8092–8105 *Nucleic Acids Research, 2014, Vol. 42, No. 12* *Published online 16 May 2014*
doi: 10.1093/nar/gku445

Structural and functional implications of the QUA2 domain on RNA recognition by GLD-1

Gerrit M. Daubner[1], Anneke Brümmer[2,†], Cristina Tocchini[3,†], Stefan Gerhardy[1], Rafal Ciosk[3], Mihaela Zavolan[2] and Frédéric H.-T. Allain[1,*]

[1]Institute of Molecular Biology and Biophysics, Eidgenössische Technische Hochschule (ETH) Zürich, 8093 Zürich, Switzerland, [2]Biozentrum, University of Basel, 4056 Basel, Switzerland and [3]Friedrich Miescher Institute for Biomedical Research, 4002 Basel, Switzerland

Received March 23, 2014; Revised May 4, 2014; Accepted May 6, 2014

ABSTRACT

The STAR family comprises ribonucleic acid (RNA)-binding proteins that play key roles in RNA-regulatory processes. RNA recognition is achieved by a KH domain with an additional α-helix (QUA2) that seems to extend the RNA-binding surface to six nucleotides for SF1 (*Homo sapiens*) and seven nucleotides for GLD-1 (*Caenorhabditis elegans*). To understand the structural basis of this probable difference in specificity, we determined the solution structure of GLD-1 KH-QUA2 with the complete consensus sequence identified in the *tra-2* gene. Compared to SF1, the GLD-1 KH-QUA2 interface adopts a different conformation resulting indeed in an additional sequence-specific binding pocket for a uracil at the 5′end. The functional relevance of this binding pocket is emphasized by our bioinformatics analysis showing that GLD-1 binding sites with this 5′end uracil are more predictive for the functional response of the messenger RNAs to *gld-1* knockout. We further reveal the importance of the KH-QUA2 interface *in vitro* and that its alteration *in vivo* affects the level of translational repression dependent on the sequence of the GLD-1 binding motif. In conclusion, we demonstrate that the QUA2 domain distinguishes GLD-1 from other members of the STAR family and contributes more generally to the modulation of RNA-binding affinity and specificity of KH domain containing proteins.

INTRODUCTION

Ribonucleic acid (RNA)-binding proteins are essential for cellular physiology and especially for messenger RNA (mRNA) metabolism. A multitude of different RNA-binding domains (RBDs) exist and the most abundantly represented in proteins are the RNA recognition motif (RRM) (1,2), followed by the K homology (KH) domain (3). An extended version of the latter is found within the signal transduction and activation of RNA (STAR) protein family. In a simplified version of the phylogenetic tree, the STAR protein family consists of the three branches SF1, Sam68 and the Quaking (Qk) related proteins, comprised of QK1, HOW (*Drosophila melanogaster*) and GLD-1 (*Caenorhabditis elegans*) (4). All its members play key roles in all stages of RNA metabolism and in developmental processes across species (4). Consequently, disease associations have been reported, particularly in cancer (5,6) and neurological diseases such as human inherited ataxia, multiple sclerosis or schizophrenia (7,8). To understand their multitude of functions and role in disease, it is necessary to decipher the molecular basis of their RNA recognition.

STAR protein family members share a very high sequence similarity in their extended KH domain (KH-QUA2) that binds RNA. Structural characterization of this domain was first accomplished in solution for SF1 with RNA (9), followed by the isolated KH-QUA2 domain of QK1 (*Xenopus laevis*) (10). The conventional type 1 KH domain has a three-dimensional $\beta_1\alpha_1\alpha_2\beta_2\beta_3\alpha_3$ topology with a conserved GxxG loop between the first two α-helices. RNA binding is accomplished by a cleft that can accommodate four nucleotides, mainly recognized by van der Waals forces, hydrophobic and electrostatic interactions (3). The fourth α-helix (QUA2) enlarges this RNA-binding surface by two to three nucleotides (9). In addition, all members, except SF1, also contain an N-terminal dimerization domain (QUA1) (4). Structures of the isolated QUA1 domains of GLD-1, Qk1 and Sam68 show that dimerization is constituted by a helix-turn-helix motif that is in a perpendicular orientation to the other protomer and that way forms a hydrophobic zipper stabilizing the dimer (11–13). The composition of the complete RNA-binding domain was finally visualized within recent crystal structures of QK1 and GLD-1 in complex with RNA (14).

*To whom correspondence should be addressed. Tel: +41 44 633 39 40; Fax: +41 44 633 12 94; Email: allain@mol.biol.ethz.ch
†The authors wish it to be known that, in their opinion, the second and third authors should be considered as Joint Second Authors.

GLD-1 is a germline-specific STAR protein and essential to control germline development in *C. elegans* (15). In the germline, regulation of gene expression is achieved to a great extent through post-transcriptional mechanisms, particularly acting at the level of translation. This is emphasized by the observation that most genes expressed in the oogenic germline are regulated through their 3′ untranslated region (UTR) rather than at the promoter level (16). Furthermore, the temporally and spatially tightly regulated germline contains approximately four times the amount of RNA-binding proteins than somatic cells (17). Since the germline-specific translational repressor GLD-1 associates with a substantial proportion of germline mRNAs (18), it is seen as the key germline regulatory factor. It regulates the spermatogenesis-to-oogenesis switch (19), the mitosis-to-meiosis decision (20), promotes meiotic prophase progression, maintains germ cell identity (21) and inhibits apoptosis (22). GLD-1 regulates this broad range of functions by adopting a dual role in both translational repression and transcript stabilization (18).

The importance of RNA binding for the function of GLD-1 is emphasized by several critical mutations within the KH and QUA2 domain leading to severe phenotypes *in vivo* (23). To gain insights into RNA binding by GLD-1, the consensus sequence 5′-(U>C>G/A)ACU(C/A)A(C/U)-3′ of GLD-1 was initially dissected by gel electrophoretic mobility shift assays (EMSAs) through mutations of a binding site identified within the 3′UTR of the *tra-2* target gene (24). Subsequently, a quantitative RNA code to predict GLD-1 binding motifs (GBMs) was proposed based on RIP-Chip combined with fluorescence polarization experiments (25). Building upon this high-throughput study, a second code appended machine learning and biophysical modeling to deduce mRNA target selection by GLD-1 (25,26). Altogether, the binding motifs inferred in these studies suggest that GLD-1 might specifically recognize seven nucleotides and thus would seem to bind one additional nucleotide at the 5′end compared to the human family member SF1 (9). Functional data further indicate that there is a strong correlation between the strength of GLD-1 binding and its degree of target regulation (25).

Recently, the crystal structure of the GLD-1 STAR domain was solved in complex with the 5′-CUAAC(AA)-3′ RNA (14). This structure provides valuable insights into the molecular recognition of the RNA by the KH domain as well as into the molecular basis of GLD-1 dimerization. However, this study may provide limited characterization of RNA recognition by the QUA2 domain, because the sequence that was used does not cover the 5′end of the consensus motif derived in the above-mentioned high-throughput studies. Thus, we set to investigate GLD-1 binding to a longer binding site with the goal to determine how RNA binding is achieved at the 5′end.

Here, we present the solution structure of the extended KH domain (KH-QUA2) of GLD-1 bound to a binding site comprising the complete consensus sequence embedded in the 3′UTR of *tra-2* (19). Our structure shows an extensive involvement of the QUA2 domain in RNA recognition and reveals additional binding pockets for two nucleotides and in particular for a uracil at the 5′end. While this binding pocket for a uracil appears conserved within the Qk-related branch of the STAR family members, it is not present in SF1 due to the different arrangement of the SF1 KH-QUA2 interface. We further demonstrate with bioinformatics approaches that uracil is the preferred nucleotide at the first position of predicted GLD-1 binding sites and that this nucleotide is functionally important, since these binding sites are most predictive for RNA stabilization and translational repression. Finally, we dissect the KH-QUA2 interface of GLD-1 and illustrate its importance for RNA binding both *in vitro* and *in vivo*. This allows us to draw conclusions about the role of the QUA2 domain in modulating RNA-binding specificity and affinity of the whole STAR protein family.

MATERIALS AND METHODS

Protein and RNA preparation

We cloned the open reading frame (ORF) of *gld-1* KH-QUA2 (aa 195–336) by using restriction sites Nde1 and Sap1 into the pTYB1 vector (NEB) containing a C-terminal Intein-Tag. The protein was overexpressed at 20°C for 22 h in minimal M9 medium (1 gl^{-1} ^{15}N-NH_4Cl, 2 gl^{-1} ^{13}C-glucose) after induction with 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Purification via the IMPACT system (NEB) required only one single chromatographic step using chitin beads. The self-cleavage activity of the intein was induced by on-column incubation over night with 50 mM dithiothreitol (DTT) and resulted in the protein without any additional residues. The sample was dialyzed against the final buffer and then passed over a Superdex75 (GE Healthcare). The final buffer was optimized by differential scanning fluorimetry (DSF) and nuclear magnetic resonance (NMR) spectroscopy (50 mM NaCl, 20 mM Na_2HPO_4/ NaH_2PO_4, 3 mM DTT, pH = 6.5). The protein could be concentrated to 0.45 mM at 4°C with a 10-kDa molecular mass cutoff membrane. The construct was temperature sensitive and was only stable between 4°C and 20°C. Therefore, the concentration had to be decreased to ∼0.2 mM for NMR measurements at 30°C.

RNA was purchased from Dharmacon, deprotected according to the manufacturer's protocol, lyophilized and resuspended in the final buffer.

NMR samples of protein–RNA complexes were prepared at a ratio of 1:1 in a volume of 250 μl. In contrast to the free protein, the complex was stable at a concentration of 0.45 mM at 30°C.

NMR measurements and resonance assignments

All the measurements were conducted in the same buffer (50 mM NaCl, 20 mM Na_2HPO_4/ NaH_2PO_4, 3 mM DTT, pH = 6.5) at a temperature of 30°C. We recorded our experiments on Bruker AVIII 500 MHz, AVIII 600 MHz, AVIII 700 MHz, AVIII 750 MHz and AVIII HD 900 MHz spectrometers. Except for the 750 MHz, all spectrometers were equipped with cryo-probes.

We used for the backbone, aliphatic and aromatic side-chain assignments: 2D [^{15}N-^{1}H] HSQC, 2D [^{13}C-^{1}H] HSQC, 3D HNCA, 3D HN(CO)CA, 3D HNCO, 3D HN(CA)CO, 3D CBCA(CO)NH, 3D HC(C)H TOCSY, 3D (H)CCH TOCSY, 15N-resolved and 13C-resolved [^{1}H,^{1}H] NOESY, all recorded in H_2O (27).

For the stereospecific assignment of isopropyl groups of valine and leucine we used a constant time 2D [^{13}C-^{1}H] HSQC with a 10% ^{13}C-labeled sample (28). The tautomeric form of the histidine imidazole side chain was determined through a long range 2D [^{15}N-^{1}H] HSQC (29). Protein backbone φ and ψ constraints were predicted by talos+ (30) and as additional criteria we calculated the φ angle from homonuclear $^{3}J_{HNH\alpha}$ coupling constants, measured by 3D HNHA (31). Residual dipolar couplings (RDCs) for H_N-N were measured with a spin state-selective 2D [^{15}N-^{1}H] HSQC (32) with partial alignment induced by 14 mg/ml Pf1 phage (strain LP11–92, ASLA) in the final buffer. The heteronuclear {^{1}H}-^{15}N values were measured as proposed employing water flip-back pulse (33).

To assign the resonances of the unlabeled RNA, we recorded: 2D [^{1}H-^{1}H] TOCSY, 2D [^{1}H-^{1}H] NOESY, 2D F1-filtered F2-filtered [^{1}H-^{1}H] NOESY and natural abundance 2D [^{13}C-^{1}H] HSQC, all in D_2O (34). Intermolecular nuclear overhauser effects (NOEs) were obtained by using a 2D [^{1}H-^{1}H] NOESY, 2D F2 filtered [^{1}H-^{1}H] NOESY and 3D F1-filtered F2-edited [^{1}H-^{1}H] NOESY (35), all recorded in D_2O. We used a mixing time of 120 and 150 ms for NOESY spectra, 23 ms for 3D TOCSY spectra and 6, 30 and 50 ms for 2D TOCSY spectra.

Structure calculation and refinement

Data processing was performed with topspin 3.0 (Bruker) and for data evaluation we applied sparky 3.114 (36). To automatically generate peak lists from 2D [^{1}H,^{1}H] NOESY, ^{15}N-resolved and ^{13}C-resolved [^{1}H,^{1}H] NOESY spectra, we used the AtnosCandid software (37,38). After manually refinement of each peak list, intramolecular NOE distance constraints were automatically assigned through seven cycles using CYANA 'noeassign' (38). We included additional hydrogen-bond constraints derived from hydrogen-deuterium exchange experiments on the amide protons. Furthermore, backbone torsion angle constraints from talos+ (30) were only added, if the phi value was consistent with the one calculated from the measured homonuclear $^{3}J(H_NH_\alpha)$ coupling constants. Intramolecular RNA and intermolecular distance restraints were manually assigned and calibrated based on fixed inter-atomic distances. The RDC restraints were included from the beginning with an estimated error of 0.5 Hz and 1 Hz, respectively. The weighting factor was 0.025 for CYANA 'noeassign' and 0.05 for the subsequent CYANA 3.0 calculation. Starting from random structures, 250 preliminary structures were calculated and the 50 structures having the lowest target function were selected for further refinement. This was achieved by a restrained simulated annealing run in implicit water with the SANDER module of AMBER 9 (39) using the ff99SB force field (40,41). The final best 20 structures were selected based on lowest energy and NOE violations. Originally ranked 15, the most representative structure used for all figures in this study is on position 1 in the final pdb. We analyzed the structure with the program suit iCING (42), which includes PROCHECK (43) and WHAT IF (44). Analysis of RDCs was carried out with PALES (45). The electrostatic potential was calculated using PDB2PQR (AMBER force field) and the APBS software package (46).

Isothermal titration calorimetry

Measurements were conducted on a VP-ITC instrument (MicroCal) that was calibrated according to the manufacturer's protocol. Concentrations of RNA and protein were calculated based on their optical density absorbance at 260 nm or 280 nm, respectively. The sample cell was loaded with 1.4 ml of 10 μM RNA and the syringe with 150 μM of protein. For binding affinities above 1 μM we used 300 μM of protein. Measurements were executed at 20°C in the final buffer (with 3 mM β-mercaptoethanol instead of DDT) using 35 consecutive injections of protein (6 μl). Data were integrated and normalized using the Origin 7.0 software according to a 1:1 (RNA to protein) ratio binding model. Standard deviation is based on three measurements for the wild type and two measurements for the mutants. Titrating of RNA into protein resulted in the same binding affinity within the standard deviation.

Computational analysis of GLD-1 binding

Binding sites were predicted with the framework described in the Materials and Methods section of (26) using the binding motif that was inferred from measured GLD-1 binding affinities to 43 oligonucleotides (25) and the probability for the accessibility of a 13 mer (for the 7-mer binding motif and three additional nucleotides on each side). Only predicted binding sites with a posterior probability above 0.0005 were considered and the binding score of a transcript was calculated as the sum of the probabilities of its binding sites. Separate binding scores were calculated for each transcript by summing up only binding sites that had a uracil at the first position of the binding motif or one of the other nucleotides. The z-score of enrichment of binding sites was calculated as $(n\text{-}\mu)/\sigma$, where n is the number of predicted binding sites containing a certain nucleotide at position 1, μ is the mean and σ is the standard deviation of the number of binding sites with a certain nucleotide that would be expected given the total number of binding sites in a transcript region and the mono-nucleotide composition of that transcript region.

Computational analysis of high-throughput data sets

GLD-1 RIP-Chip data sets were taken from the supplement of (25) and (18). Translational repression and transcript stabilization by GLD-1 was calculated from log2 array expression levels of total mRNA, mRNA in gonads and polysomal mRNA, each measured in wild-type and *gld-1* null mutant worms and taken from the supplement of (18). The strength of the translational repression of an mRNA was calculated as: log2(polysomal mRNA in wild type) – log2(total mRNA in wild type) – log2(polysomal mRNA in *gld-1* null mutant) + log2(total mRNA in *gld-1* null mutant) and the stabilization of a transcript as: log2(gonad mRNA in wild type) – log2(gonad mRNA in *gld-1* null mutant).

Nematode culture, mutants and reporter lines

The allele *gld-1(rrr1)* were isolated after an ethyl methane sulfonate (EMS) mutagenesis screen (47). The creation of

plasmids with the 3′UTR germline reporters and corresponding transgenic strains are described in (18,25). Standard procedures (47) were used to maintain the different strains and animals were grown at 20°C. Immunostaining experiments against GLD-1 (18) were performed as previously described (48). Goat anti-rabbit IgG alexa-568 (dilution 1:750, Invitrogen) was used as secondary antibody. Pictures were taken with a Zeiss AxioImager Z1 microscope equipped with an Axiocam MRm REV 2 CCD camera. Images were exported into Adobe Photoshop CS4 to be processed and equal changes were applied to any picture.

RESULTS

GLD-1 binding studies with the TRA-2/GLI element (TGE) repeat of *tra-2*

We first aimed at characterizing RNA binding of the GLD-1 KH-QUA2 domain to one of its best characterized target genes *tra-2*, which regulates the spermatogenesis-to-oogenesis switch (19). The 3′UTR of the *tra-2* mRNA comprehends two identical, so-called TGE repeats each composed of two GBMs (Figure 1A). The recombinant protein (amino acids 195–336) was expressed in *Escherichia coli* and purified using the IMPACT system (49). Although the 2D [^{15}N-^{1}H] HSQC spectrum indicated a folded protein with well-dispersed resonances (Supplementary Figure S1A), we could only assign 77% of the backbone resonances of the free protein due to chemical exchange broadening in the variable loop and parts of α-helices 2 and 4 (Supplementary Figure S1B). Under such conditions, structure determination of the free protein was not possible using NMR spectroscopy.

Therefore, we focused our investigations on the interaction between the KH-QUA2 domain and the two GBMs of *tra-2* using isothermal titration calorimetry (ITC) measurements and NMR titrations. While we measured a dissociation constant (K_d) of 0.061 μM for the 5′-CUACUCAUAU-3′ RNA (Supplementary Figure S2A, predicted binding site underlined), the K_d was 5-fold increased for binding to the 5′-AUUUAAUUU-3′ RNA (Supplementary Figure S2C). NMR titration resulted in well-dispersed spectra for both RNA, saturated at a ratio of 1:1 (Figure 1B). This allowed us to assign almost completely (98%) the backbone resonances of both protein–RNA complexes. We observed the largest chemical shift perturbations within α-helix 1, β-strand 2 and α-helix 4 (QUA2 domain) (Supplementary Figure S2B and D). In addition, {^{1}H}-^{15}N NOE data showed that the QUA2 domain becomes more rigid upon RNA binding (Supplementary Figure S2E and F). Comparison of the extent of chemical shift perturbations revealed that both RNA sequences are recognized in the same way by the KH domain (Figure 1C), but that the involvement of the QUA2 domain differed considerably between the two RNAs. First, the chemical shift perturbations within α-helix 4 were substantially smaller for GLD-1 bound to the weak binding site (Figure 1C). Second, as evidenced by the comparison of {^{1}H}-^{15}N NOE data, the residues at the end of α-helix 4 and in α-helix 1 were less rigid in the weak than in the strong binding site (Supplementary Figure S2G). These data strongly suggest that the QUA2 domain is primarily responsible for the difference in RNA-binding affinity between the two GBMs. Since the crystal structure of GLD-1 with the 5′-CUAAC(AA)-3′ RNA represents binding to the weak binding site within the TGE repeat (14), we focused on solving the structure of GLD-1 KH-QUA2 bound to the strong binding motif 5′-CUACUCAUAU-3′.

Overview of the GLD-1 KH-QUA2 bound to 5′-CUACUCAUAU-3′ RNA

We calculated the solution structure of GLD-1 KH-QUA2 in complex with 5′-CUACUCAUAU-3′ based on 2899 NOE-derived distance restraints including 152 intermolecular ones. In addition, 36 RDC-derived restraints were used from backbone amides measured in 14 mg/ml Pf1 phage with an excellent signal/noise ratio (Supplementary Figure S3). The final ensemble of the 20 energy-best conformers presents a precise structure with an RMSD of 0.79 Å for all heavy atoms (Table 1 and Figure 2A). Only the variable loop, due to a lack of long-range NOEs, and the two nucleotides at the 5′ and the 3′end, due to a lack of intramolecular and intermolecular NOEs, are not well defined in this ensemble.

Consistent with Cα and Cβ chemical shifts, the protein adopts a $\beta_1\alpha_1\alpha_2\beta_2\beta_3\alpha_3\alpha_4$ conformation (Figure 2B). In contrast to the crystal structure of GLD-1, the variable loop between β2 and β3 does not fold into two defined α-helices in solution (14). In agreement with this, Cα and Cβ chemical shifts in the loop have values typical for unstructured regions and {^{1}H}-^{15}N NOE values indicate that the loop is less rigid than structured parts of the protein (Supplementary Figure S2E). We investigated if an isoform of GLD-1 featuring a tripeptide insertion (Leu-Leu-Lys) in the variable loop changes the protein behavior (23). While the RNA-binding affinity remains the same (Supplementary Figure S4B), the variable loop becomes even less rigid as evidenced by {^{1}H}-^{15}N NOE data (Supplementary Figure S4C). A similar observation was made for the Wilms Tumor suppressor protein, where insertion of a tripeptide (Lys-Thr-Ser) in the linker between zinc fingers 3 and 4 also increased flexibility (50).

The extended, single-stranded RNA is located in a hydrophobic cleft reaching from β-strand 2 to the GxxG loop down to the QUA2 domain. While the nucleotides at the 5′end are bound by the QUA2 domain, the KH domain recognizes the nucleotides at the 3′end (Figure 2B). The sugar puckers of A22, U23 and A24 adopt a C3′-endo conformation and all others are in C2′-endo as observed by a strong H1′–H2′ correlation in a 2D [^{1}H-^{1}H] TOCSY spectrum. The negative charge of the phosphate backbone is neutralized by the positive potential created by, among others, the side chains of arginine 229, 314, 328 and lysine 234 and 243 (Figure 2C).

Specific interactions of the KH domain with the 5′-CUACUCAUAU-3′ RNA

The GLD-1 KH domain binds the five nucleotides at the 3′end in a nearly identical manner than observed in the KH–RNA complexes of Nova-1, Nova-2 and SF1 (9,51–52).

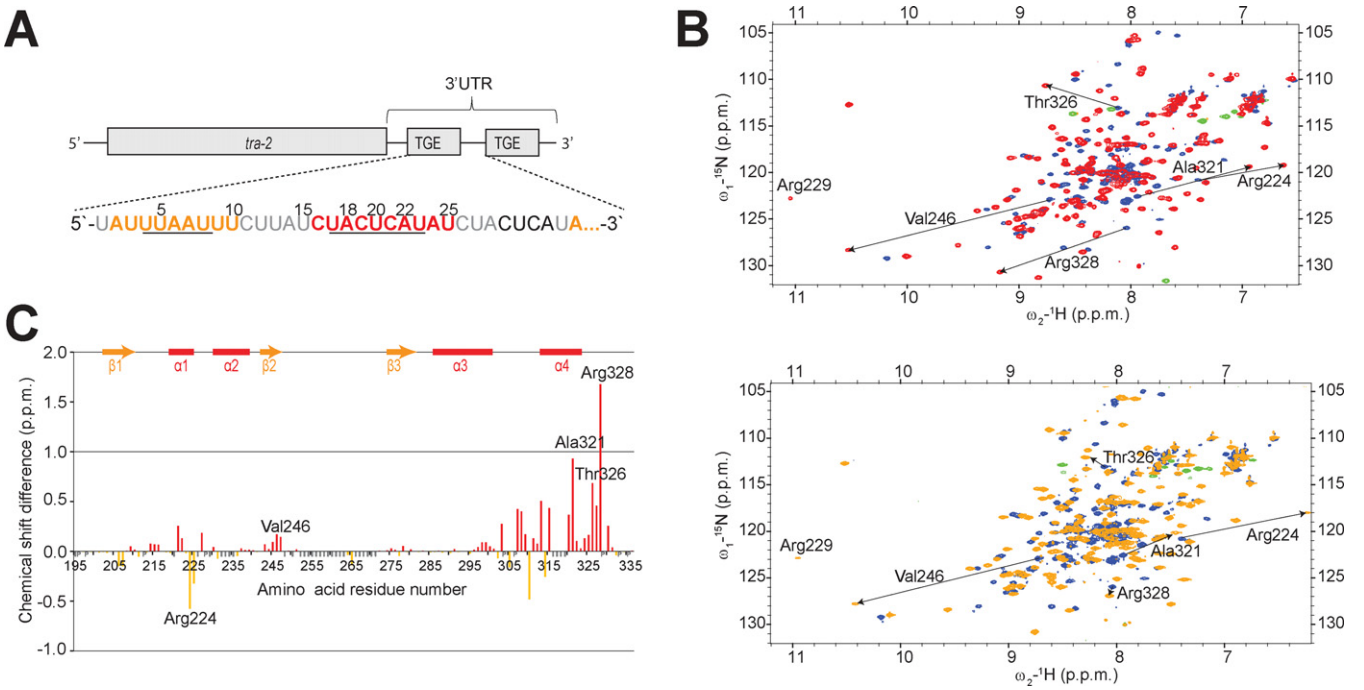

Figure 1. Overview of GLD-1 KH-QUA2 binding to the RNA sequence found in the 3′UTR of the *tra-2* mRNA. (**A**) Schematic representation of the *tra-2* gene. The sequence of both TGE repeats within the 3′UTR is shown in detail with the higher affinity target site of GLD-1 in red, the lower affinity target site in orange and additional nucleotides of the TGE repeat in gray. (**B**) Top: NMR titration of GLD-1 KH-QUA2 with the 5′-CUACUCAUAU-3′ RNA. Depicted is an overlay of a 2D [^{15}N-^{1}H] HSQC of the free form in blue and the bound form in a 1:1 ratio in red. Folded peaks are colored in light green (free) and green (bound). The five residues with the largest chemical shift perturbations are highlighted with an arrow. Bottom: NMR titration of GLD-1 KH-QUA2 with the 5′-AUUUAAUUU-3′ RNA. Depicted is an overlay of a 2D [^{15}N-^{1}H] HSQC of the free form in blue and the bound form (1:1 ratio) in orange. Folded peaks are colored in light green (free) and green (bound). Residues to compare with the higher affinity binding sites are highlighted with an arrow. (**C**) Difference of the chemical shift perturbations between GLD-1 KH-QUA2 bound to 5′-CUACUCAUAU-3′ (Supplementary Figure S2B) and bound to 5′-AUUUAAUUU-3′ RNA (Supplementary Figure S2D). Positive values (red) mean that this residue shows a higher perturbation with 5′-CUACUCAUAU-3′. Negative values (orange) mean that this residue shows a higher perturbation with 5′-AUUUAAUUU-3′. Missing assignments are marked with gray bars and proline with a gray P.

The last three nucleotides A22, U23 and A24 form an A-helical conformation that can accommodate any nucleotide at its 3′end but favors a pyrimidine at its center according to data from fluorescence polarization experiments (Figure 2D) (25). A typical feature of KH domains are the two hydrogen bonds that the base of A22 forms with the main chain of Val246 on β-strand 2, thus mimicking a Watson–Crick base pair interaction (Figure 2D). The interaction with the amide is supported by hydrogen-exchange data (Supplementary Figure S5A), the large chemical shift perturbation upon RNA binding (Supplementary Figure S2B) and the typical characteristic shift of the backbone NH at 10.4 pm as observed in other KH domains (9,53). Consequently, this interaction appears in all the conformers of the ensemble as calculated by the hydrogen bond analysis tool (H-BAT) (Supplementary Figure S5C) (54). Together with van der Waals contacts between A22 and Leu226, these two hydrogen bonds determine the highly specific recognition of an adenine (24).

The base of the preceding C21 is located on Val222 at the exactly same position as the adenine in the crystal structure of GLD-1 (Figure 2D) (14). The phosphate backbone of C21 forms a hydrogen bond with the Arg229 backbone amide leading to its characteristic chemical shift at 11 ppm as observed in other KH domains (Figure 1B) (9,53). Some of our structures (30%) show a hydrogen bond between Arg247 and the O2 of C21, similar to contacts found in Nova-2 (Supplementary Figure S5C) (51). Although the intermolecular contacts indicate specific recognition of the cytosine, an adenine is still preferred at this position according to previously published data (25,55). The reason is probably the larger interaction surface of the purine over the pyrimidine base with Val222 and Leu226.

U20 is located at the interface between α-helix 1 and α-helix 4 in a cleft-like pocket and its base is located between Gly223 on α-helix 1 and Leu317 on α-helix 4 (Figure 2E). Hydrogen bonds are formed between the uracil H3 and the side chain of Gln316, and between the 2′ hydroxyl group and the backbone carbonyl of Gly227. The very small cleft within the binding interface and the specific hydrogen bond of the imino group might be the reason for the high specificity of a uracil at this position. Altogether, our solution and the crystal structure support together well the U(C/A)A(C/U) consensus found for this part of the recognition sequence of GLD-1 binding (24).

Specific interactions of the QUA2 domain with the 5′-CUACUCAUAU-3′ RNA

The GLD-1 QUA2 domain sequence specifically recognizes the three remaining nucleotides at the 5′end of the consensus sequence. While the binding mode for C19 and A18 is

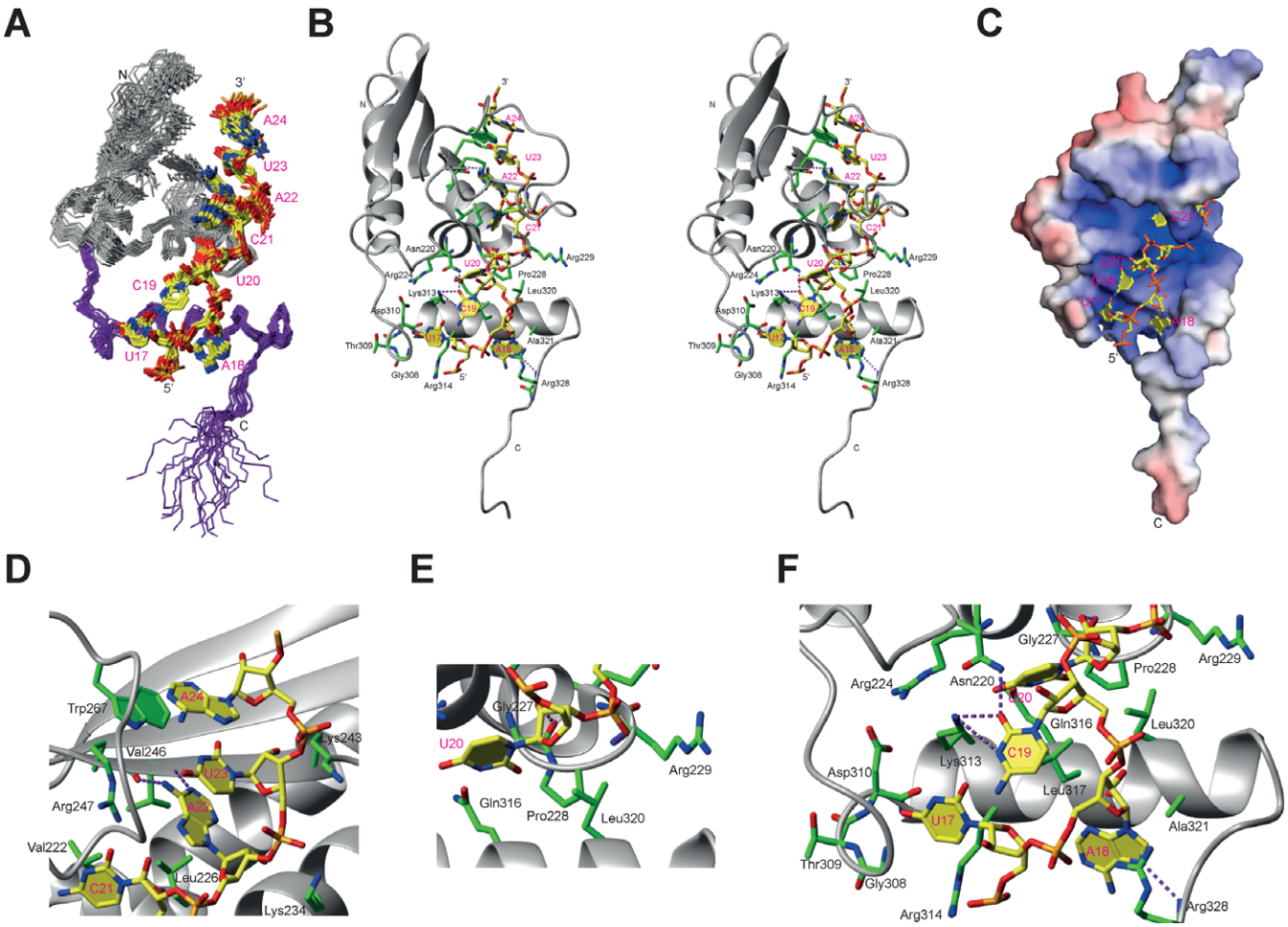

Figure 2. Overview of the solution structure of GLD-1 KH-QUA2 bound to 5′-CUACUCAUAU-3′ centered on the QUA2 domain. (**A**) Backbone traces (N, Cα, C') of the 20 energy-best structures superimposed on the backbone of the structured part (198–248, 274–328). The not-well-defined variable loop and nucleotides C18 and U25 are omitted for a better overview. The RNA is shown in stick representation with the carbon atoms in yellow, nitrogen in blue, phosphate in orange and oxygen in red. (**B**) Most representative structure in stereo view, centered on the QUA2 domain. The backbone is in ribbon, the amino acids with important roles in RNA binding are shown in stick representation, with the carbon atom in green. All other atoms have the same color code as in Figure 2A. (**C**) Surface potential of the energy-best structure.

similar as observed in QK1 and SF1 (9,14), the most 5′ nucleotide U17 was so far not structurally characterized in any other family member.

The base of C19 is localized on Leu317 and the amide group of Asn220 forms a hydrogen bond with its O2 and the phosphate backbone (Figure 2F). Furthermore, the side chain amine of Lys313 forms two hydrogen bonds with the O2 and N3 of C19. The importance of this residue for RNA binding is emphasized by a 65-fold decrease in RNA-binding affinity upon mutation to an alanine (Supplementary Table S1B).

The subsequent base of A18 contacts α-helix 4 and adopts a *syn* conformation (Figure 2F). The base forms van der Waals contacts with Ala321 and the Arg328 side chain counteracts the negative charge of the phosphate backbone. In addition, all conformers of our ensemble show a hydrogen bond between N7 and the backbone amide of Arg328, supported by a large chemical shift perturbation (Supplementary Figure S2B). In 30% of our structures, we also observe a hydrogen bond between the backbone carbonyl of Arg328 with the A18 amino protons (Supplementary Figure S5C), which would explain the preference of an adenine over a guanine at this position (55).

Finally, we also observe a specific binding pocket for U17 as identified by several unambiguously assigned intermolecular NOEs (Supplementary Figure S5B). The uracil forms van der Waals contacts with Lys313 and Arg314 and the O4 is pointing into a small pocket formed by the linker residues Gly308, Thr309 and Asp310 (Figure 2F). The backbone amides of those three residues are all possible candidates to form a hydrogen bond with U17 O4 (Supplementary Figure S5C). In addition, U17 O2 and its 2′ hydroxyl group form intramolecular hydrogen bonds with the amino group of C19 (Figure 2F). Although other nucleobases like cytosine are able to form this intramolecular hydrogen bond with C19, only a uracil can protrude into the small pocket formed by the linker residues. In conclusion, our structure explains also well the 5′-UAC-3′ consensus found at the 5′end of the GBM (24,25).

Table 1. Structural statistics for GLD1 KH-QUA2 with 5′-CUACUCAUAU-3′ RNA

	Protein	RNA
NMR distance and dihedral constraints		
Distance constraints		
Total NOE	2899	
Intra-residue	589	80
Inter-residue		
Sequential ($\|i-j\|=1$)	804	37
Non-sequential ($\|i-j\|>1$)	1235	2
Hydrogen bonds[a]	38	0
Protein–RNA intermolecular	152	
Intermolecular hydrogen bonds	0	
Total dihedral angle restraints[b]	67	8
RDCs[c]	36	0
Structure statistics		
Constraint violations (mean and SD)		
Average number of NOE violations >0.3 Å	6 ± 2.3	
Max. NOE violation (Å)	0.45 ± 0.10	
Number of dihedral angle violations >5°	0.1 ± 0.2	
Max. dihedral angle violation (°)	1.33 ± 1.75	
Deviations from idealized geometry		
Bond lengths (Å)	0.0036	
Bond angles (°)	1.637	
RMS deviation from mean structure (Å)[d]		
Protein		
Heavy	0.80 ± 0.07	
Backbone	0.36 ± 0.08	
RNA		
Heavy	0.59 ± 0.16	
Protein + RNA		
Heavy	0.79 ± 0.08	
Ramachandran statistics[e]		
Most favored regions	89.8%	
Additional allowed regions	10.1%	
Generously allowed regions	0.1%	
Disallowed regions	0.0%	

Calculated for the ensemble of the 20 violation best structures selected out of the 30 lowest energy structures.
[a]Based on slow exchanging amide protons in D_2O.
[b]Based on TALOS+ and measurement of $^3J(H_NH_\alpha)$.
[c]For statistics on Q-factor and correlation coefficient see Supplementary Figure S3.
[d]Protein range: 201–247 + 275–328; RNA range: 17–24.
[e]Protein range: 201–247 + 275–328.

In order to get an idea about the degree of specificity for U17, we measured the binding affinity of every possible nucleobase at this position by ITC measurements. In agreement with earlier published data (24,25), we observe a 2- to 3-fold decrease in RNA-binding affinity upon replacement of the uracil with one of the other three nucleobases (Supplementary Table S1A). Although differences in affinity at the first position appear quite small, we next investigated if sequences with a uracil at the 5′end might be functionally important.

Location bias and functional role of a 5′uracil-containing site

First, we wanted to analyze if sequences with U1 (= U17 in the *tra-2* sequence) are preferred targets of GLD-1. In order to evaluate GLD-1 binding sites in the transcripts of *C. elegans*, we used our previously described computational predictions (26), which are based on a binding motif that

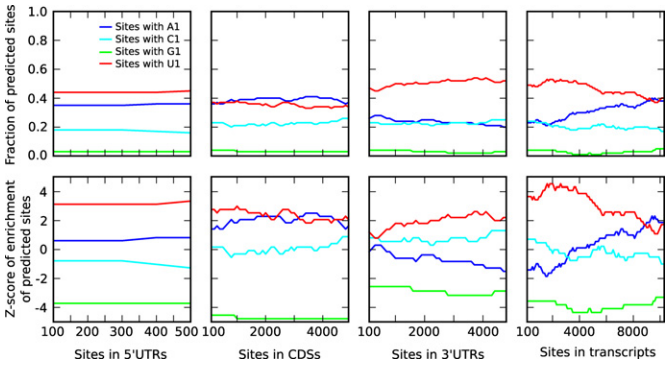

Figure 3. Location of GLD-1 binding sites with different nucleotides at position 1. Top: fraction of predicted binding sites for GLD-1 in *C. elegans* 5′UTRs, CDS, 3′UTRs and transcripts that have U (red), A (blue), C (cyan) or G (green) at position 1. The predicted binding sites were sorted according to the prediction score (left to right corresponds to the highest to lowest score) and divided into bins of 100 sites. Bottom: the z-score of the enrichment of sites with a particular nucleotide at position 1 (colors as in the top panels) within a bin, relative to what is expected given the mono-nucleotide composition of the corresponding transcript region.

was inferred from measured affinities of GLD-1 binding to 43 short oligonucleotides and on the accessibility of binding sites in a folded RNA molecule. Although the binding motif itself showed no strong preference for a nucleotide at the first position, we observed that more than 40% of the top predicted binding sites contain a uracil at position 1 (Figure 3). The fraction of binding sites containing U1 was smaller in the coding region (30–40%) compared to the 3′UTR (∼50%). Compared to the frequency that would be expected from the relative abundance of U nucleotides in different transcript regions, there is a significant enrichment of predicted sites with a U at position 1 in all transcript regions (Figure 3, bottom panels). This may hint at a specific role of U1-containing binding sites, most likely connected to the observed increase in binding affinity.

We then sought to determine whether we can improve the power of the original binding model in predicting functionally relevant GLD-1 targets by distinguishing sites with uracil from sites with another nucleotide at the first position. We used transcripts from two previously published RIP-Chips (18,25) that are expressed in the germline (germline tag >4) (17) and calculated the sum of posterior probabilities for the two categories of binding sites for all transcripts (26). We found that enrichment of transcripts in RIP-Chip relates better to the binding score computed based on sites that contain U as opposed to any other nucleotide at position 1 (Figure 4A). The Spearman correlation between the prediction rank and the rank in RIP-Chip (calculated for median ranks computed from 10 transcripts at a time, in the order of decreasing prediction score) was higher when the prediction score was computed from sites with U at the first position as opposed to any other sites (0.88 versus 0.74 for the first RIP-Chip and 0.84 versus 0.83 for the second RIP-Chip experiment). We therefore investigated the possible functions of these sites.

To this end, we ranked putative GLD-1 targets with our GLD-1 target prediction model (26) and then determined the rank of these putative targets in terms of either the de-

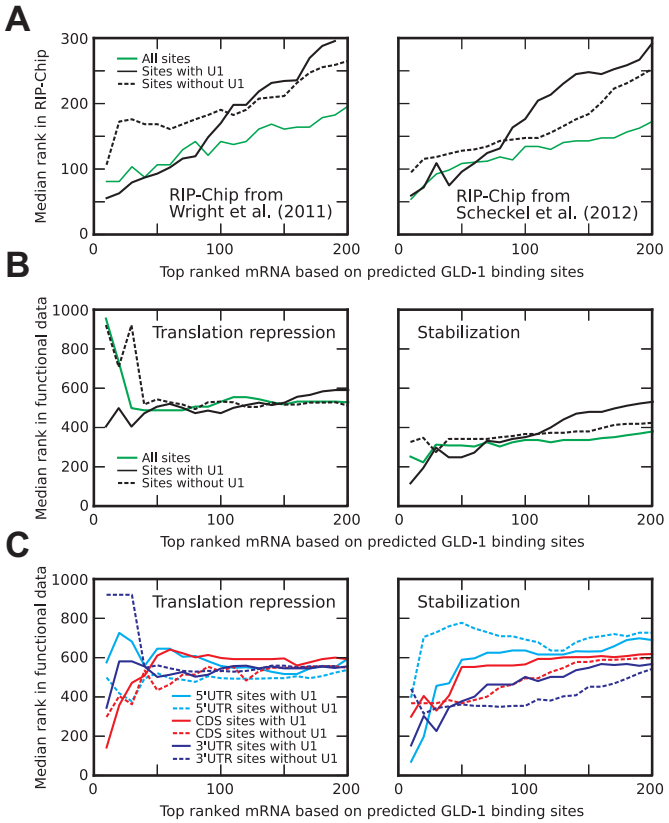

Figure 4. Relationship between the affinity and type of binding sites in different transcript regions and their functional effects. (**A**) Correlation between the rank of transcripts in the computational model of GLD-1 binding (considering binding sites with a posterior probability above 0.0005) and their rank in the GLD-1 RIP-Chip experiments by (26) and (18). The top 200 transcripts predicted based on either sites with U at position 1 (black solid line) or other nucleotides at position 1 (black dashed line) or both (green) were sorted from the highest to the lowest prediction score. The median rank in the RIP-Chip enrichment was then computed for the top x predicted transcripts, where x increased from 10 to 200, in steps of 10 transcripts. (**B**) Correlation between the rank of transcripts in the computational binding model (considering binding sites with a posterior probability above 0.0005) and their rank in translational repression or mRNA stabilization [data from (18)]. The graphs were constructed as described for panel (A), but using the estimated degree of translational repression or mRNA stabilization that the transcripts underwent in *gld-1* mutant compared to wild type worms, as opposed to the enrichment of the transcripts in RIP-Chip. (**C**) Correlation between the rank of transcripts in the computational binding model (considering binding sites with a posterior probability above 0.0005) and their rank in the translational repression or mRNA stabilization [data from (18)]. The graphs were constructed as described for panel (B), but separate prediction scores were calculated for different transcript regions based on binding sites located in 5′UTRs (cyan), CDS (red) or 3′UTRs (blue).

gree of translational repression or that of mRNA stabilization that were measured in *gld-1* mutant compared to wild-type worms. We found that binding sites with U1 were more predictive than other sites for the degree of translational repression (Spearman correlation 0.51 versus −0.14, respectively) and stabilization (Spearman correlation 0.69 versus 0.66, respectively) of their corresponding transcripts (Figure 4B). It was recently proposed that transcript stabilization by GLD-1 is mainly mediated by high-affinity binding sites in the 3′UTR, while translational repression could also be induced by binding sites of intermediate affinity not only in 3′-UTRs but also in coding regions (26). Since sequences with a uracil at position 1 show an increased binding affinity and are preferentially located in the 3′UTR, we sought to determine whether these binding sites are preferentially active in transcript stabilization. Indeed top ranked transcripts with U1-containing sites underwent stronger stabilization, especially when the sites were located in UTRs (Figure 4C). U1-containing sites were also more predictive than other sites for the degree of translational repression when they were located in the 3′-UTR or in the coding sequence (CDS). It should be noted here that this finding is based only on indirect measurements of mRNA stabilization and translation repression (see the Materials and Methods section) and that especially the functional relevance of predicted GLD-1 binding sites in the coding sequence still awaits validation *in vivo*. Overall, these computational considerations reinforce the observation that GLD-1 specifically recognizes a heptameric consensus sequence and that U1-containing sequences are more predictive for the functional response of the mRNAs to *gld-1* knockout.

Conservation of specific residues within the QUA2 domain

It was proposed that only GLD-1 is able to bind seven nucleotides, while the other QK-related proteins and SF1 recognize only six nucleotides and show no preference for a uracil at the 5′end (56). To test the conservation of RNA binding by the QUA2 domain within the STAR protein family, we tested the effect of mutations within the RNA-binding surface of GLD-1 by ITC measurements.

The only residue involved in RNA binding not conserved among the QK-related members GLD-1, QK1 and HOW is Arg314 (Supplementary Figure S6B). Upon mutation to a lysine as found in QK1 and HOW, we observe a small 1.2-fold decrease in RNA-binding affinity and the same preference for a uracil (Supplementary Table S1B). This can be explained by the positively charged side chain of both residues that is able to compensate for the negative charge of the U17 phosphate backbone. Since the residues of the KH-QUA2 interface are highly conserved in the Qk-related family members, we predict that they both should also have specificity for a uracil at the 5′end of the consensus sequence. Indeed, when examining the crystal structure of QK1, we find an identical preformed binding pocket that should be able to accommodate a uracil similar to GLD-1 (see the Discussion section).

Besides a mutation of Arg314 to lysine, the adjacent residue (Lys313 in GLD-1) is an arginine in SF1 (Supplementary Figure S6B). The transposition of those two residues was already identified to be responsible for the difference in RNA-binding affinity between SF1 and its yeast homolog BBP (57). There, mutations of R240K and K241R lead to a significant increase in RNA-binding affinity of SF1. Accordingly, inverse mutations K313R and R314K in GLD-1 lead to an 11-fold decrease in RNA-binding affinity in our ITC measurements (Supplementary Table S1B). An arginine is not able to form the same hydrogen bonds with C19 in comparison with lysine at this position and its larger size probably prevents the intramolecular hydrogen bond formed between C19 and U17. This could explain the 10-

fold difference in RNA-binding affinity measured by ITC for binding of SF1 (9) and GLD-1 (Supplementary Table S1A) to the 5′-UACUAAC-3′ consensus RNA. However, analysis of both solution structures indicates that the KH-QUA2 interfaces are very different in both proteins and that the difference in RNA recognition cannot be reduced to a single-point mutation (see the Discussion section).

Effect of a disruption of the KH-QUA2 interface on RNA binding

Our structure reveals that the interface between the KH and QUA2 domain is to a large degree stabilized by intramolecular protein interactions. We were especially intrigued about the contact between Leu320 of α-helix 4 and Pro228 of the GxxG loop (Figure 2E), since it is the only conserved inter-domain contact between GLD-1 and SF1 in the KH-QUA2 interface. In order to test the importance of the conserved Pro228 for this interface, we first mutated this residue to an alanine and measured a 19-fold decrease in RNA-binding affinity by ITC (Supplementary Table S1B). Since the hydrophobic contact should be maintained with this mutant, this underlines the importance of the proline backbone on the KH-QUA2 interface. To investigate only the importance of the hydrophobic contact between Pro228 and Leu320, we then mutated Leu320 to a serine and observe a 26-fold decrease in RNA-binding affinity (Supplementary Table S1B). Finally, the combined effect on the protein backbone and on the contact with the QUA2 domain was tested with mutation of Pro228 to serine leading to a much higher, 89-fold decrease in RNA-binding affinity (Supplementary Table S1B). Those mutagenesis studies demonstrate that both the cyclic structure of the proline and the hydrophobic contact with Leu320 are essential to maintain the integrity of the KH-QUA2 interface and thus for RNA binding.

Next, we investigated the P228S mutant by NMR spectroscopy. The protein is folded as evidenced by a 2D [^{15}N-^{1}H] HSQC (Supplementary Figure S7A), but the chemical shifts of residues within α-helices 1, 2 and at the end of α-helix 3 differ substantially compared to the wild type (Supplementary Figure S7C). This demonstrates that the cyclic structure of the proline indeed leads to a very specific backbone conformation of the KH domain. In contrast, the chemical shifts for all residues within the QUA2 domain are unchanged compared to the wild type, which emphasizes that the contact between Pro228 and Leu320 and more generally the KH-QUA2 interface are not present in the free state of GLD-1. This is consistent with solution data for QK1 that showed that the additional α-helix 4 is induced by RNA binding (10).

Upon NMR titration of the P228S mutant with the strong binding site 5′-CUACUCAUAU-3′ (Supplementary Figure S7D), the chemical shift perturbations within the KH domain show only a very small decrease compared to the wild type (Supplementary Figure S7F). However, most of the amides within the KH-QUA2 interface disappear probably due to a change in their chemical exchange rates, which hints at an impaired KH-QUA2 interface. Altogether, our NMR and ITC data substantiate the importance of Pro228 and its intramolecular interaction with Leu320 for the integrity of the KH-QUA2 interface and thus RNA binding.

Phenotype of the P228S mutation in *C. elegans*

We identified the same Pro228 to serine mutation in *C. elegans* through a genetic screen using the mutagen EMS. The phenotype of this mutation, named *gld-1(rrr1)*, did not show a null phenotype characterized by a tumorous germline (class A) (58), where potential germ cells return to mitotic proliferation (23). Instead, we observe small abnormal stacked oocytes and a feminization of the germline (class E) (Figure 5A). Interestingly, the same phenotype was already identified as the result of a deletion of amino acids 322–331 (23). Since Pro228 is contacting the residue Leu320 right in front of this deletion, it demonstrates that in both class E phenotypes the end of the QUA2 domain cannot participate the same way in RNA binding as in the wild type. This reiterates the importance of the Pro228 to Leu320 contact at the KH-QUA2 interface for RNA binding *in vivo*.

Since the phenotype hints at residual activity of GLD-1 function, we first verified by fluorescence micrographs stained for GLD-1 that the protein stability is not affected *in vivo* (Supplementary Figure S8). Then we tested the change in repression level for the three previously described 3′UTR germline reporters for *rme-2*, *egg-1* and *oma-2* (18,25). The reporter for *rme-2* showed the same expression pattern and fluorescence intensity in both wild-type and *gld-1(rrr1)* worms (Figure 5B, compare the continuous red lines). In contrast, the *egg-1* and *oma-2* reporter showed a small (Figure 5C) and a quite strong (Figure 5D) germline derepression, respectively, when compared to wild-type worms. These results are highly consistent with the rank that the 3′UTRs of these genes are assigned by our beforehand applied GLD-1 target prediction model (2520 for *rme-2*, 286 for *egg-1* and 7 for *oma-2*) (26) and according to the GBM predictor (0.8 for *rme-2*, 3.1 for *egg-1* and 4.0 for *oma-2*) (25,26). This is thus consistent with earlier reports indicating that translational repression correlates with the amount and strength of GBMs within the 3′UTR of its target transcripts (25). Our data illustrate how the KH-QUA2 interface, altered in *gld1 (rrr1),* modulates the RNA-binding affinity of GLD-1 and depending on the sequence of the mRNA GLD-1 binding site affects the level of translational repression.

DISCUSSION

In solving the solution structure of the GLD-1 KH-QUA2 domain with the complete consensus sequence, we revealed the sequence specific binding of two additional nucleotides compared to the crystal structure (14) and the important implications of a uracil at the 5′end of the consensus for GLD-1 function. Specifically, we showed that binding sites with such a 5′end uracil are enriched in all transcript regions and are particularly abundant in the UTRs, and that they are more predictive of the functional behavior of GLD-1 target transcripts. While U1-containing sites located in the UTRs are more likely involved in transcript stabilization, those that are located in the CDS are more active in translational repression than other sites. We also investigated the

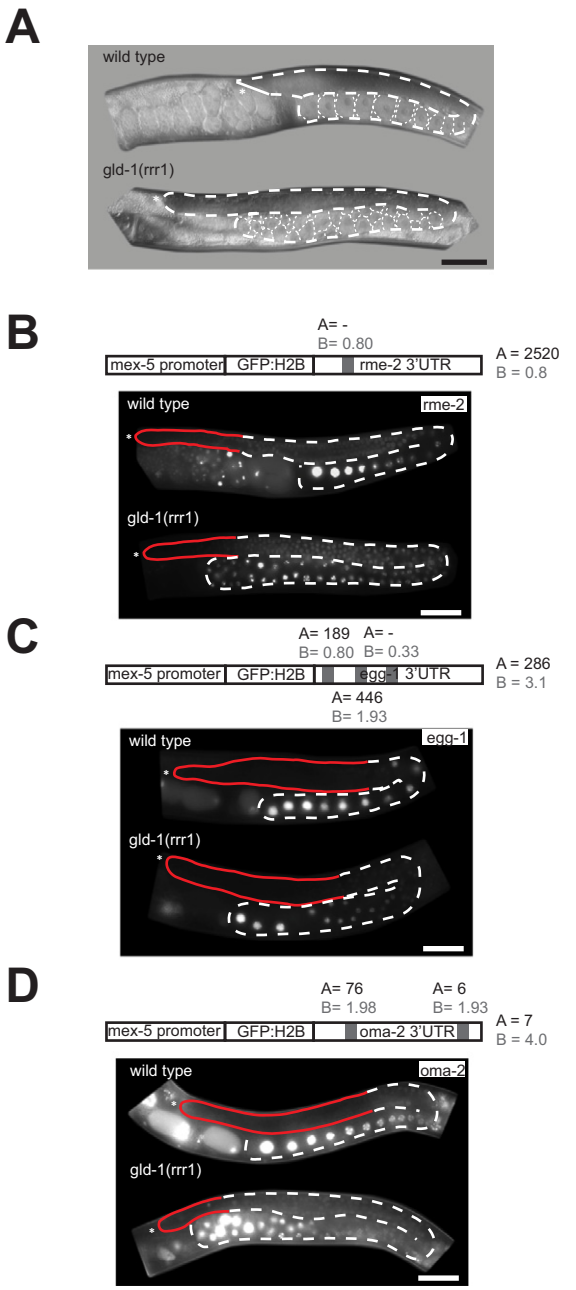

Figure 5. Phenotype of the Pro228 to serine mutation in *C. elegans*. (**A**) Differential interference contrast (DIC) micrographs of *gld-1(rrr1)* and wild-type animals grown at 20°C. The gonads are outlined with a dashed line. *gld-1(rrr1)* exhibits an abnormal oocyte phenotype, consisting of small and abnormal stacked oocytes and absence of sperm [= feminization of the germline (Fog) phenotype]. These features make it belong to 'class E' of *gld-1* mutants (58). (**B**)–(**D**) Photomicrographs of gonads from live wild-type and mutant worms expressing 3′UTR reporters for *rme-2*, *egg-1* and *oma-2*. The gonads are outlined with a dashed line, whereby a red continuous line marks the area of target repression. A scheme of the 3′UTR reporter is shown above each photomicrograph. The GFP:H2B-tagged reporters are driven by a constitutive-expressed germline promoter (mex-5) and fused to the 3′UTR of the respective endogenous gene. The predicted GBMs according to the biophysical model (A) (26) and (B) according to the GBM predictor (25) are shown as gray square. The predicted strength of GLD-1 binding to the whole 3′-UTR based on the sum of their binding sites is shown at the end of the respective reporter. The 3′UTR reporters show different levels of repression correlating with the number and strength of GLD-1 binding sites.

critical role of the KH-QUA2 interface in GLD-1 for RNA binding and the phenotypic consequence of modifying this interface *in vivo*. Yet the *in vivo* consequences of mutating specifically U1 in some GLD-1 targets remain to be investigated. Based on our results, we are now able to envisage RNA binding of the dimer to the whole TGE repeat, interpret the effect of previously reported point mutations within the QUA2 domain and finally discuss the impact of the QUA2 domain on RNA binding for the whole STAR protein family.

Comparison of solution and crystal structure and implications for TGE-repeat binding

Both the crystal (14) and solution structure (this work) of GLD-1 in complex with RNA were solved using different methodologies, protein design, RNA sequences and buffer conditions. On the RNA site, we used a longer RNA for the solution structure and subsequently observed sequence-specific binding of two additional nucleotides at the 5′end by the QUA2 domain. On the protein site, we notice two main differences: one is in the variable loop which is flexible in solution and does not show defined α-helices as in the X-ray structure and another is in the orientation of the QUA2 domain that differs by 12° between the two structures. The reason for the latter is that our solution structure lacks the contacts with the QUA1 domain and that the crystal structure is missing the two most 5′ nucleotides bound by the QUA2 domain. Nevertheless, the similarity between the solution structure of the isolated KH-QUA2 domain and the crystal structure of the dimer is still very high. If we refine the solution structure without RDCs, we gain a similar quality factor (31%) as for the crystal structure (27%) (Supplementary Figure S3C) (59). Moreover, structural alignment of the defined protein backbone of the two KH-QUA2 domains produces a low RMSD of 0.83 Å. Accordingly we observe only small deviations in all important protein–RNA contacts.

It was shown by EMSAs that one GLD-1 dimer can bind to a single TGE repeat composed of two GBMs (60). The first binding site 5′-UUAAU-3′ lacks the specifically recognized nucleotides on positions 1 and 2, while the second binding site 5′-UACUCAU-3′ represents the complete GLD-1 consensus sequence (Figure 1A). We showed that both sequences are able to induce binding between the KH and QUA2 domains, but our chemical shift perturbation experiments clearly showed that the extent of binding differs between the two RNA-binding sequences. In fact, while our solution structure comprises the complete consensus sequence 5′-CUACUCAUAU-3′, the RNA sequence 5′-CUAACAA-3′ used for crystallization misses the two nucleotides at the 5′end and reflects well binding of GLD-1 to the first binding site within the TGE repeat. Since we have now structures of the GLD-1 KH-QUA2 bound to both half-site and complete consensus sequence, we can envisage binding of the dimer to a bipartite sequence like a single TGE repeat.

To get an idea about the distance between the two binding sites of the TGE repeat, we overlaid the complete consensus sequence on one of the two STAR domains of the crystal structure and measured a distance of either 48 Å or

62 Å to the remaining half-site (Supplementary Figure S9A and B). Assuming an inter-phosphate distance of ∼7 Å for nucleotides in C2′-endo conformation (5.9 Å for C3′-endo), the conformation observed in the crystal structure is indeed able to bridge the eight nucleotide linker of the TGE repeat. This however is only possible if the TGE repeat is recognized in a specific direction by the GLD-1 dimer, the 5′end being recognized by protomer B and the 3′end by protomer A (Supplementary Figure S9B). Note that the different orientation of the QUA2 domain relative to the KH domain observed in our solution structure would be in agreement with the thesis that one GLD-1 dimer can bind one TGE repeat, since overlaying the dimer with our solution structure on the QUA2 domain further reduces the distance between the two protomers to 45 Å (Supplementary Figure S9C). In summary, our data support the conformation of the dimer observed in the crystal structure and we expect only small structural changes upon binding to a bipartite sequence, mainly caused by the QUA2 domain. Future studies will have to investigate binding of the GLD-1 dimer to a single TGE repeat in solution to finally prove and visualize our observations.

Understanding the impact of mutations within the QUA2 domain

Dimerization is not affecting the intermolecular protein–RNA interactions of the RBD, since we observe conserved contacts between the dimer of the crystal structure and the isolated KH-QUA2 domain in solution. Therefore, the solution structure qualifies to interpret the effect of previously published point mutations both *in vitro* and *in vivo*. Mutagenesis *in vivo* was shown to have critical effects on the whole organism, leading to tumor formation or masculinization and feminization of the hermaphrodite germline (23,61). In addition, the importance of amino acids for RNA binding was also tested *in vitro* by EMSAs (55) and here by ITC measurements (Supplementary Table S1B). While mutations within the KH domain have already been well interpreted with the help of the crystal structure (14), the solution structure enables us to interpret all the GLD-1 mutations localized in the QUA2 domain (Supplementary Figure S9E).

The first category of mutations comprises mutations that directly affect RNA binding. This is the case for a mutation of Arg328 and Arg314 to glutamate, since the negatively charged side chain will create a repulsive effect on the phosphate backbone of A18 and U17, respectively. If Arg314 is mutated to a serine instead of a glutamate, an only small decrease in RNA-binding affinity was observed (55). This is consistent with our structure, since a serine has a smaller side chain with less negative charge. Another huge decrease in RNA-binding affinity is reported for a mutation of Ala321 to aspartate (55), because the larger and charged side chain disrupts the binding pocket for the base of A18. Furthermore, a mutation of Gly227 to either serine or aspartate showed a null phenotype A2 in live worms (23), since it blocks binding of U19 and therefore probably prevents the involvement of the QUA2 domain in RNA binding. In addition, we report here the critical mutation of Lys313 to alanine. Although this mutation diminishes van der Waals contacts with U17, the larger impact on RNA-binding affinity is due to a loss of hydrogen bonds with the C19 base.

A second category of mutations indirectly affects RNA binding through a disruption of the KH-QUA2 interface. The Asp310 to asparagine mutation leads to a null A2 phenotype (23), because it removes the interaction with the guanidinium group of Arg224 stabilizing the KH-QUA2 interface. In addition, a Gly308 to glutamate mutation leads to a less severe phenotype D (23), since it probably decreases the flexibility of the polypeptide chain and thereby disrupts the characteristic conformation of the loop at the beginning of the QUA2 domain. Although this leads to a loss of the U17 binding pocket, the larger effect on RNA-binding affinity can be contributed to changes within the KH-QUA2 interface. Finally, we reported here the phenotype E for a Pro228 to serine mutation. This mutations leads to conformational changes within the KH-QUA2 interface and thereby disrupts RNA binding.

Overall, the severe effect of mutations within the QUA2 domain emphasizes its critical role for GLD-1 high affinity RNA binding. Due to the high conservation of the RNA-binding domain within the STAR family, we wondered about the uniqueness of the KH-QUA2 interface and especially the additional binding pocket formed by the QUA2 domain.

The QUA2 domain modulates the binding specificity of the STAR protein family

So far, the protein-RNA structures of the STAR family members GLD-1, QK1 and SF-1 have been solved. In terms of conservation, GLD-1 shares 67% overall similarity with QK1 and 34% with SF1 (Supplementary Figure S6A). Accordingly, the conservation of all RNA-contacting residues is very high (Supplementary Figure S6B). Furthermore, the crystal structure of QK1 shows a low RMSD (1.07 Å) and the solution structure of SF1 a high RMSD (3.01 Å) to our solution structure of GLD-1. This discrepancy between SF1 and GLD-1 can be primarily deduced to differences at the KH-QUA2 interface.

In SF1, the KH-QUA2 interface is stabilized exclusively by hydrophobic interactions (Supplementary Figure S10D). But in GLD-1 we observe hydrophobic contacts only between Pro228 of the GxxG loop and Leu320 of the QUA2 domain, which we proved to be essential for RNA binding *in vitro* and *in vivo*. The hydrophobic interface between α-helices 1, 3 and 4 in SF1 is replaced by a hydrogen bond between Arg224 and Asp310 in GLD-1 (Supplementary Figure S10C). The different conformation of the GLD-1 KH-QUA2 interface creates a binding pocket for a uracil at the 5′end that does not exist in SF1 (Supplementary Figure S10E). Besides the difference in sequence specificity, SF1 has also a more than 10-fold smaller RNA-binding affinity than GLD-1. Our mutagenesis data suggest that the transposition of two amino acids within the QUA2 domain could be primarily responsible for this difference in addition to other structural rearrangements. This is supported by previous reports, where the difference in RNA-binding affinity between SF1 and its yeast homolog BBP is the result of

the same transposition (57). Consequently, SF1 recognizes a hexameric RNA sequence and GLD-1 a heptameric consensus sequence.

Interestingly, it was proposed that only GLD-1 recognizes a heptameric consensus sequence, but not the close homolog QK1 (56). This observation is based on fluorescence polarization experiments with QK1 that showed no specificity at the 3′end (62). However, a later published SELEX selected 72% of sequences with a uracil at position 1 (63) and this preference was then also verified by competition fluorescence polarization (64). The crystal structure of QK1 with 5′-ACUAAC(AA)-3′ RNA supports the later observations (14), since the KH-QUA2 interface is very similar to the one of GLD-1 resulting in the presence of the same binding pocket for an additional uracil (Supplementary Figure S10H and I). We also verified that the only none conserved residue of all RNA-contacting residues in GLD-1 and QK1 is not responsible for any significant change in RNA-binding affinity or specificity. Therefore, we can conclude that the other Qk-related family members QK1 and HOW share with GLD-1 sequence-specific recognition of seven nucleotides.

Another member of the STAR protein family is Sam68, prominent for its ability to link signal transduction pathways to RNA metabolism (65). Surprisingly, its SELEX-derived 5′-UAAA-3′ consensus sequence only comprises the nucleotides bound by the KH domain, but not the three nucleotides at the 5′end recognized by α-helix 4 (66). While the RNA-contacting residues show a high degree of conservation in the KH domain, it is very low in the QUA2 domain (Supplementary Figure S6B). Ala321, which forms van der Waals contacts with an adenine in GLD-1, is a serine. Arg328, which compensates for the negative charge of the phosphate backbone in GLD-1, is a glutamate. Lys313, which forms two hydrogen bonds with C19 in GLD-1, is a cysteine. And Arg314, which compensates for the U17 phosphate backbone, is a glutamine. Therefore, we assume that the QUA2 domain of Sam68 either recognizes RNA in a different way or has lost the ability to recognize RNA as it contains even RNA repulsive point mutations. The functional implications of this observation certainly constitute a very interesting open question.

In summary, we could show here that within the STAR protein family, QUA2 is the protein domain primarily modulating the RNA-binding specificity and affinity. While the KH domain shows a high degree of conservation and is less adaptable, the QUA2 domain provides RNA-binding versatility and constitutes the main factor that can explain the different RNA-binding modes observed in SF1, Sam68 and the Qk-related family members.

ACCESSION NUMBER

Atomic coordinates and NMR restraints for the structure have been deposited in the Protein Data Bank under accession code 2MJH.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We would like to thank D. Subasic for providing us with the *gld-1* cDNA; M. Blatter for his help with structure calculations; M. Schubert, F. Damberger and T. Stahel for their expertise in NMR spectroscopy; and all the members of the SNF Sinergia grant for helpful discussions.
Author contributions: F.H.-T.A. and G.M.D. designed the project; G.M.D. prepared samples for structural studies, performed NMR experiments, evaluated NMR data, setup structure calculations and conducted ITC measurements; F.H.-T.A. helped with data analysis; S.G. established the protein purification protocol and conducted the NMR data analysis of the free protein; A.B. performed the computational analysis; C.T. conducted the *C. elegans* experiments; G.M.D., A.B. and F.H.-T.A. wrote the manuscript; all authors discussed the results and approved the manuscript.

FUNDING

Swiss National Science Foundation [CRSII3_127454 to M.Z., F.H.-T.A.; 31003A_149402 to R.C]. Funding for open access charge: Swiss National Science Foundation [CRSII3_127454 to M.Z., F.H.-T.A.; 31003A_149402 to R.C].
Conflict of interest statement. None declared.

REFERENCES

1. Daubner,G.M., Clery,A. and Allain,F.H. (2013) RRM-RNA recognition: NMR or crystallography…and new findings. *Curr. Opin. Struct. Biol.*, **23**, 100–108.
2. Maris,C., Dominguez,C. and Allain,F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.*, **272**, 2118–2131.
3. Valverde,R., Edwards,L. and Regan,L. (2008) Structure and function of KH domains. *FEBS J.*, **275**, 2712–2726.
4. Artzt,K. and Wu,J.I. (2010) STAR trek: an introduction to STAR family proteins and review of quaking (QKI). *Adv. Exp. Med. Biol.*, **693**, 1–24.
5. Novikov,L., Park,J.W., Chen,H., Klerman,H., Jalloh,A.S. and Gamble,M.J. (2011) QKI-mediated alternative splicing of the histone variant MacroH2A1 regulates cancer cell proliferation. *Mol. Cell. Biol.*, **31**, 4244–4255.
6. Yang,G., Fu,H., Zhang,J., Lu,X., Yu,F., Jin,L., Bai,L., Huang,B., Shen,L., Feng,Y. *et al.* (2010) RNA-binding protein quaking, a critical regulator of colon epithelial differentiation and a suppressor of colon cancer. *Gastroenterology*, **138**, 231–240.e5.
7. McInnes,L.A. and Lauriat,T.L. (2006) RNA metabolism and dysmyelination in schizophrenia. *Neurosci. Biobehav. Rev.*, **30**, 551–561.
8. Bockbrader,K. and Feng,Y. (2008) Essential function, sophisticated regulation and pathological impact of the selective RNA-binding protein QKI in CNS myelin development. *Future Neurol.*, **3**, 655–668.
9. Liu,Z., Luyten,I., Bottomley,M.J., Messias,A.C., Houngninou-Molango,S., Sprangers,R., Zanier,K., Kramer,A. and Sattler,M. (2001) Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science*, **294**, 1098–1102.
10. Maguire,M.L., Guler-Gane,G., Nietlispach,D., Raine,A.R., Zorn,A.M., Standart,N. and Broadhurst,R.W. (2005) Solution structure and backbone dynamics of the KH-QUA2 region of the Xenopus STAR/GSG quaking protein. *J. Mol. Biol.*, **348**, 265–279.
11. Beuck,C., Qu,S., Fagg,W.S., Ares,M. Jr and Williamson,J.R. (2012) Structural analysis of the quaking homodimerization interface. *J. Mol. Biol.*, **423**, 766–781.
12. Beuck,C., Szymczyna,B.R., Kerkow,D.E., Carmel,A.B., Columbus,L., Stanfield,R.L. and Williamson,J.R. (2010) Structure of the GLD-1 homodimerization domain: insights into STAR protein-mediated translational regulation. *Structure*, **18**, 377–389.

13. Meyer,N.H., Tripsianes,K., Vincendeau,M., Madl,T., Kateb,F., Brack-Werner,R. and Sattler,M. (2010) Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. *J. Biol. Chem.*, **285**, 28893–28901.
14. Teplova,M., Hafner,M., Teplov,D., Essig,K., Tuschl,T. and Patel,D.J. (2013) Structure-function studies of STAR family Quaking proteins bound to their *in vivo* RNA target sites. *Genes Dev.*, **27**, 928–940.
15. Lee,M.H. and Schedl,T. (2010) C. elegans star proteins, GLD-1 and ASD-2, regulate specific RNA targets to control development. *Adv. Exp. Med. Biol.*, **693**, 106–122.
16. Merritt,C., Rasoloson,D., Ko,D. and Seydoux,G. (2008) 3′ UTRs are the primary regulators of gene expression in the *C. elegans* germline. *Curr. Biol.*, **18**, 1476–1482.
17. Wang,X., Zhao,Y., Wong,K., Ehlers,P., Kohara,Y., Jones,S.J., Marra,M.A., Holt,R.A., Moerman,D.G. and Hansen,D. (2009) Identification of genes expressed in the hermaphrodite germ line of *C. elegans* using SAGE. *BMC Genomics*, **10**, 213.
18. Scheckel,C., Gaidatzis,D., Wright,J.E. and Ciosk,R. (2012) Genome-wide analysis of GLD-1-mediated mRNA regulation suggests a role in mRNA storage. *PLoS Genet.*, **8**, e1002742.
19. Jan,E., Motzny,C.K., Graves,L.E. and Goodwin,E.B. (1999) The STAR protein, GLD-1, is a translational regulator of sexual identity in Caenorhabditis elegans. *EMBO J.*, **18**, 258–269.
20. Kadyk,L.C. and Kimble,J. (1998) Genetic regulation of entry into meiosis in Caenorhabditis elegans. *Development*, **125**, 1803–1813.
21. Biedermann,B., Wright,J., Senften,M., Kalchhauser,I., Sarathy,G., Lee,M.H. and Ciosk,R. (2009) Translational repression of cyclin E prevents precocious mitosis and embryonic gene activation during *C. elegans* meiosis. *Dev. Cell*, **17**, 355–364.
22. Schumacher,B., Hanazawa,M., Lee,M.H., Nayak,S., Volkmann,K., Hofmann,E.R., Hengartner,M., Schedl,T. and Gartner,A. (2005) Translational repression of *C. elegans* p53 by GLD-1 regulates DNA damage-induced apoptosis. *Cell*, **120**, 357–368.
23. Jones,A.R. and Schedl,T. (1995) Mutations in gld-1, a female germ cell-specific tumor suppressor gene in Caenorhabditis elegans, affect a conserved domain also found in Src-associated protein Sam68. *Genes Dev.*, **9**, 1491–1504.
24. Ryder,S.P., Frater,L.A., Abramovitz,D.L., Goodwin,E.B. and Williamson,J.R. (2004) RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1. *Nat. Struct. Mol. Biol.*, **11**, 20–28.
25. Wright,J.E., Gaidatzis,D., Senften,M., Farley,B.M., Westhof,E., Ryder,S.P. and Ciosk,R. (2011) A quantitative RNA code for mRNA target selection by the germline fate determinant GLD-1. *EMBO J.*, **30**, 533–545.
26. Brümmer,A., Kishore,S., Subasic,D., Hengartner,M. and Zavolan,M. (2013) Modeling the binding specificity of the RNA-binding protein GLD-1 suggests a function of coding region-located sites in translational repression. RNA, **19**, 1317–1326.
27. Sattler,M., Schleucher,J. and Griesinger,C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spectrosc.*, **34**, 93–158.
28. Schubert,M., Manolikas,T., Rogowski,M. and Meier,B.H. (2006) Solid-state NMR spectroscopy of 10% 13C labeled ubiquitin: spectral simplification and stereospecific assignment of isopropyl groups. *J. Biomol. NMR*, **35**, 167–173.
29. Pelton,J.G., Torchia,D.A., Meadow,N.D. and Roseman,S. (1993) Tautomeric states of the active-site histidines of phosphorylated and unphosphorylated IIIGlc, a signal-transducing protein from Escherichia coli, using two-dimensional heteronuclear NMR techniques. *Protein Sci.*, **2**, 543–558.
30. Shen,Y., Delaglio,F., Cornilescu,G. and Bax,A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, **44**, 213–223.
31. Vuister,G.W. and Bax,A. (1993) Quantitative J correlation: a new approach for measuring homonuclear three-bond $J(H_NH_{alpha})$ coupling constants in 15N-enriched proteins. *J. Am. Chem. Soc.*, **115**, 7772–7777.
32. Cordier,F., Dingley,A.J. and Grzesiek,S. (1999) A doublet-separated sensitivity-enhanced HSQC for the determination of scalar and dipolar one-bond J-couplings. *J. Biomol. NMR*, **13**, 175–180.
33. Grzesiek,S. and Bax,A. (1993) Measurement of amide proton exchange rates and NOEs with water in 13C/15N-enriched calcineurin B. *J. Biomol. NMR*, **3**, 627–638.
34. Peterson,R.D., Theimer,C.A., Wu,H. and Feigon,J. (2004) New applications of 2D filtered/edited NOESY for assignment and structure elucidation of RNA and RNA-protein complexes. *J. Biomol. NMR*, **28**, 59–67.
35. Lee,W., Revington,M.J., Arrowsmith,C. and Kay,L.E. (1994) A pulsed field gradient isotope-filtered 3D 13C HMQC-NOESY experiment for extracting intermolecular NOE contacts in molecular complexes. *FEBS Lett.*, **350**, 87–90.
36. Goddard,T.D. and Kneller,D.G. (2007) SPARKY 3.
37. Herrmann,T., Guntert,P. and Wuthrich,K. (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR*, **24**, 171–189.
38. Herrmann,T., Guntert,P. and Wuthrich,K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.*, **319**, 209–227.
39. Case,D.A., Cheatham,T.E. III, Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
40. Wang,J.M., Cieplak,P. and Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.
41. Lindorff-Larsen,K., Piana,S., Palmo,K., Maragakis,P., Klepeis,J.L., Dror,R.O. and Shaw,D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**, 1950–1958.
42. Doreleijers,J.F., Vranken,W.F., Schulte,C., Markley,J.L., Ulrich,E.L., Vriend,G. and Vuister,G.W. (2012) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res.*, **40**, D519–D524.
43. Laskowski,R.A., Rullmannn,J.A., MacArthur,M.W., Kaptein,R. and Thornton,J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
44. Vriend,G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, **8**, 52–56, 29.
45. Zweckstetter,M., Hummer,G. and Bax,A. (2004) Prediction of charge-induced molecular alignment of biomolecules dissolved in dilute liquid-crystalline phases. *Biophys. J.*, **86**, 3444–3460.
46. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10037–10041.
47. Brenner,S. (1974) The genetics of Caenorhabditis elegans. *Genetics*, **77**, 71–94.
48. Navarro,R.E., Shim,E.Y., Kohara,Y., Singson,A. and Blackwell,T.K. (2001) cgh-1, a conserved predicted RNA helicase required for gametogenesis and protection from physiological germline apoptosis in C. elegans. *Development*, **128**, 3221–3232.
49. Chong,S., Montello,G.E., Zhang,A., Cantor,E.J., Liao,W., Xu,M.Q. and Benner,J. (1998) Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step. *Nucleic Acids Res.*, **26**, 5109–5115.
50. Laity,J.H., Dyson,H.J. and Wright,P.E. (2000) Molecular basis for modulation of biological function by alternate splicing of the Wilms' tumor suppressor protein. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 11932–11935.
51. Lewis,H.A., Musunuru,K., Jensen,K.B., Edo,C., Chen,H., Darnell,R.B. and Burley,S.K. (2000) Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell*, **100**, 323–332.
52. Teplova,M., Malinina,L., Darnell,J.C., Song,J., Lu,M., Abagyan,R., Musunuru,K., Teplov,A., Burley,S.K., Darnell,R.B. *et al.* (2011) Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. *Structure*, **19**, 930–944.
53. Braddock,D.T., Baber,J.L., Levens,D. and Clore,G.M. (2002) Molecular basis of sequence-specific single-stranded DNA

recognition by KH domains: solution structure of a complex between hnRNP K KH3 and single-stranded DNA. *EMBO J.*, **21**, 3476–3485.
54. Tiwari,A. and Panigrahi,S.K. (2007) HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures. *In Silico Biol.*, **7**, 651–661.
55. Lehmann-Blount,K.A. and Williamson,J.R. (2005) Shape-specific nucleotide binding of single-stranded RNA by the GLD-1 STAR domain. *J. Mol. Biol.*, **346**, 91–104.
56. Ryder,S.P. and Massi,F. (2010) Insights into the structural basis of RNA recognition by STAR domain proteins. *Adv. Exp. Med. Biol.*, **693**, 37–53.
57. Garrey,S.M., Cass,D.M., Wandler,A.M., Scanlan,M.S. and Berglund,J.A. (2008) Transposition of two amino acids changes a promiscuous RNA binding protein into a sequence-specific RNA binding protein. *RNA*, **14**, 78–88.
58. Francis,R., Barton,M.K., Kimble,J. and Schedl,T. (1995) gld-1, a tumor suppressor gene required for oocyte development in Caenorhabditis elegans. *Genetics*, **139**, 579–606.
59. Cornilescu,G., Marquardt,J.L., Ottiger,M. and Bax,A. (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.*, **120**, 6836–6837.
60. Galarneau,A. and Richard,S. (2009) The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. *BMC Mol. Biol.*, **10**, 47.
61. Jones,A.R., Francis,R. and Schedl,T. (1996) GLD-1, a cytoplasmic protein essential for oocyte differentiation, shows stage- and sex-specific expression during Caenorhabditis elegans germline development. *Dev. Biol.*, **180**, 165–183.
62. Ryder,S.P. and Williamson,J.R. (2004) Specificity of the STAR/GSG domain protein Qk1: implications for the regulation of myelination. *RNA*, **10**, 1449–1458.
63. Galarneau,A. and Richard,S. (2005) Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat. Struct. Mol. Biol.*, **12**, 691–698.
64. Carmel,A.B., Wu,J., Lehmann-Blount,K.A. and Williamson,J.R. (2010) High-affinity consensus binding of target RNAs by the STAR/GSG proteins GLD-1, STAR-2 and Quaking. *BMC Mol. Biol.*, **11**, 48.
65. Sanchez-Jimenez,F. and Sanchez-Margalet,V. (2013) Role of Sam68 in post-transcriptional gene regulation. *Int. J. Mol. Sci.*, **14**, 23402–23419.
66. Lin,Q., Taylor,S.J. and Shalloway,D. (1997) Specificity and determinants of Sam68 RNA binding. Implications for the biological function of K homology domains. *J. Biol. Chem.*, **272**, 27274–27280.