcal/mol)** | **Y = A**       | **C**           | **U**           | **G**           | **NN**           |
| Flipped residue position \( f \) | **ΔΔG_k^f (kcal/mol)** | **Y = A**       | **C**           | **U**           | **G**           | **NN**           |
| 3/4              | 1.63            | 1.51            | 1.60            | 1.21            | >1.50 \(^a\)     |
| 4/5              | >2.00           | 1.36            | 1.75            | >2.00 \(^a\)     |
| 5/6              | >2.00           | 1.83            | 1.68            | >2.00 \(^a\)     |
| 6/7              | ~1.15           | ~1.07           | ~1.34           | ~0.35           | 0.67            |

\(^{a}\) indicates a lower limit.

2-flip of any sequence.
large datasets and the models used to fit them call for additional scrutiny. Indeed, whenever variable terms are added, fits will improve and one can, at least some of the time, get excellent fits from incorrect models. Misfitting can occur even with far more data than the number of model variables, because the data at hand typically will not equally probe each parameter. Our designed libraries provide some insurance against this limitation, relative to random or natural variants; nevertheless, one cannot design a library that equally tests all model parameters without already knowing the values for those parameters. For example, the inclusion of the coupling terms in the PUM2 model was necessary to account for a handful of clear outliers but their inclusion had a negligible effect on the overall RMSE because the terms apply to a miniscule fraction of the total RNAs and provide modest adjustments even for those. Indeed, while there is likely some degree of local coupling for PUF4 as well, we did not include coupling terms in the PUF4 model because there was too little systematic data to confidently add such terms. Nevertheless, the absence of these terms may account for the somewhat higher RMSE for the final PUF4 model relative to replicate measurements and relative to the prior PUM2 model (Fig. 3d; Supplementary Fig. 3b, 34; 0.55, 0.30, and 0.34 kcal/mol, respectively).

To directly address these concerns, we carried out several additional tests of our PUF4 model. We calculated the sensitivity of the fit to each parameter (Figs. 5, 6), which showed that the values obtained were well defined. We then further demonstrated the similarity of the individual site free energy terms ($\Delta\Delta G_{b}$) for PUF4 and PUM2 by comparing their single mutant penalties (Fig. 4 and Supplementary Fig. 6). Sensitivity analysis was carried out by varying the $\Delta\Delta G$ value of each of the 56 terms in the PUF4 model individually and determining the effect of that variation on the overall RMSE of the fit. Figures 5, 6 compares the best fit value (the minima of the curve) and sensitivity of the fit (i.e., change in RMSE) to variation in each parameter of the PUF4 and PUM2 models. Focusing first on the individual site terms ($\Delta\Delta G_{b}$, Fig. 5), the residue preferred at each site was the same in all cases for PUF4 and PUM2 (Fig. 5, red box), and the penalties for variation from the preferred residue were also highly similar. For both proteins, there were generally larger deleterious effects for residues at the 5′ region of the binding site. (The physical origins of these specificity differences are not known). As expected, the sensitivity was asymmetric, especially for large deleterious substitutions where their effects often lead to difficult-to-measure or unmeasurable binding such that the minimum value of the penalty is better defined. In addition, because these substitutions are rarely found in the RNAs with measurable binding, they have a smaller impact on the overall RMSE. Nevertheless, reasonable estimates are obtained in most instances as binding is observed for some RNAs with each substitution (see also Supplementary Fig. 6).
Fig. 5 | Dependence of the overall fit (RMSE) on the individual fit bound parameters in the PUF4 and PUM2 models. The parameter sensitivity of each bound term for the individual sites ($\Delta\Delta G^b_{\text{bX}}$; see Fig. 3a) was determined by varying the $\Delta\Delta G$ value (x axis) of one parameter while keeping the other parameters constant (see “Methods” for details). PUM2 parameter sensitivities (blue curves, left axis) were replotted from ref. 34 and compared to parameters for the PUF4 model (green curves, right axis) following the same fitting procedure. Vertical dashed blue and green lines indicate the best fit values for each model (PUM2 and PUF4, respectively) and were similar to the minima obtained in the sensitivity analysis. Red boxes indicate the consensus or reference sequence at each position.
PUM2, with the marked exception of the flips at the 6/7 position as noted above (Fig. 6). Flipping at this position is clearly favorable (negative $\Delta \Delta G$) for PUF4 and clearly unfavorable (positive $\Delta \Delta G$) for PUM2. Similarly, flipping two residues at the 6/7 position for PUF4 is less unfavorable than for PUM2 for two-residue flips (Table 1). The sensitivity analysis supports the conclusions above of highly similar energetic contributions that are independent of the other residues.

To further dissect the individual site free energy terms ($\Delta \Delta G_i$), we compared single mutations for PUF4 and PUM2 relative to their respective consensus sequences. Each single mutation was made in four different scaffolds to control for structural effects, although not all scaffold variants were present on the analyzed chip due to bottlenecks in library generation (see "Methods"; ref. 34). This comparison again showed similar effects for PUF4 and PUM2, with average observed differences of 0.44 ± 0.4 kcal/mol (Supplementary Fig. 6b). The library also contained variants in the two residue 5′ and 3′ of the consensus sequence; these gave no 5′ preferences, as for PUM2, and modest preferences for the first position 3′ of the consensus site that were also similar to what was observed for PUM2 (Supplementary Fig. 6d and ref. 34). Thus, a more direct look at a subset of the data provides additional support for the derived PUF4 binding model and its similarity to that for PUM2. These data also support similar energetics for the individual PUM2 and PUF4 sites, consistent with their structural and sequence similarity (Fig. 1b, c).

**Discussion**

We have gone beyond motif descriptions for RNA recognition and provided a mathematical model that is quantitative and predictive for RNA occupancy by the RNA binding protein (RBP) PUF4 from *S. cerevisiae*. Comparison of this model to that previously derived for PUM2 provides insights into RNA recognition by these RBPs and identifies the origin of their specificity differences.

The motif descriptions for PUF4 RNA recognition are more complex than those for PUM2 (Fig. 1a). However, the mathematical descriptions for their binding thermodynamics are of similar complexity. The extensive differences that are observed in their binding specificities arise essentially from a difference in a single term that accounts for flipping out a residue between two Pumilio recognition sites (6 and 7; Fig. 3a); this term is favorable for PUF4 but unfavorable for PUM2 (Fig. 4c and Table 1). The benefit of the thermodynamic and mathematical approach is further underscored by its ability to account for the change in specificity of an engineered PUM1 variant, where a change in a (different) single model parameter provides a quantitative description for binding by the engineered variant despite binding differences of >1 kcal/mol for one in five RNA sequences.

Binding motifs have been obtained for many RBPs via traditional genomic experiments or variations thereof. But many are highly degenerate and thus have limited predictive power, and machine learning approaches to date have also yielded motifs with low information content. The superiority of mathematical modeling based on accurate thermodynamic data over motif descriptions is apparent at two levels. First, a motif description is, by definition, a linear, energetically additive model—i.e., binding of residues in a set order with site energetic contributions that are independent of the other residues that are present. But even the Pumilio proteins that appear to have modular recognition elements require terms for non-linear binding ($\Delta \Delta G_i$ terms, Fig. 3a) to accurately account for their specificities (ref. 34 and this work). Although one can expand motif models by defining multiple motifs, this approach leaves unclear how to weight the relative affinities across them. Second, while our Pumilio binding models require little-to-no coupling to be highly predictive,
mathematical models can be expanded, as demanded by current or future data, to include coupling terms and ensure predictive power. RBPs with less regular recognition elements than the Pumilio proteins and with less defined motifs (Fig. 1a, b) will likely require more coupling terms (Fig. 3a, ΔΔG2), reflecting RNA-RNA interactions in bound complexes and interconnected binding modes.

As implied above, the form of the mathematical model needed to describe RNA binding is rooted in the fundamental physical properties of these interactions. The free energy terms for each base binding to each site in PUM1/2 and PUF4 are nearly the same, consistent with the high degree of conservation in their binding interfaces (Fig. 1, Supplementary Fig. 1 and ref. 34). In this way, mathematical models reveal thermodynamic properties that require biophysical studies to uncover molecular explanations.

Our mathematical model showed that recognition specificity is the same or nearly the same within each of the Pumilio proteins. Instead, the specificity arises from the energetic terms for binding continuously (PUM1/2) versus non-continuously (PUF4) from repeat 6 to repeat 7 (Figs. 1b and 4c, d). These observations are consistent with prior motif assignments that showed that PUF4 tends to bind longer RNA motifs than PUM2 (Fig. 1a)9,10,26,27,36. Instead, the specificity landscapes for RNA binding proteins (RBPs). a The landscape of RNA affinity for two RBPs, where the color coding indicates the specificity (bar under plot). b–d Mock comparison of affinities for two RBPs binding, RBP #1 (orange) and RBP #2 (green), to all RNAs in a transcriptome. Three plots are shown that represent a continuum of possibilities, ranging from fully distinct targets (b) to identical specificities (d) with e showing an intermediate case with specificity overlap. Plots in a–d are in terms of ΔΔG, which provides affinities relative to a reference RNA. e, f PUF4/PUF3 affinity–specificity landscapes for the highest affinity binding site for the respective RBP of all S. cerevisiae 3UTRs in the Saccharomyces cerevisiae database (e) and for PUF3 and PUF4 3’UTR targets identified from previous in vivo experiments (f). The mathematical binding models from this and prior work56 were used to calculate the highest affinity binding site within the 3’UTR of each transcript. To obtain estimates for PUF3 using the highly homologous PUM2 data, an additional term was added to account for PUF3’s favorable interaction with C at position ~2 (see Supplementary Table 2 and “Methods”). Values are shown as ΔΔG relative to the respective consensus sequences. Points are given for each binding site (N = 5089) and shaded based on the density of points with the corresponding value. In f, PUF3 targets are depicted by an orange triangle; targets that bind PUF3 and PUF4 with a yellow square, and PUF4 targets with a green circle. N is the number of known targets for each. g Thermodynamic affinity predictions compared to PUF4 RNA–IP enrichment from Saccharomyces cerevisiae. The enrichment (log2 ratio) for PUF4 targets was reported in ref. 9 for significantly enriched targets (from four independent experiments). The enrichment was estimated as 0 for sites in the yeast genome with predicted affinities <4 kcal/mol that were not identified in the target list. The mean of the enrichment was taken for bins of 0.5 kcal/mol with error bars indicate 95% CIs of the mean.
flipping for PUF4 and flipping penalties that are similar to those for PUM2 at other sites suggests that altered curvature, if responsible for the specificity difference, manifests its thermodynamic effects at a particular site. Additional large-scale thermodynamic specificity studies with additional Pumilio protein variants and mutants will be required to test models for their specificity. Specificity differences across similar Pumilio repeats, and how flipping of residues to yield non-contiguous binding is favored and discouraged at different positions and in different Pumilio proteins.

We next turn to broader biological implications and introduce a specificity–evolvability paradox. High specificity in biology is typically considered to be a favorable trait, in line with how one would engineer a high-precision system; the more specific an RBP, the better defined its targets and the more precise its regulation can be. However, increased specificity lessens the likelihood that new RBP functions will evolve, as random mutations in RNA sequence are less likely to significantly increase RBP occupancy at the newly mutated sites. In this scenario, the selective advantage needed to guide the evolution of a protein for a new regulatory role may require multiple deleterious RNA mutations and thus be of low probability. Thus, there is a tension between specificity and evolvability, where evolvability describes the ability to effectively traverse a fitness landscape to find alternative fitness maxima.

To further explore this paradox, consider the binding specificities of two hypothetical RBPs that we describe by the affinity–specificity landscape in Fig. 7a. In one extreme (Fig. 7b), the RBPs have highly distinct specificities, so that the strong binders to RBP1 bind weakly to RBP2 (points in orange region) and vice versa (points in green region); there is no significant overlap in binding (yellow) until binding is negligible for both. In this case, the transition of RNA targets between the RBPs is unlikely, as multiple simultaneous changes in the target sequence are likely needed to provide a selective advantage. However, the Pumilio proteins PUM2/PUF3 maintain a common binding specificity throughout evolution, but their target sequences have changed, with at least five events in eukaryotic evolution leading to distinct RNA target sets28,34. The observed near-continuum of binding specificities across sequences close in Hamming space has likely been key in allowing these transitions28,34. In the other extreme (Fig. 7d) the affinities fall along the diagonal and there is no specificity—all RNAs bind RBP1 and RBP2 with the same affinity; in this case, there is regulation by both RBPs rather than by RBP1 or RBP2. This scenario appears to hold with human PUM1 and PUM234 and likely other Pumilio proteins49,50; presumably multiple RBPs with shared specificity allows same or similar RNAs to be regulated at different times or locations, or in different ways.

In reality, there is a continuous distribution of possible RBP specificities, with the scenario shown in Fig. 7c representing an intermediate between those in Fig. 7b, d, and this distribution might be expected to be representative of most RBPs in nature. Our mathematical models allowed us to explore this conjecture with affinity–specificity landscapes for natural RBPs, and the affinity–specificity landscape for PUF3/PUF4 fits the intermediate regime (Fig. 7e). (We use PUF3, a close homolog of PUM1/2, because it is present along with PUF4 in S. cerevisiae UTRs that share similar high affinity for PUF3 and PUF4, as is important for biological specificity and consistent with their largely distinct target sets28. Nevertheless, there are numerous 3′UTRs that share similar high affinities for PUF3 and PUF4, implying that there is the potential to share binding and potentially use that overlap to shift specificities in the face of a selective pressure. Indeed, a remarkable switch occurred in a branch of fungal evolution, where >100 PUF3 targets transitioned to control by PUF444.32. The change in regulation from one Pumilio RBP to another is predicted to be more probable than a change in regulation by an unrelated RBP because of their overlapping specificity (Fig. 7e). Indeed, the highest affinity PUF3 and PUF4 sites within a given 3′UTR are sometimes the same and sometimes different (see also Supplementary Fig. 7f, g), and the high density of sites of similar and intermediate affinity (centered on the diagonal at around −1.5 kcal/mol in Fig. 7e) arises because of the high probability of finding a UGUA sequence (the common Pumilio strong binding motif) within a 3′UTR.

Figure 7f shows the affinities and specificities for the PUF3 and PUF4 RNA targets identified from in vivo experiments, calculated using our mathematical model. On average, the enrichment of PUF4 targets follows the thermodynamic affinity predictions of this model (Fig. 7g), analogous to what was reported for PUM234. These results also meet our general expectation that the strongest PUF3 (orange) and PUF4 (green) binders tend to be targets of each (Fig. 7f). However, there is a set of PUF3 targets with similar predicted affinities for PUF3 and PUF4 (orange; Pred. ∆ΔG ~0 kcal/mol), and a small number of the PUF3 and PUF4 targets are predicted to bind weakly (upper right quadrant). Also, of the three targets that appear to be common, one has similar strong affinity, but the others are predicted to prefer PUF3 or PUF4 (yellow squares). These outliers may represent limitations in quantitating occupancy from genomic pulldown and cross-linking approaches and/or cellular features that alter binding specificities. These observations underscore the power of comprehensive thermodynamic measurements, as carried out herein, to provide clear predictions and the future challenge to develop approaches to accurately and quantitatively report RNA/RBP occupancies in cells.

**Methods**

**Library design**

An RNA library was designed as outlined in Fig. 2a and Supplementary Fig. 2a. Complete sequence information of each variant is provided in the Source data.

**Library preparation and sequencing**

Methods for amplification, assembly, and sequencing of the library are reported in ref. 34. Sequences for primers and oligos referenced below can be found in ref. 34. See also Supplementary Fig. 2b.

In brief, oligonucleotides were first amplified using emulsion PCR (ePCR) to give a diverse library pool (64–130 nt length). Libraries were then fractionated by size on an 8% polyacrylamide gel into six fractions (based on UV visualization of marker lanes). Purified fractions were re-amplified using the Read2 and RNAFlaInt adapt primers. Final constructs for each library fraction were assembled by PCR using C_read1_bc_RNA, D_read2, OligoC, and OligoD primers, as illustrated in Supplementary Fig. 2a. The C_read1_bc_RNA primer incorporates a unique molecular identifier (UMI), a randomized 15 nucleotide barcode, to check for sequencing errors31. The library was then diluted by bottlenecks to ~700,000 total UMI, a randomized 15 nucleotide barcode, to check for sequencing errors31. The library was sequenced using MiSeq Reagent Kit v3 on Illumina MiSeq instrument.

**Protein expression and purification**

The RNA-binding domain of S. cerevisiae PUF4 (537–888) was first cloned into a PET28a-based expression vector (New England Biolabs) between a His-tag and TEV cleavage site at the N terminus and a SNAP tag at the C-terminus. For protein expression, constructs were transformed into E. coli BL21 (DE3) strain (Agilent) and induced with 1 mM IPTG, between an OD600 of 0.6–0.8 and at 22 °C for 18–20 h. Cells were pelleted by centrifuging at 5000 × g at 4 °C for 15 min and resuspended in lysis buffer A (20 mM NaHEPES (pH 7.4, 500 mM KOAc (Spectrum), 5% glycerol (ThermoScientific), 0.2% Tween-20 (Sigma), 10 mM imidazole (Sigma-Aldrich)), 2 mM dithiothreitol (DTT, Sigma-Aldrich), 1 mM phenylmethylsulfonyl fluoride (PMSF, ThermoScientific), and 1X Complete Mini protease inhibitor cocktail (Roche)). To ensure efficient lysis, cells were passed through four EmulSifex (Avestin) cycles. The mixture was centrifuged at 20,000 × g, 4 °C for 20 min, separating the supernatant from the cell
debris. Polyethylene imine (Sigma) was added dropwise to the supernatant to a final concentration of 0.21% (w/v) over 30 min at 4 °C and constant agitation to precipitate nucleic acids. The mixture was centrifuged at 20,000 × g and 4 °C for 20 min. The resulting supernatant was filtered through a 0.25 μm syringe filter. FPLC (Bio-Rad) was used to purify the PUF4 construct from the cleared lysate. The solution was loaded onto a Nickel-chelating HisTrap HP column (GE) and washed using Buffer A excluding protease inhibitors. The protein was eluted over a 10–500 mM imidazole gradient over 1 h using Buffer A and B (20 mM Na-HEPES, pH 7.4, 500 mM KOAc (Spectrum), 5% glycerol (ThermoScientific), 0.2% Tween-20 (Sigma), 500 mM imidazole (Sigma-Aldrich)). Protein fractions were pooled and desalted using a HiPrep 26/10 desalting column (GE) into Buffer C (20 mM Na-HEPES, pH 7.4, 100 mM KOAc, 5% glycerol, 0.1% Tween-20, 2 mM DTT). The His-tag was removed by a second HisTrap HP column purification, eluting over a 0–500 mM imidazole gradient using Buffer C and B to separate protein from cleaved adapter. The purified protein fractions were again pooled and desalted into Buffer C using a HiPrep 26/10 desalting column (GE). The purified protein was next loaded onto a HiQ column and eluted over a gradient of 50 mM–1 M KOAc using Buffer C and D (20 mM Na-HEPES, pH 7.4, 1 M KOAc, 5% glycerol, 0.1% Tween-20, 2 mM DTT). Protein fractions were pooled and desalted into Buffer E (20 mM Na-HEPES, pH 7.4, 100 mM KOAc, 5% glycerol, 0.1% Tween-20 and 2 mM DTT) and concentrated using Amicon Ultra-0.5 filters (Millipore). The PUF4 solution was diluted two-fold with Buffer E with 80% glycerol (final concentration 43%) and stored at −20 °C until use.

Cy3B-labeling of SNAP-tagged proteins
PUF4 was labeled with Cy3B-BG (product of coupling Cy3B-NHS, GE, and BG-NH2, NEB, as described in ref. 34) for RNA-MaP binding experiments following a similar protocol to that described by ref. 34. SNAP-tagged PUF4 was labelled with Cy3B in Buffer C by combining 20 μM Cy3B-BG and 10 μM of purified SNAP-PUF4 stock at 4 °C for 12–14 h in the dark while rotating. Free dye was removed with a 7 kDa Zeba Spin Desalting Columns (Thermo Fisher Scientific). The fluorescently-labelled protein was concentrated using an Amicon Ultra-0.5 filter (10 kDa, Millipore). For storage at −20 °C, the solution was diluted two-fold with Buffer C with 80% glycerol as described above.

RNA-MaP equilibrium binding experiments
Equilibrium binding experiments were performed as previously described34. In total, three equilibrium binding experiments were performed with PUF4 protein. Two measurements were taken on chip 1, testing the effects of salt concentration on binding and an additional replicate experiment of high salt on chip 2. We use the high salt to precipitate nucleic acids. The mixture was centrifuged at 20,000 × g and 4 °C for 20 min. The resulting supernatant was filtered through a 0.25 μm syringe filter. FPLC (Bio-Rad) was used to purify the PUF4 construct from the cleared lysate. The solution was loaded onto a Nickel-chelating HisTrap HP column (GE) and washed using Buffer A excluding protease inhibitors. The protein was eluted over a 10–500 mM imidazole gradient over 1 h using Buffer A and B (20 mM Na-HEPES, Spectrum, pH 7.4, 500 mM KOAc (Spectrum), 5% glycerol (ThermoScientific), 0.2% Tween-20 (Sigma), 500 mM imidazole (Sigma-Aldrich)). Protein fractions were pooled and desalted using a HiPrep 26/10 desalting column (GE) into Buffer C (20 mM Na-HEPES, pH 7.4, 100 mM KOAc, 5% glycerol, 0.1% Tween-20, 2 mM DTT). The His-tag was removed by a second HisTrap HP column purification, eluting over a 0–500 mM imidazole gradient using Buffer C and B to separate protein from cleaved adapter. The purified protein fractions were again pooled and desalted into Buffer C using a HiPrep 26/10 desalting column (GE). The purified protein was next loaded onto a HiQ column and eluted over a gradient of 50 mM–1 M KOAc using Buffer C and D (20 mM Na-HEPES, pH 7.4, 1 M KOAc, 5% glycerol, 0.1% Tween-20, 2 mM DTT). Protein fractions were pooled and desalted into Buffer E (20 mM Na-HEPES, pH 7.4, 100 mM KOAc, 5% glycerol, 0.1% Tween-20 and 2 mM DTT) and concentrated using Amicon Ultra-0.5 filters (Millipore). The PUF4 solution was diluted two-fold with Buffer E with 80% glycerol (final concentration 43%) and stored at −20 °C until use.

Cy3B-labeling of SNAP-tagged proteins
PUF4 was labeled with Cy3B-BG (product of coupling Cy3B-NHS, GE, and BG-NH2, NEB, as described in ref. 34) for RNA-MaP binding experiments following a similar protocol to that described by ref. 34. SNAP-tagged PUF4 was labelled with Cy3B in Buffer C by combining 20 μM Cy3B-BG and 10 μM of purified SNAP-PUF4 stock at 4 °C for 12–14 h in the dark while rotating. Free dye was removed with a 7 kDa Zeba Spin Desalting Columns (Thermo Fisher Scientific). The fluorescently-labelled protein was concentrated using an Amicon Ultra-0.5 filter (10 kDa, Millipore). For storage at −20 °C, the solution was diluted two-fold with Buffer C with 80% glycerol as described above.

RNA-MaP equilibrium binding experiments
Equilibrium binding experiments were performed as previously described34. In total, three equilibrium binding experiments were performed with PUF4 protein. Two measurements were taken on chip 1, testing the effects of salt concentration on binding and an additional replicate experiment of high salt on chip 2. We use the high salt to define replicate error, which is similar to that previously found for PUM234. We derive the binding model using the low salt data as that provides the largest range of affinities and thus the most information for developing our model. Data from all experiments are available in Source data.

Two salt concentrations were tested varying the MgCl2 and KOAc concentrations in the binding buffer; low (0.75 mM MgCl2, 30 mM KOAc) and high (2 mM MgCl2, 100 mM KOAc). All other reagents remained constant in each buffer (20 mM Na-HEPES, pH 7.4, 0.1% Tween-20, 5% glycerol, 0.1 mg/mL BSA, and 2 mM DTT). The library on the chip was sequentially equilibrated with increasing concentrations of Cy3B-labelled PUF4. Two-fold serial dilutions were prepared with PUF4 in binding buffer at concentrations ranging from 0.0125 to 409.6 nM in addition to a no-protein control. Dilutions were stored at 4 °C in the dark until use. Each solution was pumped into the flow cell sequentially and incubated at 25 °C for 33 min at the lowest protein concentration to 19 min for the highest protein concentration, taking fluorescent measurements after each addition.

Determining active protein fraction by titration
Cy3B-labeled, SNAP-tagged PUF4 (10–100 nM final) was incubated with a saturating concentration of unlabeled consensus RNA (20 nM; ATUGUGUAGUGU: Integrated DNA Technologies (IDT), Coralville, IA) and trace 32P-labeled RNA of the same sequence (<0.1 nM) for 35 min (25 °C) under high salt conditions (2 mM MgCl2, 100 mM KOAc, 20 mM Na-HEPES, pH 7.4, 0.2% Tween-20, 5% glycerol, 0.1 mg/mL BSA and 2 mM DTT). After equilibration, 7.5 μL aliquots were transferred to 5 μL of ice-cold loading buffer, containing 6.25% Ficoll PM 400 (Sigma), 0.075% BPB, and 2.5 μM unlabeled consensus RNA in binding buffer. The samples were loaded on a pre-chilled 20% native acrylamide gel and run at 42 V/cm constant voltage at 5 °C, in 0.5x Tris/Borate/EDTA (TBE) running buffer (44.5 mM Tris-borate, 1 mM Na2EDTA, pH 8.3). The gel was dried, exposed to a phosphorimagery screen, and scanned with a Typhoon 9400 Imager. The binding signal was quantified using TotalLab Quant (TotalLab, Newcastle-Upon-Tyne, UK), and fitting was performed with KaleidaGraph 4.1 (Synergy Software, Reading, PA; RRID:SCR_014980). The active protein fraction (69%) was determined based on the intersection of lines fit through protein concentrations above and below the breakpoint.

Sequencing data processing
Sequencing data was processed as previously described in ref. 34. Briefly, the tile identifier and x- and y-locations for each cluster were computationally derived from the fastq sequencing output as outlined in refs. 40,53. Sequencing clusters were divided into three categories (1) RNA library of interest, (2) fiducial marker, and (3) PhiX background DNA based on alignment of the read1 sequence to (1) RNA polymerase initiation site and stall sequence TTTATGCT ATATTTTTCATGTAGTAAGGAGGTTGTATGGAGCAGTCTCGAT CC, (2) fiducial marker sequence CTGGGTCACAGGACCTC TGTCCTTCC, or (3) neither, respectively. Only RNA-encoding clusters were used for K0 fitting. Library variants were assigned to sequencing data clusters using a unique molecular identifier (UMI, barcode in Supplementary Fig. 2b) as described in ref. 34.

Fluorescence normalization and determination of the free energy of binding
The protein-bound fluorescence (green channel) was normalized to the amount of RNA transcribed (red channel) for a given cluster. The normalized fluorescence values at varied protein concentration were used to estimate the equilibrium dissociation constant Kd through a multi-step fitting procedure to allow robust fitting over a range of affinities as previously described in ref. 34. First, the normalized fluorescence values were fit to a binding curve to obtain best-fit values for Kd for each variant and variant-independent parameters fmax,fmax and KD,NNS. Next, the distribution of variant-independent values was used to refine these parameters for all variants. Finally, the variant-independent variables were used to refine an estimate for Kd for each variant.

The binding model includes a term for non-specific binding where a second protein monomer can bind to the RNA–protein complex at high protein concentrations, as was observed previously for PUM2 (ref. 34, Eq. 1).

\[ R + P \rightleftharpoons R \cdot P \rightleftharpoons P \cdot P \]

(1)

where R is RNA, P is protein, Kd is the dissociation constant (Kd = eGRT), and KD,NNS is the non-specific dissociation constant (KD,NNS = eGmaxRT).
The normalized fluorescence is related to the protein concentration \([P]\) by the following equation:

\[
f = f_{\text{min}} + f_{\text{max}} \frac{[P]}{[P] + K_D} \left(1 + \frac{[P]}{[P] + K_{D,\text{NS}}}ight)
\]  

(2)

**Single-cluster fitting**

Python’s least-squares fitting package (v0.8.3) was used to fit the above binding model to the normalized fluorescence values of each individual cluster. During fitting all parameters \((K_D, f_{\text{min}}, f_{\text{max}},\) and \(K_{D,\text{NS}}\)) were allowed to vary to find the optimal fit. The initial estimates and constraints are as follows: \(f_{\text{min}}\) was initialized as the median fluorescence values of all clusters at the lowest applied protein concentration and constrained to be greater than or equal to zero; \(f_{\text{max}}\) was initialized as the maximum observed fluorescence value for each individual cluster and constrained to be greater than or equal to zero; \(K_D\) was initialized as the highest observed protein concentration; \(K_{D,\text{NS}}\) was initialized as five-fold the highest applied protein concentration.

**Defining variant-independent parameters**

To increase the confidence of the \(K_D\) initially fit above, the parameters \(f_{\text{min}}, f_{\text{max}}\) and \(K_{D,\text{NS}}\) were estimated based on the per-variant values for each individual dataset, where the per-variant values are the median of the single-cluster values associated with the same molecular variant. The value for \(f_{\text{min}}\) was largely consistent across variants (Supplementary Fig. 2d). Thus, the median value across all variants was used as an estimate for \(f_{\text{min}}\). The values for \(f_{\text{max}}\) are similar for different salt concentrations, with the following values: 0.0134 for the low salt concentration and 0.0159 and 0.0388 for the high salt replicate concentrations. The value of \(f_{\text{max}}\) varied widely for variants with weak affinities where protein binding was not saturating at the concentrations tested. In contrast, the \(f_{\text{max}}\) value was consistent for high-affinity variants (Supplementary Fig. 2e). To find common values for \(f_{\text{max}}\) a subset of variants with low and precisely measured \(K_D\) values based on the single-cluster fits were used to fit a gamma distribution as outlined in ref. 34. Variants with low \(K_D\) values were defined as precise if they had a standard error of \(\Delta G < 1\) kcal/mol, a standard error on the fit \(f_{\text{max}}\) less than \(f_{\text{max}}\) and an initial \(K_D\) value that was <15% of the highest protein concentration tested. The significance of the data filtering was assessed by rejecting the null hypothesis that 25% of clusters would pass all of these filters by chance alone (binomial \(p\) value < 0.01). \(K_{D,\text{NS}}\) values do not vary considerably for non-saturating variants, similar to \(f_{\text{min}}\) (Supplementary Fig. 2f). The values for \(\Delta G_{\text{NS}}\) are similar for different salt concentrations, with the following values: \(\sim 8.49\) kcal/mol for the low salt concentration and \(\sim 8.30\) and \(\sim 8.58\) kcal/mol for the high salt replicate concentrations.

**Determining \(K_D\)**

A refined estimate of \(K_D\) was determined by refitting a single binding curve to the median fluorescent cluster values at each protein concentration per variant, using the fixed variant-independent parameters for \(f_{\text{max}}\) and \(K_{D,\text{NS}}\) from above. For variants that did not achieve near-saturation, defined as the median fluorescence values at the highest concentration of protein being less than the 99.7% confidence interval of the fit \(f_{\text{max}}\) distribution, \(f_{\text{max}}\) was resampled from the variant-independent distribution. For variants that achieved saturation, \(f_{\text{max}}\) was allowed to vary. To obtain error estimates for \(K_D\), we resampled clusters of a given variant 100 times, fitting the median values with each iteration to determine a 95% confidence interval. The confidence interval was used to determine an error:

\[
\Delta G_{\text{error}} = \Delta G_{\text{upper bound}} - \Delta G_{\text{lower bound}}
\]  

(3)

**Data filtering**

The \(K_D\) values were used to calculate the free energy of binding, \(\Delta G\), where \(\Delta G = RT\ln(K_D)\) (Source data). Variants were included in our analysis if they met the following criteria: (1) Variants with \(\Delta G_{\text{error}} < 1.0\) kcal/mol, (2) Five or more clusters per variant in each experiment and replicate, and (3) observed \(\Delta G\) values less than \(-7.69\) kcal/mol, where more than 15% of RNA was bound at the highest protein concentration.

**Combining replicates**

A small shift (0.1 kcal/mol) in the mean difference between the highest affinity binders was observed between replicate datasets. This was corrected by subtracting the offset from the second replicate dataset. Duplicates were then combined as described by computing the error-weighted mean:

\[
\Delta G_{\text{comb}} = \left(\frac{\Delta G_1}{\sigma_1^2} + \frac{\Delta G_2}{\sigma_2^2}\right)^{-1}
\]  

(4)

and the combined error computed via error propagation:

\[
\sigma_{\text{comb}} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}
\]  

(5)

**Filtering out structured variants**

PUF protein binding is affected by RNA structure34.41. In the initial derivation of the PUM2 thermodynamic model, sequences with predicted structure were filtered out. For this we used Vienna RNAfold (v. 2.4.14) command “RNAfold-pO -T25 -C” and constrained the full length of the designed RNA variants to be single stranded with the exception of the stem when using Scaffold S2a and S2b, which was truncated, e.g.:

\[
\text{UUCUUUCUGUAUAUAAUUCUUCUU}
\]

which was compared with the fully unconstrained sequences equally truncated:

\[
\text{UUCUUUCUGUAUAUAAUUCUUCUU}
\]

.................

Variants were then removed based on their structural stability:

\[
\Delta G_{\text{fold}} = \Delta G_{\text{unconstrained}} - \Delta G_{\text{constrained}}
\]  

(6)

with a cutoff of \(\Delta G_{\text{fold}} < -0.5\) kcal/mol, giving 6180 high precision affinity variants.

**Determining relative affinities**

Affinities were determined relative to the linearly bound UGUAUAUAU sequence. For the reference sequence, an additional structural cutoff of \(\Delta G_{\text{fold}} < -0.2\) kcal/mol was used before choosing the median value of all occurrences across the library (\(n = 237\)) as reference affinity. The \(\Delta \Delta G\) for all variants were calculated, where:

\[
\Delta \Delta G_{\text{obs}} = \Delta G_{\text{variant}} - \Delta G_{\text{UGUAUAUAU,median}}
\]  

(7)

**Assessing alternative binding registers in single mutants**

PUF4 can bind in any available register of the RNA variant. To assess the binding affinity of each alternative registers, the full sequence of
each single mutant was split into overlapping 9-mer windows shifted by a single base. Affinity of each window was determined via the equation below using an additive competitive model, where $X$ is the residue (A, C, U, or G), $b$ is the position of the bound residue (1–9), and $\Delta \Delta G_b$ the penalty ascribed by each mutation relative to the PUF4 consensus UGUAUAUA. Alternative binding registers were determined by comparing the sequence of the window with highest affinity with the sequence of the designed register. All alternative registers and their associated affinities are listed in Supplementary Fig. 6c. Alternative binding registers were only observed for mutants at position 2, each of which had predicted affinities within 0.5 kcal/mol of experimentally observed values and are within experimental error.

$$\Delta \Delta G_{\text{pred}} = -k_B T \ln \left( \sum_{b=1}^{9} e^{-\Delta \Delta G_b} \right)$$

(8)

Development, testing, and evaluation of PUF4 thermodynamic binding models

PUF4 binding was predicted for all 6180 filtered sequences by calculating $\Delta \Delta G_{\text{pred}}$ using the $\Delta \Delta G_b$ and $\Delta \Delta G_f$ penalties from the PUM2 additive non-competitive model described in ref. 34. Due to the high stability of the hairpin stem for variants with scaffold S2a and S2b, only predicted by computing the relative ensemble affinity of each individual variant.

$$\Delta \Delta G_{\text{ensemble}} = -RT \ln \left( \sum_{b=1}^{9} e^{-\Delta \Delta G_b} \right)$$

(9)

The relative ensemble affinity is given by the partition function, summing over all possible registers and modes. Modes include: additive consecutive binding ($\sum_{b=1}^{9} e^{-\Delta \Delta G_b}$), single flips ($\sum_{b=1}^{9} e^{-\Delta \Delta G_f}$), two consecutive flips ($\sum_{b=1}^{9} e^{-2 \Delta \Delta G_f}$), and two independent flips ($\sum_{b=1}^{9} e^{-2 \Delta \Delta G_f}$). The predicted values (Pred. $\Delta \Delta G$) were compared to the experimentally observed values (Obs. $\Delta \Delta G$) in Fig. 3b.

Next, the PUM2 model was refined by allowing all flip parameters to vary and optimize by minimizing the sum of the squared error at a time while the other values were fixed. The value was then varied within the chosen upper and lower bound and the RMSE of Obs. $\Delta \Delta G$ versus Pred. $\Delta \Delta G$ calculated. Parameters that did not show a distinct minima were reset to the value where the initial plateau was reached.

BLAST alignment

The DNA sequence for the yeast PUF4 (Uniprot P25339) homology domain was used as the query sequence for a protein-protein alignment with BLAST (blastp). The PUF4 sequence was aligned to the homology domains for yeast PUF3 (Uniprot Q07507) and human PUM2 (Uniprot Q5T7B2).

Predicting PUF3 and PUF4 binding sites in the yeast genome

The S. cerevisiae genome was obtained from the Saccharomyces Genome Database (www SGD archive.yeastgenome.org/sequence/S288C_reference/genome_releases/, version R62-1). The forward and reverse strands were each divided into 11 nt sliding windows, shifting by a single base. PUF3 (an ortholog of PUM2) and PUF4 binding affinities were predicted for each 11 nt window using the PUM2 or PUF4 models, respectively, to determine the relative ensemble affinity, $\Delta \Delta G_{\text{ensemble}}$ to each consensus sequence (UGUAUAUA for PUF3/PUM2 and UGUAUAUAU for PUF4). Each ensemble affinity is the partition function of binding in every possible register consecutively or flipped, following the equation below:

$$\Delta \Delta G_{\text{ensemble}} = -RT \ln \left( \sum_{b=1}^{9} e^{-\Delta \Delta G_b} + \sum_{b=1}^{9} e^{-\Delta \Delta G_f} + e^{-2 \Delta \Delta G_f} + e^{-2 \Delta \Delta G_f} \right)$$

(10)

The highest affinity register for each 11 nt windows was sorted with all of the windows from highest affinity to lowest. For any window with an overlap in sequence to a window higher on the list, the lower window was removed. Windows with $\Delta \Delta G > 4$ kcal/mol were kept for further analysis, resulting in 407,329 and 419,354 binding sites for PUF3/PUM2 and PUF4, respectively. Comparisons across models were made by matching the affinities for each model for the same window.

Predicting PUF3 and PUF4 binding sites in the S. cerevisiae genome with a longer sequence motif

PUF3 has a similar binding motif to PUM2 with an additional preference for bases CN upstream of the 5′ end of the UGUAUAUAU consensus. To more accurately predict binding to PUF3, two additional positions were included in the mathematical model at position −2 and −1 to the 5′ end, yielding an additional 8 parameters, 4 single mutants at each position. Parameters were estimated based on PUF3 binding affinities calculated in ref. 51. The $K_D$ values were converted to $\Delta G$ using $\Delta G = -RT \ln K_D$ and relative affinities computed with respect to the published UGUAUAUA affinity (add references). The final single mutant parameter for C at position −2 was set to −2 kcal/mol with all other additional parameters set to 0 kcal/mol (see Supplementary Table 2).

The upstream preference for PUF3 requires an extension of the genome analysis window by two bases. The method described above was repeated with a 13 nt sliding window. To allow for a comparison to the PUF4 model, its motif was also extended by two bases upstream with all 8 single mutant parameters set to zero. After filtering, 51936 and 494332 binding affinities with $\Delta \Delta G > 4$ kcal/mol were calculated for PUF3 and PUF4, respectively.

3′UTR annotation for yeast transcriptome

The 3′UTR annotations for yeast transcriptome were downloaded from the Saccharomyces Genome Database (www SGD archive.yeastgenome.org/sequence/S288C_reference). The annotation was analyzed in Python and chromosome, start, stop, and sequence information collated into a BED file and compared to the yeast genome version R62.
Duplicates and overlapping annotations were removed by selecting the longest annotated sequence. Binding sites for PUF3 and PUF4 were determined by counting all filtered binding windows within the annotated sequence or selecting the highest affinity binding window as required.

**Predicting the binding affinity of known PUF4 and PUF3 targets in S. cerevisiae**

A list of PUF4 and PUF3 targets was taken from ref. 9. For target sites in the 3′ UTR, the gene location was compared to predicted binding affinities from the 13 nt window analysis described above. The highest affinity (lowest ΔΔG) site was determined in each 3′ UTR for PUF3 and PUF4 from the windows described above.

**Reporting summary**

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

**Data availability**

Source data for Figs. 2c, 3b–d, including the thermodynamic binding data generated in this study are available in Source data. All other data supporting the findings of this study are available from the corresponding author on request. Source data are provided with this paper.

**Code availability**

Key scripts for the analyses reported here have been deposited at https://github.com/HerschlagLab/PUF4Model.git.

**References**

1. Mitchell, S. F. & Parker, R. Principles and properties of euaryotic mRNPs. Mol. Cell 54, 547–558 (2014).
2. Müller-McNeill, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat. Rev. Genet. 14, 275–287 (2013).
3. Keene, J. D. RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. 8, 533–543 (2007).
4. Keene, J. D. RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. 8, 533–543 (2007).
5. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. Nat. Rev. Genet. 15, 829–845 (2014).
6. Jankowsky, E. & Harris, M. E. Specificity and nonspecificity in RNA–protein interactions. Nat. Rev. Mol. Cell Biol. 16, 533–544 (2015).
7. Singh, G., Pratt, G., Yeo, G. W. & Moore, M. J. The clothes make the mRNA: past and present trends in mRNP fashion. Annu. Rev. Biochem. 84, 325–354 (2015).
8. Baltz, A. G. et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell 46, 674–690 (2012).
9. Gerber, A. P., Herschlag, D. & Brown, P. O. Extensive association of functionally and cytologically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol. 2, 342–354 (2004).
10. Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D. & Brown, P. O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol. 6, 2297–2313 (2008).
11. Darnell, R. B. HTS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip. Rev. RNA 1, 266–286 (2010).
12. Ule, J. et al. CLIP identifies Nova-regulated RNA networks in the brain. Science 320, 1212–1215 (2003).
13. Hafner, M. et al. CLIP and complementary methods. Nat. Rev. Methods Prim. 1, 20 (2021).
14. Wheeler, E. C., Van Nostrand, E. L. & Yeo, G. W. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley Interdiscip. Rev. RNA 9, e1436 (2018).
15. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499, 172–177 (2013).
16. Xue, Y. et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol. Cell 36, 996–1006 (2009).
17. Bradley, T., Cook, M. E. & Blankett, M. SR proteins control a complex network of RNA-processing events. RNA 21, 75–92 (2015).
18. Lebedeva, S. et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Mol. Cell 43, 340–352 (2011).
19. Kuhn, M. & Johnson, K. Applied Predictive Modeling (Springer, 2013).
20. Zamore, P. D., Williamson, J. R. & Lehmann, R. The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. RNA 3, 1421–1433 (1997).
21. Zhang, B. et al. A conserved RNA-binding protein that regulates sexual fates in the C. elegans hermaphrodite germ line. Nature 390, 477–484 (1997).
22. Lu, G. & Hall, T. M. T. Alternate modes of cognate RNA recognition by human PUMILIO proteins. Structure 19, 361–367 (2011).
23. Wang, X., Zamore, P. D. & Tanaka Hall, T. M. Crystal structure of a Pumilio homology domain. Mol. Cell 7, 855–865 (2001).
24. Gupta, Y. K., Nair, D. T., Wharton, R. P. & Aggarwal, A. K. Structures of human Pumilio with noncognate RNAs reveal molecular mechanisms for binding promiscuity. Structure 16, 549–557 (2008).
25. Wang, X., Mclachlan, J., Zamore, P. D. & Tanaka Hall, T. M. Modular recognition of RNA by a human Pumilio-Homology Domain. Cell 110, 501–512 (2002).
26. Fazlollahi, M. et al. Harnessing natural sequence variation to dissect posttranscriptional regulatory networks in yeast. G3, Genes, Genomes, Genet. 4, 1539–1553 (2014).
27. Lapointe, C. P. et al. Architecture and dynamics of overlapped RNA regulatory networks. RNA 23, 1636–1647 (2017).
28. Hogan, G. J., Brown, P. O. & Herschlag, D. Evolutionary conservation and diversification of Puf RNA binding proteins and their mRNA targets. PLoS Biol. 13, e1002307 (2015).
29. Leibovich, L. & Yakhini, Z. Efficient motif search in ranked lists and applications to variable gap motifs. Nucleic Acids Res. 40, 5832–5847 (2012).
30. Campbell, N. H. et al. Molecular basis of structure–activity relationships between Sulphef metal complexes and human telomeric DNA quadruplexes. J. Med. Chem. 55, 209–222 (2012).
31. Galgano, A. et al. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the mRNA regulatory system. PLoS ONE 3, e3164 (2008).
32. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141, 129–141 (2010).
33. Spassov, D. S. & Jurecic, R. Cloning and comparative sequence analysis of PUM1 and PUM2 genes, human members of the Pumilio family of RNA-binding proteins. Gene 299, 195–204 (2002).
34. Jarmoskaite, I. et al. A quantitative and predictive model for RNA binding by human Pumilio proteins. Mol. Cell 74, 986–981 (2019).
35. Cheong, C.-G. & Tanaka Hall, T. M. Engineering RNA sequence motifs to alter protein and microRNA target sites by PAR-CLIP. Mol. Cell 141, 129–141 (2010).
36. Spassov, D. S. & Jurecic, R. Cloning and comparative sequence analysis of PUM1 and PUM2 genes, human members of the Pumilio family of RNA-binding proteins. Gene 299, 195–204 (2002).
37. Jarmoskaite, I. et al. A quantitative and predictive model for RNA binding by human Pumilio proteins. Mol. Cell 74, 986–981 (2019).
38. Cheong, C.-G. & Tanaka Hall, T. M. Engineering RNA sequence motifs to alter protein and microRNA target sites by PAR-CLIP. Mol. Cell 141, 129–141 (2010).
39. Wilinski, D. et al. RNA regulatory networks diversified through curvature of the PUF protein scaffold. Nat. Commun. 6, 8213 (2015).
40. She, R. et al. Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome. Proc. Natl Acad. Sci. USA 114, 3619–3624 (2017).
41. Jarmoskaite, I., Alsadhan, I., Vaidyanathan, P. P. & Herschlag, D. How to measure and evaluate binding affinities. Elife 9, 1–34 (2020).
42. Moore, K. S. & ‘t Hoen, P. A. C. Computational approaches for the analysis of RNA-protein interactions: a primer for biologists. J. Biol. Chem. 294, 1–9 (2019).
43. Pan, X., Rijnbeek, P., Yan, J. & Shen, H.-B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics 19, 511 (2018).
44. Pan, X., Yang, Y., Xia, C., Mirza, A. H. & Shen, H. Recent methodology progress of deep learning for RNA–protein interaction prediction. WIREs RNA 10, e1544 (2019).
45. Miller, M. T., Higgin, J. J. & Tanaka Hall, T. M. Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p. Nat. Struct. Mol. Biol. 15, 397–402 (2008).
46. Wang, Y., Opperman, L., Wickens, M. & Tanaka Hall, T. M. Structural basis for specific recognition of multiple mRNA targets by a PUF regulatory protein. Proc. Natl Acad. Sci. USA 106, 20186–20191 (2009).
47. Valley, C. T. et al. Patterns and plasticity in RNA-protein interactions enable recruitment of multiple proteins through a single site. Proc. Natl Acad. Sci. USA 109, 6054–6059 (2012).
48. Kirschner, M. & Gerhart, J. Evolvability. Proc. Natl Acad. Sci. USA 95, 8420–8427 (1998).
49. Wang, Y. & Voronina, E. Diverse roles of PUF proteins in germline stem and progenitor cell development in C. elegans. Front. Cell Dev. Biol. 8, 29 (2020).
50. Pan, X. et al. The PUF protein family: overview on PUF RNA targets, biological functions, and post transcriptional regulation. Int. J. Mol. Sci. 19, 410 (2018).
51. Zhu, D., Stumpf, C. R., Krahn, J. M., Wickens, M. & Tanaka Hall, T. M. A 5’ cytosine binding pocket in Puf3p specifies regulation of mitochondrial mRNAs. Proc. Natl Acad. Sci. USA 106, 20192–20197 (2009).
52. Gerber, A. P., Luschnig, S., Krasnow, M. A., Brown, P. O. & Herschlag, D. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. Proc. Natl Acad. Sci. USA 103, 4487–4492 (2006).
53. Denny, S. K. et al. High-throughput investigation of diverse junction elements in RNA tertiary folding. Cell 174, 377–390 (2018).
54. Kershaw, C. J. et al. Integrated multi-omics analyses reveal the curvature of the PUF protein scaffold. Proc. Natl Acad. Sci. USA 114, e2816–e2825 (2017).
55. Wilinski, D. et al. Recurrent rewiring and emergence of RNA regulatory networks. Proc. Natl Acad. Sci. USA 114, e2816–e2825 (2017).
56. Engel, S. R. et al. The reference genome sequence of Saccharomyces cerevisiae: then and now. G3: Genes, Genomes, Genet. 4, 389–398 (2014).

Acknowledgements
Financial support was provided by the National Institutes of Health (R01 GM132899 to D.H.). We thank Greg Hogan and Traci Hall for the thoughtful comments on the manuscript and the Herschlag laboratory for discussion and review of the manuscript.

Author contributions
D.H., W.J.G., I.J., and P.P.V. conceived the project and experimental design. I.J. and P.P.V designed the RNA library, performed the RNA-MaP experiment, and purified the PUF4 protein. S.K.D. derived the fluorescent values to determine binding affinities. C.S., S.K.D., and W.R.B. wrote the code for obtaining binding affinities. C.S. performed the model fitting and parameter sensitivity analysis with help from W.R.B. L.D.H. aligned and analyzed the PUF4 and PUM2 crystal structures and sequences. C.S. applied the Pumilio models to the yeast transcriptome. L.D.H. compared the Pumilio models to the in vivo Pumilio targets. D.H., L.D.H., and C.S. wrote the manuscript with input from all authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31968-z.

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

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022