Specificity and nonspecificity in RNA–protein interactions

Eckhard Jankowsky[1–3] and Michael E. Harris[2]

Abstract | To fully understand the regulation of gene expression, it is critical to quantitatively define whether and how RNA-binding proteins (RBPs) discriminate between alternative binding sites in RNAs. Here, we describe new methods that measure protein binding to large numbers of RNA variants and that reveal **[Au: ok? 'measure.. binding patterns' did not seem correct]** the binding patterns they produce, including affinity distributions and free energy landscapes. We discuss how the new methodologies and the associated concepts enable the development of inclusive, quantitative models for RNA–protein interactions that transcend the traditional binary classification of RBPs as either specific or nonspecific.

RNA–protein interactions are critical for the regulation of gene expression[1]. Research over recent decades has shown that RNA is invariably bound and often altered by proteins in cells, and that in biological environments RNAs generally function together with proteins as RNA–protein complexes known as ribonucleoproteins (RNPs)[2,3]. It has also become clear that cellular RNA–protein interactions form a highly complex network involving numerous RNAs and proteins[4]. In addition, a multitude of diseases have been linked to misregulation or malfunction of proteins that interact with RNA[5–7]. Thus, deciphering RNA–protein interactions on both molecular and cellular scales is central to understanding human physiology and disease.

Typical eukaryotic cells contain thousands of different RNAs[8]. For every protein that interacts with RNA, it is critical to understand the molecular characteristics that define whether and how the protein discriminates between different potential binding sites in these RNAs. For this purpose, proteins that interact with RNA are traditionally classified as either 'specific' or 'nonspecific'. Specific proteins associate preferentially with defined RNA sequence or structure motifs, or a combination thereof. Nonspecific proteins associate with RNA sites that seem to be devoid of sequence or structure motifs. Roughly half of all proteins that interact with RNA fall into the nonspecific category. Examples include translation elongation and initiation factors, and proteins involved in RNA degradation[9,10]. Binding to diverse RNA sites is critical for the biological function of nonspecific proteins.

Although the terms specific and nonspecific are widely used, numerous studies that mapped RNA–protein interactions in cells or measured RNA–protein associations for large numbers of sequences *in vitro* indicated that specificity, or the lack thereof, is considerably more nuanced than suggested by the binary specific versus nonspecific classification. As descriptions of cellular RNA–protein interaction networks move towards systems-level quantitative models[4,11,12], and as other lines of research attempt to engineer novel RNA-binding proteins (RBPs)[13–16], a comprehensive, quantitative view on specificity and nonspecificity is required. In this Review, we discuss emerging approaches aimed at this goal. We start with a brief overview of the tremendous complexity of RNA–protein interactions *in vivo* (in the cell). We then discuss new methods that enable quantitative measurement of protein binding to large numbers of RNA variants, as well as description of the resulting binding distributions: **[Au: colon OK here?]** binding models and free energy landscapes. Finally, we review the insights and potential provided by these new methods and associated concepts that contribute towards devising a nuanced, inclusive description of specific and non-specific RNA–protein interactions. **[Au: ok?]**

RNA–protein interaction complexity

In mammalian cells, more than 1,000 diverse proteins interact with RNA[1,17–19]. For the purpose of this Review, we refer to these proteins as RBPs, although only a subset of these proteins function solely to bind RNA. In humans, a certain set of RBPs is expressed in all tissues investigated thus far[1]. For other RBPs, expression can vary considerably, and some are expressed exclusively in certain tissues[1,5,20,21]. Many RBPs have a modular structure, often containing multiple and different **[Au: ok?**

[1]*Center for RNA Molecular Biology, Case Western Reserve University.*
[2]*Department of Biochemistry, Case Western Reserve University.*
[3]*Department of Physics School of Medicine, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA.*
e-mails: exj13@case.edu; meh2@case.edu

doi:10.1038/nrm4032
Published online
XXXX 2015

Table 1 | **Classification of common protein domains that interact with RNA***

Domain class	Subclass (superfamily): Family
Nucleotidyltransferases	PAPs: Canonical PAPs and non-canonical PAPs
	Terminal uridylate transferases
	CCA-adding enzyme
	Guanylyltransferases
	RNA ligases
	2′–5′ PAPs
	RNA-dependent RNA polymerases
Ribonucleases	α/β[‡] **[Au: does the / mean ‘and’, ‘or’ or ‘and/or’? Please avoid the solidus]**: RNase A, RNase H, 3'→5' exo, **[Au: addition of prime symbols ok?] [Au: can ‘exonuclease’ be written out or defined below?]** RNase II, RNase R **[Au:OK?]**, RNase E, RNase PH and metallolactamase
	α and β: RNase T2 and XRN1
	α[‡]: RNase III
	Decapping enzyme
RNA-modifying enzymes	tRNA synthetases: class I and class II
	Deaminases: ADAR, APOBEC, TadA and CDA
	Pseudouridine synthases
	Methyltransferases: RMT, SPOUT, Radical SAM-dependent methyltransferase and FAD/NAD(p) **[Au: what does the / mean here? Please spell out. And why is there a p in brackets? Does it mean NAD and/or NADP (best spelled out)?]**
Helicases	Superfamily 1: Upf1-like
	Superfamily 2: Ski2-like, RIG-I-like, DEAD-box, DEAH, RHA **[Au: OK?]**, Viral superfamily 2 and Cas3
	Superfamily 3
	Superfamily 4
	Superfamily 5
GTPase	EF-Tu, EF-G **[Au: OK?]**, BMS1, SNU114**[Au: OK?]**
RNA-binding domains	RRM, KH, S1, OB-fold, PUF, sRBD, zinc-fingers, PAZ, PIWI, LSM, KOW, MIF4G, NTF2, GAR, HEAT repeat, homeodomain and CSD

[Au: we usually use all upper case for genes and proteins from no particular organism (when talking in general). Should we use upper case throughout for the proteins here? Or are certain proteins above and below only found in particular organisms (e.g. yeast or fly) or are the papers they are from on different organisms?] ADAR, adenosine deaminase that acts on RNA; APOBEC, apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like; BMS1, BMh-sensitive 1; Cas3, CRISPR-associated protein 3; CDA, cytidine deaminase; CSD, cold shock domain; dsRBD, double-stranded RNA-binding domain; EF-G, elongation factor G; EF-Tu, elongation factor thermo unstable; FAD/NAD(p), flavin adenine dinucleotide/nicotinamide adenine dinucleotide phosphate **[Au: please define the solidus (/) here. Please clarify why the p is in brackets]**; GAR, glycine arginine rich; HEAT, Huntington, elongation factor 3, protein phosphatase 2A and TOR1; KH, K homology; KOW, Kyprides–Onzonis–Woese; LSM, like Sm; MIF4G, MA-3 and eIF4G; NTF2, nuclear transport factor 2; OB-fold, oligonucleotide/oligosaccharide-binding fold; PAPs, poly(A) polymerases; PAZ, PIWI, Argonaute and Zwille; PIWI, P-element induced wimpy testis; PUF, Pumilio and FBF; RHA, RNA helicase A; RIG-I, retinoic acid-inducible gene I; RMT, ribomethyltransferase; RRM, RNA-recognition motif; S1, similarity to ribosomal protein 1; Ski2, superkiller 2; SNU114, small nuclear ribonucleoprotein-associated 114; SPOUT, spoU and trmD RNA methylase; TadA, tight adherence protein A; Upf1, up-frameshift suppressor 1; XRN1, 5′–3′ exoribonuclease 1. *The classification of ribonuclease domains is based on data from REF. 17, of helicase domains on data from REF. 25 and of methyltransferase domains on data from REF. 134. The compilation of RBDs is based on data from REF. 1. [‡]The classification of nucleases is based on protein folds: α, α-helices; β, β-sheets.

or ‘many different’?] RNA-interacting domains[1,22,23]. RNA-interacting domains are traditionally called RNA-binding domains (RBDs), but these domains often also harbour functions other than RNA binding (TABLE 1). For the purpose of this Review, we keep the RBD designation. The main RBD classes include enzymatic domains that chemically alter RNA (such as nucleotidyltransferases, ribonucleases and RNA-modifying enzymes) and those that couple nucleotide binding or hydrolysis to RNA binding or structural remodelling (such as GTPases and helicases) (TABLE 1). **[Au: If the examples in brackets are the only examples of these types of domains then please stet the ‘such as’]** In addition, there are numerous RBDs that only bind RNA. Some RBDs are found in large numbers of proteins[1,5,17]. The most frequently occurring is the RNA-recognition motif (RRM), an RNA-binding module present in several hundred mammalian proteins[24]. The most common enzymatic domain is the helicase

Nucleotidyltransferases
Enzymes that catalyse the transfer of a phosphorylated nucleoside from one compound to another.

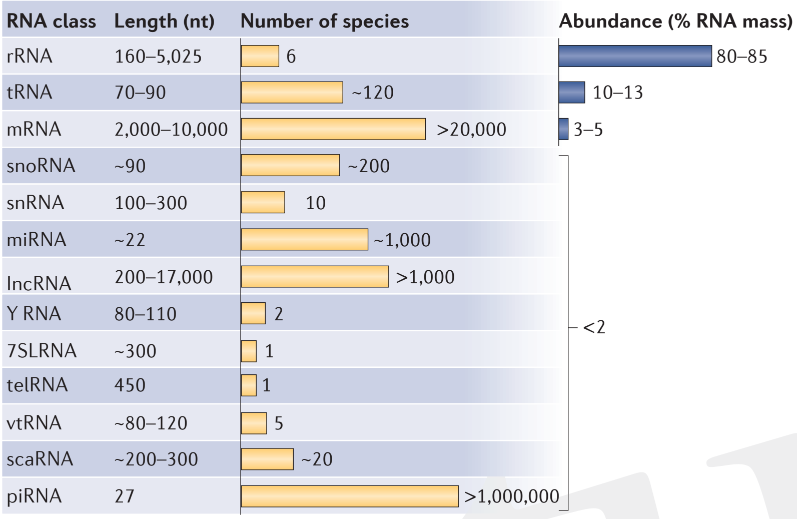


Figure 1 | **The major classes of eukaryotic RNAs.** For each class of RNA, approximate length, number of different species and abundance are indicated. For more detailed information, see REF. 133. The length of mRNAs is given for mature, processed species; the number of species refers to putative mRNA-coding genes. Long non-coding RNAs (lncRNAs) include all RNAs that do not explicitly belong to another RNA class and that exceed 200 nucleotides (nt) in length. 7SLRNA refers to the RNA component of the signal-recognition particle (SRP). PIWI-interacting RNAs (piRNAs) are expressed only at specific stages of germ cell development and are not included in calculations of cellular RNA abundances. miRNA, microRNA; rRNA, ribosomal RNA; scaRNA, small Cajal body-specific RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNAs; telRNA, telomeric RNA; vtRNA, vault RNA **[Au: Does 'vtRNA' refer to 'vault RNA' or to 'viral tRNA'>].**

domain, which is found in roughly 70 human proteins that interact with RNA[17,25]. By contrast, other domains — for example RNA guanyltransferase — **[Au: not guanylyltransferases, to match table?]** are found in only a single protein per organism[26]. Finally, proteins that interact with RNA vary widely in their abundance, ranging from few to 100,000 molecules per cell[27].

RNA binding is not restricted to proteins with domains that are traditionally viewed as RBDs. Recent work has revealed extensive RNA association of considerable numbers of metabolic enzymes lacking previously identified RBDs[18,19,28,29]. Other studies show association of (mostly long non-coding) RNAs with transcription factors[30–32]. The number of proteins that demonstrably interact with RNA is thus likely to grow in the future.

The number of RNA species far exceeds the number of RBPs in typical eukaryotic cells. Human cells encode more than 20,000 different mRNAs (FIG. 1); most cell types express between 11,000 and 15,000 at any time[33]. The diversity of mRNAs is further increased by alternative splicing[34] and by chemical modifications[35–38]. In addition to mRNAs, metazoan cells can express thousands of species of long non-coding RNAs and hundreds of microRNAs (miRNAs), tRNAs and small nucleolar RNAs (snoRNAs). At certain stages of germ cell development, large numbers of PIWI-interacting RNAs (piRNAs) are expressed[39]. Conversely, there are only a few ribosomal RNA (rRNA) and small nuclear RNA (snRNA) species. In addition, cleaved RNA fragments are emerging as potential regulatory molecules[40–42]. The various RNA types differ dramatically in their abundance. In most eukaryotic cells, rRNAs account for roughly 80–85% of the cellular RNA mass, followed by tRNAs, mRNAs and snoRNAs; all other RNAs together account for less than 2% of the mass (FIG. 1). At certain stages of germ cell development, these RNA mass ratios might change owing to the expression of piRNAs[39]. Even within each RNA class, abundance varies widely. The expression levels for mRNAs range over four orders of magnitude[33]. A small number of mRNA species often accounts for 50% of the cellular mRNA mass. For example, 50% of the mRNA mass is contributed by only 250 mRNA species (~4%) in yeast, by 900 mRNA species (~7%) in the human cerebellum and by fewer than 10 mRNA species (~0.01%) in **[Au: human?]** liver tissue[33]. Another factor contributing to the disparity in cellular RNA mass is that RNAs vary greatly in their length, ranging from more than 10,000 nucleotides (mRNAs and long non-coding RNAs) to only 22 nucleotides (miRNAs) (FIG. 1).

Any individual RNA is usually bound by multiple proteins[3,4]. Different proteins can bind simultaneously, subsequently **[Au: does this mean in succession? If not, please clarify what it is subsequent to]** or in a mutually exclusive manner[3,4]. Conversely, most proteins can bind multiple RNAs[43]. Some proteins, such as the mRNA-export factors, need to contact many diverse mRNAs[44], and the translation elongation factor thermo unstable (EF-Tu) binds all charged tRNAs[45]. Given the number of RNAs and RBPs, the number of possible RNA–protein interactions is extremely large. Further variation is added by proteins that do not directly contact the RNA but modulate the binding of RNA by RBPs; for example, through post-translational modifications or through interactions with RBPs[46,47]. RNAs can also interact with one another, as illustrated most prominently by the interactions between mRNAs, miRNA and competing endogenous RNAs (ceRNAs)[48,49]. Given the simultaneous presence of large numbers of RNAs and RBPs and the layers of modulation of their interactions by other proteins, cellular RNA–protein interactions represent a massive set of interdependent interactions. Most RBDs recognize sites made of only 3 to 8 nucleotides and often tolerate a high degree of sequence variation in these binding sites[3]. Thus, the number of potential interactions of even highly selective proteins in organisms with small transcriptomes, such as yeast, can be extraordinarily large.

Every interaction between an individual protein and a specific RNA site is dictated by the inherent affinity of the protein for the RNA site, the concentration of the protein, the concentration of the RNA, the competition from other RNAs for association with the protein, and the competition among **[Au: 'among' ok?]** other proteins for the RNA's binding site. In addition, proteins that interact with or modify RBPs can profoundly affect RNA-binding patterns. Therefore, it is not surprising that substrate selection by a given protein rarely conforms to a binary specific versus nonspecific binding model. Yet, the challenge remains to devise models that describe RNA–protein interactions in sufficient

Charged tRNAs
tRNA molecules that are chemically bonded by a 2′ or 3′ aminoacyl linkage to its cognate amino acid.

Competing endogenous RNAs
(ceRNAs). RNAs that regulate other RNA transcripts by competing for shared micro RNAs.

quantitative detail to allow predictions of the RNA-binding pattern of individual proteins under a defined set of parameters. A critical first step towards this goal is addressed by approaches that quantitatively assess the binding of proteins to many different RNA sites.

Measuring protein binding to many RNAs

Several methods have been developed to determine protein-binding sites on RNAs on a transcriptome-wide scale[50,51]. The techniques rely either on the covalent crosslinking of proteins to RNA by ultraviolet irradiation followed by immunoprecipitation (crosslinking and immunoprecipitation (CLIP) and its derivatives)[52–55] or on immunoprecipitation of RNA-bound proteins with a chemical crosslinker (RNA–protein immunoprecipitation in tandem (RIPiT)[56]) or without[57]. The crosslinked RNA fragments are identified by next-generation sequencing or microarray analysis. These methods represent a quantum leap forward with respect to visualizing protein-binding patterns on RNAs, often revealing binding to numerous different sites on large numbers of RNAs. The binding sites often allow the definition of consensus motifs for protein binding[43]. Although powerful and highly instructive, these techniques do not currently provide the quantitative data necessary to assess affinity or binding and dissociation kinetics of RNA–protein interactions.

Other, novel approaches aim to quantitatively measure protein binding to large numbers of RNA variants *in vitro*. Recently, *in vitro* selection by systematic evolution of ligands by exponential enrichment (SELEX) was combined with high-throughput sequencing[13,58,59]. SELEX has traditionally been used to identify the few RNA species most preferentially bound by RBPs[59]. The combination with next-generation sequencing allows the analysis of much larger numbers of sequences and thus provides insight into the RNA-binding preferences of RBPs beyond the tightest-bound species[13,58,59]. SELEX has also been used to determine the binding affinities **[Au:OK?]** of RBPs in the cell[60]. However, even when combined with next-generation sequencing, SELEX approaches produce a bias in the RNA-binding analysis towards the highest-affinity targets.

To avoid this bias, other techniques have been developed that directly analyse interactions of proteins with large populations of diverse RNAs (BOX 1). These methods bypass the selection and amplification cycles of the SELEX procedure and allow measurements of both weakly and tightly bound RNA species. Some of these techniques are analogous to high-throughput methods for investigating the binding of transcription factors to large numbers of DNA sequences[61]. All of these approaches measure differences in protein binding to a pool of diverse RNA substrates (BOX 1).

RNA–protein affinity distributions

Many studies that map RNA–protein interactions on a transcriptome-wide scale show that RBPs often bind to RNA sites that vary considerably in sequence or structure[43,62]. This is expected for proteins considered to be nonspecific binders but seems to contradict the notion of sequence-specific binding proteins. Similar observations have been made for DNA binding by sequence-specific transcription factors[63]: *in vitro* measurements of intrinsic affinities of transcription factors for all possible sequence variants of DNA oligomers showed that each protein had a wide range of binding affinities to different sequence variants[64–66]. Differences between equilibrium dissociation constants for low- and high-affinity sites are often considerable, and they can span several orders of magnitude[63–66].

To describe the entire range of affinities seen for a given DNA-binding protein or RBP towards all possible DNA or RNA species, it is useful to use affinity distributions[67], which are histogram plots of substrate variants with similar affinities (FIG. 2). Affinity distributions have revealed incremental contributions **[Au: is it possible to clarify 'incremental' here?]** of the nucleotides in the binding site to the equilibrium binding free energy, rather than a drastic difference between nucleotide composition in preferred and non-preferred sites (FIG. 2). For sequence-specific transcription factors, physiologically preferred binding sites cluster at the high-affinity region of the distribution[61,67,68].

A complete, quantitative RNA-affinity distribution has so far been reported only for C5, the protein subunit of RNase P from *Escherichia coli*[67] (FIG. 2). Distributions of ranked binding preferences, which are related but not identical to an affinity distribution, have been measured by the RNAcompete method for a larger number of RBPs[69]. The shape of the observed affinity distributions is similar to those seen for transcription factors[67], also suggesting incremental contributions of the nucleotides in the binding site to the binding free energy.

Incremental differences between sequence variants explain why proteins can bind with similar affinity to a range of seemingly divergent sequences (FIG. 2a), which is particularly significant for RBPs because cognate sites for most RBDs encompass only 3–8 nucleotides[3]. Potential binding sites of this size occur (with a few substitutions) at very high frequency even in small genomes. Inconstant protein-binding preferences may be attributable to the varying expression levels of the RNAs[33]. **[Au: we were not 100% clear on your intended meaning. Please check edits and clarify if the above is not correct. In what way is 'ambivalence' amplified?]** At limiting concentrations of protein, low-affinity, non-consensus sites in highly expressed RNAs can compete efficiently for protein binding with high-affinity consensus sites in RNAs expressed at a lower level. This might be one of the reasons that, in cells, proteins that are considered specific often bind to sites with relatively poor matches to their consensus motifs[62]. Whether binding to such degenerate sites has biological consequences other than protein sequestration is an open question.

Affinity distributions also provide the means to comprehensively quantify the specificity of a given RBP or RBD, although to our knowledge this type of quantification has not yet been reported. The width of the distribution provides an objective measure for the capacity of an RBP or RBD to globally discriminate between RNA substrate variants. Multimodal distributions are also conceivable and would describe different binding modes, which could arise, for example, through the formation of stable RNA structures by a subset of substrates.

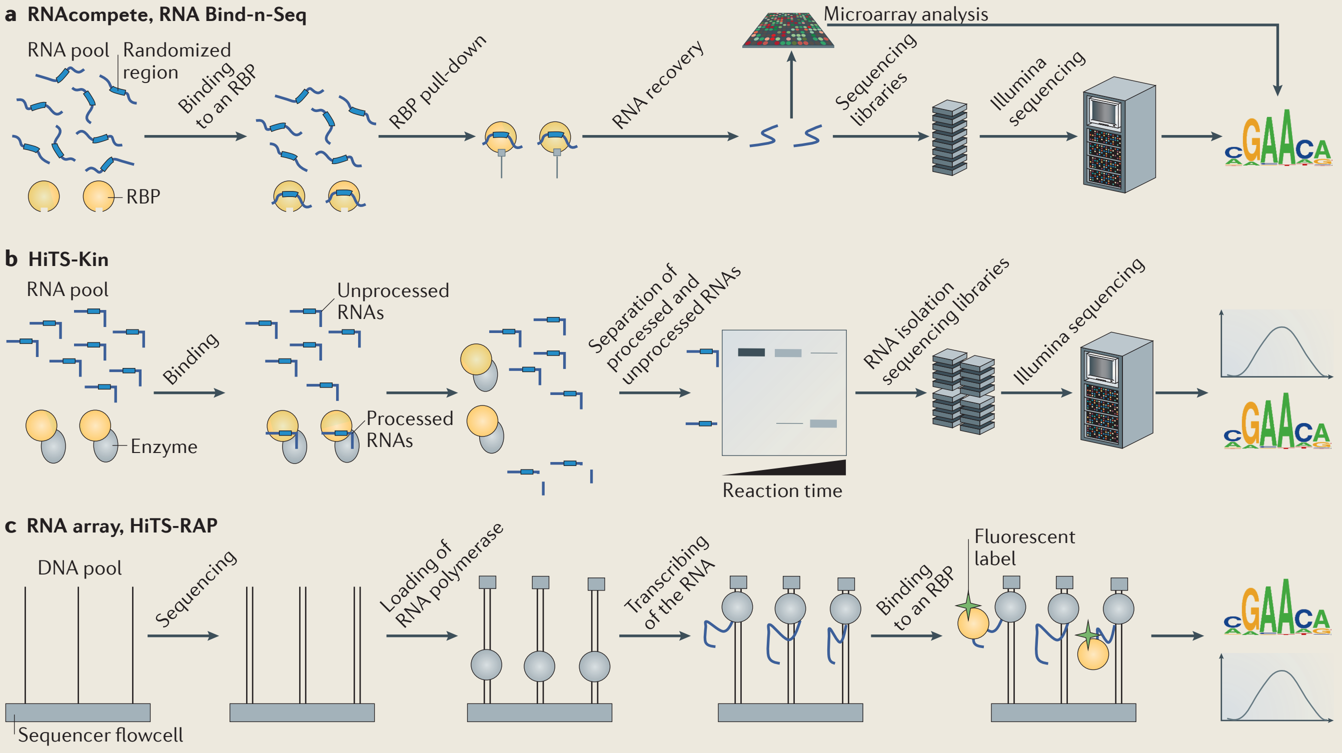
Equilibrium binding free energy
The Gibbs free energy (ΔG), typically measured in units of kcal per mol, **[Au: 'per' ok?]** for an equilibrium binding reaction that is related to the equilibrium dissociation constant, K_d.

Box 1 | Techniques for measuring protein binding to many RNA sequences *in vitro*

In RNAcompete and RNA Bind-n-Seq, a pool of RNA species, each containing a region of randomized sequence, is incubated *in vitro* with a specific RNA-binding protein (RBP). The RBP is pulled down, bound RNAs are recovered and their sequences are determined by microarrays (RNAcompete)[130] or next-generation sequencing (Bind-n-Seq; see the figure, part **a**)[104]. These methods have been used to determine sequence motifs for RNAs that bind tightest to a given protein[69,130]. The number of sequences that can be measured simultaneously is currently limited to between approximately 2.5×10^8 and 5×10^8 RNAs, corresponding to 9–10 randomized nucleotides.

High-throughput sequencing kinetics (HiTS-Kin; see the figure, part **b**) follows the enzymatic processing of RNA in a reaction that depends on an RBP, thus measuring functional RBP binding to RNA[67]. Processed and non-processed RNA species are separated (for example, using gel electrophoresis). The ratios of processed versus non-processed RNAs over time are analysed by next-generation sequencing, providing kinetic information[67]. HiTS-Kin can be adapted to different experimental systems and to reactions *in vivo*, provided that reactive and unreactive RNA species can be separated.

RNA array and high-throughput sequencing–RNA-affinity profiling (HiTS-RAP; see the figure, part **c**) directly visualize the RNA–protein interactions in the Illumina next-generation sequencer[91,131]. A pool of diverse DNA sequences is immobilized in the sequencer flowcell. Each respective sequence is identified by a round of sequencing. Subsequently, each DNA serves as template for RNA polymerase to transcribe RNA. Following transcription, the polymerase is stalled (squares mark the block, which in RNA array is biotin-streptavidin at the terminus of the DNA and in HiTS-RAP is the binding of the protein terminus utilization substance to a terminator site in the DNA). **[Au:OK? Tus and Ter removed as they are not in the figure]** Transcribed RNA remains bound to the stalled polymerase, thus allowing identification of each RNA species. A fluorescently labelled RBP is added and its interaction with the RNA is directly monitored by measuring fluorescence changes at the positions that correspond to the different RNAs. Proteins can be flowed in and out the flowcell multiple times and at different concentrations, providing readouts of binding and dissociation kinetics in real time[91]. RNA array and HiTS-RAP are conceptually similar to techniques for measuring the kinetics of protein–RNA interactions by single-molecule fluorescence via total internal reflection[132].



How affinity distributions are related to RNP structures is currently not understood. A recent pioneering study investigated the structural basis for a range of affinities that the bacterial RBP ribosomal RNA small subunit methyltransferase E (RsmE) shows towards different substrate variants[70], although no complete affinity distribution was measured. For RsmE, conformational adaptation of protein side chains and of RNAs is responsible for the range of affinities[70].

Binding models

As noted, affinity distributions are useful because they represent a non-biased description of protein binding to unstructured RNA, to a defined RNA structure, or to a combination of both. For proteins that bind to unstructured RNA, sequence variants in the high-affinity region of the distribution share a consensus sequence motif[67] that can be expressed as a sequence probability logo[61]; other regions of the distribution do not share a sequence consensus (FIG. 2b). Consensus sequences describe the probability by which a given nucleotide is present at a given position in the binding site for a subset of all sequence variants[61,68]. The larger the number of sequence variants in a given subset of the distribution, the weaker the consensus (FIG. 2b). There are several approaches to delineate consensus motifs from binding-site data obtained either *in vitro* or *in vivo*[71–78]. A consensus motif can guide a qualitative prediction of whether or not a

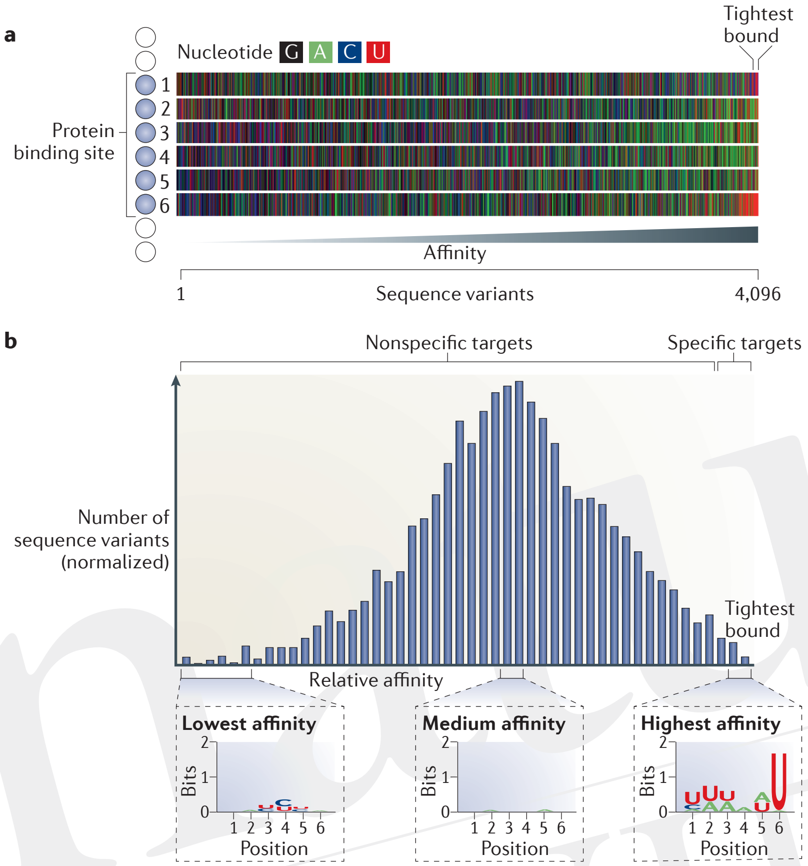


Figure 2 | **RBP affinity distributions.** **a** | Ranked affinities for an RNA-binding protein (RBP) with a binding site of six nucleotides (C5 from *Escherichia coli*) to all possible RNA variants[67]. The numbers on the left indicate the nucleotide position in the binding site. **b** | Histogram of relative affinities (on a logarithmic scale) for the sequence variants shown in part **a**. Relative affinities are calculated in relation to a standard variant, which can be chosen freely[67]. Specifically bound RNA variants cluster in the high-affinity region of the distribution and produce a binding consensus sequence (motif), shown as a logo underneath the plot. The remainder of the distribution consists of nonspecific RNA variants, which do not produce a consensus motif. **[Au: perhaps explain the lowest-affinity logo here too?]**

protein binds well to a certain motif. However, in most cases, consensus motifs do not allow affinity calculations for different sequence variants, and they only rarely provide information on the characteristics of the entire affinity distribution[61,68].

The simplest model to describe the binding of a protein to all RNA sequence variants is the position weight matrix (PWM)[61,68]. A PWM is a score calculated for each nucleotide at each position in the binding site (FIG. 3a). The sum of the individual nucleotide scores for a given sequence provides a score for this sequence variant[61]. The PWM can also be visualized as a logo[68], but it is important to note that a PWM logo differs from the probability logo discussed above. If affinities are expressed as binding free energies, a PWM becomes an energy score, describing the energetic contribution of each nucleotide at each position to the binding free energy[61,68]. A PWM assumes **[Au: OK?]** that the nucleotides at each position contribute independently of each other to the binding of the protein[61,68,79]. PWMs often explain only a subset of the observed experimental variance in affinities[61,66,68,80], and they frequently fail to accurately explain the highest and lowest observed affinities[67]. To more accurately describe observed affinities of DNA-binding proteins, **[Au: should a link be given in next sentence suggesting similar >1 PWM use for RBPs?]** it has been suggested to use more than one PWM for a single protein[63]. This would imply multiple binding modes of an RBP.

A significant improvement in the description of the experimental variance is often obtained by considering coupled contributions from multiple positions in the binding site[67,81,82] (FIG. 3b). Couplings are incorporated by assigning a score for each combination of nucleotides and then summing the score for the combinations present in a given sequence[67]. The incorporation of even a modest number of pairwise couplings (called either a pairwise interaction matrix (PIM) or a dinucleotide weight matrix (DWM)) often improves the binding model[67,81,82]. However, it is critical to carefully evaluate that an improved fit does not result simply from the incorporation of more variables in the model. Of note, only roughly 20–30% of the entire sequence space is needed to produce an unambiguous binding model, provided that the sequences cover the entire range of the affinity distribution[67]. Interdependencies between neighbouring nucleotides in binding sites of DNA-interacting proteins have also been described by hidden Markov models[83,84], and these models are applicable to RBPs as well.

Although accounting for interdependencies between two nucleotides often improves the binding model, further improvements can be accomplished by considering higher-order couplings between more than two nucleotides[79]. Various approaches to accomplish this goal have been developed for DNA-binding proteins, including higher-order hidden Markov models[85], neural network analyses[86], decision tree-guided approaches[87], higher-order Bayesian networks[88] and approaches that incorporate protein structural information[89]. Neural network analysis has been applied to RNA–protein interactions measured *in vitro* with the high-throughput sequencing kinetics (HiTS-Kin) approach[67]. In this case, neural network analysis did not markedly improve the fit of the model to the data for the C5 protein, suggesting that pairwise couplings between mostly adjacent nucleotides are the major contributor to protein binding in the tested case[67]. Hidden Markov models were also used to improve rules predicting RNA-binding patterns for the splicing factor polypyrimidine tract-binding protein 1 (PTBP1) *in vivo*[90].

Free energy landscapes

Substrate affinities for proteins that interact with RNA usually refer to equilibrium binding constants, which express the energetic difference between ground state (protein and RNA are unbound) and product state (protein and RNA are bound) in a one-step binding reaction (FIG. 4a). However, differences in equilibrium binding affinity between substrate variants can arise from alterations in ground, transition or product state energies, or from combinations thereof (FIG. 4a). These alterations can be assessed only through measurements of association rate

Sequence space
The set of all possible nucleotide sequence combinations of *N* length defined by 4^N.

Hidden Markov models
Probabilistic models used in molecular biology to describe the binding specificity of a protein or ligand derived from a set of bound sequences assuming a Markov process with unobserved (hidden) states.

Neural network analyses
Any of a family of pattern-recognition algorithms that use approximate nonlinear functions using sets of adaptive weights.

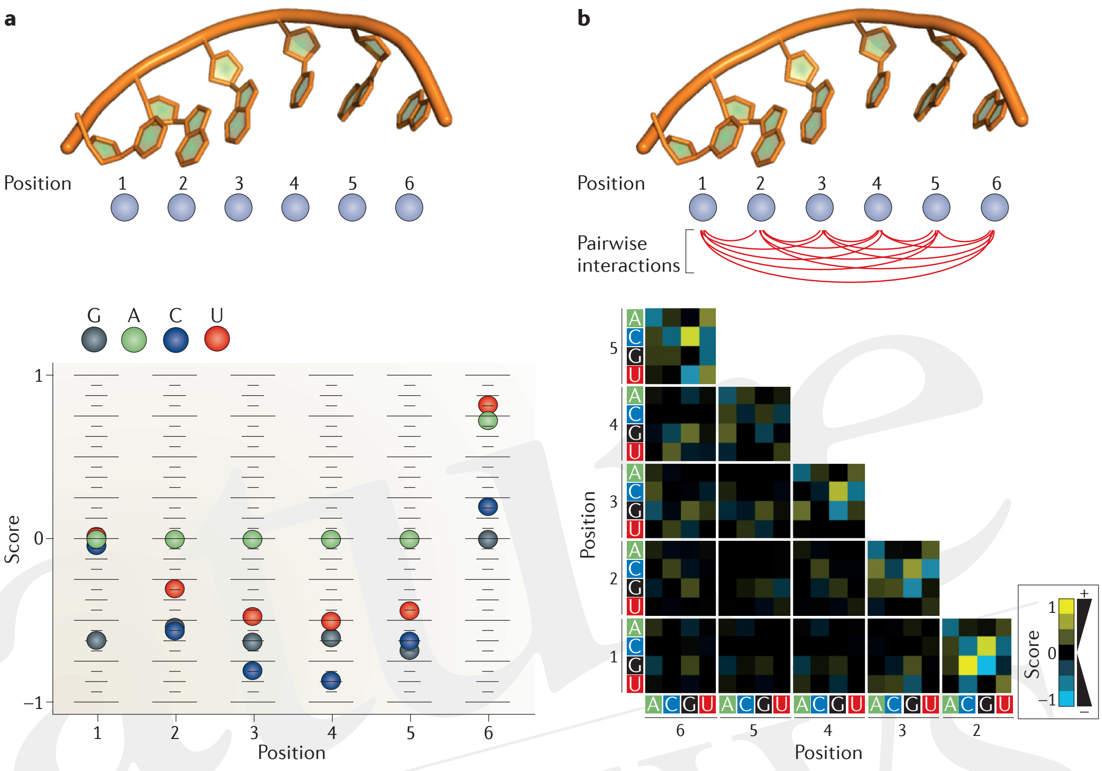


Figure 3 | **RBP binding models. a** | Position weight matrix (PWM). The structure denotes a hypothetical RNA-binding protein (RBP) RNA-binding site comprised of six nucleotides. The plot (coloured circles) depicts the score (linear coefficient) for each base at each position. The score is calculated from affinity distributions such as the one shown in FIG. 2b. The score for each base corresponds to the contribution of the indicated nucleotide at each position to the overall binding free energy (a higher the score indicates tighter binding). **b** | A binding model considering interactions between two bases (pairwise interaction matrix (PIM) or dinucleotide weight matrix (DWM)). The structure denotes a hypothetical RBP RNA-binding site with six nucleotides; lines show the possible pairwise (energetic) couplings between two nucleotides. Coloured fields correspond to the score for each of the 16 pairwise nucleotide permutations at each two positions. Scores are calculated from affinity distributions such as that shown in FIG. 2b. A yellow field (denoting a high score) indicates that a given dinucleotide combination promotes binding (that is, increases the overall PWM score). A blue field (denoting a low score) indicates inhibition of binding by a given dinucleotide combination. A black field indicates no significant contribution either way. Figure, part **b**, **[Au: which part of panel b?]** from REF. 67, Nature Publishing Group.

constants and dissociation rate constants for the substrate variants. To date, few studies have reported rate constants for many substrates[67,91], and to our knowledge only one[91] reported both association and dissociation rate constants — for the binding of the bacteriophage MS2 coat protein to a large set of variants of the cognate RNA hairpin[91]. In this study, differences in substrate preferences were mainly due to variations in substrate association rate constants, with comparably small contributions by dissociation rate constants. These observations suggest that, for the MS2–substrate system, differences in RNA binding are mainly due to variations in ground state energies, most probably reflecting the significance of RNA structure for substrate binding by MS2 (REF. 91).

Although comparable data for other RNA–protein interactions have not yet been reported, RNA structure is likely to affect even those proteins that bind to presumably unstructured sites (FIG. 4b). A subset of a randomized substrate population will form at least transient secondary structures[92,93], and the unfolding of even relatively unstable structures will affect the substrate's ground state and thereby the overall affinity distribution. Although it is known that sequestration of protein-binding sites by RNA structures affects protein binding *in vitro*[94] and *in vivo*[95], it has not been explored to which degree more subtle changes in substrate ground state free energies contribute to binding specificity. The potential effect of even transient RNA structures on substrate specificity emphasizes the importance of the RNA sequences surrounding protein-binding sites.

Specific versus nonspecific interactions

As noted above, affinity distributions of RBPs measured *in vitro* and RNA-binding patterns of numerous RBPs measured in cells have raised questions regarding the widely used classification of RBPs as either specific or nonspecific. A more nuanced, quantitative view on specificity and nonspecificity is emerging, based on recent technical advances in measuring RNA–protein binding *in vitro* (BOX 1).

Specific versus nonspecific RBPs. The majority of studies on RNA–protein interactions have focused on specific RBPs, even though nonspecific proteins are numerous

Decision tree-guided approaches
Any of a family of statistical classifying methods for sorting data according to attributes (for example, nodes) that form a hierarchy encoded as a tree.

Higher-order Bayesian networks
A type of statistical model that represents a set of random variables and their conditional dependencies expressing the quantitative strength of the connections between variables.

Equilibrium binding affinity
A quantitative description of the energetic strength of the interaction between two molecules, typically expressed by the dissociation constant, K_d.

Association rate constants
The second-order rate constants, with unit $Mol^{-1}s^{-1}$, that describe the binding (association) of two molecules in solution.

Dissociation rate constants
The first-order rate constants, with unit s^{-1}, that describe the dissociation of a complex between two molecules in solution.

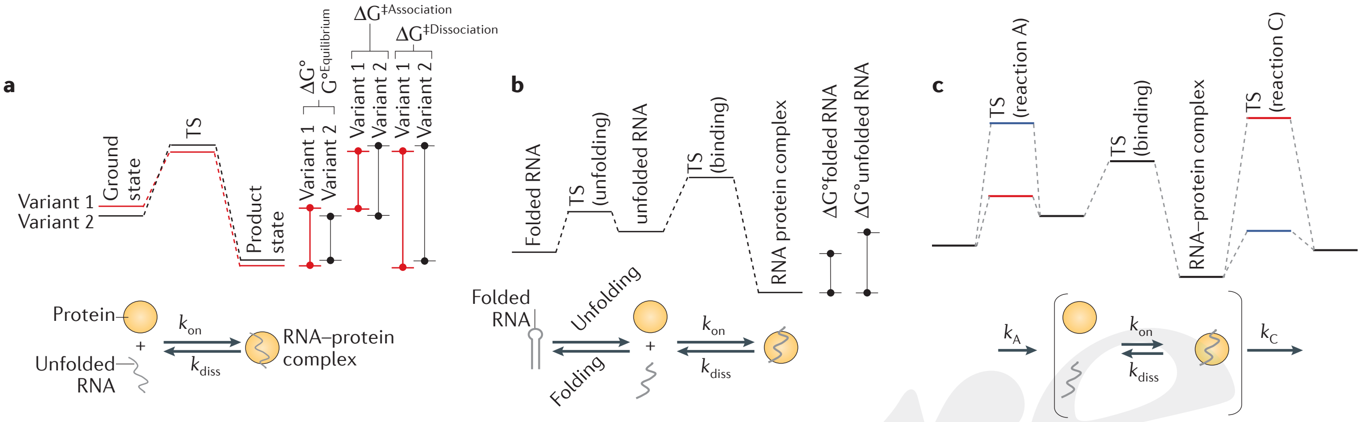


Figure 4 | **Free energy landscapes of RNA–protein interactions. a** | The free energy landscape for a single step of a reversible binding reaction between a protein and two RNA variants is shown. The different binding affinities of sequence variants are reflected in different equilibrium free energy changes ($\Delta G^{\circ \text{Equilibrium}}$). Different equilibrium binding affinities can result from differences in ground-state, transition-state (TS) or product-state free energies, which correspond to changes in activation free energy for association ($\Delta G^{\ddagger \text{Association}}$), dissociation ($\Delta G^{\ddagger \text{Dissociation}}$) or both. **[Au: for simplicity should the unfolded RNA be the same way up in parts a, b and c (and in complex)?] b** | Free energy landscape of a binding reaction between a protein and a structured RNA. Only one RNA variant is shown for simplicity. In this example, the hairpin RNA binds in its unfolded state; however, the depicted process also applies to structural transitions that are more complex than hairpin unfolding. The unfolding step affects the equilibrium free energy change (ΔG°), and thus the binding affinity. **c** | The kinetic context of an RNA–protein binding reaction. Only one RNA variant is shown for simplicity. The scheme shows ground state and TS for three consecutive reactions. Protein–RNA binding is the middle step. Intrinsic specificity can translate into biological specificity only for the scenario indicated by the red coloured transitions state energies for reaction A (rate constant k_A) and C (rate constant k_C). **[Au: bottom of part c, should there be a plus sign between the protein and unfolded RNA?]** All other combinations of transition state energies reduce the intrinsic specificity that is provided by the isolated binding reaction. k_{diss}, rate of dissociation; k_{on}, rate of association. **[Au: are these definitions correct? but why is it k_{on} but not k_{off}?] [Au: please ensure that single-letter variables are in italics, otherwise in roman]**

and perform many important biological functions. A recent study determined the affinity distribution for a nonspecific *E. coli* protein, the C5 subunit of RNase P[67]. C5 binds, in conjunction with the catalytic RNA unit of RNase P, to all cellular tRNA precursors at a degenerate binding site[67,96,97]. Despite the lack of a consensus binding motif in its physiological substrates, the affinity distribution for C5 was extremely similar to those seen for highly specific proteins[67]. As in the case of specific RBPs, **[Au: ok?]** the high-affinity region of the affinity distribution of C5 revealed a consensus sequence, indicating that C5 is inherently specific towards certain sequences. In contrast to specific RBPs, the physiological substrates of C5 do not fall in the high-affinity region of the distribution but in the median region, which does not have a consensus, and in which large differences in sequence have only small effects on affinity (FIG. 2b). Defined binding models could be readily derived from the C5 affinity distribution[67], suggesting that the differences between specific and nonspecific RBPs are not inherent features of the proteins. Rather, specific and nonspecific binding modes represent different parts of the affinity distribution (FIG. 2b).

It is perhaps not surprising that even nonspecific RBPs have intrinsic specificity, given that protein and RNA surfaces at the binding interface have irregular features. Some RNA species are thus more likely to form favourable interactions with a protein than others. This notion probably applies to the vast majority of RNA–protein interactions. A possible exception is proteins that bind exclusively to the backbone of an RNA A-form helix, because the backbone of an A-form RNA helix is thought to be structurally similar for diverse sequences[98,99]. Yet, helices dynamically open and close locally in a sequence-dependent manner[98,100], and they may be distorted on protein binding, as seen for double-stranded RBD–RNA complexes[101].

The kinetic context of RNA–protein interactions. The RNA-binding study of C5 also highlighted the significance of the context in which a binding reaction occurs. One critical and perhaps obvious aspect for this context is the availability of substrates in the transcriptome. For C5, most of the tightest binding sequence variants are not present in the expressed substrates. RNA structure also plays an important part in the context of a binding reaction, as discussed above (FIG. 4b).

A third, potentially highly significant, factor is the kinetic context — the kinetics of the reactions that precede and follow the binding step (FIG. 4c). This kinetic context is dictated by the concentration of the protein, by the concentration of the RNA, by the rate constants for substrate binding and dissociation, and by how these rate constants compare with those of the steps that precede and follow the binding step. The intrinsic specificity of the protein for any given RNA substrate is given by the ratio of rate constants for substrate binding and dissociation (FIG. 4a). However, intrinsic specificity translates

A-form helix
A right-handed double helix formed by nucleic acids, primarily RNA, with characteristic numbers of base pairs per turn, a deep major groove and a shallow minor groove.

Maximal specificity
An optimal mode of molecular recognition resulting in the largest difference in binding free energy between cognate and non-cognate ligands.

into near-maximal specificity only if [Au: OK? (moved 'only')] the step preceding the binding is fast compared with the binding step and the step following the binding step is slow compared with both binding and dissociation. All other scenarios neutralize intrinsic specificity to various degrees (FIG. 4c). Therefore, an inherently highly specific protein can readily operate under an entirely nonspecific regime, or a protein can be toggled between nonspecific and specific modes, solely through changes in the rate constants of steps unrelated to binding or through changes in RNA or protein concentrations. The kinetic context is dictated by proteins that may or may not directly interact with the RBP in question. Although we are not aware of studies that have explicitly tested the kinetic context for RBPs, this context is likely to contribute to the wide range of binding sites seen during the transcriptome-wide mapping of RNA-binding sites for many proteins.

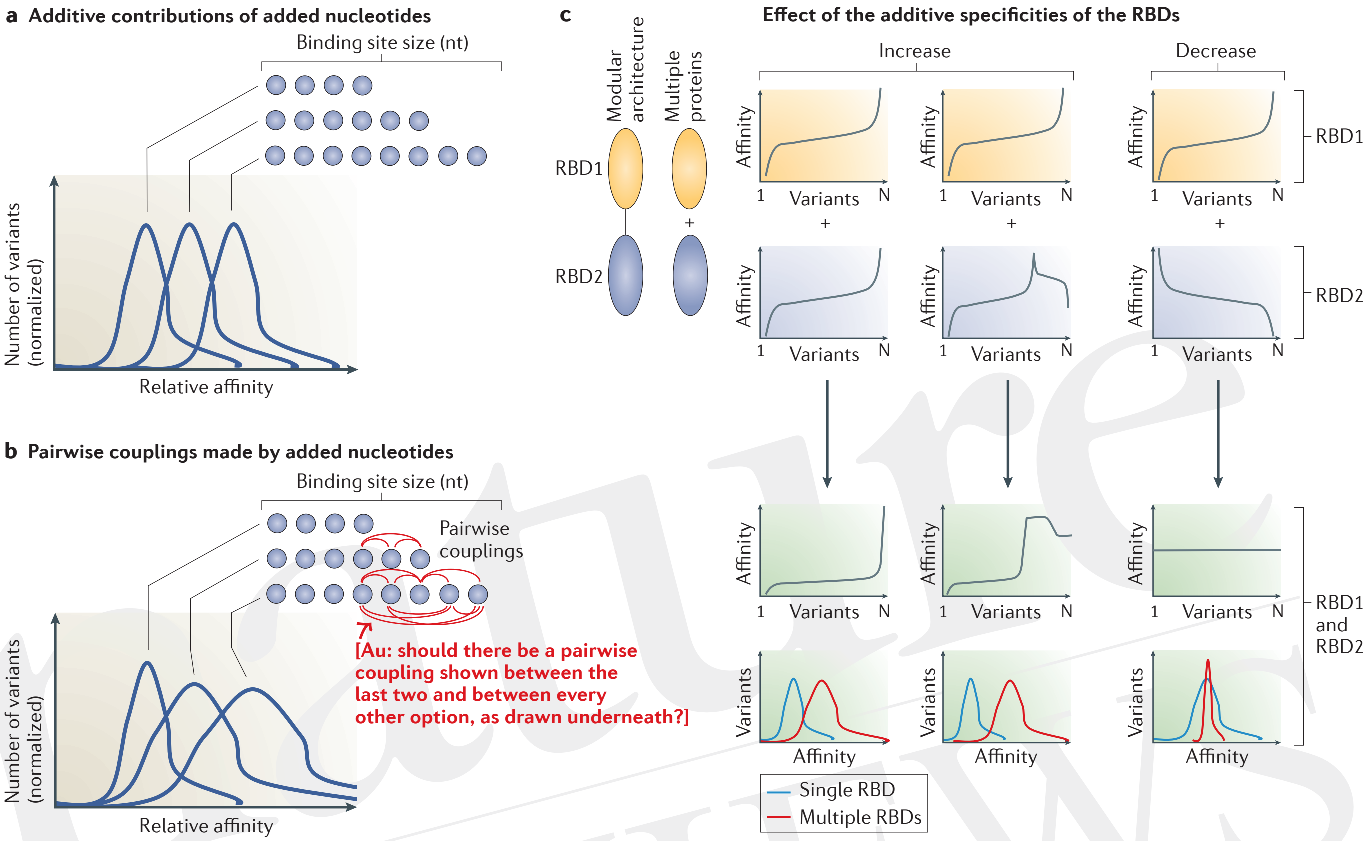
Given the significance and the ubiquity of the kinetic context, we believe it is useful to distinguish between the biological specificity and the intrinsic specificity of a protein towards substrate variants. The biological specificity is the preference of a protein for sequence variants *in vivo*, as revealed by techniques like CLIP. The intrinsic specificity, reflected in the affinity distribution, is the preference of a protein for sequence variants when only the binding reaction is examined *in vitro*. Intrinsic specificity is equivalent to the classical definition of specificity for enzymatic reactions: $\mathrm{F}^{\mathrm{Specificity}} = (k_{\mathrm{cat}}/K_{\mathrm{m}})^{\mathrm{Substrate1}}/(k_{\mathrm{cat}}/K_{\mathrm{m}})^{\mathrm{Substrate2}}$, where $\mathrm{F}^{\mathrm{Specificity}}$ **[Au: please check all single-letter variables are in italics, otherwise in roman]** refers to the factor by which the enzyme prefers substrate 1 over substrate 2, k_{cat} to the turnover number and K_{m} to the Michaelis constant[102]. An obvious challenge is to quantitatively define the connection between intrinsic specificity and biological specificity for RBPs. The first attempts in this direction have integrated *in vivo* and *in vitro* specificity measurements, which aided the identification of cellular regulatory protein-binding sites from CLIP data[103,104]. In addition, matching of preferred *in vitro* binding motifs for RBPs with CLIP data is a mark of progress towards integrating *in vitro* and *in vivo* data[69].

Modulating intrinsic specificity of RBPs. In many RBPs, the intrinsic specificity of individual RBDs seems to be insufficient to accomplish the biologically required specificity of the RBP[3,22,23,105]. Mechanisms have therefore evolved that enhance the intrinsic specificity of RBPs to better discriminate between cognate and non-cognate binding sites. Conversely, proteins that need to interact with diverse RNA sites in an indiscriminate fashion must use mechanisms to compensate for the unavoidable intrinsic specificities of their RBDs.

Intrinsic specificity of an RBD can be enhanced by increasing the size of the RNA-binding site, to recognize more nucleotides. A larger RBD binding site is expected to bind the target variant tighter than a small site would, but a larger binding site can also bind non-target variants tighter, and the discrimination between target and non-target sites will not necessarily increase[106] (FIG. 5a). However, discrimination between target and non-target sites depends on whether additional nucleotides contribute independently to overall affinity. Independent contributions of nucleotides result in only modest increases of discrimination with increasing binding site size (FIG. 5a). By contrast, energetic coupling between nucleotides can result in large increases in selectivity (FIG. 5b). An increase in binding site size that is thought to lead to enhanced specificity is seen for RRMs[24,107], in which changes in binding site size are accomplished through the use of alternative RNA-binding modes by the core RRM fold[24].

A widely observed mechanism that affects the intrinsic specificity of RBPs is the inclusion of multiple RBDs in a single protein (FIG. 4c). As noted, a large fraction of the proteins that interact with RNA contain multiple RBDs[1,22,23]. This modular architecture results in proteins with affinity distributions that combine the affinity distributions of their individual RBDs (FIG. 5c). These combinations can enhance binding specificity if the affinity distributions of the different RBDs favour similar sequence variants or if they favour different sequence variants in a non-compensatory fashion. Modular protein architectures can also enable proteins to recognize non-contiguous sequences[23] and thus intervening RNA sequences can become important contributors to specificity[108–110]. In addition, protein regions that link different RBDs can modulate the contribution of each RBD to the protein's overall RNA affinity and even promote cooperativity between RBDs[111]. Multiple RBDs with different inherent affinity distributions can also compensate for each other in a given protein and lead to uniform binding of an RBP to a wide range of diverse substrates (FIG. 5c). This is seen for EF-Tu, which binds to all charged tRNAs with similar affinity[112]. EF-Tu contains a binding site for the tRNA and one for the cognate amino acid[113]. The binding energies for tRNAs and amino acids at each site differ, but they compensate for their respective differences, thereby resulting in nearly uniform binding for all correctly charged tRNAs[112].

Multiple RBDs do not necessarily need to be part of the same protein; they can be encoded by different, yet interacting, proteins (FIG. 5c). This is widely seen in RNPs[114–116], including large RNA–protein assemblies such as the spliceosome[117–121] or the eukaryotic translation initiation machinery[122–124]. Moreover, multiple modular RBDs can assemble on the same RNA substrate, further increasing selectivity[23]. An advantage of combining different proteins for binding to a given RNA site is the possibility of regulating their interactions through variations in the concentration and post-translational modifications of the individual proteins[46]. Different proteins can bind cooperatively or anti-cooperatively, **[Au: is there another term that means anti-cooperatively (or is it possible to put a few words in brackets for total clarity for any non-specialist readers)?]** and these modes of protein–protein interactions can further amplify intrinsic specificity or provide compensation for the intrinsic specificity of individual RBDs. Multiple identical RBDs can also assemble in homo-oligomers of RNA-interacting proteins[70] and can thereby enhance selectivity for longer target sites[110,125].



[Au: should there be a pairwise coupling shown between the last two and between every other option, as drawn underneath?]

Figure 5 | **Mechanisms to increase or decrease the intrinsic specificity of RBPs.** **a** | Increases in the size of the RNA-binding site of an RNA-binding protein (RBP), with additive contributions to the binding energy being made by the additionally bound nucleotides. Extra nucleotides (nt) in the binding site shift the affinity distribution towards higher affinities but do not significantly broaden the distribution and thus do not lead to large increases in the inherent specificity of the RBP. **b** | Increases in the size of the RNA-binding site of an RBP, with contributions of pairwise energetic coupling to the binding energy being made by the additionally bound nucleotides. Extra nucleotides in the binding site shift the affinity distribution towards higher affinities and broaden the distribution, thus increasing the inherent specificity of the RBP. **c** | Increases or decreases in the intrinsic specificity of an RBP through the use of multiple RNA-binding domains (RBDs). Multiple RBDs (RBD1 and RBD2) can be part of the same protein or separate proteins. The panels in the first and second rows show ranked affinity distributions (according to the same sequences for both RBDs) for each RBD. The panels in the third row show the ranked affinity distribution of the combination of both RBDs; the corresponding affinity distribution is shown in the fourth row and colour coded as indicated. Inherent protein specificity can be increased by the additive specificities of the additional RBD or decreased by compensatory specificities. Intrinsic specificities for individual RBDs can vary. Binding preferences of individual RBDs do not need to be strictly additive, but can be synergistic, either through interactions between the RBDs or through cooperative binding of multiple proteins. **[Au: should N be in italics (x9) and should '1' and 'N' be present in the bottom row graphs in part c?]**

Future perspective

High-throughput sequencing methods have opened new possibilities to measure and understand specificity and nonspecificity in RNA–protein interactions, both *in vitro* and *in vivo*. It is now possible to directly determine affinities for all, or at least for a large number of, possible binding site sequence variants for most RBDs *in vitro*, and to derive comprehensive binding models. These new tools have already provided important insight into principles that underlie binding specificity and nonspecificity. Although not all of these techniques can yet be readily applied in every laboratory, it is likely that binding models will emerge for many more RBDs and RBPs over the next years.

Future challenges include the integration of quantitative binding models with structural data. To date, most structures of RBPs have been solved with only a single RNA substrate, usually representing a high-affinity target; only in very few cases do structures exist for low-affinity targets[70] or alternative substrates[126]. Yet these data, combined with comprehensive binding models, will be the most instructive for linking structure and intrinsic specificity[126]. It will be equally important to determine binding models for proteins with mutated RBDs and, if possible, to integrate structural and binding models for such mutant proteins. Comparisons of binding models for wild-type and mutated proteins might also be an inroad to understanding the virtually unexplored effects of transient RNA structure and kinetic context on RNA–protein interactions in the cell.

Ultimately, we want to devise models that accurately describe and possibly predict the RNA-binding patterns for proteins *in vivo*. This will require quantitative models that integrate RNA binding *in vitro* and *in vivo* with other aspects of RNA biology. An important step towards such models has been recently made using techniques that assess RNA secondary structures *in vivo*[127–129]. A critical but unconquered barrier for the development of quantitative models of RNA–protein interactions is the lack of methods to determine the kinetics of RNA–protein binding *in vivo* for individual RNA sites.

1. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
2. Moore, M. J. From birth to death: the complex lives of eukaryotic mRNAs. *Science* **309**, 1514–1518 (2005).
3. Mitchell, S. F. & Parker, R. Principles and properties of eukaryotic mRNPs. *Mol. Cell* **54**, 547–558 (2014).
4. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).
5. Gerstberger, S., Hafner, M., Ascano, M. & Tuschl, T. in *Systems Biology of RNA Binding Proteins* (ed. Yeo, G. W.) 1–55 (Springer, 2014).
6. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet.* **29**, 318–327 (2013).
7. Scheper, G. C., van der Knaap, M. S. & Proud, C. G. Translation matters: protein synthesis defects in inherited disease. *Nat. Rev. Genet.* **8**, 711–723 (2007).
8. McGettigan, P. A. Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* **17**, 4–11 (2013).
9. Parker, R. & Song, H. The enzymes and control of eukaryotic mRNA turnover. *Nat. Struct. Mol. Biol.* **11**, 121–127 (2004).
10. Aitken, C. E. & Lorsch, J. R. A mechanistic overview of translation initiation in eukaryotes. *Nat. Struct. Mol. Biol.* **19**, 568–576 (2012).
11. Janga, S. C. & Mittal, N. Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins. *Adv. Exp. Med. Biol.* **722**, 103–107 (2011).
12. Gerstberger, S., Hafner, M. & Tuschl, T. Learning the language of post-transcriptional gene regulation. *Genome Biol.* **14**, 130 (2013).
13. Campbell, Z. T., Valley, C. T. & Wickens, M. A protein-RNA specificity code enables targeted activation of an endogenous human transcript. *Nat. Struct. Mol. Biol.* **21**, 732–738 (2014).
14. Wang, Y., Wang, Z. & Tanaka Hall, T. M. Engineered proteins with Pumilio/*fem-3* mRNA binding factor scaffold to manipulate RNA metabolism. *FEBS J.* **280**, 3755–3767 (2013).
15. Chen, Y. & Varani, G. Engineering RNA-binding proteins for biology. *FEBS J.* **280**, 3734–3754 (2013).
16. Choudhury, R., Tsai, Y. S., Dominguez, D., Wang, Y. & Wang, Z. Engineering RNA endonucleases with customized sequence specificities. *Nat. Commun.* **3**, 1147 (2012).
17. Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* **30**, 1427–1464 (2002).
18. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
19. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
20. Thomson, T. & Lin, H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu. Rev. Cell Dev. Biol.* **25**, 355–376 (2009).
21. Brook, M., Smith, J. W. & Gray, N. K. The DAZL & PABP families: RNA-binding proteins with interrelated roles in translational control in oocytes. *Reproduction* **137**, 595–617 (2009).
22. Singh, R. & Valcárcel, J. Building specificity with nonspecific RNA-binding proteins. *Nat. Struct. Mol. Biol.* **12**, 645–653 (2005).
23. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).
24. Cléry, A., Blatter, M. & Allain, F. H. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18**, 290–298 (2008).
25. Fairman-Williams, M. E., Guenther, U.-P. & Jankowsky, E. SF1 and SF2 helicases: family matters. *Curr. Opin. Struct. Biol.* **20**, 313–324 (2010).
26. Ghosh, A. & Lima, C. D. Enzymology of RNA cap synthesis. *Wiley Interdiscip. Rev. RNA* **1**, 152–172 (2010).
27. Firczuk, H. *et al.* An *in vivo* control map for the eukaryotic mRNA translation machinery. *Mol. Syst. Biol.* **9**, 635 (2013).
28. Hentze, M. W. & Preiss, T. The REM phase of gene regulation. *Trends Biochem. Sci.* **35**, 423–426 (2010).
29. Mitchell, S. F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* **20**, 127–133 (2013).
30. Ng, S. Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA *RMST* interacts with SOX2 to regulate neurogenesis. *Mol. Cell* **51**, 349–359 (2013).
31. Di Ruscio, A. *et al.* DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503**, 371–376 (2013).
32. Hudson, W. H. & Ortlund, E. A. The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.* **15**, 749–760 (2014).
33. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, e1000598 (2009).
34. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
35. Liu, N. *et al.* Probing N^6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* **19**, 1848–1856 (2013).
36. Dominissini, D. *et al.* Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature* **485**, 201–206 (2012).
37. Pan, T. N^6-methyl-adenosine modification in messenger and long non-coding RNA. *Trends Biochem. Sci.* **38**, 204–209 (2013).
38. Carlile, T. M. *et al.* Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515**, 143–146 (2014).
39. Weick, E. M. & Miska, E. A. piRNAs: from biogenesis to function. *Development.* **141**, 3458–3471 (2014).
40. Thompson, D. M., Lu, C., Green, P. J. & Parker, R. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA* **14**, 2095–2103 (2008).
41. Ivanov, P., Emara, M. M., Villen, J., Gygi, S. P. & Anderson, P. Angiogenin-induced tRNA fragments inhibit translation initiation. *Mol. Cell* **43**, 613–623 (2011).
42. Saikia, M. *et al.* Angiogenin-cleaved tRNA halves interact with cytochrome *c*, protecting cells from apoptosis during osmotic stress. *Mol. Cell. Biol.* **34**, 2450–2463 (2014).
43. Milek, M., Wyler, E. & Landthaler, M. Transcriptome-wide analysis of protein–RNA interactions using high-throughput sequencing. *Semin. Cell Dev. Biol.* **23**, 206–212 (2012).
44. Müller-McNicoll, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA–protein complexes. *Nat. Rev. Genet.* **14**, 275–287 (2013).
45. Agirrezabala, X. & Frank, J. Elongation in translation as a dynamic interaction among the ribosome, tRNA, and elongation factors EF-G and EF-Tu. *Q. Rev. Biophys.* **42**, 159–200 (2009).
46. Jangi, M. & Sharp, P. A. Building robust transcriptomes with master splicing factors. *Cell* **159**, 487–498 (2014).
47. Zhou, Z. & Fu, X. D. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* **122**, 191–207 (2013).
48. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
49. Tay, Y., Rinn, J. & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344–352 (2014).
50. König, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein–RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* **13**, 77–83 (2012).
51. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol.* **15**, 203 (2014).
52. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **14**, 1212–1215 (2003).
53. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
54. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
55. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
56. Singh, G., Ricci, E. P. & Moore, M. J. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. *Methods* **65**, 320–332 (2014).
57. Singh, G. *et al.* The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell* **151**, 750–764 (2012).
58. Campbell, Z. T. *et al.* Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Rep.* **1**, 570–581 (2012).
This study combines *in vitro* selection and high-throughput sequencing to measure and characterize protein binding to a large number of RNA variants.
59. Ozer, A., Pagano, J. M. & Lis, J. T. New technologies provide quantum changes in the scale, speed, and success of SELEX methods and aptamer characterization. *Mol. Ther. Nucleic Acids* **5**, e183 (2014).
60. Lorenz, C. *et al.* Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts. *Nucleic Acids Res.* **38**, 3794–3808 (2010).
61. Stormo, G. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.* **11**, 751–760 (2010).
62. Sanford, J. R. *et al.* Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.* **19**, 381–394 (2009).
63. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
64. Rowe, W. *et al.* Analysis of a complete DNA–protein affinity landscape. *J. R. Soc. Interface* **7**, 397–408 (2009).
65. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
66. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
67. Guenther, U. P. *et al.* Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature* **502**, 385–388 (2013).
This paper introduces the HiTS-Kin method and measures affinity distributions for an RBP lacking canonical specificity.
68. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quant. Biol.* **1**, 115–130 (2013).
69. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
This study describes a large-scale effort to define RNA-binding motifs for many RBPs with the RNAcompete technique.
70. Duss, O., Michel, E., Diarra dit Konté, N., Schubert, M. & Allain, F. H. Molecular basis for the wide range of affinity found in Csr/Rsm protein–RNA recognition. *Nucleic Acids Res.* **42**, 5332–5346 (2014).
This study investigates differences in the structures of high- and low-affinity targets of an RBP.
71. Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15**, R17 (2014).
72. Pancaldi, V. & Bähler, J. *In silico* characterization and prediction of global protein–mRNA interactions in yeast. *Nucleic Acids Res.* **39**, 5826–5836 (2011).

73. Livi, C. M. & Blanzieri, E. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. *BMC Bioinformatics* **15**, 123 (2014).
74. Puton, T., Kozlowski, L., Tuszynska, I., Rother, K. & Bujnicki, J. M. Computational methods for prediction of protein–RNA interactions. *J. Struct. Biol.* **179**, 261–268 (2012).
75. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373 (2006).
76. Tran, N. T. & Huang, C. H. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol. Direct* **9**, 4 (2014).
77. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
78. Reyes-Herrera, P. H. & Ficarra, E. Computational methods for CLIP-seq data processing. *Bioinform. Biol. Insights.* **8**, 199–207 (2014).
79. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014).
80. Zhao, Y. & Stormo, G. Jury remains out on simple model of transcription factors. *Nat. Biotechnol.* **6**, 480–483 (2011).
81. Zhao, Y., Ruan, S., Pandey, M. & Stormo, G. D. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **191**, 781–790 (2012).
82. Siddharthan, R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS ONE* **5**, e9722 (2010).
83. Bulyk, M. L., Johnson,P. L. & Church, G. M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30**, 1255–1261 (2002).
84. Mathelier, A. & Wasserman, W. W. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* **9**, e1003214 (2013).
85. Grau, J., Posch, S., Grosse, I. & Keilwagen, J. A general approach for discriminative *de novo* motif discovery from high-throughput data. *Nucleic Acids Res.* **41**, e197 (2013).
86. Zhou, Q. & Liu, J. S. Extracting sequence features to predict protein–DNA interactions: a comparative study. *Nucleic Acids Res.* **36**, 4137–4148 (2008).
87. Hooghe, B., Broos, S., van Roy, F. & De Bleser, P. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Res.* **40**, e106 (2012).
88. Ben-Gal, I. *et al.* Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**, 2657–2666 (2005).
89. Liu, L. A. & Bradley, P. Atomistic modeling of protein–DNA interaction specificity: progress and applications. *Curr. Opin. Struct. Biol.* **22**, 397–405 (2012).
90. Han, A. *et al. De novo* prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Comput. Biol.* **10**, e1003442 (2014).
91. Buenrostro, J. D. *et al.* Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
This work introduces RNA array technology and measures binding and dissociation kinetics for large numbers of RNA sequence variants.
92. SantaLucia, J. J. & Turner, D. H. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers* **44**, 309–319 (1997).
93. Forsdyke, D. R. Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. *J. Theor. Biol.* **248**, 745–753 (2007).
94. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Res.* **40**, e54 (2012).
95. Maenner, S., Müller, M., Fröhlich, J., Langer, D. & Becker, P. B. ATP-dependent roX RNA remodeling by the helicase maleless enables specific association of MSL proteins. *Mol. Cell* **51**, 174–184 (2013).
96. Smith, J. K., Hsieh, J. & Fierke, C. A. Importance of RNA-protein interactions in bacterial ribonuclease P structure and catalysis. *Biopolymers* **87**, 329–338 (2007).
97. Rueda, D., Hsieh, J., Day-Storms, J. J., Fierke, C. A. & Walter, N. G. The 5' leader of precursor tRNAAsp bound to the *Bacillus subtilis* RNase P holoenzyme has an extended conformation. *Biochemistry* **44**, 16130–16139 (2005).
98. Snoussi, K. & Leroy, J. L. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry* **40**, 8898–8904 (2001).
99. Faustino, I., Pérez, A. & Orozco, M. Toward a consensus view of duplex RNA flexibility. *Biophys. J.* **99**, 1876–1885 (2010).
100. Bothe, J. R. *et al.* Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. *Nat. Methods* **8**, 919–931 (2011).
101. Masliah, G., Barraud, P. & Allain, F. H. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell. Mol. Life Sci.* **70**, 1875–1895 (2013).
102. Cornish-Bowden, A. Enzyme specificity: its meaning in the general case. *J. Theor. Biol.* **108**, 451–457 (1984).
103. Li, J. *et al.* Identifying mRNA sequence elements for target recognition by human Argonaute proteins. *Genome Res.* **24**, 775–785 (2014).
104. Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell.* **54**, 887–900 (2014).
This paper introduces the RNA Bind-n-Seq method.
105. Zearfoss, N. R. *et al.* A conserved three-nucleotide core motif defines Musashi RNA-binding specificity. *J. Biol. Chem.* **289**, 35530–35541 (2014).
106. Herschlag, D. Implications of ribozyme kinetics for targeting the cleavage of specific RNA molecules *in vivo*: more isn't always better. *Proc. Natl Acad. Sci. USA* **88**, 6921–9625 (1991).
107. Auweter, S. D., Oberstrass, F. C. & Allain, F. H. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.* **34**, 4943–4959 (2006).
108. Lamichhane, R. *et al.* RNA looping by PTB: evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proc. Natl Acad. Sci. USA* **107**, 4105–4110 (2010).
109. Mickleburgh, I. *et al.* The organization of RNA contacts by PTB for regulation of FAS splicing. *Nucleic Acids Res.* **42**, 8605–8620 (2014).
110. Zhang, W. *et al.* Crystal structures and RNA-binding properties of the RNA recognition motifs of heterogeneous nuclear ribonucleoprotein L: insights into its roles in alternative splicing regulation. *J. Biol. Chem.* **288**, 22636–22649 (2013).
111. Romanelli, M. G., Diani, E. & Lievens, P. M. New insights into functional roles of the polypyrimidine tract-binding protein. *Int. J. Mol. Sci.* **14**, 22906–22932 (2013).
112. LaRiviere, F. J., Wolfson, A. D. & Uhlenbeck, O. C. Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. *Science* **294**, 165–168 (2001).
This article introduces the concept of thermodynamic compensation for RBDs, which enables EF-Tu to achieve the nearly uniform affinity for diverse RNAs.
113. Nilsson, J. & Nissen, P. Elongation factors on the ribosome. *Curr. Opin. Struct. Biol.* **15**, 349–354 (2005).
114. Hennig, J. *et al.* Structural basis for the assembly of the Sxl–Unr translation regulatory complex. *Nature* **515**, 287–290 (2014).
115. Wasmuth, E. V., Januszyk, K. & Lima, C. D. Structure of an Rrp6–RNA exosome complex bound to poly(A) RNA. *Nature* **511**, 435–439 (2014).
116. Andersen, C. B. *et al.* Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. *Science* **313**, 1968–1972 (2006).
117. Wahl, M. C., Will, C. L. & Lührmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell.* **136**, 701–718 (2009).
118. Zhou, L. *et al.* Crystal structures of the Lsm complex bound to the 3′ end sequence of U6 small nuclear RNA. *Nature* **506**, 116–120 (2014).
119. Weber, G., Trowitzsch, S., Kastner, B., Lührmann, R. & Wahl, M. C. Functional organization of the Sm core in the crystal structure of human U1 snRNP. *EMBO J.* **29**, 4172–4184 (2010).
120. Kondo, Y., Oubridge, C., van Roon, A. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5″splice site recognition. *eLife* **4**, e04986 (2015).
121. Montemayor, E. J. *et al.* Core structure of the U6 small nuclear ribonucleoprotein at 1.7-Å resolution. *Nat. Struct. Mol. Biol.* **21**, 544 –551 (2014).
122. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731–745 (2009).
123. Erzberger, J. P. *et al.* Molecular architecture of the 40S•eIF1•eIF3 translation initiation complex. *Cell* **158**, 1123–1135 (2014).
124. Marintchev, A. *et al.* Topology and regulation of the human eIF4A/4G/4H helicase complex in translation initiation. *Cell* **136**, 447–460 (2009).
125. Antson, A. A. *et al.* Structure of the *trp* RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **401**, 235–242 (1999).
126. Cieniková, Z., Damberger, F. F., Hall, J., Allain, F. H. & Maris, C. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *J. Am. Chem. Soc.* **136**, 14536–14544 (2014).
This study correlates structures of multiple substrates with specificity information from high-throughput studies of the RNA targets of an RBP *in vivo*.
127. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* **505**, 701–705 (2014).
128. Ding, Y. *et al. In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
129. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
130. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27**, 667–670 (2009).
This paper introduces the RNAcompete technology to define high-affinity motifs for large numbers of RBDs.
131. Tome, J. M. *et al.* Comprehensive analysis of RNA-protein interactions by high-thoughput sequencing–RNA affinity profiling. *Nat. Methods* **11**, 683–688 (2014).
This paper introduces the HiTS-RAP technique.
132. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).
133. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).
134. Motorin, Y. & Helm, M. RNA nucleotide methylation. *Wiley Interdiscip. Rev. RNA* **2**, 611–631 (2011).

Acknowledgements

The authors thank the members of the Jankowsky and the Harris groups for discussions on the subject of this Review. Work in the author's laboratories is supported by the US National Institute of General Medical Sciences.

Competing interests statement

The authors declare no competing interests.

Author biographies

Eckhard Jankowsky is a professor in the Center for RNA Molecular Biology at Case Western Reserve University, Cleveland, Ohio, USA. His laboratory focuses on understanding molecular mechanisms of RNA–protein interactions using biochemical, biophysical and high-throughput sequencing approaches. His group has been instrumental in elucidating biological roles and molecular mechanisms of RNA helicases. Together with Michael Harris, he developed the high-throughput sequencing kinetics (HiTS-Kin) technique to quantitatively analyse specificity for RNA–protein interactions. Eckhard Jankowsky's homepage. http://jankowskylab.org/

Michael E. Harris is Professor of Biochemistry at Case Western Reserve University, Cleveland, Ohio, USA. His research focuses on transition-state analyses of phosphoryl-transfer reactions and RNA molecular recognition. He and his colleagues developed methods to measure kinetic isotope effects on RNA hydrolysis and transphosphorylation reactions. Together with Eckhard Jankowsky, he uses novel methods for high-throughput analysis of RNA specificity and enzymology aimed at achieving a comprehensive understanding of specificity and catalysis by RNA-binding proteins and RNA-processing enzymes. **[Au: would you like to add a homepage?]**

Online summary

- Mammalian cells encode tens of thousands of RNA species and more than 1,000 proteins that interact with them. Many of these proteins can bind to multiple RNAs, and any given RNA can interact with many proteins, giving rise to highly complex networks of cellular RNA–protein interactions.
- New approaches based on high-throughput sequencing technologies have been developed to quantitatively measure the interaction of proteins simultaneously with large numbers of RNAs.
- These approaches have revealed that specificity in RNA–protein interactions represents a continuum from low-affinity to high-affinity RNA substrate variants. This continuum is quantitatively described by affinity distributions and comprehensive binding models.
- Affinity distributions for RNA-binding proteins (RBPs) that are considered specific RNA binders do not differ fundamentally from affinity distributions for nonspecific RBPs, indicating that even the latter have inherent binding specificity. However, physiological targets of specific proteins fall into the high-affinity range of the affinity distribution, whereas physiological targets of nonspecific proteins do not.
- The biological specificity of RBPs is affected by RNA structure, other proteins, RNA and protein concentrations, and the kinetics of reactions that precede or follow the RNA–protein binding steps.
- Mechanisms have evolved to amplify or compensate for inherent specificities of RNA-binding domains. These include changes in the size of the RNA-binding site of proteins, the combination of multiple RNA-binding domains in a single RBP and the coordinated binding of multiple RBPs.

TOC

Subject categories

Biological sciences / Chemical biology / Proteins / RNA-binding proteins [URI /631/92/612/1230