**Article**

# KH-RNA interactions: back in the groove.

*Giuseppe Nicastro[1], Ian A. Taylor[1] and Andres Ramos[2,1*]*

*[1]Division of Molecular Structure, MRC National Institute for Medical Research, London, UK*
*[2] Research Department of Structural and Molecular Biology, University College London, London, UK*

**Abstract**

The hnRNP K-homology (KH) domain is a single stranded nucleic acid binding domain that mediates RNA target recognition by a large group of gene regulators. The structure of the KH fold is well characterised and some initial rules for KH-RNA recognition have been drafted. However, recent findings have shown that these rules need to be revisited and have now provided a better understanding of how the domain can recognize a sequence landscape larger than previously thought as well as revealing the diversity of structural expansions to the KH domain. Finally, novel structural and functional data show how multiple KH domains act in a combinatorial fashion to both allow recognition of longer RNA motifs and remodelling of the RNA structure. These advances set the scene for a detailed molecular understanding of KH selection of the cellular targets.

## Introduction

The hnRNP K-homology (KH) domain is a single stranded, sequence specific, nucleic acid binding domain present in proteins that regulate gene expression in eukaryotes and prokaryotes. The domain comprises three α-helices that pack onto the surface of a central antiparallel β-sheet (**Figure 1a**). Eukaryotic type I and prokaryotic type II KH domains share a minimal βααβ core, with two additional α and β elements positioned either C-terminal (type I, eukaryote) or N-terminal (type II, prokaryote) to this core motif **[1-3]** (**Figure 1a**). In all KH domain-nucleic acid structures the nucleic acid backbone interacts with a conserved GxxG loop that links the two helices of the minimal KH core (**Figure 1b**). This orients four nucleic acid bases towards a groove in the protein structure where hydrophobic interactions and a network of mainchain and sidechain hydrogen bonds mediate nucleobase recognition (**Figure 1c**) **[4-6]**. So far, domains with a classical KH fold but lacking the conserved GxxG motif have shown no nucleic acid-binding activity - although they interact with other nucleic acid binding domains and can modulate their RNA binding activity **[7]**.

KH domains are found in arrays of up to 15 repeats (**Figure 1d**) and neighbouring domains can be structurally decoupled or form two-domain structural units. Individual domains often cooperate in the selection of the cellular targets and the combinatorial action of multiple domains allows recognition of longer sequences to increase specificity. Alternatively, longer RNA sequences can be recognised by expanding the classical KH fold with the addition of secondary structure elements and by dimerisation of KH containing proteins, such as the one reported for the STAR family of RNA regulators **[8]**.

A set of rules for KH-nucleic acid recognition has been devised based on the structures of KH domains in complex with RNA and DNA targets and on functional and biochemical information regarding their physiological targets **[9]**. However, recent data suggests that some of these rules should be re-visited. Further, new information has come to light regarding how KH domains can act combinatorially to recognise longer RNA sequences, a necessity to select targets within the large pool of cellular RNAs. Below we review recent key studies on KH-RNA interactions and discuss how these studies have changed our understanding of recognition and of the role played by KH domains in RNA regulation.

## An expanded sequence specificity landscape for KH domains

The structure of several KH domain-nucleic acid complexes including, amongst others, Nova-1 KH3 **[4]**, SF1 **[10]**, hnRNPK **[9]**, FBP **[11]**, and PCBP2 **[12]** indicated that, in the protein-RNA complexes, the two central nucleobases of the tetranucleotide recognition sequence (positions 2 and 3) were either adenine or cytosine, while a pyrimidine was preferred in the first and fourth position. The recognition of a cytosine in position 2 is mediated by hydrophobic contacts or stacking interactions and by two H-bonds between an arginine side chain located on the central β-sheet of the domain and the O2 and N3 groups of the base. Instead, when adenine is selected at position 2 the bulky arginine side chain is replaced by a smaller lysine that better accommodates the purine base (**Figure 2a**). In position 3, the Watson-Crick edge of the nucleobase is recognised specifically by two Hydrogen-

bonds with the backbone amide and carboxyl moieties of same amino acid from the second (third for type II domains) strand of the KH β-sheet. Only an A or a C can make the double hydrogen bond, which was thought to be an intrinsic feature of KH-nucleic acid recognition. Discrimination between A and C can be attributed to an additional direct or water-mediated contact between the nucleobase and a protein side chain on α2 (**Figure 2a**).

Recently, it was shown that the third KH domain of KSRP recognises a G-rich sequence in the precursor of the tumour suppressor Let-7 miRNA **[13]** rather than a canonical A/C rich sequence. This recognition is necessary to enhance the processing of the precursor to the mature miRNA. The structural basis of G-recognition was revealed by the solution structure of the KH3-AGGGU complex **[14,15]**. In KH3 of KSRP the hydrophobic groove is wider than that of other KH-nucleic acid complexes (**Figure 2b**) allowing G bases in position 1, 2 and 3 to enter the groove. Moreover, in positions 2 and 3, guanine is not only tolerated but also selected for. In position 2, lysine 368 makes a specific hydrogen bond with the O6 of Guanine-2. When the lysine is substituted by arginine the specificity of position 2 changes from a purine to a pyrimidine, as predicted, but the RNA binding affinity of the domain remains unchanged **[15]**. In position 3, the broader groove facilitates a shift of the base edge along β2 and the formation of three intermolecular hydrogen bonds between the Watson-Crick edge of the base and the carboxy and amide groups of two residues, isoleucine 356 and phenylalanine 358. A fourth hydrogen bond is formed with the side-chain of glutamine 349, completing a recognition pattern that is very specific, but also different from the canonical one (**Figure 2b**). Although no other structures of KH domains in complex with G-rich sequences are available, in vitro and in cell data have suggested that the fourth KH domains of the protein IMP1 recognises a G-rich RNA sequence **[16]**. Overall, the data show that sequence specific recognition in the KH hydrophobic groove relies both on specific contacts and on the overall shape of the groove, which is governed by the conformation of multiple side chains. Several of the key side chain determinants of recognition have been identified and we expect on-going work to result in further advances in deciphering the contact-specific sequence recognition code of KH domains. A future more complex challenge is to understand how the overall side chain composition defines the shape of the binding-groove and provides a template for nucleobase discrimination.

## Recognition of longer RNA sequences by expanded KH domains

The hydrophobic groove of canonical KH domains can recognize up to four nucleotides. Specific recognition of longer sequences by individual domains is achieved by expanding the classical KH fold with additional secondary structure elements that elongate the RNA recognition surface **[17,18]**. The best studied of these expanded KH folds is the Signal Transduction and Activation of RNA (STAR) fold, which mediates RNA recognition in a family of proteins important to splicing, mRNA shuttling and translational repression (**Figure 3a**) **[8,19-21]**. The STAR domain includes an N-terminal protein dimerisation motif (the QUA1 motif), a central KH domain and a C-terminal motif involved in RNA binding (the QUA2 motif). The structure of the SF1-RNA complex (**Figure 3b**) has shown that QUA2 comprises a long loop followed by an amphipathic helix that folds back to contact the α1 and α3 helices and the GxxG loop of the KH domain. This extends the hydrophobic surface of the KH groove to allow recognition of an additional 2 nucleotides.

Three recent structures, two crystallographic and one in solution, have expanded our understanding of STAR-RNA recognition. The crystal structures of Qk1 and GLD-1

**[22,23]** in complex with the RNA Recognition element (RRE) target RNA, have provided an overall structural description of STAR domain-RNA interactions and shown that direct contacts with the RNA are made exclusively by the KH and QUA2 domains. Nevertheless, QUA1 dimerization increases the RNA binding affinity of the individual KH-QUA2 units within the homodimer, possibly as a result of QUA1-QUA2 contacts that stabilize the KH-QUA2 arrangement **[22,24]**. A major factor that determines the KH-QUA2 structural arrangement is the interaction with RNA. In the absence of RNA, the QUA2 helix is poorly ordered, as shown by the analysis of backbone motions of free Qk1 **[23,25]**. An important role of RNA binding in shaping the protein interaction surface is also suggested by the finding that, despite a network of conserved QUA2-RNA interactions, the orientation of the QUA2 helix varies between two complexes of GLD1 with RNA targets of different length **[22,23]**. In the solution structure of the GLD-1-RNA complex three nucleotides rather than two make contact with the QUA2 motif and the nucleobase of the third nucleotide docks into a shallow pocket in the QUA2. It has been proposed that this additional contact results in an approximate 10° change in the orientation of the QUA2 helix with respect to the X-ray structure. An important open question is whether the structural flexibility observed in the arrangement of the QUA2 motif is related to a functional flexibility in the recognition of different of RNA targets.

The GLD-1 and Qk1 proteins are dimers in the cell and indeed recognise bi-partite RNA sequences. The structure of a (dimeric) QUA1-KH-QUA2 STAR protein in complex with a complete bi-partite RNA sequence will help clarify how dimerisation is related to the arrangement of the two recognition sequences. Interestingly, the expanded KH domain of FILIA, a protein important in RNA regulation during mouse oogenesis, also creates a strong inter-molecular dimer **[26]**. However, this domain has a different structural arrangement with respect to the STAR domain (Figure 3B) and its specific target sequence is yet to be identified.

Finally, structural extensions have also been reported for type II KH domains. In the ERA-RNA complex an extended type II KH domain interacts with nine nucleotides - five of which make contacts outside the hydrophobic groove. The additional contacts however are made within the core KH fold and not with the structural extension (**Figure 3B**). The structure of the ERA-RNA complex shows that the KH fold can interact with RNA using residues outside the hydrophobic groove and highlights the differences between eukaryotic and prokaryotic KH domains **[27]**.

## Combinatorial binding and architectural role by multiple KH domains

As observed for other RNA binding domains, combinatorial binding of multiple KH domains within the same protein is often key to high affinity and high specificity interaction with the RNA target **[28,29]**. Individual KH domains bind their short single stranded RNA target sequences with dissociation constants in the micromolar range ($10^{-6}$ to $10^{-4}$ M), while the dissociation constant of multi-KH protein–RNA complexes is in the low nanomolar range **[14,28,29,30]**. The contribution of individual domains to RNA binding and protein function has been evaluated using domain deletions and, more recently, conservative double mutations in the GxxG loop that eliminate RNA binding of individual domains **[30]**. Additionally, structural and biophysical studies have characterized inter-domain interactions. In several instances, flexible linkers join individual domains establishing some degree of domain coupling in the binding, while also allowing the protein to adapt the inter-domain arrangement to different RNA (and DNA) targets. However, other arrangements have been reported where stable inter-domain association of neighbouring KH domains produces a continuous binding site proposed to play a role in the folding of the RNA targets.

The most common inter-domain arrangement in di-domain KH units is a side by side anti-parallel configuration where the three β-strands of each KH orient to form an extended six-stranded β-sheet (**Figure 4a**). This arrangement is stabilised by contacts between the two KH α3 helices. NMR and X-ray crystallography studies of the KH di-domains of the Nova-1 **[31]**, IMP1 **[16,32]** and PCBP2 **[33,34]** proteins, indicate that the dimer does not undergo significant conformational rearrangements upon RNA (or DNA) binding. Instead the two domains form a rigid structural unit and an architectural function has been proposed where binding of the same nucleic acid chain to the anti-parallel hydrophobic grooves of each KH domains results in a 180 ° looping of the RNA **[16,32]**. In order to validate this observation in a cellular context, the distribution and 5'-to-3' orientation of the cognate sequences of IMP1 KH3 and KH4 was analysed in a set of physiological targets **[16]**. The analysis showed that both RNA sequences were present in the majority of targets at a distance within the boundaries expected from the structure of the di-domain and *in vitro* binding data. However, the sequences were found both in a 5-3' and a 3'-5' arrangement, in equal shares, indicating that the RNA chain can loop on either side of the di-domain.

Other inter-domain KH-KH arrangements have also been reported implying that different architectural constraints are imposed on the RNA chain. For example, the orthogonal arrangement of the KH2-KH3 domains of KSRP **[35]** (**Figure 4a**) has been proposed to bend the RNA chains with ~90 ° angle while the NusA binding of the type II KH di-domain to one contiguous stretch of 11 nucleotides **[36]** (**Figure 4a**) also results in a change in direction of the RNA but with no requirement for looping. Other inter-domain arrangements have been reported, for example for the two KH domains of the Fragile X mental retardation protein (FMRP) but it is unclear whether both domains bind to specific RNA targets **[37,38]**.

KH-domains that do not contain the conserved GxxG motif – and do not interact with RNA – can assemble in multi-domain units where more than one KH-KH interface is observed, exemplified by the structure of the type I KH tetra-domain unit of Bicaudal-C (Bic-C) (**Figure 4b**) **[39]**. Type II KH domains lacking the GxxG motif have also been recently reported in proteins of the metallo-β-lactamase superfamily. As seen for type I KH, type II KH domains lacking the GxxG loop also do not participate in the recognition of the RNA targets **[40-42]**.

## Perspectives

It is an exciting time for the study of KH domains. Recent findings have highlighted their capacity to recognize a diverse range of sequences and to expand and combine sequence recognition of individual domains to select longer RNA targets. In addition, di-domain KH units likely play a general role in re-shaping the structure of the bound RNA. It is now important to understand how the architectural role of di-domain units relates to the structure and function of large protein-RNA particles. A further set of questions concerns the complex relationship between nucleobase specificity of individual KH units, the use of multiple KH domains and the targeting of specific RNAs in the cell. A combined structural and cell biology approach, including the genomic analysis of the targets, will be required to understand not only KH recognition of high affinity sequences but, also target selection in a cellular context where binding to intermediate affinity targets is functionally relevant for many KH domains.

## Acknowledgements

The research of A. R. is supported by the UK Medical Research Council grant MC_PC_13051 and by University College London. The research of I.A.T. and G.N. is supported by UK Medical Research Council grant U117565647.

**References and recommended reading**


[1] Musco G, Stier G, Joseph C, Castiglione Morelli MA, Nilges M, Gibson TJ, Pastore A: **Three-dimensional structure and stability of the KH domain: molecular insights into the fragile X syndrome.** *Cell* 1996, **85**:237-245.

[2] Grishin NV: **KH domain: one motif, two folds.** *Nucleic Acids Res* 2001, **29**:638-643.

[3] Valverde R, Edwards L, Regan L: **Structure and function of KH domains**. *FEBS J* 2008, **275**:2712-2726.

** [4] Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, Burley SK: **Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome.** *Cell* 2000, **100**:323-332.
The first structure of a KH domain in complex with the RNA target reveals the core features of KH-RNA interaction.

[5] Jensen KB, Musunuru K, Lewis HA, Burley SK, Darnell RB: **The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain.** *Proc Natl Acad Sci U S A* 2000, **97**:5740-5745.

[6] Yu Q, Ye W, Jiang C, Luo R, Chen HF: **Specific Recognition Mechanism between RNA and the KH3 Domain of Nova-2 Protein.** *J Phys Chem B*. 2014 Oct 21, DOI: 10.1021/jp5079289.

[7] Oddone A, Lorentzen E, Basquin J, Gasch A, Rybin V, Conti E, Sattler M: Structural and biochemical characterization of the yeast exosome component Rrp40. *EMBO Rep* 2007, **8**:63-69.

[8] Ryder SP, Massi F:**Insights into the structural basis of RNA recognition by STAR domain proteins.** *Adv Exp Med Biol* 2010, **693**:37-53. Review.

[9] Backe PH, Messias AC, Ravelli RB, Sattler M, Cusack S: **X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids.** *Structure*. 2005, **13**:1055-1067.

** [10] Liu Z, Luyten I, Bottomley MJ, Messias AC, Houngninou-Molango S, Sprangers R, Zanier K, Krämer A, Sattler M: **Structural basis for recognition of the intron branch site RNA by splicing factor 1.** *Science* 2001, **294**:1098-1102.
The SF1 QUA2 structural element extends the core RNA interaction surface of the KH domain and is pivotal to the higher affinity and specificity recognition of the branch site sequence.

* [11] Braddock DT, Louis JM, Baber JL, Levens D, Clore GM: **Structure and dynamics of KH domains from FBP bound to single-stranded DNA.** *Nature* 2002, **415**:1051-1056.
The structure highlights how the long flexible linkers often present between KH domains allow the targeting of non-adjacent RNA sequences.

[12] Du Z, Lee JK, Tjhen R, Li S, Pan H, Stroud RM, James TL: **Crystal structure of the first KH domain of human poly(C)-binding protein-2 in complex with a C-rich strand of human telomeric DNA at 1.7 A.** *J Biol Chem* 2005, **280**:38823-38830.

[13] Trabucchi M, Briata P, Garcia-Mayoral M, Haase AD, Filipowicz W, Ramos A, Gherzi R, Rosenfeld MG: **The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs.** *Nature* 2009, **459**:1010-1014.

[14] García-Mayoral MF, Díaz-Moreno I, Hollingworth D, Ramos A: **The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets.** *Nucleic Acids Res* 2008, **36**:5290-5296.

** [15] Nicastro G, García-Mayoral MF, Hollingworth D, Kelly G, Martin SR, Briata P, Gherzi R, Ramos A: **Noncanonical G recognition mediates KSRP regulation of let-7 biogenesis.** *Nat Struct Mol Biol* 2012, **19**:1282-1286.
This work redefines our understanding of how the sequence specificity of a KH domain is determined and how it can be changed, *in vitro* and in the cell.

* [16] Patel VL, Mitra S, Harris R, Buxbaum AR, Lionnet T, Brenowitz M, Girvin M, Levy M, Almo SC, Singer RH, Chao JA: **Spatial arrangement of an RNA zipcode identifies mRNAs under post-transcriptional control.** *Genes Dev* 2012, **26**:43-53.
The 5'-3'positioning of the target sequences of IMP1 KH3 and KH4 in the ensemble of physiological targets is related to RNA looping and to the structure of the two-domain unit.

[17] García-Mayoral MF, Hollingworth D, Masino L, Díaz-Moreno I, Kelly G, Gherzi R, Chou CF, Chen CY, Ramos A: **The structure of the C-terminal KH domains of KSRP reveals a noncanonical motif important for mRNA degradation.** *Structure* 2007, **15**:485-498.

[18] Wong AG, McBurney KL, Thompson KJ, Stickney LM, Mackie GA: **S1 and KH domains of polynucleotide phosphorylase determine the efficiency of RNA binding and autoregulation.** *J Bacteriol* 2013, **195**:2021-2031.

[19] Feng Y, Bankston A: **The star family member QKI and cell signaling.** *Adv Exp Med Biol* 2010, **693**:25-36.

[20] Beuck C, Qu S, Fagg WS, Ares M Jr, Williamson JR: **Structural analysis of the quaking homodimerization interface.** *J Mol Biol* 2012, **423**:766-781.

[21] Brümmer A, Kishore S, Subasic D, Hengartner M, Zavolan M: **Modeling the binding specificity of the RNA-binding protein GLD-1 suggests a function of coding region-located sites in translational repression.** *RNA* 2013, **19**:1317-26.

* [22] Teplova M, Hafner M, Teplov D, Essig K, Tuschl T, Patel DJ: **Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites.** *Genes Dev* 2013, **27**:928-940.
This study provides a first and important structural insight on how the QUA1 dimerisation motif of STAR proteins contacts the proteins' RNA binding units.

* [23] Daubner GM, Brümmer A, Tocchini C, Gerhardy S, Ciosk R, Zavolan M, Allain FH: **Structural and functional implications of the QUA2 domain on RNA recognition by GLD-1.** *Nucleic Acids Res* 2014, **42**:8092-8105.
In a related study, the details of QUA2-RNA recognition are explored to connect the orientation of the QUA2 helix with differences in the RNA target.

[24] Chen T, Damaj BB, Herrera C, Lasko P, Richard S: **Self-association of the single-KH-domain family members Sam68, GRP33, GLD-1, and Qk1: role of the KH domain.** *Mol Cell Biol* 1997, **17**:5707-5718.

[25] Maguire ML, Guler-Gane G, Nietlispach D, Raine AR, Zorn AM, Standart N, Broadhurst RW: **Solution structure and backbone dynamics of the KH-QUA2 region of the Xenopus STAR/GSG quaking protein.** *J Mol Biol* 2005, **348**:265-279.

[26] Wang J, Xu M, Zhu K, Li L, Liu X: **The N-terminus of FILIA forms an atypical KH domain with a unique extension involved in interaction with RNA.** *PLoS One* 2012, **7**:e30209.

[27] Tu C, Zhou X, Tropea JE, Austin BP, Waugh DS, Court DL, Ji X: **Structure of ERA in complex with the 3' end of 16S rRNA: implications for ribosome biogenesis.** *Proc Natl Acad Sci U S A* 2009, **106**:14843-14848.

[28] Lunde BM, Moore C, Varani G: **RNA-binding proteins: modular design for efficient function.** *Nat Rev Mol Cell Biol* 2007, **8**:479-490.

[29] Zhang C, Lee KY, Swanson MS, Darnell RB: **Prediction of clustered RNA-binding protein motif sites in the mammalian genome.** *Nucleic Acids Res* 2013, 41:6793-6807.

* [30] Hollingworth D, Candel AM, Nicastro G, Martin SR, Briata P, Gherzi R, Ramos A: **KH domains with impaired nucleic acid binding as a tool for functional analysis.** *Nucleic Acids Res* 2012, **40**:6873-6886.
This manuscript describes a non-invasive strategy for assessing the role of individual KH domains in RNA binding by the full-length protein and its relationship with protein function.

* [31] Teplova M, Malinina L, Darnell JC, Song J, Lu M, Abagyan R, Musunuru K, Teplov A, Burley SK, Darnell RB, Patel DJ: **Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1.** *Structure* 2011, **19**:930-944.
The two KH in this di-domain structural unit assume an antiparallel orientation with both nucleic acid binding grooves available for RNA binding.

[32] Chao JA, Patskovsky Y, Patel V, Levy M, Almo SC, Singer RH: **ZBP1 recognition of beta-actin zipcode induces RNA looping.** *Genes Dev* 2010, **24**:148-158.

[33] Du Z, Lee JK, Fenn S, Tjhen R, Stroud RM, James TL: **X-ray crystallographic and NMR studies of protein-protein and protein-nucleic acid interactions involving the KH domains from human poly(C)-binding protein-2.** *RNA* 2007, **13**:1043-1051.

[34] Du Z, Fenn S, Tjhen R, James TL: **Structure of a construct of a human poly(C)-binding protein containing the first and second KH domains reveals insights into its regulatory mechanisms.** *J Biol Chem* 2008, **283**:28757-28766.

* [35] Díaz-Moreno I, Hollingworth D, Kelly G, Martin S, García-Mayoral M, Briata P, Gherzi R, Ramos A: **Orientation of the central domains of KSRP and its implications for the interaction with the RNA targets.** *Nucleic Acids Res* 2010, **38**:5193-5205.

The relative orientation of the two KH domains in this structure implies a 90° change in the direction of the RNA chain upon binding.

** [36] Beuth B, Pennell S, Arnvig KB, Martin SR, Taylor IA: **Structure of a Mycobacterium tuberculosis NusA-RNA complex.** *EMBO J* 2005, **24**:3576-3587.
The tight packing of the NusA type I KH1 and KH2 domains creates a two-domain continuous nucleic acid binding groove that interacts with RNA with high affinity and specificity and remodels the E.coli anti-terminator element. This study provides a direct structural insight in the architectural role of a multi-KH unit.

* [37] Valverde R, Pozdnyakova I, Kajander T, Venkatraman J, Regan L: **Fragile X mental retardation syndrome: structure of the KH1-KH2 domains of fragile X mental retardation protein.** *Structure* 2007, **15**:1090-1098.
The two KH domains of the protein FMRP bind ssRNA with low affinity, but this structure shows that both hydrophobic grooves are in a canonical conformation and potentially available for RNA binding.

[38] Ascano M Jr, Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M, Williams Z, Ohler U, Tuschl T**: FMRP targets distinct mRNA sequence elements to regulate protein expression.** *Nature* 2012, **492**:382-386.

[39] Nakel K, Hartung SA, Bonneau F, Eckmann CR, Conti E: **Four KH domains of the C. elegans Bicaudal-C ortholog GLD-3 form a globular structural platform.** *RNA* 2010, **16**:2058-2067.

[40] Nishida Y, Ishikawa H, Baba S, Nakagawa N, Kuramitsu S, Masui R: **Crystal structure of an archaeal cleavage and polyadenylation specificity factor subunit from Pyrococcus horikoshii.** *Proteins* 2010, **78**:2395-2398.

[41] Silva AP, Chechik M, Byrne RT, Waterman DG, Ng CL, Dodson EJ, Koonin EV, Antson AA, Smits C: **Structure and activity of a novel archaeal β-CASP protein with N-terminal KH domains.** *Structure* 2011, **19**:622-632.

[42] Mir-Montazeri B, Ammelburg M, Forouzan D, Lupas AN, Hartmann MD: **Crystal structure of a dimeric archaeal cleavage and polyadenylation specificity factor.** *J Struct Biol* 2011, **173**:191-195.

## Figure Legends

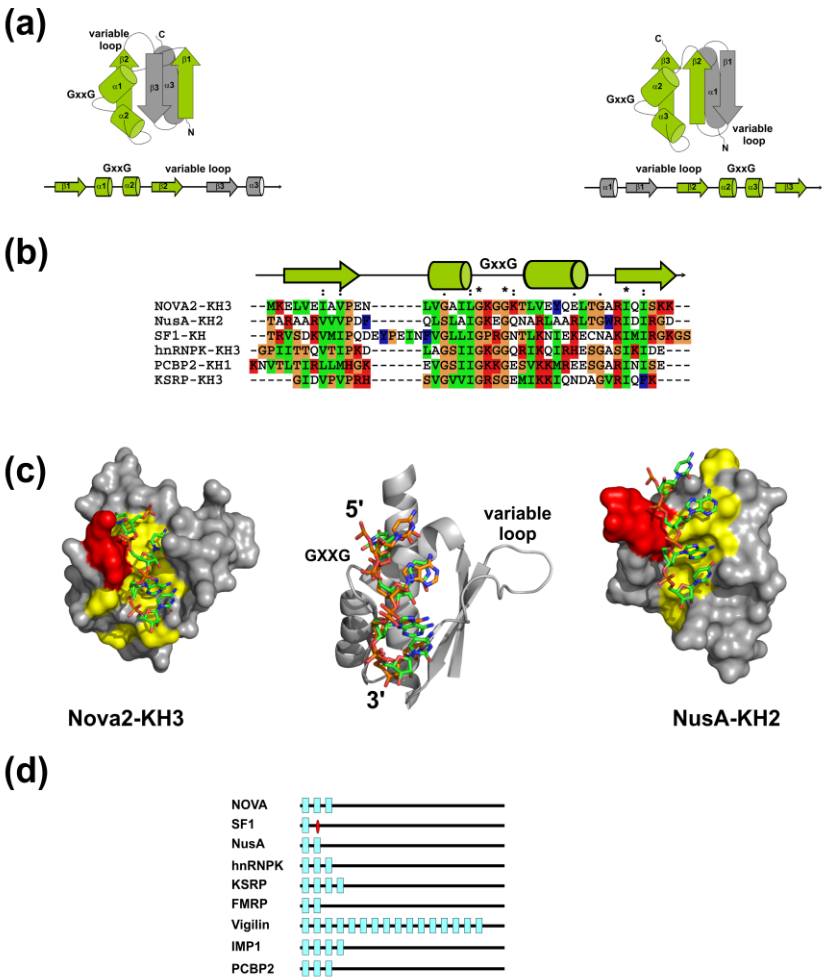



**Figure 1.**

**KH domain structure and global features of KH-RNA interactions.**
(**a**) Cartoon representation of the secondary structure and fold of type I and type II KH domains. Core secondary structure elements are shown in green, additional elements in grey. (**b**) Sequence alignment of the minimal βααβ core of type I and type II KH domains. Secondary structure elements are drawn above the alignment. (**c**) Molecular surfaces of the Type I Nova2 KH3 (left) and Type II NusA KH2 (right) in complex with RNA. Hydrophobic residues are shown in yellow, while the conserved GxxG loop is highlighted in red. The four nucleotides recognized in the KH3 hydrophobic groove are displayed as sticks, color-coded by atom type. In the middle is a ribbon representation of the Nova-2 KH3-RNA complex. The four nucleotides are displayed by sticks, colour-coded by atom type, with carbon atoms in yellow. The equivalent nucleotides in the superimposed structure of NusA-KH2-RNA complex are also similarly displayed, but with carbon atoms in orange. (**d**) Schematic representation of the multiplicity and arrangement KH domains (light blue boxes) in a selection of eukaryotic proteins.

**Figure 2.**

**RNA sequence recognition in the KH domain hydrophobic groove.**
(**a**) KH domain canonical and non-canonical (far right) recognition of nucleobases in positions 2 and 3 of the target tetra-nucleotide. The two nucleotides are displayed by sticks, colour-coded by atom type, with carbon atoms in green, while interacting amino acid are similarly colour coded but with carbon atoms in magenta. Hydrogen bonds are shown as black dashed lines and the red sphere represents a water molecule. (**b**) Molecular surfaces of the four RNA-bound KH domains shown above. Hydrophobic residues are coloured red. The bound nucleic acids are drawn using a stick representation and are coloured yellow.

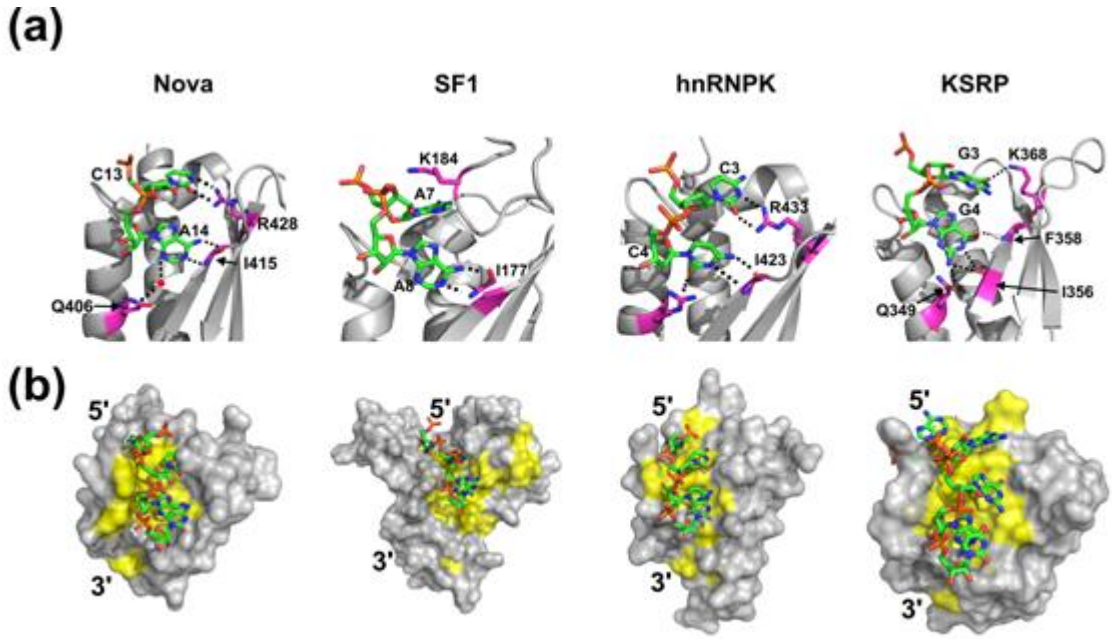
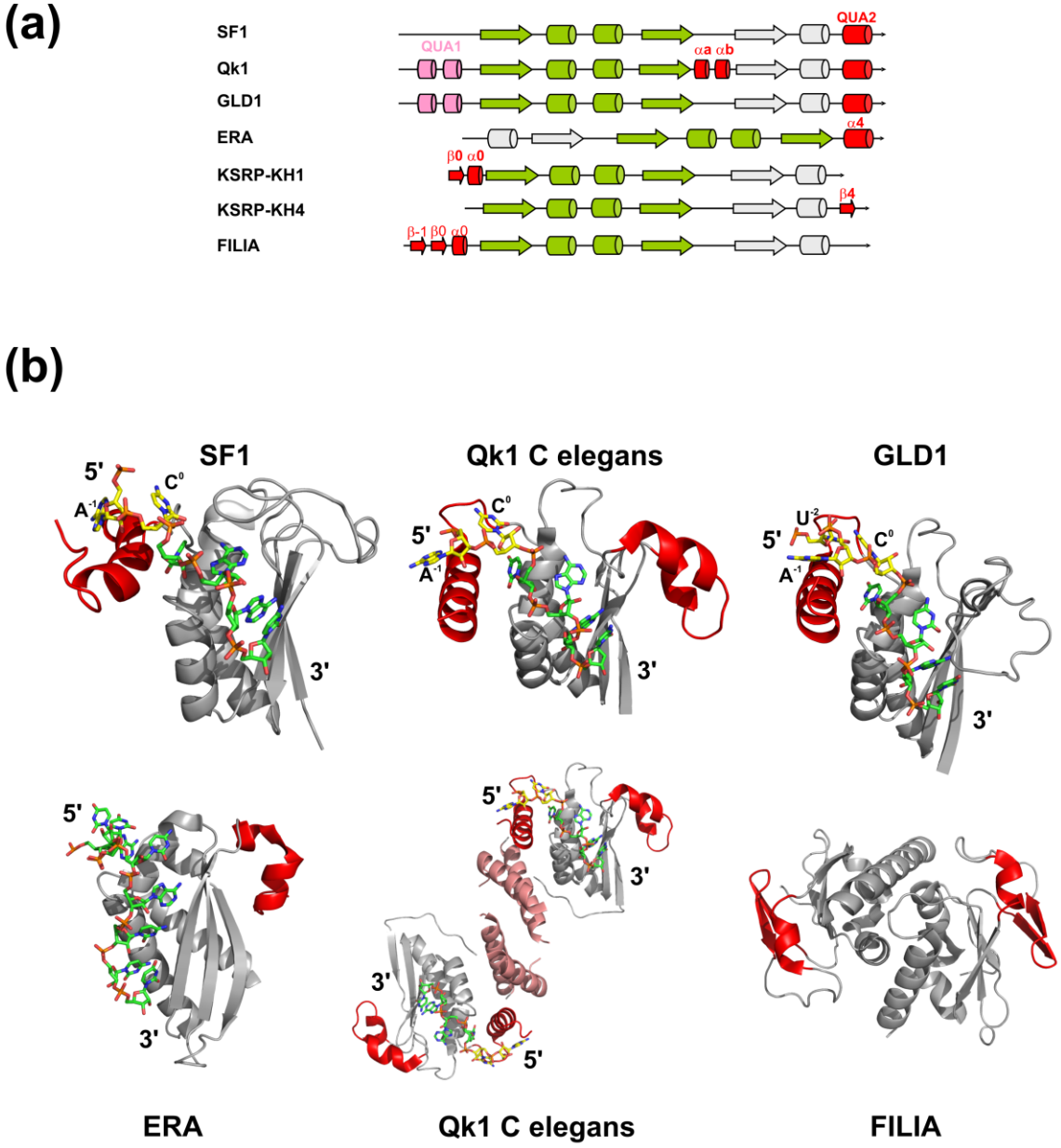


**Figure 3. Extended KH domains: structure, dimerisation and RNA binding.** (**a**) Schematic of the arrangement of secondary structure elements of the extended KH domains. Elements belonging to the common KH βααβ motif are shown in green, the type I/type II specific secondary structures are shown in white and the secondary structure elements that extended KH domains are shown in pink and red. (**b**) Ribbon representations of expanded KH domains, either free or in complex with RNA. The core KH domain is in grey and the RNA is displayed as sticks, colour coded by atom type with the carbon atoms in green (for the four nucleotides bound in positions 1-to-4 of the hydrophobic groove) or yellow (for the additional nucleotides that make contact with the QUA2 element). The nucleotides making contact with the QUA2 element are numbered starting from the nucleobase 5' to position 1 of the hydrophobic groove (position 0). The RNA bound to the ERA protein has all of the carbon atoms in green, as the contacts made are different than in canonical KH domains. Nucleotides that are not making contacts with the protein are not displayed in any of the structures. Additional secondary structure elements are colour-coded as in **a**, the QUA1 dimerisation region of Qk1 in pink and QUA2 extra RNA binding elements in red.



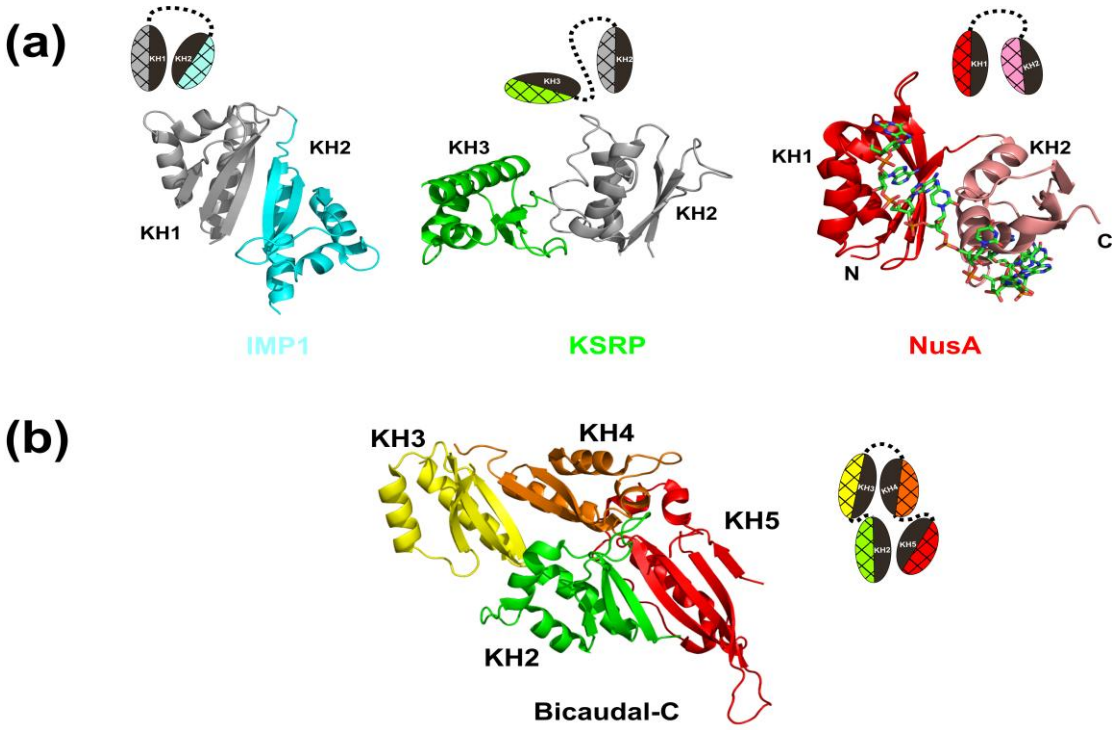



**Figure 4.**

**Multi-KH structural units.**
Ribbon representation of the multi-KH units of (**a**) Bicaudal-C and (**b**) IMP1, KSRP and NusA-RNA. The NusA bound RNA is displayed as sticks, colour coded by atom type. For each structure a schematic arrangement of the two domains is drawn, with the β-sheet face shaded black area and the helical face hatched, colour coded as the corresponding domain.