

Structural mechanisms of RNA recognition: sequence-specific and non-specific RNA-binding proteins and the Cas9-RNA-DNA complex

Ting Ban · Jian-Kang Zhu · Karsten Melcher ·
H. Eric Xu

Received: 26 June 2014/Revised: 28 October 2014/Accepted: 10 November 2014/Published online: 29 November 2014
© Springer Basel 2014

Abstract RNA-binding proteins play crucial roles in RNA processing and function as regulators of gene expression. Recent studies have defined the structural basis for RNA recognition by diverse RNA-binding motifs. While many RNA-binding proteins recognize RNA sequence non-specifically by associating with 5' or 3' RNA ends, sequence-specific recognition by RNA-binding proteins is typically achieved by combining multiple modular domains to form complex binding surfaces. In this review, we present examples of structures from different classes of RNA-binding proteins, identify the mechanisms utilized by them to target specific RNAs, and describe structural principles of how protein–protein interactions affect RNA recognition

specificity. We also highlight the structural mechanism of sequence-dependent and -independent interactions in the Cas9-RNA-DNA complex.

Keywords RNA recognition · RNA-binding protein · Tandem repeat module · Sequence-specific recognition · Protein–protein interaction

Abbreviations

| | |
|----------|---|
| PPR | Pentatricopeptide repeat |
| Pre-mRNA | Precursor messenger RNA |
| RBDs | RNA-binding domains |
| TPR | Tetratricopeptide repeat |
| PUF | Pumilio-FBF |
| TZF | TTP-like zinc finger |
| AGO | Argonaute-like |
| FBF | fem-3 binding factor |
| ssRNA | Single-stranded RNA |
| TTP | Tristetraprolin |
| PIWI | P-element-induced wimpy testes |
| PAZ | PIWI, Argonaute, and Zwillie |
| MDA5 | Melanoma differentiation-associated gene 5 |
| RIG-I | Retinoic acid-inducible gene I |
| PPP-RNA | 5' triphosphate RNA |
| RRM | RNA recognition motif |
| U2AF | U2 auxiliary factor |
| hnRNP | Heterogeneous nuclear ribonucleoprotein |
| CRISPR | Clustered regularly interspaced short palindromic repeats |

T. Ban · H. E. Xu
VARI-SIMM Center, Center for Structure and Function of Drug Targets, Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

T. Ban · J.-K. Zhu
Shanghai Center for Plant Stress Biology and Shanghai Institute of Plant Physiology and Ecology, Shanghai Institute of Biological Sciences, Chinese Academy of Sciences, 300 Fenglin Road, Shanghai 200032, China

J.-K. Zhu
Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47906, USA

K. Melcher · H. E. Xu
Laboratory of Structural Sciences and Laboratory of Structural Biology and Biochemistry, Van Andel Research Institute, 333 Bostwick Ave., N.E, Grand Rapids, MI 49503, USA

Present Address:

T. Ban (✉)
Roche Innovation Center Shanghai, 720 Cailun Road, Shanghai 201203, People's Republic of China
e-mail: ting.ban@roche.com

Introduction

RNA processing plays critical roles in regulating gene expression, including pre-mRNA splicing, RNA-

interference, and mRNA stabilization. RNA-binding proteins (RBPs) are central to this form of regulation. They function in nearly all pathways that are associated with RNA processing and promote the activity of functional and structural RNA molecules [13, 30].

Most RNA-binding proteins contain multiple RNA-binding domains, many of which are structurally and functionally modular. Normally, a single modular RNA-binding domain (RBD) does not have sufficient binding capacity to interact with RNA in a sequence-specific manner because the recognition sequences are often too short. Instead, multiple RNA-binding domains are tethered together to create a much larger binding interface that recognizes a longer sequence to increase specificity and enhance the affinity for target RNAs [19, 24]. To understand the function of RNA-binding proteins, it is important to know how these domains function together as RNA recognition units.

Structural biology has revealed the molecular basis for RNA recognition by individual domains [19]. Here, we focus on how RNA-binding proteins recognize their target RNAs, summarize the RNA recognition mechanisms by different types of RBDs and their modular arrangements, and discuss the importance of dimer formation and protein–protein interactions in precise RNA recognition. We also review the recent structures of Cas9/RNA complex, the essential component of the new genome editing technology [4, 17, 28].

Non-sequence specific recognition

A series of structures of RBPs that recognize selective target RNAs in a non-sequence-specific manner have been reported. Most of these proteins recognize target RNA via binding of marker groups at the 5' or 3' ends of RNA fragments [24]. For instance, PIWI (P-element induced wimpy testes) utilizes a highly conserved binding pocket to recognize the defining 5' phosphate group in the siRNA guide strand [26]; PAZ (PIWI, Argonaute, and Zwiille) recognizes single-stranded 3' overhangs of siRNA through stacking interactions and hydrogen bonding [33]. In this section, we review two recent complex structures of RNA-bound RBPs that have revealed the molecular basis of how this non-sequence-specific recognition occurs.

The first example is the innate immune effector IFIT5, a tetratricopeptide repeat protein that selectively binds viral RNA in a sequence-independent manner by recognizing their characteristic free 5' triphosphate ends [1]. IFIT5 contains 24 α -helices, which are connected by short linkers and which surround a highly positively charged deep pocket that is well suited for the accommodation of nucleic acids (Fig. 1a). The authors determined the crystal

structure of apo IFIT5 as well as structures of IFIT5 in complex with three different RNA oligonucleotides: IFIT5-oligo-A, IFIT5-oligo-C and IFIT5-oligo-U. All three complex structures share the exact same oligonucleotide binding pocket. The 5'-triphosphate group is buried deeply within the pocket and makes a multitude of hydrogen bond and salt-bridge interactions with residues on helix α 2 (E33, T37 and Q41) and the concave inner pocket surface (K150, Y250 and R253) (Fig. 1b). Because critical interactions are made with the γ -phosphate group, the pocket is unlikely to bind 5'-monophosphorylated or 5'-hydroxylated RNA with considerable affinity. Thus, the structure of the IFIT5 TPR domain has evolved to specifically interact with 5' triphosphate-containing RNAs, and through this binding mechanism to distinguish between host and non-self viral nucleic acids.

The second example is the human Argonaute-2 (hAgo2) protein. The crystal structure of hAgo2 bound to microRNA-20a revealed that the microRNA interacts with all four domains of hAgo2 as well as with the linkers connecting these domains (Fig. 1c). This structure provides a framework for how Ago proteins recognize microRNAs and function in inducing RNA-directed silencing [12]. The 5'-end of miR-20a is tethered to hAgo2 through interaction between the terminal monophosphate of “U” and a binding pocket composed of the Mid and PIWI domains. K533, Y529, K570, K566, and Q545 from the Mid domain as well as R812 from the PIWI domain interact with the monophosphate of U1 through salt bridges and hydrogen bonds (Fig. 1d). The following seven nucleotides (5'-AAA-GUGCU-3') form the seed sequence, which is located in a narrow RNA binding groove. Recognition of this sequence is mainly mediated through base-independent interactions with backbone phosphates and sugars, consistent with the high sequence variability of seed sequences and with the ability of Argonaute proteins to accept many small RNA sequences [37]. The structure of Ago proteins is very conserved among eukaryotic and prokaryotic homologs, indicating a conserved mode of RNA binding as the basis for 5'-nucleotide recognition of the guide strand.

Beside single-stranded RNA recognition, non-sequence specific recognition also involves binding of specific RBPs to double-stranded RNAs. MDA5 (melanoma differentiation-associated gene 5) and RIG-I (retinoic acid-inducible gene I) are related cytoplasmic viral RNA receptors in the vertebrate innate immune system. Structures of both proteins bound to double-stranded (ds) RNAs expanded our understanding of RNA recognition by RBPs and revealed complex strategies for RNA signaling [16, 38]. The crystal structure of RIG-I revealed binding of the ds RNA end to a narrow, highly basic channel, in which residues make predominantly contacts with the sugar-phosphate backbone of both strands and recognize the viral ds RNA ends. Most

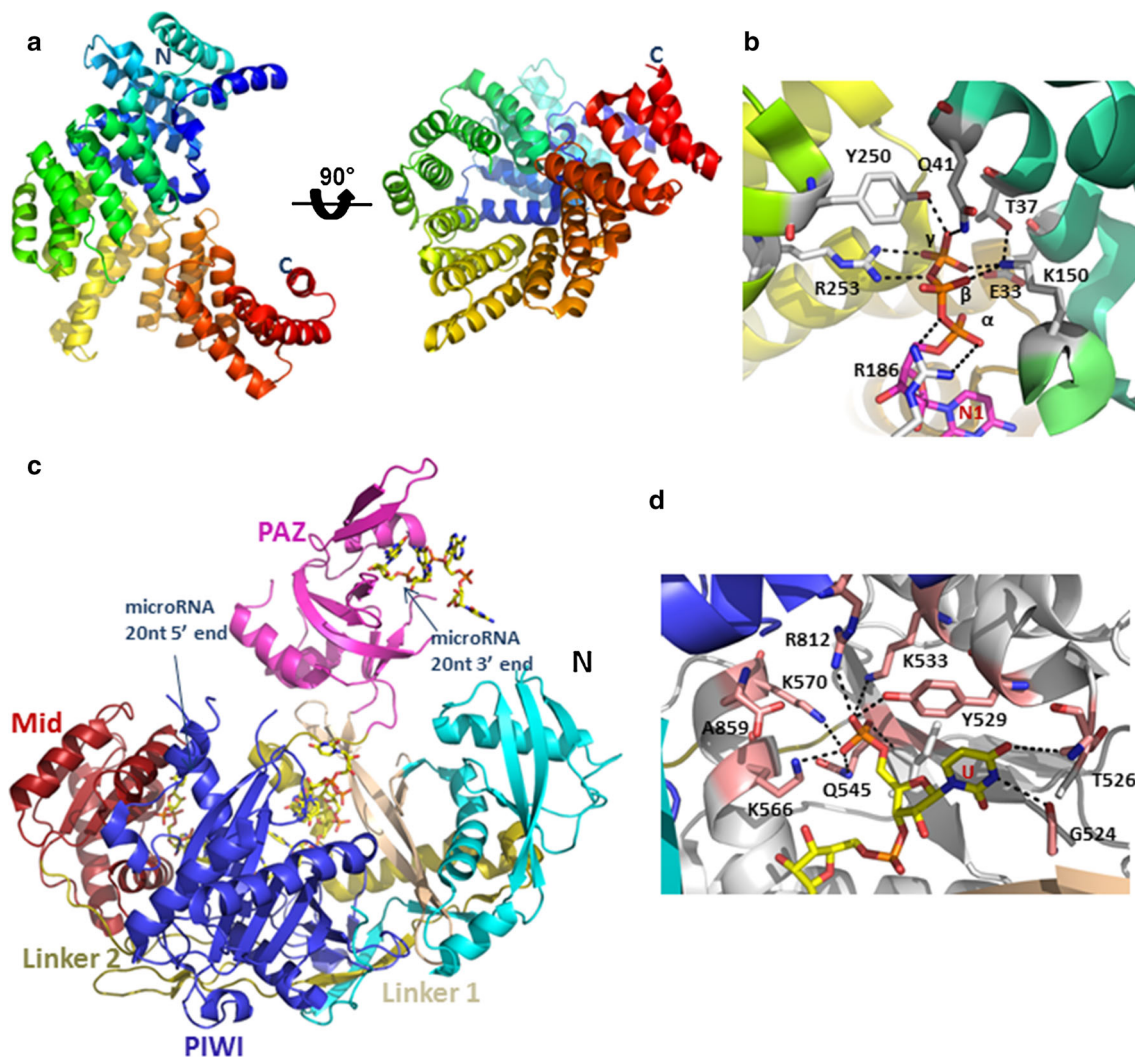


Fig. 1 RBPs can bind target RNAs in a non-sequence specific manner, which is dependent on recognition of marker groups at the 5' and 3' ends of target RNA molecules. **a** Overall structural views of IFIT5 in absence of RNA presented as cartoon models (PDB accession code: 4HOQ). **b** Close-up view of the residues making contacts with the RNA 5'-triphosphate group (PDB accession code: 4HOR). The three phosphates are labeled as α , β , γ ; RNA and the amino acids that interact with the triphosphate group are presented as stick models. The 5' nucleotide (N1) is shown in stick representation, with carbon atoms in pink, phosphor atoms in orange, nitrogens in blue and oxygens in red. **c** Overall structure of hAgo2 in complex with miR-20a (PDB accession code: 4F3T). The Mid domain is

colored in red, the PIWI domain is dark blue, the PAZ domain is magenta and the N terminal domain is light blue. The 5' end of miR-20a is trapped at the interface of the Mid domain and PIWI domain, and the 3' end of miR-20a is bound to the PAZ domain. RNA is represented as stick models. **d** Close-up view of the interactions of the first miR-20a base (U1) and the terminal monophosphate with hAgo2. Interacting residues are shown in stick representation with carbons in pink, nitrogens in blue and oxygens in red. The RNA is shown as stick model, with carbons in yellow and phosphors in orange. The Mid domain is shown in gray and the PIWI domain is dark blue. Hydrogen bond and salt bridge interactions are indicated by black dashed lines

of these contacts are formed between phosphoryl oxygens of the RNA backbone and charged amino acids, as well as between the ribose 2'-hydroxyl groups of the RNA and amide groups of conserved Q and N residues at the protein-RNA interface [16]. The ternary complex structure of MDA5 bound to a 12-bp dsRNA and an ATP analog indicated that the MDA5 RNA-binding channel recognizes the internal duplex structure, whereas RIG-I recognizes the terminus of dsRNA, through primarily RNA phosphate

backbone and 2' hydroxyl group contacts, consistent with the sequence-independent recognition of double-stranded RNAs [38].

Non-sequence-specific recognition is an essential mechanism of RBPs to target and regulate RNA metabolism by interactions that share limited relationships with those between RBPs and RNA bases. Formation of a positively charged binding surface or pocket suitable to accommodate negatively charged RNAs contributes to

strong RNA binding. Thus, it is a common theme for non-sequence-specific recognition by RBPs, which recognize “marker” groups in the 5′ or 3′ ends of target RNAs with a specific pocket, and make sequence-independent interactions with backbone phosphates of RNA through the positively charged surface on the RBPs.

Tandem domains define sequence-specific RNA-binding domains

Pentatricopeptide repeat proteins

The pentatricopeptide repeat (PPR) proteins form an extended family of RNA-binding proteins that have conserved functions in yeast, plants, and humans. A large number of PPR proteins have been identified and have been demonstrated to have important roles in RNA processing. In plants, PPR proteins are primarily involved in mitochondrial and chloroplast RNA processing, with more than 400 members in *Arabidopsis thaliana* alone [7]. Expression of the small chloroplast genome requires hundreds of nuclear gene products; among them are PPR proteins, which play a crucial role in recognizing and regulating RNA processing through their ability to bind target RNA in a sequence-specific manner [18]. PPR proteins are characterized by degenerate 35-amino acid motifs that are arranged in tandem repeats [25, 29]. The PPR proteins consist of two classes, denoted P and PLS. P-class proteins consist of one or more tandem arrays of the canonical 35 amino acid PPR motifs, whereas PLS-class proteins contain mixtures of canonical P-type motifs and variant ‘longer’ (L-) and ‘shorter’ (S-) type motifs, and mainly function in RNA splicing and editing [32]. The crystal structures of the PPR domain of human mitochondrial RNA polymerase and that of PRORP1 from *A. thaliana* revealed that each PPR motif adopts two antiparallel α -helices connected by a turn, resembling the structure of the 34-amino acid TPR motif [14, 31, 32].

P-type PPR proteins in general contain more than 10 PPR motifs, which have the unusual ability to bind long RNA segments with high affinity, thereby preventing the bound RNA from interacting with other proteins or RNA. It has been proposed that PPR proteins bind specific RNA nucleotides via the combinatorial action of two amino acids in each repeat [6]. The combinatorial amino acid code for nucleotide recognition by P-type PPR motifs was proposed to be: $T_5D_{35} = G$; $T/S_5N_{35} = A$; $N_5D_{35} = U$; $N_5N/S_{35} = C$, where residues 5 and 35 are from the 5th and 35th residues of the same PPR motif. The interactions between PPR motifs and corresponding RNAs can affect gene expression in various ways depending on their position with respect to *cis* elements and open reading frames (ORFs).

Recent crystal structures of PPR10 in RNA-free and RNA-bound states elucidated the basis of specific RNA recognition by a P-class PPR protein [39]. PPR10 forms a right-handed super-helical assembly with 19 PPR motifs (Fig. 2a). The PPR10/18-nt *PSAJ* RNA (5′-GUA-UUCUUUAAUUAUUUC-3′) complex structure revealed an asymmetric dimer that was generated in the presence of target RNA, with the RNA fragments buried in the cavities formed by two PPR10 monomers (Fig. 2b, c).

The structure is generally consistent with the combinatorial code prediction. The *PSAJ* RNA segment is specifically recognized in a modular fashion, in which each nucleotide responds to one PPR motif (Fig. 2d). Recognition of U15 and U16 of the 3′ end of the *PSAJ* RNA by PPR10 appears coordinated by repeats 16 and 17 following the proposed recognition pattern: the 5th residues of repeats 16 (N635) and 17 (N671) contact the U15 and U16 bases through hydrogen bonds, and the combination of the 5th and 35th residues of the PPR repeats define the specificity of RNA recognition (Fig. 2e). A number of direct and water-mediated hydrogen bonds are found between the backbone phosphate/ribose groups of U15-U16-U17-C18 and the polar residues on repeats 15–19 of PPR10, indicating that, in addition to the specificity-determining interactions between PPR motifs and RNA bases, non-sequence-specific interactions enhance the affinity between PPR motifs and RNA. In the structure of *PSAJ*-bound PPR10, only 6 out of 18 nucleotides in the *PSAJ* RNA element strictly follow the modular pattern. Therefore, more structural characterizations are required to further rationalize the “binding codes” for specific RNA recognition by P-class PPRs [39].

The PPR tracts of PLS-class proteins appear to function similarly to those of P-class proteins, in that they form a binding surface that recognizes the sequence in a single-stranded RNA molecule [7]. Only one structure of an RNA-bound PLS-class PPR protein has been determined to date. The mechanisms of RNA recognition by small PLS-class PPR proteins are presented by recent structural and biochemical studies of *Thylakoid assembly 8* (*THA8*) proteins [5, 20].

THA8 is a maize gene that encodes a small protein localized to chloroplasts. It is required for the splicing of the *ycf3-2* and *trnA* group II introns, whose sequences contain multiple *THA8*-binding sites [21]. The apo-structures of the *THA8* homolog from *Arabidopsis thaliana* and *Brachypodium distachyon*, denoted as *THA8L* and *THA8*, respectively, indicated five tandem PPR repeats arranged in a pattern similar to that of P-class PPR and TPR proteins. *THA8* and *THA8L* share a similar structure and RNA binding specificity even though they only have 26 % sequence identity, suggesting a conserved function of the *THA8* homologs (Fig. 3a) [20].

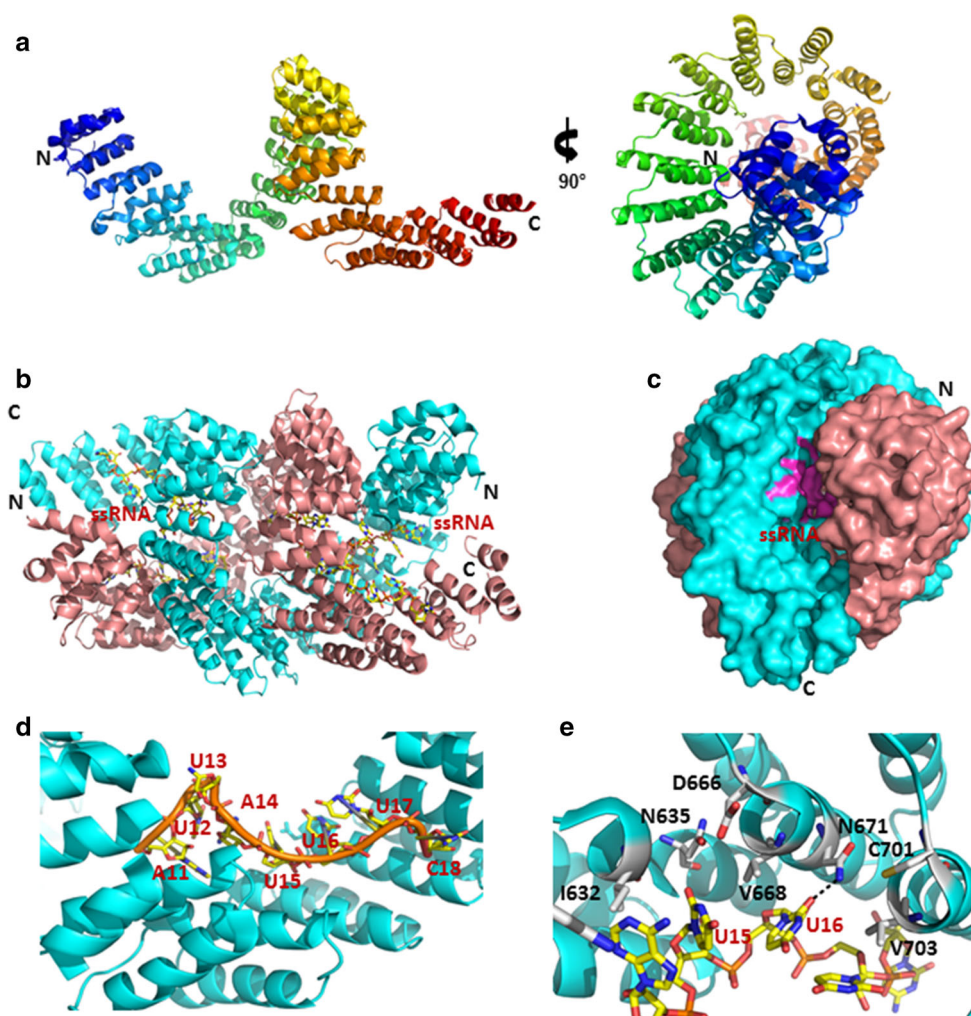


Fig. 2 P-Class PPR proteins bind specific RNA nucleotides via the combinatorial action of two amino acids in each PPR repeat. **a** Overall structure of apo-PPR10 shown in cartoon representation (PDB accession code: 4M57). **b** Structure of the PPR10-PSAJ RNA complex (PDB accession code: 4M59). Two PPR10 monomers are shown in *cyan* and *pink*, respectively. Single-stranded RNAs are shown as stick models and colored in *magenta*. **c** Surface representation of the binding pocket for ssRNA binding, with the region contributing to RNA binding shown in *magenta*. **d** The eight nucleotides at the 5' end of the PSAJ RNA segment are specifically

recognized by PPR10 in a modular fashion. RNA is shown as stick and cartoon representation with carbons in *yellow*, nitrogens in *blue*, and oxygens in *red*. The RNA backbone is shown in *orange* and PPR repeats in *cyan*. **e** Recognition of the bases U15 and U16 by PPR10 follows the “binding code”. The 5th residues of repeats 16 (N635) and 17 (N671) contact U15 and U16 through hydrogen bonds that are indicated by *black dashed lines*. RNA is shown in stick representation and colored the same as in **(d)**. Residues that define the specificity of RNA recognition are presented as stick, with carbons in *white*, nitrogens in *blue* and oxygens in *red*

The complexes of THA8 bound to two short RNA fragments revealed that RNA-binding induces formation of an asymmetric THA8 dimer, as further confirmed by biochemical analysis. The RNA is bound at the dimer interface assembled by the C-terminal part of one monomer and the N-terminal part of the other monomer (Fig. 3b). The dimer association creates a concave surface at the interface with strong positive charge potential, which is complementary to the negatively charged RNA molecule (Fig. 3c). A conserved G-nucleotide of the bound RNAs interacts with both monomers and is stacked between two tyrosine residues (Y169 and Y205) from position 3 of PPR

motifs 4 and 5 (Fig. 3d). The G base is nearly buried within a pocket formed by the THA8 dimer. The residues that contact the G-base (T172 and D203) are the 5th and last residues from motif 4, thus the mode of G-base contact is consistent with the proposed model of the combinatorial amino acid code. The 2' hydroxyl group of the G-nucleotide ribose makes a direct hydrogen bond with R176 and the 2' hydroxyl of the preceding A-nucleotide is contacted by a water-mediated hydrogen bond with R58 from the neighboring THA8. These contacts ensure the strong preference of THA8 for single-stranded RNA over double-stranded RNA or DNA.

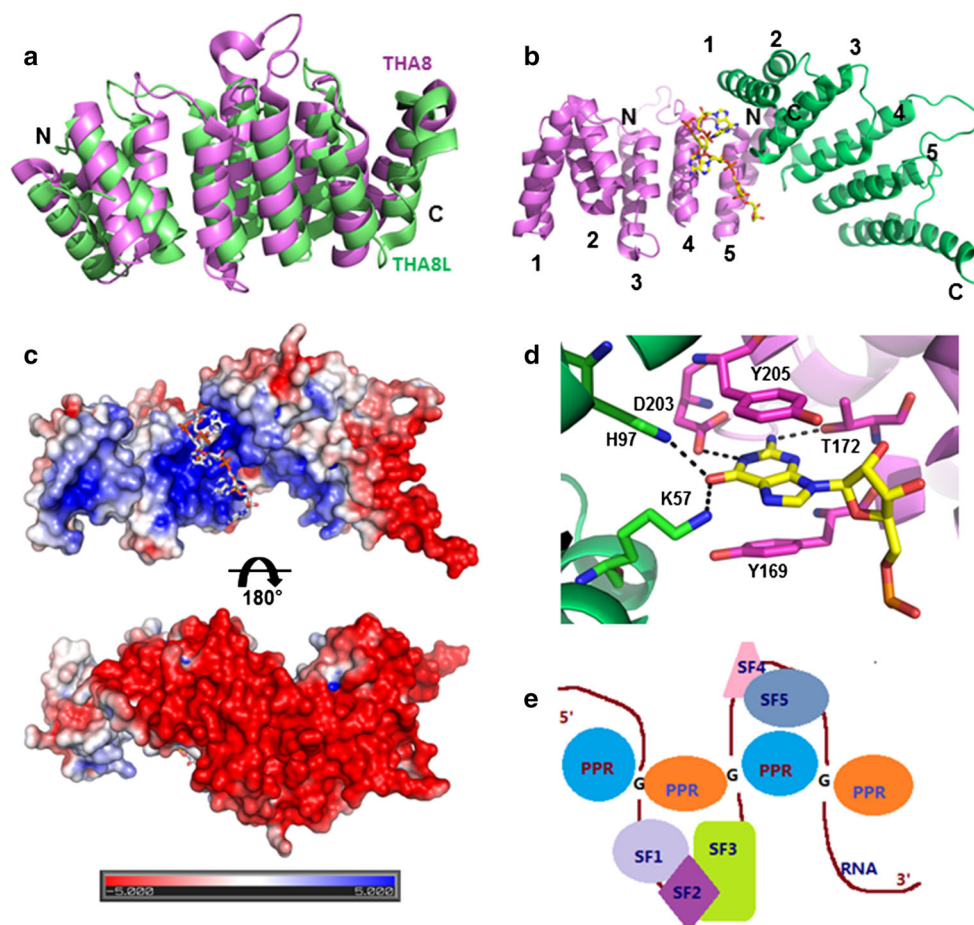


Fig. 3 The THA8 homolog recognizes target RNAs through formation of a dimer or oligomer. **a** Superposition of apo-THA8 (magenta, PDB accession code: 4ME2) and THA8L (green, PDB code: 4LEU) structures in cartoon representation. **b** The structure of THA8 in complex with a 13-nucleotide Zm-4 RNA shown in cartoon representation, with the two monomers colored in green and magenta (PDB accession codes: 4N2Q). The bound Zm-4 RNA fragment (AGAAA) is shown as stick model at the dimer interface. **c** Surface charge distribution of the two different sides of the THA8 dimer. The bound RNA fragment is shown as stick model. **d** Close-up view of the THA8-dimer interactions with the G nucleotide of the AGAAA motif. The carbon atoms are colored in green and magenta for two monomers that interact with Zm4 RNA. Residues that interact with

RNA are shown in stick representation, using the same color as the PPR repeat to which they belong. The G nucleotide is shown in stick representation, with carbons in white, nitrogens in blue, and oxygens in red. Hydrogen-bond interactions are indicated by black dashed lines. **e** A model for the regulation of RNA recognition by short PPR proteins. The single-stranded pre-mRNA induces the oligomerization of PPR proteins, thus bringing several discontinuous regions of RNA into proximity to be recognized by other splicing factors to facilitate the alternative splicing of introns. PPR proteins are presented as blue and orange ovals; SF splicing factors that are recruited by mRNA are colored individually, G G nucleotides that are responsible for RNA recognition

Based on the structures and biochemical analyses, the authors proposed a new model of RNA recognition by RNA-induced formation of an asymmetric dimer of a PPR protein. Dimerization or oligomerization of short PPR proteins could facilitate the recruitment of other splicing factors to discontinuous regions of pre-mRNA for efficient spliceosome assembly (Fig. 3e). For small PPR proteins, the capacity for RNA binding is inherently small. Each THA8 monomer has a roughly flat, rectangular shape that is not optimal for binding RNA. Dimerization creates a concave surface with strong positive charge potential that is complementary to short RNA fragments with negative

charges. The dimer formation also increases the binding affinity of THA8 for RNA by engaging more residues from both molecules for contacting the G nucleotide [20].

Pumilio homolog proteins

Pumilio-FBF (Puf) domain proteins have been reported in organisms ranging from unicellular yeast to higher multicellular eukaryotes, including humans and plants. They are involved in a variety of post-transcriptional RNA processing reactions, including RNA decay, RNA transport, rRNA processing and translational repression [2, 11].

Pumilio and FBF (fem-3 binding factor) together comprise the founding members of the PUF family of RNA-binding proteins. Most Pumilio proteins bind RNAs in a sequence-specific fashion. They have distinct functions defined by phenotypic differences, mRNA target specificity, or expression patterns [19]. In this section, we review a crystal structure of a PUF protein and highlight the modular function that defines differences in RNA recognition in this family.

The crystal structure of Pumilio revealed an architecture consisting of eight repeat motifs that each comprises three alpha helices. The repeats pack against one another to form an extended curved structure that resembles a banana [36]. The structure of human Pum1 bound to *hnNRE* RNA demonstrates that the concave surface forms the RNA-binding interface with each repeat recognizing a single nucleotide. This pattern of one nucleotide-one binding repeat is the same as that for RNA recognition by PPR motif modules. The N-terminal repeat (R1) binds to the 3'-nucleotide (N8) of the target sequence, while the C-terminal repeat (R8) binds to the 5'-nucleotide (N1) (Fig. 4a). Residues at positions 12 and 16 in each repeat directly interact with the RNA base through formation of hydrogen bonds (Fig. 4b), whereas the residue at position 13 is involved in a stacking interaction between two adjacent bases [35]. The structure suggests a "recognition code", where residues at positions 12 and 16 in each repeat contribute to the specific recognition of a base, with N₁₂Q₁₆ recognizing uracil, C₁₂Q₁₆ adenine, and S₁₂E₁₆ guanine. A8 stacks between R936 and H972 and U3 stacks between Y1123 and H1159 from repeat 8, indicating that RNA base recognition by PUF proteins is mainly mediated by hydrogen bonds and stacking interactions (Fig. 4c, d).

This structure suggests that RNA recognition by PUF protein is highly modular, indicating the potential for designing proteins with predictably altered RNA binding specificity. While the residue combination S₁₂R₁₆ was engineered to recognize cytosine, engineered PUF domains were successfully fused to a post-transcriptional regulator that sequence-specifically repressed a reporter as well as an endogenous gene in human cell lines, demonstrating the potential of the PUF domain assembly method for RBP engineering [3].

Tristetraprolin-type tandem zinc finger domains

Tristetraprolin (TTP), the best known member of a family of tandem (R/K)YKTELX₈CX₅CX₃H zinc finger proteins, can destabilize target mRNAs by binding to AU-rich elements (AREs) in their 3'-untranslated regions (UTRs) and subsequently promote deadenylation, ultimately leading to the destruction of those mRNAs. Structural studies

demonstrate that recognition of the ARE by TTP is based entirely on its sequence.

The NMR structure of the complex of the TTP-like Zinc-Finger (TZF) protein TIS11d bound to 5'-UUAUUUAUU-3' RNA presented evidence that TZF domains can interact with specific RNA sequences [15]. The structure identified two CCCH zinc finger domains connected by a linker sequence, in which each finger bound to separate 5'-UAUU-3' subsites (Fig. 5a). The two finger domains are structurally highly conserved (Fig. 5b). The protein structure is further stabilized by hydrogen bonds from main chain amino acids to zinc-bound cysteine sulfur atoms and by backbone hydrogen bonds. In addition, the conserved aromatic side chains of F162 (finger 1) and F200 (finger 2) are stacked against the side chains of H178 (finger 1) and H216 (finger 2). The interface between the protein and the 5'-UAUUUAUU-3' ARE is dominated by hydrophobic packing and hydrogen-bonding interactions. Q175 and E157 interact with the U6 base and K160 contacts the A7 base through hydrogen bonds to define the specificity of RNA sequences (Fig. 5c). Thus, specific RNA recognition by TZF is predominantly conferred by a network of intermolecular hydrogen bonds between the tandem repeat modules and RNA bases.

RNA-recognition motif domains

The RNA-recognition motif (RRM) is by far the most common and best-characterized RNA-binding module. The RRM is composed of 80–90 amino acids that form a four-stranded anti-parallel β -sheet and two helices that are arranged into a $\beta\alpha\beta\beta\alpha\beta$ fold [24]. While RRM recognize stretches of a few nucleotides, typically several RRM within an RNA-binding protein are required to provide sufficient sequence specificity. Three subclasses of RRM were identified in eukaryotes. The structures of these three subclasses of RRM bound to their target RNAs revealed the different mechanisms that they use to recognize and regulate the alternative splicing of target RNAs.

RRM proteins bind RNA through their β -sheet surface. The mechanism of RNA recognition by canonical RRM was first identified in the crystal structure of RRM of hnRNP (heterogeneous nuclear ribonucleoprotein) A1 (UP1) in complex with single-stranded telomeric DNA, rather than RNA [9]. Binding is mediated in most cases by three conserved residues: an R or K residue that forms a salt bridge with the phosphodiester backbone, and two aromatic residues that form stacking interactions with the bases. These three amino acids reside in the two highly conserved motifs, RNP (ribonucleoprotein) motif-1 (RNP1) and RNP2, in the central β -strand region, and they determine the binding to a 5'-AGG-3' fragment (Fig. 6a).

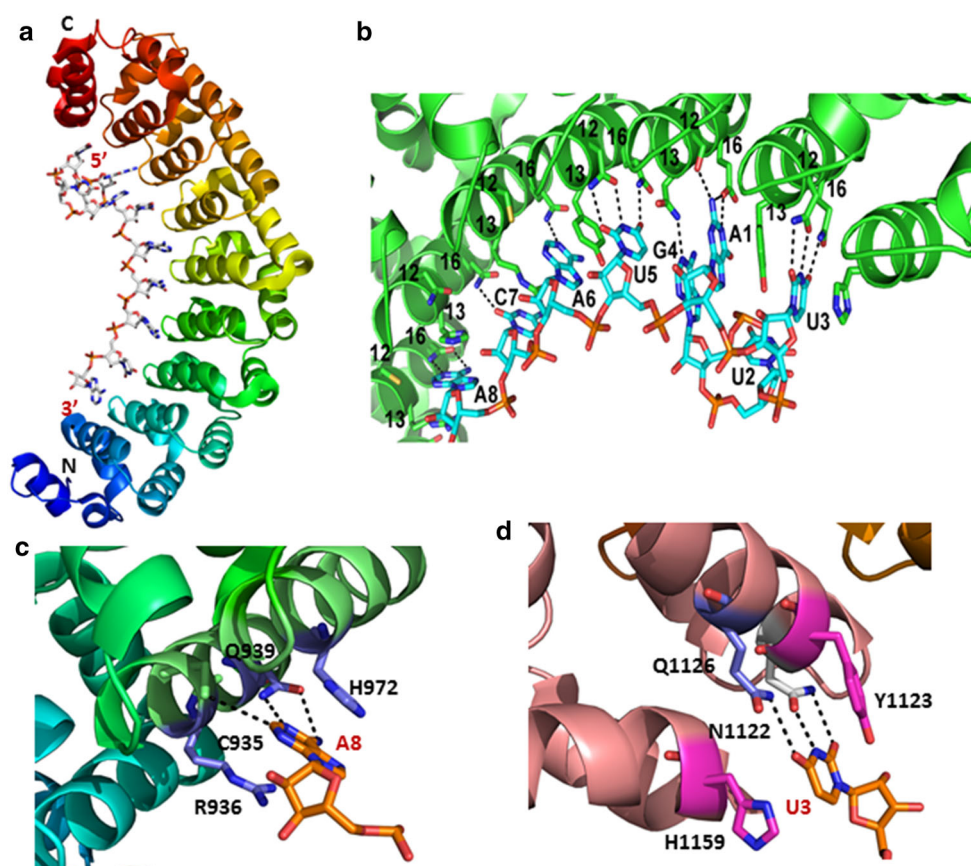


Fig. 4 Residues at positions 12 and 16 in each PUF repeat contribute to the recognition of specific RNA bases. **a** Side view of the overall structure of the human Pumilio1 PUF domain bound to NRE2-10 RNA (PDB accession code: 1M8Y). Pum1 is presented as cartoon model and RNA nucleotides are shown in stick representation, with carbons in *white*, nitrogens in *blue*, and oxygens in *red*. **b**, Enlarged view of recognition of RNA bases of NRE2-10 RNA by the PUF repeats. Residues at positions 12 and 16 from each PUF repeat form hydrogen bonds with one RNA base (shown as *black dashed lines*) and residue at position 13 stack with RNA bases. PUF repeats are

colored in *green*, with residues that define the specific RNA recognition shown as stick models. **c** Close-up view of the 12th and 16th residues of PUF repeat 3 contacting the base of the A8 nucleotide through hydrogen bonds. Hydrogen bonds with Q939 and C935 are indicated with *black dashed lines*. A8 stacks between R936 and H972. **d** Y1123 and H1159 from repeat 8 form stacking interactions with the uracil base (U3). N1122 (*white*) and Q1126 (*blue*) form hydrogen bonds with the uracil base. Hydrogen bonds are indicated by *black dashed lines*

The subclass of quasi-RRMs (qRRM) uses a totally different recognition interface, which involves the loop regions of the β -sheet surface. The three qRRM structures of hnRNP F in complex with 5'-AGGGAU-3' RNA revealed how G-tract RNA is recognized by qRRM modules [10]. The qRRM domain of hnRNP F contains the classical compact $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ RRM fold with three highly conserved loops that are responsible for G-tract RNA binding. The three guanines of the tract adopt a compact conformation resembling an arch, surrounded by three conserved residues that stack each guanine base. The structure explains how each qRRM of hnRNP F specifically recognizes three consecutive guanines by efficient binding of a surface formed by specific loops.

Finally, the subclass of pseudo-RRMs is characterized by a distinctive, invariant motif, SWQDLKD, in α -helix 1. Recently, the NMR structure of the pseudo-RRM of the

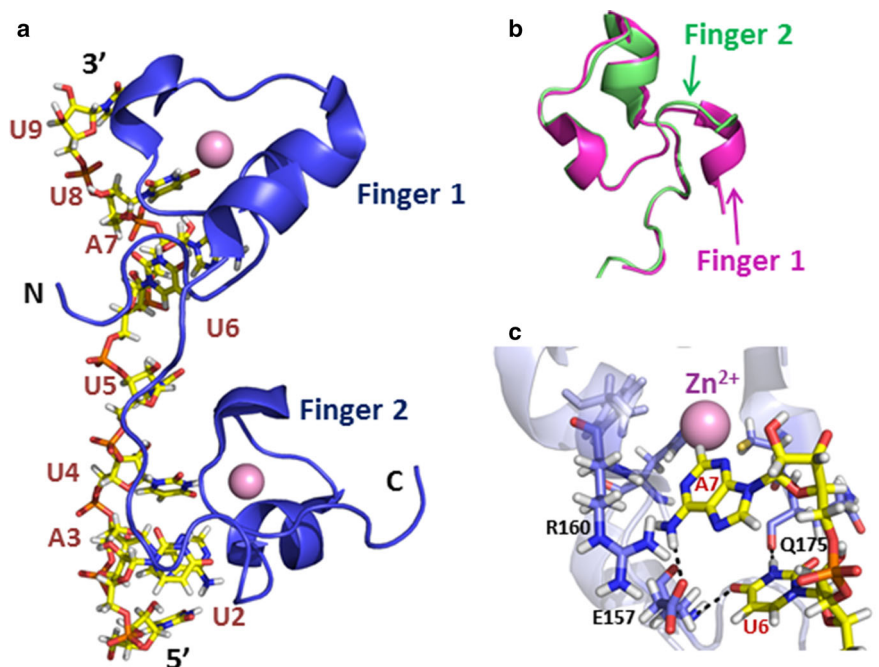
best studied SR (serine/arginine) protein, SRSF1 (SF2/ASF), bound to RNA was reported. The structure revealed that the domain specifically binds a GGA motif, primarily using the conserved residues located in α -helix 1 (Fig. 6c). This demonstrated the striking use of three alternative RNA-binding surfaces of RRM, formed by either an α -helix (pseudo-RRMs), β sheet (canonical RRM), or loops (quasi-RRMs), illustrating the RNA-binding variability of a very simple protein module [8, 27].

Protein–protein interactions can define RNA specificity

Dimerization plays a role in RNA recognition

In addition to expanding the ways in which RNA can be recognized, multiple modules also allow RNA-binding

Fig. 5 The structural basis for RNA binding by the TZF RBD. **a** Structure of the TIS11d/RNA complex (PDB accession code: 1RGO). Each Zn^{2+} finger domain (blue) is bound to one “UAUU” subsite. The RNA is shown in stick representation with carbons shown in yellow, nitrogens in blue, oxygens in red, and phosphorous in orange. The Zn^{2+} ions are presented as pink spheres. **b** Superposition of two finger domains of TIS11d, colored in green and magenta. **c** Close-up view of the interactions between finger 1 and the U6 and A7 nucleotide bases. Residues contacting nucleotides are shown as stick model with carbons in blue; RNA is presented as stick model using the same color code as in (a). Hydrogen bonds are indicated as black dashed lines



proteins to interact simultaneously with RNA and with other proteins. The simplest example of this is dimerization. Dimerization presents two recognition sites for RNA binding and can therefore provide a cooperative interaction that strengthens the affinity of the protein for the RNA [24].

For instance, PPR10 and the THA8 homolog enhance their affinities for RNA through formation of a dimer. Dimerization of THA8 generates a binding pocket at the interface of two THA8 monomers that is highly positively charged for accommodation of negatively charged RNA. PPR10 and THA8 exist as monomers in RNA-free state, but both of them assemble into dimers in the presence of target RNAs, indicating that RNA binding to both monomers induces the dimerization or oligomerization of PPR proteins (Figs. 2b, 3b). Among dsRNA binding proteins, dimerization or oligomerization of RIG-I is triggered by RNA and is dependent on RNA length, which is essential for stimulation of RNA signal transduction [16].

These two examples illustrate the role of dimerization in RNA recognition, but there are other examples of RBPs that function by dimerizing or by forming protein–protein interactions. The X-ray structure of the human core U2AF (U2 auxiliary factor) heterodimer, consisting of the U2AF35 central domain and a proline-rich region of U2AF65 [22], revealed an atypical RNA recognition motif (RRM), in which U2AF35 and the U2AF65 polyproline segment interact via tryptophan residues. Biochemical experiments demonstrated that the core U2AF heterodimer

binds RNA, and that the interacting tryptophan side chains are essential for U2AF dimerization.

Atypical RRM in splicing factors such as U2AF may serve both as protein–protein interaction motifs as well as in protein–RNA recognition.

Dimerization is also a conserved mechanism exploited by viral proteins to increase their RNA-binding surface and to facilitate RNA recognition by establishing the relative positions of RNA-interacting amino acids. For example, NS5A is a single-stranded RNA-binding protein, which participates in maintenance of the Hepatitis C virus life cycle, and which is associated with RNA replication [23]. The structure of NS5A bound to virus RNA indicated that a zinc finger domain from one NS5A molecule could interact with the zinc finger motif of an adjacent NS5A molecule, with the virus RNA located in a pocket formed by the interaction of the two zinc finger domains. This dimerization of NS5A is essential for RNA binding and for efficient and precise virus RNA replication.

Heteromeric protein–protein interactions define RNA specificity

RBDs from different proteins can cooperate to recognize RNA through a combination of protein–protein interactions. The recent dissection of a complex from the spliceosome demonstrates this principle and illustrates how even small alterations in RBDs can indirectly modulate the

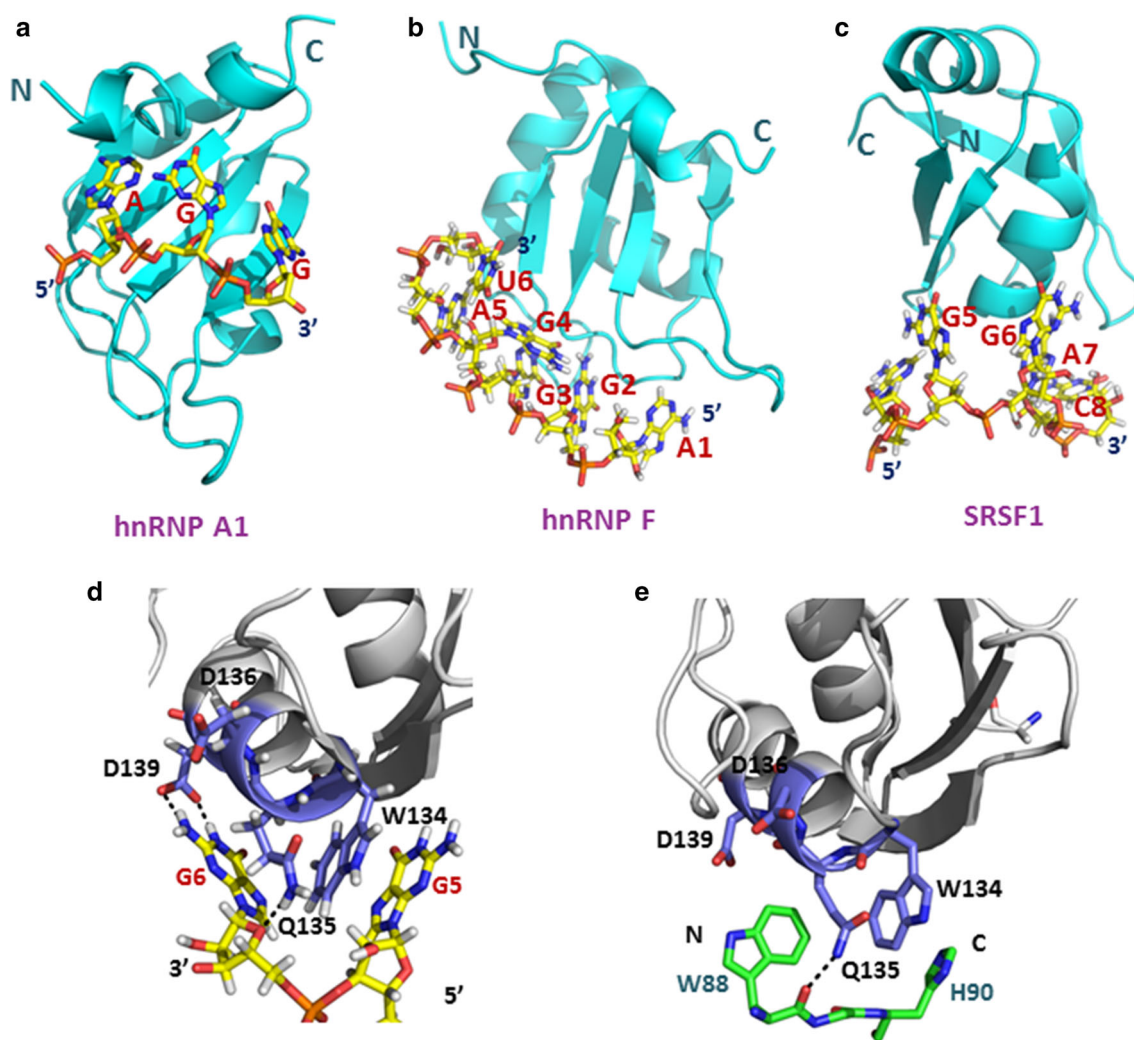


Fig. 6 Three subclasses of RRMs use different mechanisms for RNA binding. **a** Structure of hnRNP A1 RRM1 bound to single-stranded telomeric DNA (PDB accession code: 2up1). **b** Structure of hnRNP F qRRM2 bound to 5'-AGGGAU-3' RNA (PDB accession code: 2KG0). **c** Structure of the SRSF1 pseudo-RRM bound to 5'-AGGAC-3' RNA (PDB accession code: 2M8D). DNA and RNAs are shown in stick representations with carbons in yellow, nitrogens in blue, oxygens in red, and phosphorus in orange. RRM motifs are shown as cartoon models in cyan. **d** Close-up view of interactions of the SRSF1 pseudo-RRM bound to the GG dinucleotide of the 5'-

UGAAGGAC-3' RNA. **e** Close-up view of the structure of the SRSF1 pseudo-RRM bound to the Trp-Gly-His tripeptide of SRPK1 (PDB accession code: 3BEG). The side-chains of W88 and H90 occupy the same sites as G6 and G5, respectively. Both the G6 base and the side-chains of W88 and H90 could interact with SRSF1 via hydrogen bond formation. The RRMs are shown in gray with residues interacting with the RNA or peptide presented as stick models in dark blue. The Trp-Gly-His tripeptide is presented as stick model in green, with hydrogen bonds are indicated as black dashed lines

RNA-recognition properties of RBDs by altering protein-protein interactions.

During the initial steps in spliceosome assembly, SF1 (splicing factor 1) and U2AF proteins cooperatively bind to sequences at the 3'-splice site and upstream of it. The structure of a complex of the N-terminal region of SF1 with the C-terminal U2AF homology motif domain of U2AF65 revealed that the helix hairpin domain of SF1 is essential for cooperative formation of the ternary SF1-U2AF65-RNA complex [40]. The importance of the cooperative interaction was demonstrated by mutational disruption of

the SF1 and U2AF65 interaction, which dramatically decreased the affinity for RNA.

Not all proteins that interact with RBPs enhance their RNA binding affinity; some of them could decrease or abolish the interaction between RBPs and RNA by competitively binding the RNA interaction surface. Comparison of the structures of the SRSF1 pseudo-RRM bound to RNA with that of SRSF1 bound to the protein kinase SRPK1 illustrates that the mode of binding of the α -helix 1 is very similar for both the RNA and the protein (Fig. 6d, e). Two nucleotides share the same SRSF1

binding interface with two amino acids from SRPK1. This double involvement of α -helix 1 also implies that SRSF1 phosphorylation by SRPK1 and RNA binding must be mutually exclusive events. SRPK1 binding could prevent RNA recognition by occupying the shared binding surface. Pseudo-RRMs function in splicing regulation by competing with other splicing factors for binding to the same surface, rather than by recruiting splicing factors to the corresponding exon region [8].

Recognition of sgRNA and target DNA by Cas9

A unique mode of RNA recognition has been identified for Cas9 proteins. The Type II CRISPR (clustered regularly interspaced short palindromic repeats)—Cas (CRISPR-associated) system has been exploited in numerous gene-targeting applications, in which its sequence specificity is programmed by either dual crRNA–tracrRNA guides or chimaeric single-molecule guide RNAs (sgRNAs) [34, 41]. The crystal structures of *Streptococcus pyogenes* Cas9 and *Actinomyces naeslundii* Cas9, as well as the recently reported crystal structure of *S. pyogenes* Cas9 in complex with sgRNA and its target DNA, revealed the basis of target nucleic acid recognition mechanism by Cas9 endonuclease [4, 17, 28].

The crystal structure revealed that Cas9 consists of two lobes: a recognition (REC) lobe and a nuclease (NUC) lobe. The NUC lobe consists of the RuvC (residues 1–59, 718–769, and 909–1,098), HNH (residues 775–908), and PAM-interacting (PI) (residues 1,099–1,368) domains. The negatively charged sgRNA:target DNA heteroduplex is accommodated in a positively charged groove at the interface between the REC and NUC lobes (Fig. 7a). The crystal structure revealed that the sgRNA binds the target DNA to form a T-shaped architecture comprising a guide:target heteroduplex, a repeat:anti-repeat duplex, and stem loops 1–3 (Fig. 7b). The guide RNA and target DNA form the guide:target heteroduplex via 20 Watson–Crick base pairs, but within the repeat:anti-repeat duplex region, G27, A28, A41, A42, G43, and U44 are unpaired, with A28 and U44 flipped out from the duplex. These unpaired nucleotides influence the accurate architecture of the heteroduplex via stacking with nucleotide base pairs or formation of hydrogen bonds with other nucleotides nearby (Fig. 7b, f). Different domains of Cas9 bind sgRNA/target DNA duplex and stem loops via both non-sequence-specific recognition and sequence-specific recognition mechanisms [28].

The conserved arginine cluster of the bridge helix is responsible for sgRNA:DNA recognition. R66, R70, and R74 form multiple salt bridges with the sgRNA backbone, whereas R78 and R165 form a single salt bridge with the

sgRNA backbone. Cas9 recognizes the sgRNA guide region in a sequence-independent manner, except for the U16–R447 and G18–R71 interactions. The hydrogen bonds or salt bridges that mediate the interactions between the residues and U/G may thus define the sequence recognition specificity and preference (Fig. 7c). The REC1 and RuvC domains recognize the 20-bp guide:target heteroduplex in a sequence-independent manner.

In contrast to the sequence-independent recognition of the sgRNA guide region, sequence-dependent interactions exist between Cas9 and the repeat:anti-repeat duplex. The bases of U23/A49 and A42/G43 form hydrogen bonds with the side chain of R1122 and the main-chain carbonyl group of F351, respectively (Fig. 7d, e). The base of the flipped U44 is trapped between Y325 and H328 by stacking interactions, and D364 interacts with the unpaired G43 base by forming hydrogen bond and stacking interactions with Y359, which contribute to G43 nucleotide-specific recognition (Fig. 7e). The base-specific recognition of G43 and U44 by Cas9 plays a very important role in efficiency endonuclease activity.

Beside the guide:target heteroduplex and the repeat:anti-repeat duplex, sgRNA stem loops 1–3 are primarily recognized by the REC lobe and the NUC lobe. The interactions between the RNA backbone and Cas9 make the main contribution to Cas9–sgRNA binding, whereas the hydrogen bonds formed by particular residues of Cas9 define the RNA recognition specificity. Intra-nucleic acid interactions affect the architecture of the sgRNA and influence its orientation and recognition by the different domains of Cas9 (Fig. 7e, f).

The electron microscopic (EM) reconstructions of *S. pyogenes* Cas9:RNA and Cas9:RNA:DNA complexes showed that guide RNA binding results in a conformational rearrangement and formation of a channel to facilitate target DNA binding [17]. The crystal structures of the Cas9–sgRNA–target DNA ternary complex provide a critical step toward understanding the molecular mechanism of RNA-guided DNA targeting by Cas9 [4, 28], and these studies establish a framework for the rational engineering of Cas9 enzymes.

Conclusions

RBPs can selectively recognize target RNAs either non-sequence specificity or sequence specificity. Non-sequence-specific recognition occurs mainly by the formation of hydrogen bond and salt-bridge interactions with marker groups at the 5' or 3' end of target RNAs. These interactions involve predominantly the backbone phosphate groups of nucleotides and amino acid residues that are distributed on the binding surface of the RBPs, and they are

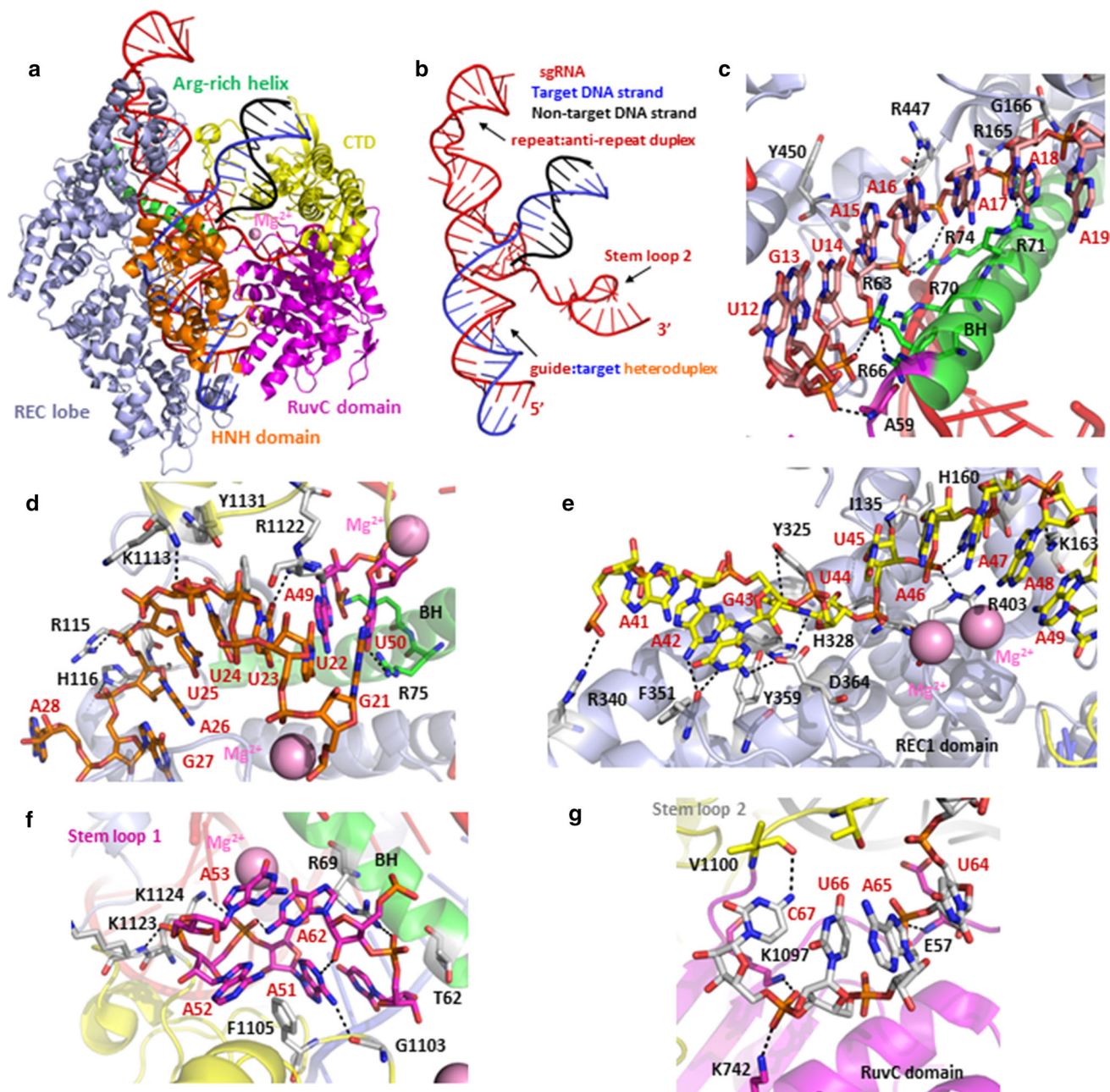


Fig. 7 Structural mechanism of sgRNA:target DNA recognition by CRISPR-associated endonuclease Cas9. **a** Overall structure of the Cas9-sgRNA-DNA ternary complex (PDB accession code: 4UN3). **b** Structure of the sgRNA: target DNA complex. The sgRNA, target DNA strand, and non-target DNA strand are colored *red*, *blue*, and *black*, respectively. **c** Close-up view of the interaction between the sgRNA guide region and the conserved arginine cluster of the bridge helix; U16-R447 and G18-R71 interactions define the specificity of RNA recognition by Cas9. **d**, **e** Sequence-dependent interactions between Cas9 and the repeat:anti-repeat duplex. The RNA bases of

U23/A49 and A42/G43 form hydrogen bonds with the side chain of R1122 and the main-chain carbonyl group of F351, respectively; the base of the flipped U44 is trapped between Y325 and H328 mediated by stacking interactions, and D364 interact with the nucleobase of the unpaired G43 by forming hydrogen bonds. **f**, **g** Close-up view of interactions between Cas9 and stem loops 1 and 2 of the sgRNA, respectively. The BH domain is colored in *green*, the RuvC domain in *magenta*, and the C-terminal domain in *yellow*. RNAs and the residues that are responsible for recognition are shown in stick representation; hydrogen bonds are indicated as *black dashed lines*

characterized by limited association between RBPs and RNA bases. Sequence-specific RNA recognition typically occurs by RNA-binding proteins that are composed of several conserved RNA-binding modules and by the shape

complementation as seen in the case of Cas9-RNA recognition. By combining these motifs in various structural arrangements, proteins are generated that can recognize RNA selectively, which is required to regulate RNA

processing. Structural biology has provided the molecular details about how individual domains recognize RNA and how these modules coordinate with each other for sequence-specific RNA binding. We have described structural principles of how multiple domains recognize RNAs and the significance of dimerization and protein–protein interaction in RNA recognition by RBPs. These structural principles have provided rational frameworks for designing sequence-specific RNA binding proteins, for example, the PUF-based artificial RNA-binding proteins [3]. Full understanding of these regulatory mechanisms will require more detailed structural studies, which we expect will expand the knowledge of RNA recognition and its dependence on the combination of multiple RBP domains and interactions with other proteins. Structural basis of protein–RNA recognition will help us to understand fundamental mechanisms of RNA biology and expand our ability to design RNA-binding proteins, which could be exploited as new tools to modulate gene expression.

References

1. Abbas YM, Pichlmair A, Gorna MW, Superti-Furga G, Nagar B (2013) Structural basis for viral 5'-PPP-RNA recognition by human IFIT proteins. *Nature* 494:60–64
2. Abbasi N, Park YI, Choi SB (2011) Pumilio Puf domain RNA-binding proteins in Arabidopsis. *Plant Signal Behav* 6:364–368
3. Abil Z, Denard CA, Zhao H (2014) Modular assembly of designer PUF proteins for specific post-transcriptional regulation of endogenous RNA. *J Biol Eng* 8:7
4. Anders C, Niewoehner O, Duerst A, Jinek M (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 513:569–573
5. Ban T, Ke J, Chen R, Gu X, Tan MH, Zhou XE, Kang Y, Melcher K, Zhu JK, Xu HE (2013) Structure of a PLS-class pentatricopeptide repeat protein provides insights into mechanism of RNA recognition. *J Biol Chem* 288:31540–31548
6. Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet* 8:e1002910
7. Barkan A, Small I (2014) Pentatricopeptide repeat proteins in plants. *Annu Rev Plant Biol* 65:415–442
8. Clery A, Sinha R, Anczukow O, Corrionero A, Moursy A, Daubner GM, Valcarcel J, Krainer AR, Allain FH (2013) Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. *Proc Natl Acad Sci USA* 110:E2802–E2811
9. Ding J, Hayashi MK, Zhang Y, Manche L, Krainer AR, Xu RM (1999) Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev* 13:1102–1115
10. Dominguez C, Fiset JF, Chabot B, Allain FH (2010) Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nat Struct Mol Biol* 17:853–861
11. Edwards TA, Pyle SE, Wharton RP, Aggarwal AK (2001) Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell* 105:281–289
12. Elkayam E, Kuhn CD, Tocilj A, Haase AD, Greene EM, Hannon GJ, Joshua-Tor L (2012) The structure of human argonaute-2 in complex with miR-20a. *Cell* 150:100–110
13. Glisovic T, Bachorik JL, Yong J, Dreyfuss G (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 582:1977–1986
14. Howard MJ, Lim WH, Fierke CA, Koutmos M (2012) Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing. *Proc Natl Acad Sci USA* 109:16149–16154
15. Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE (2004) Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* 11:257–264
16. Jiang F, Ramanathan A, Miller MT, Tang GQ, Gale M Jr, Patel SS, Marcotrigiano J (2011) Structural basis of RNA recognition and activation by innate immune receptor RIG-I. *Nature* 479:423–427
17. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S et al (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343:1247997
18. Johnson X, Wostrikoff K, Finazzi G, Kuras R, Schwarz C, Bujaldon S, Nickelsen J, Stern DB, Wollman FA, Vallon O (2010) MRL1, a conserved Pentatricopeptide repeat protein, is required for stabilization of rbcL mRNA in *Chlamydomonas* and *Arabidopsis*. *Plant Cell* 22:234–248
19. Kaymak E, Wee LM, Ryder SP (2010) Structure and function of nematode RNA-binding proteins. *Curr Opin Struct Biol* 20:305–312
20. Ke J, Chen RZ, Ban T, Zhou XE, Gu X, Tan MH, Chen C, Kang Y, Brunzelle JS, Zhu JK et al (2013) Structural basis for RNA recognition by a dimeric PPR-protein complex. *Nat Struct Mol Biol* 20:1377–1382
21. Khrouchtchova A, Monde RA, Barkan A (2012) A short PPR protein required for the splicing of specific group II introns in angiosperm chloroplasts. *RNA* 18:1197–1209
22. Kielkopf CL, Rodionova NA, Green MR, Burley SK (2001) A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell* 106:595–605
23. Lambert SM, Langley DR, Garnett JA, Angell R, Hedgethorne K, Meanwell NA, Matthews SJ (2014) The crystal structure of NSSA domain 1 from genotype 1a reveals new clues to the mechanism of action for dimeric HCV inhibitors. *Protein Sci* 23:723–734
24. Lunde BM, Moore C, Varani G (2007) RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* 8:479–490
25. Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B et al (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* 16:2089–2103
26. Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, Patel DJ (2005) Structural basis for 5'-end-specific recognition of guide RNA by the *A. fulgidus* Piwi protein. *Nature* 434:666–670
27. Ngo JC, Giang K, Chakrabarti S, Ma CT, Huynh N, Hagopian JC, Dorrestein PC, Fu XD, Adams JA, Ghosh G (2008) A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1. *Mol Cell* 29:563–576
28. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156:935–949
29. O'Toole N, Hattori M, Andres C, Iida K, Lurin C, Schmitz-Linneweber C, Sugita M, Small I (2008) On the expansion of the

- pentatricopeptide repeat gene family in plants. *Mol Biol Evol* 25:1120–1128
30. Pinker F, Bonnard G, Gobert A, Gutmann B, Hammani K, Sauter C, Gegenheimer PA, Giege P (2013) PPR proteins shed a new light on RNase P biology. *RNA Biol* 10. doi:10.4161/rna.25273
 31. Ringel R, Sologub M, Morozov YI, Litonin D, Cramer P, Temiakov D (2011) Structure of human mitochondrial RNA polymerase. *Nature* 478:269–273
 32. Small ID, Peeters N (2000) The PPR motif—a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem Sci* 25:46–47
 33. Song JJ, Smith SK, Hannon GJ, Joshua-Tor L (2004) Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* 305:1434–1437
 34. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* 507:62–67
 35. Wang X, McLachlan J, Zamore PD, Hall TM (2002) Modular recognition of RNA by a human pumilio-homology domain. *Cell* 110:501–512
 36. Wang X, Zamore PD, Hall TM (2001) Crystal structure of a Pumilio homology domain. *Mol Cell* 7:855–865
 37. Wang Y, Juranek S, Li H, Sheng G, Tuschl T, Patel DJ (2008) Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* 456:921–926
 38. Wu B, Peisley A, Richards C, Yao H, Zeng X, Lin C, Chu F, Walz T, Hur S (2013) Structural basis for dsRNA recognition, filament formation, and antiviral signal activation by MDA5. *Cell* 152:276–289
 39. Yin P, Li Q, Yan C, Liu Y, Liu J, Yu F, Wang Z, Long J, He J, Wang HW et al (2013) Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* 504:168–171
 40. Zhang Y, Madl T, Bagdiul I, Kern T, Kang HS, Zou P, Mausbacher N, Sieber SA, Kramer A, Sattler M (2013) Structure, phosphorylation and U2AF65 binding of the N-terminal domain of splicing factor 1 during 3'-splice site recognition. *Nucleic Acids Res* 41:1343–1354
 41. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* 509:487–491