THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 286, NO. 30, pp. 26732–26742, July 29, 2011
Printed in the U.S.A.

Specific and Modular Binding Code for Cytosine Recognition in Pumilio/FBF (PUF) RNA-binding Domains*[S]◆

Received for publication, March 30, 2011, and in revised form, May 24, 2011 Published, JBC Papers in Press, June 8, 2011, DOI 10.1074/jbc.M111.244889

Shuyun Dong[‡], Yang Wang[‡], Caleb Cassidy-Amstutz[§¶], Gang Lu[§], Rebecca Bigler[‡], Mark R. Jezyk[§], Chunhua Li[‖], Traci M. Tanaka Hall[§], and Zefeng Wang[‡1]

From the [‡]Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, the [§]Laboratory of Structural Biology, NIEHS, National Institutes of Health, Research Triangle Park, North Carolina 27709, the [¶]Program in Bioinformatics, North Carolina State University, Raleigh, North Carolina 27695, and the [‖]College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100124, China

Pumilio/*fem-3* mRNA-binding factor (PUF) proteins possess a recognition code for bases A, U, and G, allowing designed RNA sequence specificity of their modular Pumilio (PUM) repeats. However, recognition side chains in a PUM repeat for cytosine are unknown. Here we report identification of a cytosine-recognition code by screening random amino acid combinations at conserved RNA recognition positions using a yeast three-hybrid system. This C-recognition code is specific and modular as specificity can be transferred to different positions in the RNA recognition sequence. A crystal structure of a modified PUF domain reveals specific contacts between an arginine side chain and the cytosine base. We applied the C-recognition code to design PUF domains that recognize targets with multiple cytosines and to generate engineered splicing factors that modulate alternative splicing. Finally, we identified a divergent yeast PUF protein, Nop9p, that may recognize natural target RNAs with cytosine. This work deepens our understanding of natural PUF protein target recognition and expands the ability to engineer PUF domains to recognize any RNA sequence.

The specific interaction of RNA and protein plays vital roles in RNA regulation including splicing, localization, translation, and degradation. Such recognition may be directed toward unstructured RNA requiring discrimination of RNA sequences, folded RNA motifs, or some combination of sequence and structural specificity (1). Members of the PUF[2] protein family (named after *Drosophila* Pumilio and *Caenorhabditis elegans fem-3* mRNA-binding factor (FBF)) are sequence-specific RNA-binding proteins that regulate networks of mRNAs encoding proteins of related function (2–7). PUF proteins generally recognize the 3′-UTR of their target mRNAs to control the mRNA stability and translation (2–7).

The RNA-binding domain of PUF proteins, known as the Pumilio homology domain (PUM-HD) or PUF domain, can bind to unstructured RNA sequences in a distinct fashion. The PUF domain of human Pumilio 1 contains eight PUM repeats, each containing three α-helices packed together in a curved structure (8–10). RNA is bound as an extended strand to the concave surface of the PUF domain with the bases contacted by protein side chains. In general, each PUM repeat recognizes a single RNA base through the second helix (α2) in an antiparallel arrangement, *i.e.* nucleotides 1–8 are recognized by PUF repeats 8–1, respectively. The α2 helices of PUM repeats contain a 5-residue sequence, designated here as 12*XX*5, where the side chain at position 2 stacks with the recognized base and the side chains at positions 1 and 5 recognize the edge of the base (8, 11) (see Fig. 1*A*). Specific residues at these positions direct the base recognition properties of the repeat. This PUF-RNA recognition code makes it possible to modify a PUM repeat to bind a particular RNA base, producing a designed PUF domain that specifically recognizes a given 8-nucleotide RNA target. Such *de novo* designed RNA binders have been used to track RNA localization in cells (12, 13), study PUF protein function (14, 15), and modulate alternative splicing (16) and continue to provide a useful tool for biomedical research with possible therapeutic applications.

One limitation to application of designed PUF proteins is that although the modular code for recognition of RNA bases A, U, and G has been deduced, a code for cytosine recognition by a PUM repeat is unknown. Thus, recognition of a cytosine cannot be engineered in a repeat, although Pumilio 1 can accept any base including cytosine at the fifth position of the target sequence, and yeast Puf3p specifically recognizes a cytosine two bases upstream of the core PUF recognition sequence (17). Naturally occurring PUM repeats that specifically recognize a cytosine have not been identified, providing no clues to a cytosine-recognition code and uncertainty about whether such specific recognition exists or is possible. The identification of a combination of amino acid side chains in a PUM repeat that can recognize a cytosine is necessary to expand the use of designed PUF domains directed toward any RNA sequence.

* This work was supported, in whole or in part, by a grant from the Intramural Research Program of the National Institute of Environmental Health Sciences (to T. M. T. H.) and by a grant from the Beckman Foundation and the Kimmel Sidney Scholar award (to Z. W.).
◆ This article was selected as a Paper of the Week.
The atomic coordinates and structure factors (code 2YJY) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).
[S] The on-line version of this article (available at http://www.jbc.org) supplemental Figs. S1–S4 and Table S1.
[1] To whom correspondence should be addressed: CB #7365 University of North Carolina, Chapel Hill, NC 27599. Fax: 919-966-5640. E-mail: zefeng@med.unc.edu.
[2] The abbreviations used are: PUF, Pumilio/FBF; FBF, *fem-3* mRNA-binding factor; PUM, Pumilio; ss, splice site; ESF, engineered splicing factor; Y3H, yeast three-hybrid; Bis-Tris, 2-(bis(2-hydroxyethyl)amino)-2-(hydroxymethyl)propane-1,3-diol; RS-PUF, fusion protein of Arg/Ser rich domain and PUF domain; Gly-PUF: fusion protein of Gly rich domain and PUF domain.

Using a yeast three-hybrid system, we found that the 5-residue RNA interaction sequence SY*XX*R allows PUM repeats of human Pumilio 1 (hereafter referred to as PUF for simplicity) to specifically interact with cytosine. In a crystal structure of a complex between a mutant PUF (SY*XX*R) and cognate RNA, the arginine side chain interacts directly with the cytosine, and the serine side chain helps to position the arginine residue. We applied the recognition code to design new PUF domains to recognize RNA targets with multiple cytosine residues such as CUG repeats that are responsible for the pathogenesis of myotonic dystrophy. We also used the code to engineer splicing factors that modulate alternative splicing of both a splicing reporter and an endogenous gene. Furthermore, a naturally occurring yeast PUF protein, Nop9p, appears to contain a repeat with a code for cytosine and is conserved in homologs from yeast to human, suggesting that the natural target sequences of these PUF proteins may contain cytosine.

EXPERIMENTAL PROCEDURES

Generation of a Random Sequence Library—A PUF mutant library was generated through three PCR amplifications using primers with randomized regions (supplemental Fig. S1). In reaction 1, the 5′ portion of the Pumilio 1 PUF domain was amplified from wild-type PUF with primers Bam-Puf-1F (5′-GGA TCC GAG GCC GCA GCC GCC TTT TGG AA) and Puf-R6N-2R (5′-GAT TAC ATA **NNN** TCC ATA TTG ATC CTG TAC CAG). In reaction 2, the 3′ portion of the PUF domain was amplified with primers Puf-R6N-1F (5′-TAT GTA ATC **NNN** CAT GTA CTG GAG CAC GGT CG) and Puf-Xho-2R (5′-CTC GAG CCC CTA AGT CAA CAC CGT TCT TC). The Puf-R6N-2R contains 3 random nucleotides encoding the amino acid at position 1043, whereas Puf-R6N-1F contains random nucleotides encoding the residue at position 1047 (supplemental Fig. S1). The purified PCR products of reactions 1 and 2 were mixed as the template for reaction 3 with primers Bam-Puf1-1F and Puf-Xho-2R. The final PCR products encode the entire PUF domain and have the two randomized codons at positions 1043 and 1047.

Yeast expression plasmid encoding wild-type PUF fused at the N terminus to the Gal-4 activation domain was created by amplification of the coding sequence of the PUF domain from pTYB3-HsPUM1-HD (9) and subcloned into the pACT2 plasmid using BamHI and XhoI sites. Plasmids expressing target RNAs were made by annealing DNA oligonucleotides encoding the desired RNAs and subcloning into the pIIIA-MS2-2 plasmid using SmaI and SphI restriction sites.

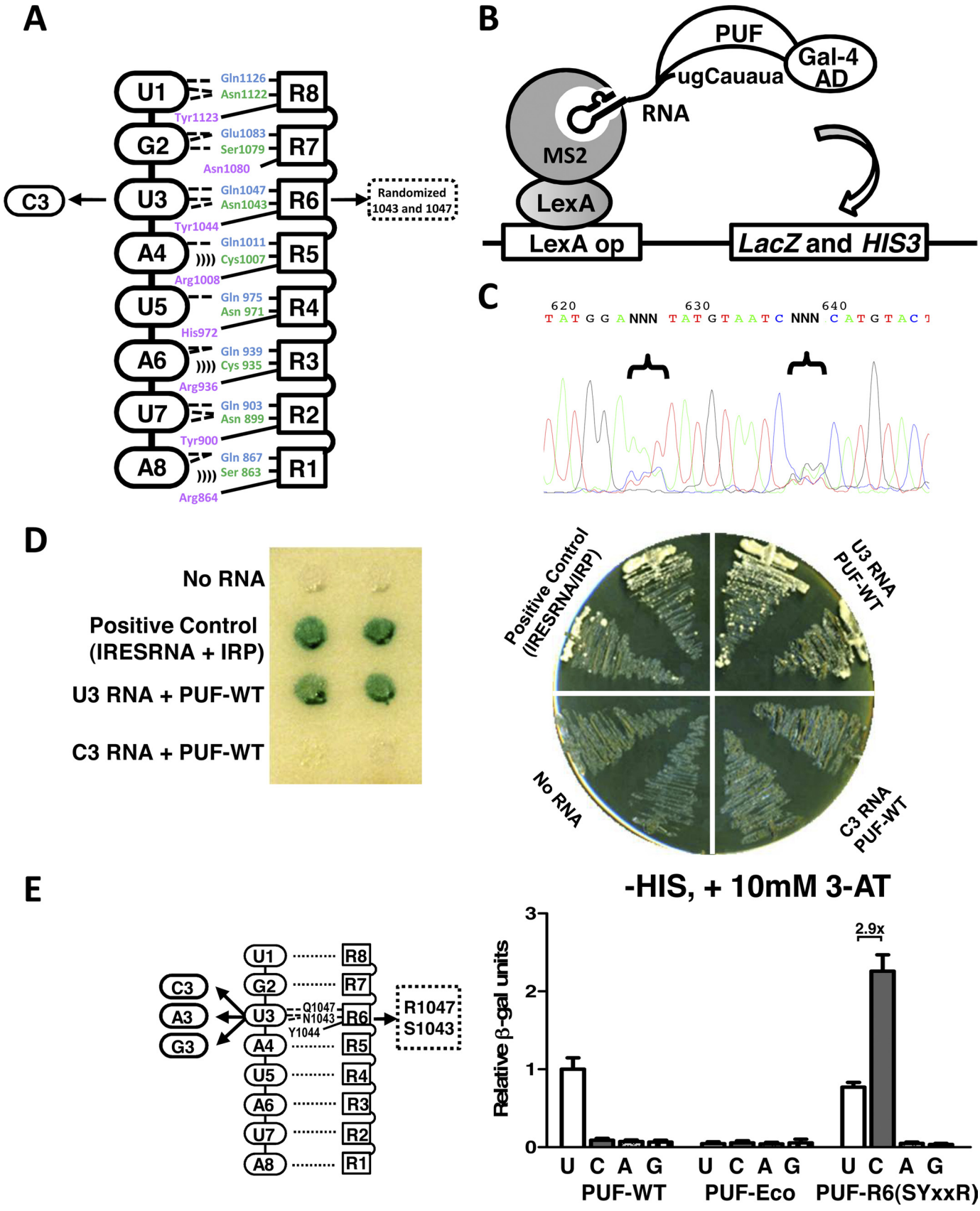
Yeast three-hybrid (Y3H) assays were performed in yeast strain YBZ-1 as described previously (18, 19). For the Y3H screen, instead of generating an *Escherichia coli* plasmid library, we generated a yeast library screening system directly through gap repair (supplemental Fig. S1). First, the pIIIA-MS2-2 plasmid carrying UGCAUAUA RNA was transformed and expressed in yeast strain YBZ-1. Second, an EcoRI site was introduced by site-directed mutagenesis into wild-type pACT2-PUF between the nucleotides encoding positions 1043 and 1047. The pACT2-PUF-EcoDNA was linearized by EcoRI and co-transformed with the random PUF PCR library at a molar ratio 1:6 into the yeast. About 50,000 yeast clones were generated, giving at least 10-fold coverage of the entire 6-nucleotide sequence space ($4^6 = 4096$). Yeast transformants were screened on plates lacking histidine and containing 10 mM 3-aminotriazole. The transformants that survived *HIS* growth selection were confirmed with *LacZ* expression. Selected yeast plasmid DNAs were sequenced and reintroduced into mother strain to confirm the interaction and specificity.

Plasmid Constructs—Additional PUF site-directed mutants carried by pACT2 were generated using the QuikChange site-directed mutagenesis kit (Agilent). The pTYB3-PUF mutants for *in vitro* protein expression were created by PCR amplification from yeast expression plasmids and subcloning into the pTYB3 plasmid using NcoI and SapI restriction sites. To generate the engineered splicing factors (ESFs) that recognize C-containing target sequences, we mutated plasmids encoding the RS-PUF or Gly-PUF fusion proteins (16).

Liquid β-Galactosidase Assays—The activity of β-galactosidase was measured using 96-well plates using 12 clones from each sample (20). The yeast colonies were randomly picked and inoculated into 12 different wells with 100 μl of culture medium in a 96-well plate. After overnight growth in 24 °C with shaking, the culture density of each well was determined by reading OD_{650} with a plate-type spectrophotometer (spectroMAX PLUS from Molecular Devices). In each clone, 25 μl of cell culture was removed and transferred into a new 96-well plate and mixed with 225 μl of assay buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 1 mM $MgCl_2$, 0.2% (w/v) Sarkosyl, and 0.4 mg/ml *O*-nitrophenol-β-D-galactopyranoside). The plate was incubated at 37 °C for 2 h, and 100 μl of 1 M carbonate solution was added into each well to stop the reaction. We measured the A_{405} with a spectrophotometer to quantify the product (nitrophenol). The β-galactosidase units were calculated as the difference of A_{405} between the sample and the background calibrated by culture densities (20).

Protein Expression, Purification, and Electrophoretic Mobility Shift Assay (EMSA)—All proteins were expressed in *E. coli* strain BL21 and purified as described previously (9, 11). Protein purity was examined with SDS-PAGE gel electrophoresis. Protein concentration was determined by Bradford assay. RNAs were generated by *in vitro* transcription and purified on denaturing gels. 50 pmol of RNAs was labeled at the 3′ end with biotinylated cytidine bisphosphate using T4 RNA ligase following the manufacturer's directions (Thermo Scientific Pierce RNA 3′ end biotinylation kit). In each sample, 20 fmol of labeled RNA (1 nM) and 4 pmol of proteins (0.2 μM) were incubated in binding buffer (10 mM HEPES, pH 7.3, 20 mM KCl, 1 mM $MgCl_2$, 1 mM DTT, and 0.1 g/liter tRNA) for 1 h at room temperature. The binding reactions were separated by electrophoresis on 6% non-denaturing PAGE run with 1× Tris-borate-EDTA at 4 °C, transferred to nylon membranes, and cross-linked to the membrane by UV. Biotin-labeled RNA was detected by chemiluminescence using the Thermo Scientific LightShift chemiluminescent RNA EMSA kit following the manufacturer's directions.

Crystallization, Structure Determination, and Refinement—Crystals of PUF-R6(SY*XX*R) mutant and C3 RNA (5′-AUUGCAUAUA) were grown by sitting drop vapor diffusion. RNA oligonucleotide was obtained from Dharmacon (Lafayette,


A
U1
G2
C3
U3
A4
U5
A6
U7
A8
R8
R7
R6
R5
R4
R3
R2
R1
Gln1126
Asn1122
Tyr1123
Glu1083
Ser1079
Asn1080
Gln1047
Asn1043
Tyr1044
Gln1011
Cys1007
Arg1008
Gln 975
Asn 971
His972
Gln 939
Cys 935
Arg936
Gln 903
Asn 899
Tyr900
Gln 867
Ser 863
Arg864
Randomized
1043 and 1047
B
PUF
Gal-4
AD
ugCauaua
RNA
MS2
LexA
LexA op
LacZ and HIS3
C
620
630
640
T A T G G A NNN T A T G T A A T C NNN C A T G T A C T
D
No RNA
Positive Control
(IRESRNA + IRP)
U3 RNA + PUF-WT
C3 RNA + PUF-WT
Positive Control
(IRESRNA/IRP)
U3 RNA
PUF-WT
No RNA
C3 RNA
PUF-WT
E
C3
A3
G3
U1
G2
U3
A4
U5
A6
U7
A8
R8
R7
R6
R5
R4
R3
R2
R1
Q1047
N1043
Y1044
R1047
S1043
-HIS, + 10mM 3-AT
Relative β-gal units
0
1
2
3
2.9x
U C A G U C A G U C A G
PUF-WT
PUF-Eco
PUF-R6(SYxxR)

CO). The protein-RNA complex was prepared by mixing a 1:1.1 molar ratio of purified protein (3.5 mg/ml) and RNA in a buffer containing 20 mM Tris-HCl, pH 7.5; 100 mM NaCl; and 1 mM DTT. One μl of complex solution was added to 1 μl of a well solution containing 30% PEG 3350, 0.2 M ammonium tartrate dibasic, and 0.1 M Bis-Tris, pH 5.5. Crystals were flash-frozen after adding an equal volume of cryoprotectant solution (32% PEG 3350, 20% ethylene glycol) to the drop. Diffraction data were collected at the Southeast Regional Collaborative Access Team (SER-CAT) beamline ID-22, Advanced Photon Source at wavelength 1.0 Å and −180 °C. All data sets were indexed, integrated, and scaled with the HKL2000 suite (21). The structure was determined by molecular replacement using the structure of human Pumilio 1 (Protein Data Bank (PDB) ID: 1M8Y) as a search model with PHASER (22). Two complexes are present in the asymmetric unit. Iterative model building was performed with COOT (23), and the resulting models were refined with PHENIX (24). All ϕ-ψ angles are within allowable regions of the Ramachandran plot. The atomic coordinates and structure factors have been deposited in the PDB (PDB ID: 2YJY).

Cell Culture, Transfection, RNA Purification, and RT-PCR—Human embryonic kidney 293T cells or breast cancer MDA-MB-231 cells were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum. Cells were seeded onto 24-well plates and transfected with Lipofectamine 2000 following manufacturer's directions. The purification of total RNA and semiquantitative RT-PCR were carried out as described previously (16).

Bioinformatic Analyses—Two PUF proteins in budding yeast, Puf2p and Nop9p, were identified by searching the Simple Modular Architecture Research Tool (SMART) database. We further searched the *Saccharomyces* Genome Database using the BLASTP program with the following queries: 1) the two natural yeast PUF repeats containing a possible C-recognition code and 2) all the yeast PUF repeats in which we replaced the native RNA recognition motifs with S*XXX*R. Only the two PUF repeats from Puf2p and Nop9p were identified. We then used the entire PUF domains of Puf2p and Nop9p as queries to search the non-redundant protein sequences using Position-Specific Iterated BLAST (PSI-BLAST) and manually inspected the positive hits to filter out repeats. A subset of representative sequences with significant matches was selected to cover a diverse range of organisms. These sequences were aligned with ClustalW, and a phylogenetic tree was generated with Phylowidget.

TABLE 1

Nucleotide sequences recovered from the Y3H screen

In total, 20 independent clones were sequenced, and the resulting codons are listed with the encoded amino acid residue in parentheses. The residue at position 1043 (position 2 in the 5-residue RNA-interaction motif) is tyrosine for all clones. Clone 18 did not have an unambiguous sequence for the first codon (indicating either an A or C in the second position) and thus was disregarded in our analyses.

	AA position 1043	AA position 1047
Wild type	AAT(Asn)	CAA(Gln)
1	AGT(Ser)	AGA(Arg)
2	AGT(Ser)	AGA(Arg)
3	AGT(Ser)	AGG(Arg)
5	AGT(Ser)	AGA(Arg)
6	AGT(Ser)	CGC(Arg)
7	AGT(Ser)	AGG(Arg)
8	TCC(Ser)	CGA(Arg)
9	AGT(Ser)	CGG(Arg)
10	AGT(Ser)	CGG(Arg)
11	AGT(Ser)	AGA(Arg)
12	AGT(Ser)	AGG(Arg)
13	TCT(Ser)	AGG(Arg)
14	TCA(Ser)	CGT(Arg)
15	AAT(Asn)	TAG (stop)
16	AGT(Ser)	AGG(Arg)
17	AGT(Ser)	AGA(Arg)
18	A(A C)T(Asn Thr)	CGG(Arg)
19	AGT(Ser)	AGA(Arg)
20	AGT(Ser)	AGG(Arg)

RESULTS

Random Library Screen for Cytosine Recognition—To select a PUM repeat that specifically recognizes cytosine, we used a Y3H system that utilizes co-expression of the PUF domain fused with the Gal-4 activation domain, an RNA target with an MS2-binding site, and an MS2-LexA fusion protein (Fig. 1*B*). This system can be used to reliably measure the relative binding affinity between RNA and protein (18, 19). For our screening, we introduced a uridine-to-cytosine mutation at the third position of a wild-type PUF target sequence (Fig. 1*B*). We generated a PUF domain library with random sequences at the first and fifth positions of the RNA interaction motif in repeat 6, which recognizes the third position of the RNA target sequence (Fig. 1, *A* and *C*, and supplemental Fig. S1). In control experiments, co-expression of wild-type PUF and its cognate target sequence (U3) resulted in activation of *HIS3* and *LacZ* reporter genes. In contrast, wild-type PUF cannot recognize the target RNA with a cytosine at the third position (C3), suggesting that our screen has a low false positive background (Fig. 1*D*). Yeast transformants were screened first for *HIS3* expression, and 200 of the resulting positive clones were reconfirmed with a *LacZ* activity assay. Plasmids encoding functional PUFs were recovered from the doubly positive yeast clones (178 clones), and a subset was

FIGURE 1. **Identification of a cytosine-recognition code by yeast three-hybrid screen.** *A*, schematic representation of the interaction between wild-type PUF and its RNA target sequence (5′-UGUAUAUA). Protein repeats are indicated by *squares*, and RNA bases are indicated by *ovals* (*dashed lines*, hydrogen bonds; *parentheses*, van der Waals contacts). For library screening, the third RNA base was mutated to cytosine (C3) and served as a new target. Nucleotides encoding positions 1043 and 1047 of the PUF were randomized in the screened library. *B*, illustration of the yeast three-hybrid assay used to screen the PUF library for binding to C3 RNA (5′-ugCauaua-3′) and to measure the PUF-RNA interaction. The interaction between Gal4-PUF and target RNA fused with MS2-binding sequence can trigger the expression of both reporter genes, *LacZ* and *HIS3*. *Gal-4 AD*, Gal-4 activation domain. *C*, sequences of the PUF library with randomized coding sequences at positions 1043 and 1047. *D*, validation of the yeast three-hybrid system. The expression of the reporter genes, *LacZ* (*left panel*) and *HIS3* (*right panel*), was measured for yeast expressing wild-type PUF and the wild-type RNA (U3) or the mutated RNA C3. Positive binding was found only when wild-type PUF and U3 RNA were expressed. The interaction between iron-responsive element RNA (*IRESRNA*) and iron regulatory protein (*IRP*) was used as positive control. *E*, measurement of specific interactions between PUF domains and RNAs with base substitutions at position 3. Positions of the mutated amino acids and RNA bases are indicated in the *left panel*. Protein-RNA binding was measured with the yeast three-hybrid system using liquid β-galactosidase assays. For each sample, 12 colonies were picked, and the experiments were performed in triplicate. The β-galactosidase activities relative to that of the wild-type PUF-U3 RNA pair were plotted to reflect the strength of protein-RNA interaction (*right*). The -fold increase in binding of the mutant protein to the cognate base *versus* the non-cognate base in the wild-type RNA is indicated above the bars. *Error bars* indicate S.D.

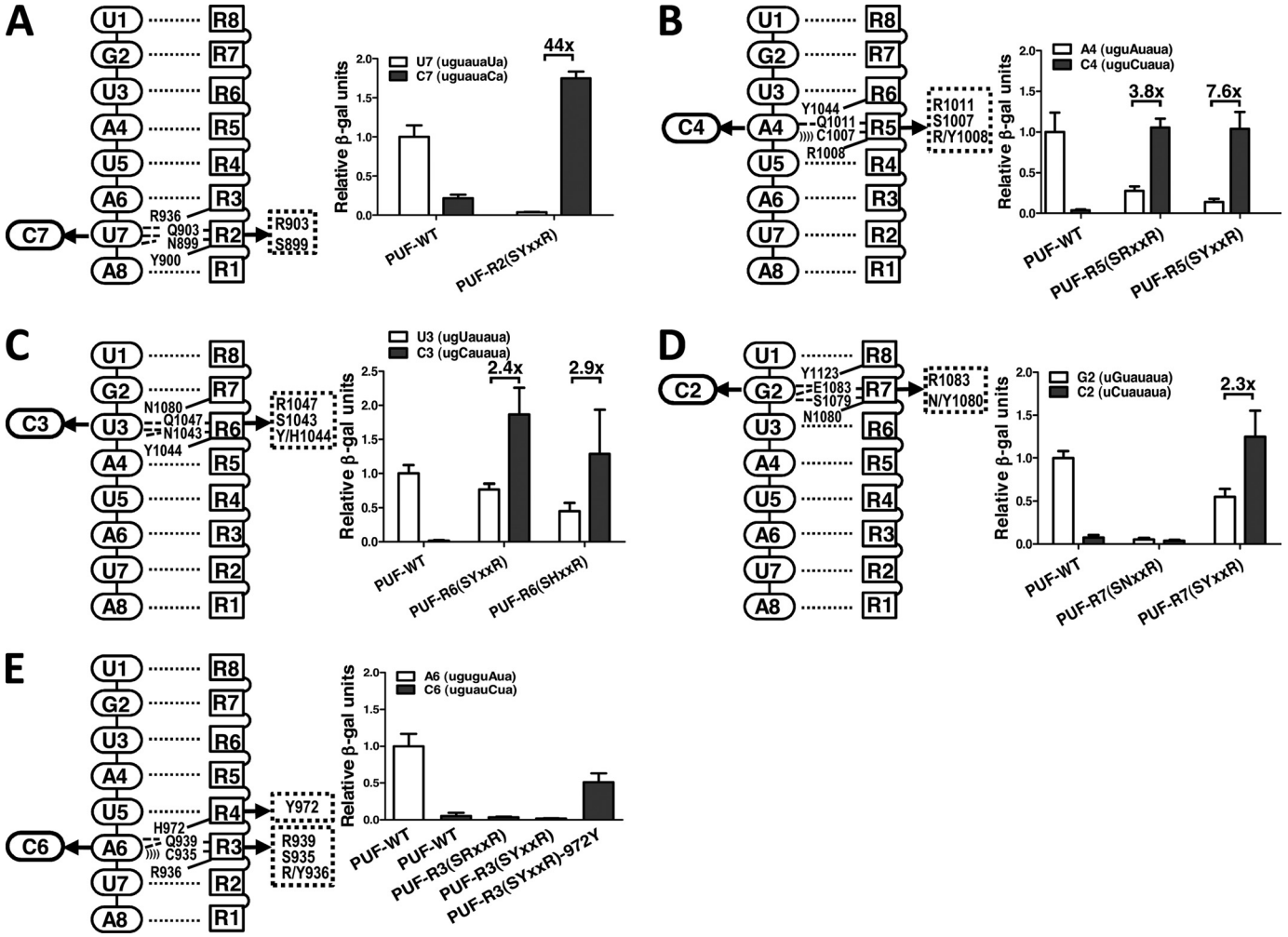


FIGURE 2. **The cytosine-recognition code can be transferred to other PUM repeats.** *A*, mutation of PUM repeat 2 to convert its binding specificity to recognize C7 RNA. Indicated mutations were introduced in repeat 2 (*left*). Protein-RNA binding measured with the yeast three-hybrid system using β-galactosidase activity is shown (*right*). Wild-type PUF and its cognate target RNA were included in all experiments as controls, and its relative activity was set to 1. *B*, mutation of PUM repeat 5 to convert its binding specificity to recognize C4 RNA. Indicated mutations were introduced in repeat 5. *C*, mutation of PUM repeat 6 to convert its binding specificity to recognize C3 RNA. Indicated mutations, including two different stacking residues, were introduced. *D*, mutation of PUM repeat 7 to convert its binding specificity to recognize C2 RNA. Indicated mutations, including two different stacking residues, were introduced. For *panels A–D*, the -fold increase in binding of the mutant protein to the cognate base *versus* the non-cognate base in the wild-type RNA is indicated above the bars. *E*, mutation of PUM repeat 3 to convert its binding specificity to recognize C6 RNA. Indicated mutations, including mutation of the base stacking residue of repeat 4, were introduced. For all panels, the experimental conditions and data analyses are similar to that in *panel A*. *Error bars* indicate S.D.

sequenced to identify amino acid combinations directing cytosine recognition.

Of the 19 unambiguous sequences we obtained, 18 coded for serine at amino acid position 1043 and arginine at amino acid position 1047, positions 1 and 5 in the 5-residue RNA interaction motif (Table 1, Fig. 1*A*). The only exception, clone 15, contained a stop codon at position 1047 and therefore is likely a false positive. The 18 clones encoding Ser-1043/Arg-1047 contained four different serine codons and six arginine codons, suggesting that our screen adequately covered sequence space. During revision of this manuscript, a study reporting the identification of a set of cytosine-specific RNA recognition side chains ((G/A/S/T/C)*XXX*R) was published (25). The more stringent conditions we used (10 mM *versus* 0.5 mM 3-aminotriazole) may have produced the dominance of the SY*XX*R sequence over other sets of side chains with arginine at the fifth position as seen in this other study (25). The relative β-galactosidase activities for the different sets of side chains suggest that the SY*XX*R combination binds most tightly (18, 25).

To examine the specificity of the newly identified C-recognition code, we measured the RNA-protein interaction between PUF domains and RNA targets containing each of the four bases at the third position (Fig. 1*E*). We found using a Y3H assay that wild-type PUF bound only to the natural target sequence with a U at the third position (U3), and a mutant protein, PUF-Eco, with an EcoRI site inserted between positions 1043 and 1047 did not recognize any of the target RNAs (Fig. 1*E*). The PUF with Ser-1043/Arg-1047 mutations in repeat 6, PUF-R6(SY*XX*R), specifically bound to the C3-containing target with similar affinity as the PUF-WT protein and U3 RNA (18) and did not recognize targets with an A3 or G3. We measured residual binding of PUF-R6(SY*XX*R) to the wild-type U3 sequence, likely due to the lack of a stacking side chain (asparagine) in repeat 7 (see below).

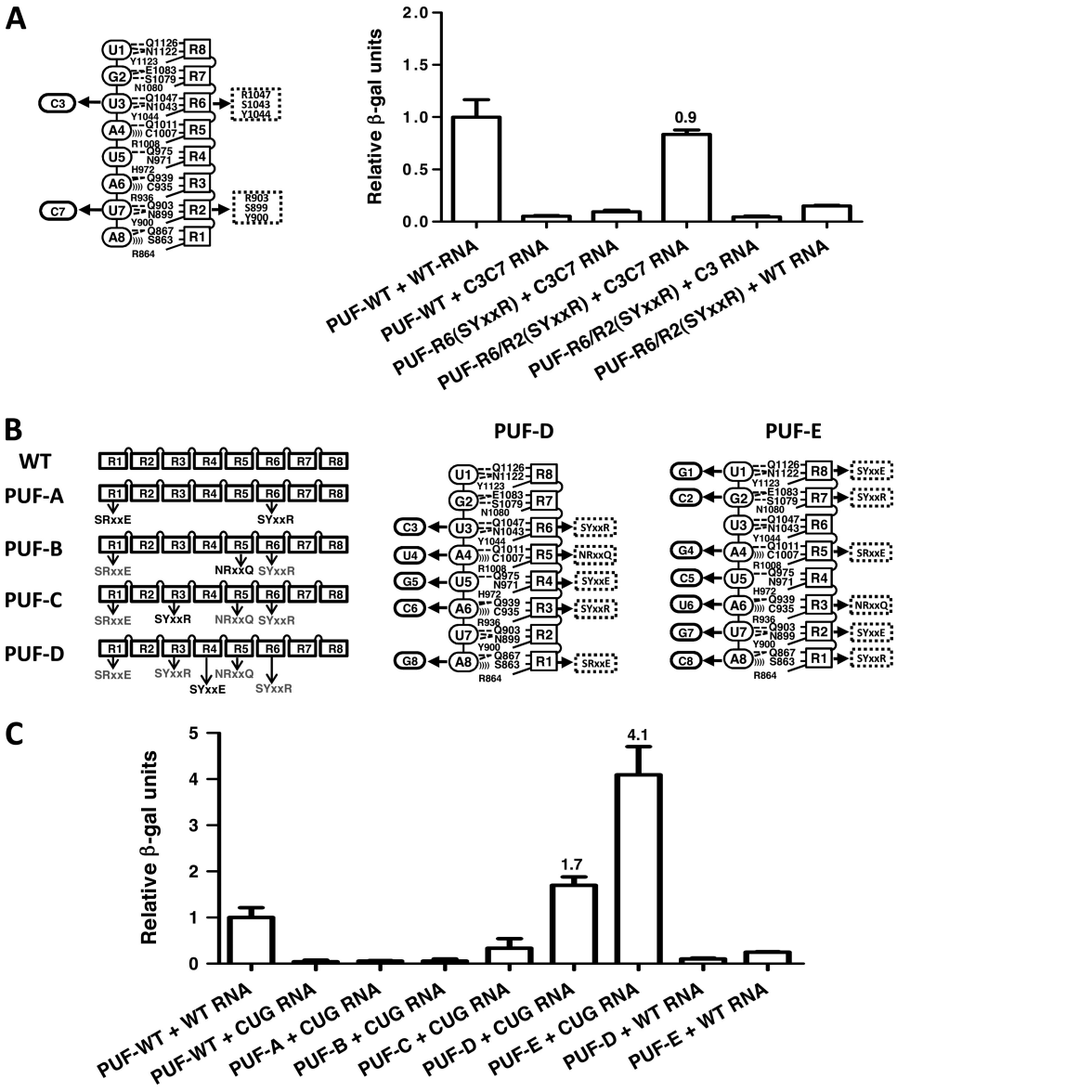


FIGURE 3. **Designed PUF domains that recognize targets with multiple cytosines.** *A*, diagram showing the mutations in PUM repeats 2 and 6 to recognize C3C7 RNA (*left*). Relative protein-RNA binding is shown as in Fig. 2 (*right*). *B*, stepwise generation of a PUF mutant (PUF-D) that can bind to $(CUG)_n$ repeat RNA. Diagrams show the mutations in PUM repeats 1, 3, 4, 5, and 6 (*center*, PUF-D) or in PUM repeats 1, 2, 3, 5, 7, and 8 (*right*, PUF-E) to recognize $(CUG)_n$ repeat RNA. *C*, relative protein-RNA binding to WT or $(CUG)_5$ RNA is shown as in Fig. 2. *Error bars* indicate S.D.

To further confirm RNA binding, we purified the recombinant PUF protein and used EMSA to demonstrate direct binding of PUF-R6(SY*XX*R) to C3 RNA (supplemental Fig. S2). Given the direct and specific interaction with this *in vitro* assay, we conclude that the expression of *LacZ* was indeed caused by the direct RNA-protein binding.

The Cytosine-recognition Code Can Be Transferred to Other PUM Repeats—To examine the modularity of the C-binding code we identified using PUM repeat 6, we applied the code to PUM repeats 2 and 5 that normally bind to U7 and A4, respectively. We then tested whether such changes specify cytosine recognition at the cognate positions (C7 for repeat 2 and C4 for repeat 5) using the Y3H assay. As predicted, mutation of the conserved RNA-interacting positions in repeat 2 (positions 899–903 becoming SY*XX*R) changed binding specificity from U7 to C7, whereas wild-type PUF did not recognize a C7 RNA target (Fig. 2*A*). Unlike PUF-R6(SY*XX*R), PUF-R2(SY*XX*R) did not recognize wild-type U7 RNA sequence.

Similarly, mutations in repeat 5 (C1007S/Q1011R or SR*XX*R) are sufficient to change the binding specificity from A4 to C4, whereas wild-type PUF does not recognize a C4 target RNA. Repeat 5 of wild-type PUF has an arginine (Arg-1008) in position to stack with the RNA base, and we found that the two mutations in the edge-interacting side chains were sufficient for cytosine recognition. Therefore, arginine can serve as the stacking amino acid residue in the C-binding code. However, introduction of a third mutation in repeat 5 (SY*XX*R in positions 1007–1011) maintained C-binding specificity and may better prevent binding to A4-containing RNA (Fig. 2*B*).

Effect of the Stacking Residue on Cytosine Recognition—It has been shown recently that the identity of the amino acid side chain that stacks with the RNA base is important for the spec-

ificity of PUM repeat-RNA interactions (26). Random and systematic mutagenesis of stacking residues of *C. elegans* FBF-2 indicated that a change from wild type in the stacking amino acid side chain of a PUM repeat can relax binding specificity to the cognate base and, to a lesser extent, the adjacent base. Tyrosine, histidine, and arginine are most commonly found at the stacking position in PUF repeats in the SMART database (accession number SM00025) and were also shown in mutagenesis screens and systematic mutagenesis experiments to maintain RNA binding of FBF-2 (26). Based on modifications of repeat 5, we established that arginine can serve as a stacking side chain for cytosine (Fig. 2*B*). We therefore further tested the effects of the identity of the stacking side chain on cytosine specificity.

We evaluated the effect of stacking side chain identity using position 1044 in repeat 6 of Pumilio 1, which has wild-type RNA interaction motif NY*XX*Q that recognizes U3. We mutated the interaction motif to SH*XX*R and measured binding of the mutant protein, PUF-R6(SY*XX*R), to wild-type U3 and mutant C3 RNA targets (Fig. 2*C*). PUF-R6(SH*XX*R) binds well to RNA containing C3 and more weakly to U3. When the stacking side chain is changed to tyrosine, PUF-R6(SY*XX*R), we see similar effects. We conclude that specific binding of cytosine can be achieved with Y/H/R as stacking residue in the cognate repeat.

Another naturally, although uncommonly, occurring side chain at the stacking position is asparagine, as seen in repeat 7 of wild-type Pumilio 1. This repeat specifically recognizes a G base with an SN*XX*E RNA interaction motif, but the side chain of Asn-1080 is not long enough to form a stacking interaction with G2 (8, 11). To change the specificity of repeat 7 to cytosine, the base-interacting residues were mutated initially to Ser-1079/Arg-1083. However, we found that PUF-R7(SN*XX*R) did not bind to target RNAs containing G2 or C2, as judged by Y3H measurement (Fig. 2*D*). When we also changed the stacking residue in repeat 7 to tyrosine (N1080Y), the resulting PUF-R7(SY*XX*R) bound strongly to C2 and more weakly to G2 (Fig. 2*D*). Thus, for cytosine recognition, a side chain forming a stacking interaction with the RNA base appears required for binding.

In addition to residues in the cognate repeat, we found that the identity of the stacking residue in the following repeat, which also contacts the RNA base, can contribute to the binding affinity at some positions. Most wild-type repeats in Pumilio 1 have tyrosine or arginine as the stacking residue in the following repeat, the only exception being repeat 3 with a histidine in repeat 4. When we tried to transfer the C-binding code to repeat 3, we found that neither SR*XX*R nor SY*XX*R introduced recognition of the cognate C6 (Fig. 2*E*). However, mutation of the following stacking residue to tyrosine (H972Y) allowed recognition of the C6 target (Fig. 2*E*).

Designed PUF Domains Recognize Targets with Multiple Cytosines—To extend our studies of the modularity of the C-recognition code, we sought to engineer new PUFs that can recognize multiple C residues in their target RNA sequences. We first created a PUF to recognize the sequence UGCAUACA (C3C7) by combining previously studied modifications in repeats 2 and 6. We found that only the PUF with both modified repeats, PUF-R6/R2(SY*XX*R), but neither the wild-type PUF nor PUFs with one modified repeat bound to the C3C7 sequence (Fig. 3*A*). This binding is specific because PUF-R6/R2(SY*XX*R) with two modified repeats did not bind to RNAs with one cytosine (C3U7) or no cytosines (wild-type U3U7) at cognate positions (Fig. 3*A*).

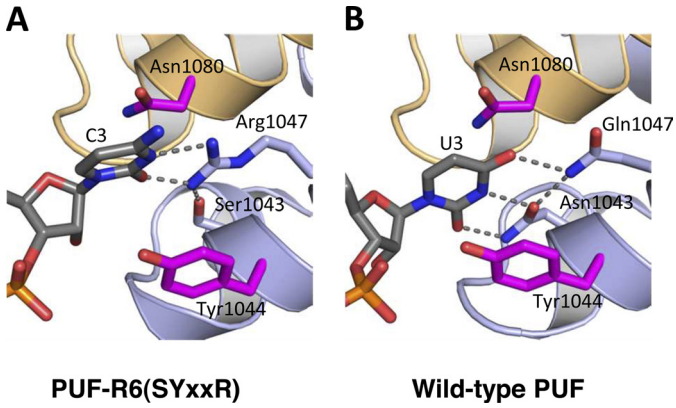


FIGURE 4. **Crystal structure of PUF-R6(SY*XX*R) in complex with C3 RNA.** *A*, interaction of PUF-R6(SY*XX*R) with C3 RNA. A ribbon diagram of interaction of repeat 6 with C3 base (complex 1 with chain A and C displayed) is shown. *B*, interaction of wild-type PUF (NY*XX*Q) with U3 RNA. A ribbon diagram of interaction of repeat 6 with U3 base is shown. RNA and base-interacting side chains are shown as *stick models* colored by atom type (*red*, oxygen; *blue*, nitrogen; *orange*, phosphorus). Carbon atoms are colored *gray* in RNA and *light blue* in RNA edge-interacting side chains, and *magenta* inside chains are in position to stack with the RNA base. Hydrogen bonds are indicated with *dashed lines*. This figure was created with PyMOL.

We next designed two PUFs that recognize 8-nucleotide signature sequences in $(CUG)_n$ RNA repeats. Expanded $(CUG)_n$ RNA repeats cause myotonic dystrophy type 1 (DM1). These toxic RNA repeats accumulate in the nucleus and sequester alternative splicing factors that normally regulate genes important for muscle and heart functions, thus leading to the pathogenesis observed in DM1 (27, 28). Through stepwise mutagenesis, we generated two PUF domains that recognize different frames of $(CUG)_n$ repeats. These proteins could be used to compete the binding of splicing factors to pathogenic $(CUG)_n$ repeats. PUF-D was designed to recognize UGCUGCUG with five mutated repeats (R1(SR*XX*E), R3(SY*XX*R), R4(SY*XX*E), R5(NR*XX*Q), and R6(SY*XX*R)), and PUF-E was designed to recognize GCUGCUGC with mutations in six repeats (R1(SY*XX*R), R2(SY*XX*E), R3(NR*XX*Q), R5(SR*XX*E), R7(SY*XX*R), and R8(SY*XX*E)) (Fig. 3*B*). We found that PUF-D and PUF-E bound strongly to a $(CUG)_5$ target RNA but not to control RNA, whereas wild-type PUF and intermediate PUFs A to C essentially had no interaction with the $(CUG)_5$ target (Fig. 3*B*). The *de novo* design of $(CUG)_n$-binding PUFs demonstrates the potential to generate new RNA-binding scaffolds that may be used for therapeutic applications.

Crystal Structure of PUF-R6(SYXXR) and Cognate C3-containing RNA—To examine how the side chains forming the C-recognition code are used to specifically recognize cytosine, we determined a crystal structure of PUF-R6(SY*XX*R) in complex with a cognate C3 RNA (5′-AUUG**C**AUAUA-3′, supplemental Table S1). In the structure, Arg-1047 contacts the O2 and N3 positions of the cytosine (Fig. 4*A*). Ser-1043 forms a hydrogen bond with an amino group of the arginine side chain, assisting in positioning Arg-1047. This interaction is similar to

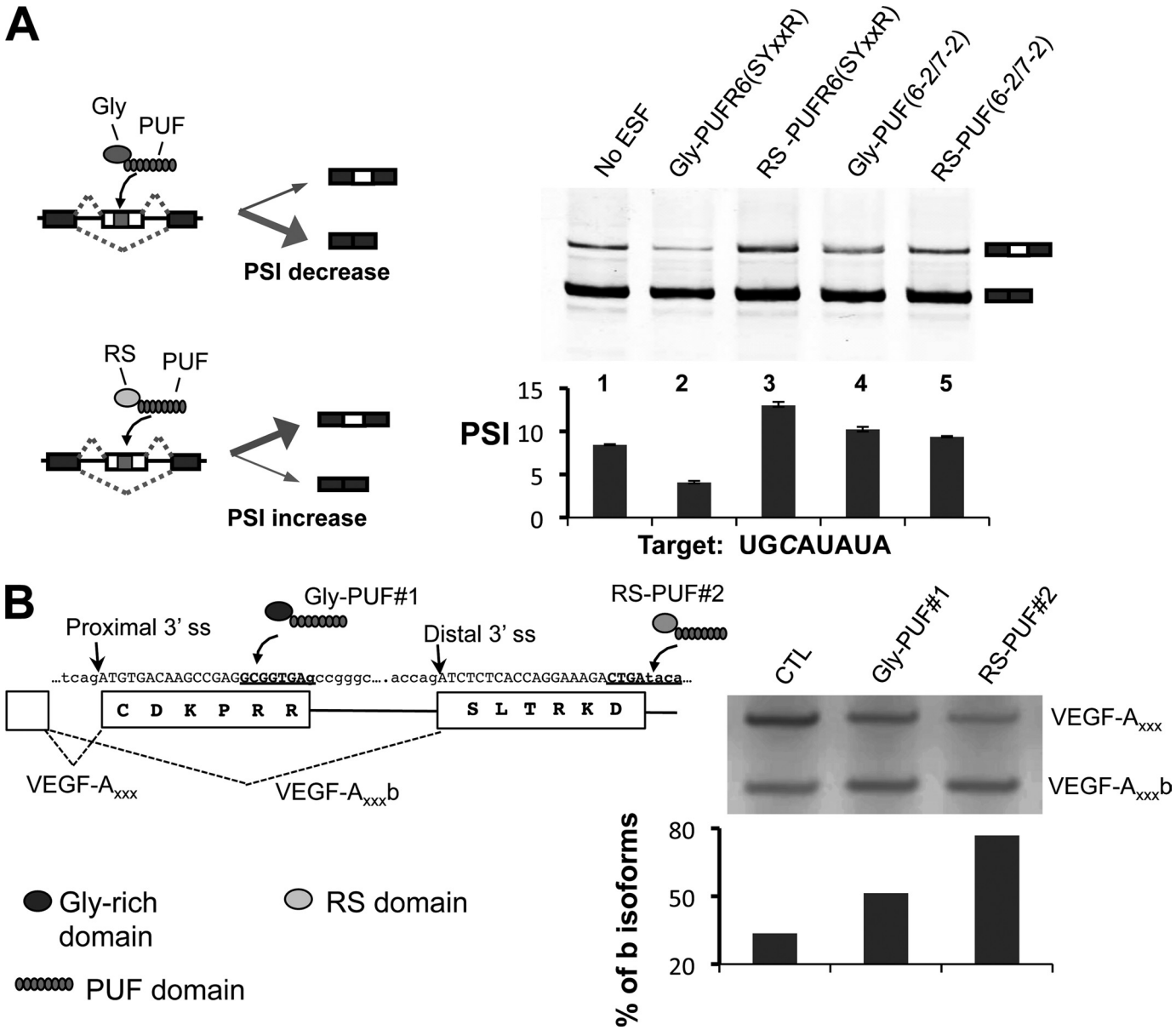


FIGURE 5. **Using the cytosine-recognition code to direct engineered splicing factors.** *A*, modulating alternative splicing of a cassette exon in a reporter RNA. *Left*, diagram of how the two types of ESFs can affect splicing of a cassette exon. Gly-PUF ESF directed to the exonic target can increase exon inclusion, whereas the RS-PUF ESF can decrease exon inclusion. *Top right*, RT-PCR products of splicing reactions; *bottom right*, quantification of splicing. The splicing reporter gene and expression vectors for different ESFs were co-transfected at 1:2 ratio into 293T cells. Total RNA was purified 24 h after transfection, and splicing of the test exon was detected with RT-PCR. The percentage of exon included isoform among all isoforms is represented with PSI value (percentage spliced in). The transfections were carried out in duplicate, and the means of the PSI value were plotted with the *error bars* indicating the data range. Significant changes (*p* values are 0.04 and 0.01 for *lanes 2* and *3* as judged by paired Student's *t* test) were observed for ESFs that recognize cognate C-containing target. *B*, design of ESFs to target endogenous VEGF-A pre-mRNA splicing. The gene and protein sequences of VEGF-A in the region near the alternative splice sites are shown with two PUFs recognizing different cytosine-containing sequences (*left panel*, *underlined sequences*). To shift the splicing toward anti-angiogenic VEGF-A isoforms, the cultured MDA-MB-231 cells were transfected with 1 μg of expression vectors of Gly-PUF#1 or RS-PUF#2. Total RNA was purified 24 h after transfection to detect VEGF-A splicing by RT-PCR. The percentages of b isoforms were quantified and are plotted below the gel (*left*). *CTL*, control.

the interaction of Asn-1043 and Gln-1047 in the wild-type protein with the Watson-Crick edge of U3 (Fig. 4*B*), although the longer arginine side chain requires the cytosine base ring position to be shifted slightly away from the RNA-binding surface. Interaction with only the known base-interacting side chains is consistent with the ability to transfer C-recognition to other PUM repeats. The crystal structure also indicates that other small side chains could occupy the position of Ser-1043 and that alternate conformations of Arg-1047 can recognize the cytosine, but the ability of the serine side chain to assist in positioning the arginine side chain may produce tighter binding (25).

Applying the Cytosine-recognition Code to Designed Artificial Splicing Factors—The PUF domain has been used as an RNA-binding scaffold to engineer novel protein factors for *in vivo* RNA localization (12, 13) and for manipulation of alternative splicing (16). Previously, we developed ESFs by combining a designed PUF domain with different splicing modulation domains to specifically regulate different types of alternative splicing events (16).

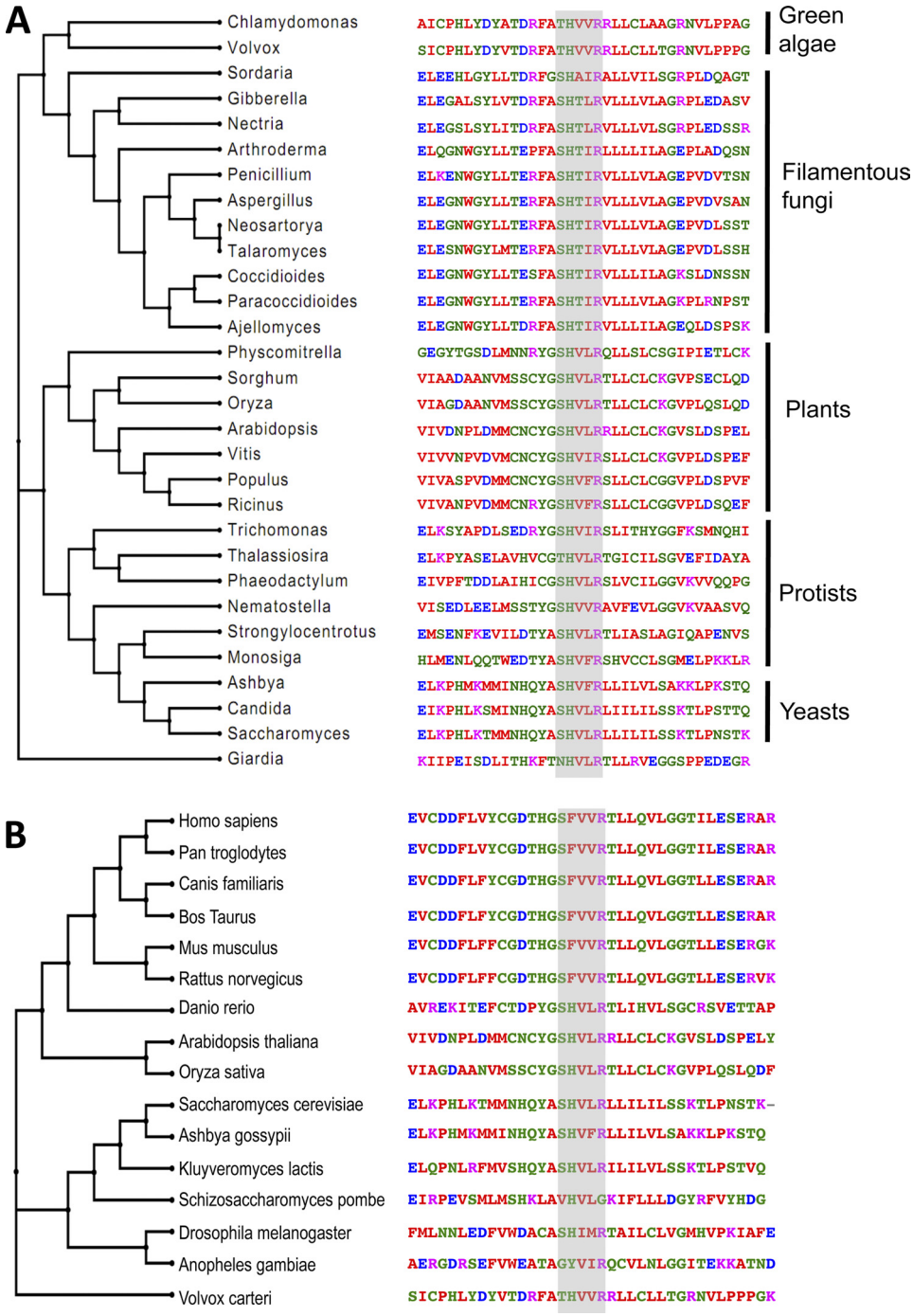


FIGURE 6. **Natural PUF proteins with putative cytosine-recognition code.** *A*, alignment and phylogenetic tree of the putative C-recognition PUM repeat in Nop9p homologs from yeast, plants, filamentous fungi, and protists. The query sequences were selected to maximize the divergence of the species but are otherwise arbitrary. The Giardia protein EES98274 was the chosen as the outgroup in the phylogenetic tree. *B*, alignment of the putative C-recognition PUM repeat in Nop9p homologs from the HomoloGene database. The homologous *Volvox carteri* protein XP_002952190 was included in the alignment as the outgroup in the phylogenetic tree. The conserved positions for cytosine recognition were *highlighted*.

To expand the application of ESFs, we created ESFs that can target C-containing elements by fusing either the Gly-rich domain of heterogeneous nuclear ribonucleoprotein A1 or the RS domain of ASF/SF2 with the PUF-R6(SY*XX*R) domain that specifically recognizes UGCAUAUA. We tested this ESF by co-transfecting 293T cells with plasmids expressing the ESF and a splicing reporter containing the cognate 8-nucleotide target sequence in an alternatively spliced cassette exon. Changes in alternative splicing were analyzed using body-labeled RT-PCR (Fig. 5*A*, *left panel*) (16). As designed, the Gly-PUF-R6(SY*XX*R) ESF repressed the inclusion of the cassette exon containing a UGCAUAUA target sequence, whereas the RS-PUF-R6(SY*XX*R) ESF increased exon inclusion (Fig. 5*A*, *lanes 2* and *3*). Splicing modulation is sequence-specific as control ESFs with non-cognate PUF domains had little effect on exon inclusion (Fig. 5*A*, *lanes 4* and *5*).

We further designed new ESFs to control the splicing of an endogenous gene using recognition of a C-containing target

sequence. We chose to manipulate the alternative splicing of VEGF-A, an important mediator of angiogenesis and a key anti-tumor target. The VEGF-A gene contains eight exons that undergo extensive alternative splicing to produce multiple isoforms (supplemental Fig. S3). One newly discovered class of isoforms (b isoforms) has anti-angiogenic activity that is opposite to canonical VEGF-A isoforms (29, 30). Most solid cancers are associated with a switch from the VEGF-A b isoforms to the pro-angiogenic a isoforms to promote angiogenesis. Thus, restoring the normal splicing balance to the b isoforms may have potential as a new anti-VEGF cancer therapy.

The two classes of VEGF-A isoforms are generated by the alternative use of a 3′ splice site (ss) in exon 8 (Fig. 5*B*). Pro-angiogenic isoforms are spliced with a proximal 3′ ss, and the anti-angiogenic b isoforms are spliced with a distal 3′ ss. The choice of alternative 3′ ss is generally controlled by regulatory *cis*-elements between the proximal and distal splice sites and/or inside the core exonic region. Therefore, we designed new PUF domains to specifically recognize sequences in these regions.

Two ESFs were designed to modulate VEGF-A alternative splicing; PUF#1 recognized the sequence GCGGUGAG between the proximal and distal 3′ ss, and PUF#2 recognized the sequence CUGAUACA downstream of the distal 3′ ss (Fig. 5*B*, *left panel*, *blue sequences*). The Gly-PUF#1 ESF should inhibit splicing of pro-angiogenic isoforms (VEGF-A_{xxx}), whereas the RS-PUF#2 ESF should promote anti-angiogenic VEGF-A_{xxx} b isoforms; thus, both should shift VEGF-A splicing toward the b isoforms. When each ESF was expressed in MDA-MB-231 cells, we indeed found that either ESF shifted splicing toward the anti-angiogenic isoforms.

DISCUSSION

The identification of a modular code to recognize cytosine makes it now possible to design PUF domains to bind any given sequence and broadens opportunities to create new research tools and therapeutic reagents. We demonstrated this application by developing new ESFs to specifically modulate the alternative splicing of VEGF-A, a key regulator of angiogenesis and cancer growth, and designing PUF domains that recognize pathogenic CUG repeats. Combined with gene delivery tools, such artificial proteins can potentially be used as new therapeutic reagents.

The identification of this C-binding motif by selection also suggests that C-binding repeats exist in natural proteins, although PUF proteins in human, *Drosophila*, and *C. elegans* with PUM repeats that recognize specifically cytosine have not been identified. The SMART database includes 4032 PUM repeats (accession number SM00025) in 600 proteins (31, 32). Among these PUM repeats, we found two *Saccharomyces cerevisiae* PUF proteins, Puf2p and Nop9p, that appear to contain a PUM repeat with RNA-interacting side chains similar to the C-binding code we identified.

Puf2p interacts preferentially with mRNAs encoding membrane-associated proteins (33). It contains a classical RNA recognition motif followed by six PUM repeats. Repeat 4 of Puf2p has an SR*XX*R RNA interaction motif, but homologs of Puf2p are restricted to the fungi and the putative C-binding code (SR*XX*R) is only found in Puf2p of *Vanderwaltozyma polyspora* and *S. cerevisiae* (supplemental Fig. S4). Other Puf2p homologs have the sequence AR*XX*R in cognate positions. Thus, it is unclear whether repeat 4 of Puf2p is a natural C-binding repeat.

The other protein with a putative C-recognition repeat, Nop9p, is involved in rRNA processing (34) and is essential for yeast survival. It has eight PUM repeats with longer intervening sequences between some repeats than are seen in more typical RNA-binding PUF proteins. Its PUM repeats are considerably divergent in sequence from those typically found in PUF proteins with known RNA recognition specificity. A search of non-redundant protein sequences with PSI-BLAST suggests that Nop9p represents an ancient class of eukaryotic proteins with homologs in fungi, plants, and protists (Fig. 6*A*). PUM repeat 3 of Nop9p possesses an SH*XX*R base recognition motif, suggesting that this repeat may recognize cytosine naturally. We found that 26 of 30 PUM repeat sequences have a putative C-binding motif (SH*XX*R) in the conserved RNA-interacting positions and that three repeats have the motif TH*XX*R, both similar to the C-binding code identified in our Y3H screen (Fig. 6*A*). One exception, the Giardia Pumilio-like protein EES98274, was least related to the others and was deliberately chosen as the outgroup in calculating a phylogenetic tree.

Nop9p homologs in the National Center for Biotechnology Information (NCBI) HomoloGene database, which has identified homologous proteins in fully sequenced genomes, indicate that Nop9p belongs to a family of proteins encoded by diverse eukaryotes, including yeast, fish, plants, flies, and mammals (Fig. 6*B*). PUM repeat 3 in 13 of 15 homologs contains a putative C-binding motif of SH*XX*R or SF*XX*R (found in mammalian homologs). *Schizosaccharomyces pombe* and mosquito homologs are more divergent and may lack this repeat or the C-binding code. The sequences of all PUM repeats in the Nop9p homologs are considerably different from typical RNA-binding PUF proteins; thus, it remains to be seen whether this family of proteins binds RNA in the same manner as Pumilio 1. Our results enrich our appreciation of the diversity of PUF proteins by identifying the Nop9p subfamily whose target RNAs remain to be discovered and whose RNA recognition mode is likely to be distinct.

Acknowledgments—We thank Dr. Marvin Wickens (University of Wisconsin) for providing the plasmids and yeast strains for the Y3H assay. We thank Dr. Rajarshi Chudhury for suggestions and help in protein purification, Dr. Chen Qiu for help with RNA binding assays, and Dr. Lars Pedersen and the staff at the SER-CAT beamline for help with X-ray data collection. Data were collected at SER-CAT 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found on-line. Use of the Advanced Photon Source was supported by the United States Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract W-31-109-Eng-38.

REFERENCES

1. Auweter, S. D., Oberstrass, F. C., and Allain, F. H. (2006) *Nucleic Acids Res.* **34,** 4943–4959
2. Crittenden, S. L., Bernstein, D. S., Bachorik, J. L., Thompson, B. E., Gallegos, M., Petcherski, A. G., Moulder, G., Barstead, R., Wickens, M., and

Kimble, J. (2002) *Nature* **417,** 660–663
3. Wickens, M., Bernstein, D. S., Kimble, J., and Parker, R. (2002) *Trends Genet.* **18,** 150–157
4. Dubnau, J., Chiang, A. S., Grady, L., Barditch, J., Gossweiler, S., McNeil, J., Smith, P., Buldoc, F., Scott, R., Certa, U., Broger, C., and Tully, T. (2003) *Curr. Biol.* **13,** 286–296
5. Schweers, B. A., Walters, K. J., and Stern, M. (2002) *Genetics* **161,** 1177–1185
6. Ye, B., Petritsch, C., Clark, I. E., Gavis, E. R., Jan, L. Y., and Jan, Y. N. (2004) *Curr. Biol.* **14,** 314–321
7. Chen, G., Li, W., Zhang, Q. S., Regulski, M., Sinha, N., Barditch, J., Tully, T., Krainer, A. R., Zhang, M. Q., and Dubnau, J. (2008) *PLoS Comput. Biol.* **4,** e1000026
8. Wang, X., McLachlan, J., Zamore, P. D., and Hall, T. M. (2002) *Cell* **110,** 501–512
9. Wang, X., Zamore, P. D., and Hall, T. M. (2001) *Mol. Cell* **7,** 855–865
10. Lu, G., and Hall, T. M. (2011) *Structure* **19,** 361–367
11. Cheong, C. G., and Hall, T. M. (2006) *Proc. Natl. Acad. Sci. U.S.A.* **103,** 13635–13639
12. Ozawa, T., Natori, Y., Sato, M., and Umezawa, Y. (2007) *Nat. Methods* **4,** 413–419
13. Tilsner, J., Linnik, O., Christensen, N. M., Bell, K., Roberts, I. M., Lacomme, C., and Oparka, K. J. (2009) *Plant J.* **57,** 758–770
14. Opperman, L., Hook, B., DeFino, M., Bernstein, D. S., and Wickens, M. (2005) *Nat. Struct. Mol. Biol.* **12,** 945–951
15. Koh, Y. Y., Opperman, L., Stumpf, C., Mandan, A., Keles, S., and Wickens, M. (2009) *RNA* **15,** 1090–1099
16. Wang, Y., Cheong, C. G., Hall, T. M., and Wang, Z. (2009) *Nat. Methods* **6,** 825–830
17. Zhu, D., Stumpf, C. R., Krahn, J. M., Wickens, M., and Hall, T. M. (2009) *Proc. Natl. Acad. Sci. U.S.A.* **106,** 20192–20197
18. Hook, B., Bernstein, D., Zhang, B., and Wickens, M. (2005) *RNA* **11,** 227–233
19. Stumpf, C. R., Opperman, L., and Wickens, M. (2008) *Methods Enzymol.* **449,** 295–315
20. Fox, J. E., Burow, M. E., McLachlan, J. A., and Miller, C. A., 3rd (2008) *Nat. Protoc* **3,** 637–645
21. Otwinowski, Z., and Minor, W. (1997) *Methods Enzymol.* **276,** 307–326
22. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) *J. Appl. Crystallogr.* **40,** 658–674
23. Emsley, P., and Cowtan, K. (2004) *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132
24. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221
25. Filipovska, A., Razif, M. F. M., Nygård, K. K. A., and Rackham, O. (2011) *Nat. Chem. Biol.*, doi: 10.1038
26. Koh, Y. Y., Wang, Y., Qiu, C., Opperman, L., Gross, L., Tanaka Hall, T. M., and Wickens, M. (2011) *RNA* **17,** 718–727
27. Wheeler, T. M., and Thornton, C. A. (2007) *Curr. Opin. Neurol.* **20,** 572–576
28. Lee, J. E., and Cooper, T. A. (2009) *Biochem. Soc. Trans.* **37,** 1281–1286
29. Harper, S. J., and Bates, D. O. (2008) *Nat. Rev. Cancer* **8,** 880–887
30. Qiu, Y., Hoareau-Aveilla, C., Oltean, S., Harper, S. J., and Bates, D. O. (2009) *Biochem. Soc. Trans.* **37,** 1207–1213
31. Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95,** 5857–5864
32. Letunic, I., Doerks, T., and Bork, P. (2009) *Nucleic Acids Res.* **37,** D229–D232
33. Gerber, A. P., Herschlag, D., and Brown, P. O. (2004) *PLoS Biol.* **2,** E79
34. Thomson, E., Rappsilber, J., and Tollervey, D. (2007) *RNA* **13,** 2165–2174