BJ www.biochemj.org

Biochem. J. (2012) **441**, 591–597 (Printed in Great Britain) doi:10.1042/BJ20111385

online data

Structure of the mRNA splicing complex component Cwc2: insights into RNA recognition

Peilong LU*†[1], Guifeng LU*‡[1], Chuangye YAN*†, Li WANG§, Wenqi LI*† and Ping YIN*†[2]

*Center of Structural Biology, Tsinghua University, Beijing 100084, China, †School of Life Sciences, Tsinghua University, Beijing 100084, China, ‡School of Medicine, Tsinghua University, Beijing 100084, China, and §School of Life Sciences, Peking University, Beijing 100084, China

The Prp19-associated complex [NTC (nineteen complex)] plays a crucial role in intron removal during premature mRNA splicing in eukaryotes. Only one component of the NTC, Cwc2, is capable of binding RNA. In the present study we report the 1.9 Å (1 Å = 0.1 nm) X-ray structure of the Cwc2 core domain, which is both necessary and sufficient for RNA binding. The Cwc2 core domain contains two sub-domains, a CCCH-type ZnF (zinc finger) and a RRM (RNA recognition motif). Unexpectedly, the ZnF domain and the RRM form a single folding unit, glued together by extensive hydrophobic interactions and hydrogen bonds. Structure-guided mutational analysis revealed that the intervening loop [known as the RB loop (RNA-binding loop)] between ZnF and RRM plays an essential role in RNA binding. In addition, a number of highly conserved positively charged residues on the β-strands of RRM make an important contribution to RNA binding. Intriguingly, these residues and a portion of the RB loop constitute an extended basic surface strip that encircles Cwc2 halfway. The present study serves as a framework for understanding the regulatory function of the NTC in RNA splicing.

Key words: RNA recognition motif, nineteen complex (NTC), spliceosome, structural biology, zinc finger.

INTRODUCTION

Intron removal in pre-mRNA (premature mRNA) is an essential step in eukaryotic mRNA processing carried out by the spliceosome, a multi-component RNP (ribonucleoprotein) assembly [1,2]. The spliceosome contains five major components, namely U1, U2, U4, U5 and U6 snRNPs (small nuclear RNPs), which collectively catalyse a two-step transesterification reaction. In addition to these five snRNPs, several protein complexes are involved in this crucial reaction. The NTC (nineteen complex) is formed by the scaffold protein Prp19 and a number of associated splicing factors. The NTC joins and stays with the spliceosome during the two-step splicing reaction, indicating an important role for this complex in pre-mRNA splicing [3,4]. For example, the NTC is required for stable association of the U5 and U6 snRNPs with spliceosome [4] and regulates the second step of the reaction [5]. Many RNA–protein and RNA–RNA interactions are specified by the NTC [1,6].

There are at least ten components in the yeast NTC [5,7,8], among which only one protein, Cwc2/NTC40 (hereafter referred to as Cwc2), is capable of binding to RNA. Cwc2 is predicted to contain two RNA-binding motifs at its N-terminus, including a CCCH-type ZnF (zinc finger) and an RRM (RNA recognition motif) [9]. Furthermore, the flexible C-terminus of Cwc2 interacts with the WD40 domain of Prp19 [10]. *In vivo* depletion of Cwc2 resulted in the destabilization of spliceosome snRNA [9]. Purified full-length Cwc2 protein exhibited normal RNA-binding capacity with low sequence specificity, whereas the RRM together with the C-terminal flexible region of Cwc2 displayed a reduced level of RNA binding [9]. However, it remains unclear how Cwc2 binds to RNA.

Both the ZnF and the RRM are among the most abundant and well-studied structural motifs in eukaryotes. The ZnF, which comes in all flavours, is capable of not only recognizing nucleic acids but also of mediating protein–protein interactions [11–14]. On the basis of structural analysis, ZnF motifs were classified into eight fold groups [13]. The TIS11d family, with $CX_8CX_5CX_3H$ (where X, any amino acid) sequence, was identified as a new subgroup of ZnF motif which contains few secondary structural elements [15]. But there is no other available structure to support this newly classified subtype. On the other hand, RRM-containing proteins play an important role in most post-transcriptional processes owing to their diverse modes of RNA binding in higher organisms. The canonical RRM binds to RNA through three aromatic residues located in the consensus sequences termed RNP1 and RNP2 [16,17]. Bioinformatic studies indicated that half of RRM-containing proteins harbour multiple copies of RRM or other domains, which are often found to be ZnF motifs [17]. RRM is involved not only in RNA recognition but also in protein–protein interactions [16]. Despite several reports on protein–protein interactions involving the RRMs [18–22], it remains unknown whether RRM can directly associate with ZnF and, if so, how this might happen.

In the present study we have determined the structure of the core domain of Cwc2 at 1.9 Å (1 Å = 0.1 nm) resolution by X-ray crystallography. Structural analysis revealed that the core domain of Cwc2 protein contains two subdomains, a CCCH-type ZnF domain and a RRM. To our surprise, the ZnF domain and the RRM are closely associated with each other through a large buried interface, appearing as a single folding unit. The extensive hydrophobic interface between ZnF and RRM is augmented by networks of hydrogen bonds. The ZnF domain

Abbreviations used: EMSA, electrophoretic mobility-shift assay; Ni-NTA, Ni^{2+}-nitrilotriacetate; NTC, nineteen complex; pre-mRNA, premature mRNA; RB loop, RNA-binding loop; rmsd, root mean squared deviation; RNP, ribonucleoprotein; RRM, RNA recognition motif; SAD, single-wavelength anomalous dispersion; snRNP, small nuclear ribonucleoprotein; ZnF, zinc finger.

[1] These authors contributed equally to this work.

[2] To whom correspondence should be addressed (email yinping@biomed.tsinghua.edu.cn).

The structure of the mRNA splicing complex component Cwc2 amino acids 1–121 + 133–227 and 1–127 alone have been deposited in the Protein Data Bank under accession numbers 3U1L and 3UIM respectively.

shows remarkable structural similarity to the TIS11d family, suggesting that it belongs to this new subtype. Interestingly, the linker loop connecting the ZnF and RRM domains, together with specific residues on several β-strands of the RRM, forms a positively charged surface strip that plays an essential role in RNA binding. The structural features, together with biochemical characterization, allowed us to propose a working model for RNA recognition.

EXPERIMENTAL

Protein preparation and crystallization

All clones were generated using a standard PCR-based cloning strategy, and mutagenesis of Cwc2 was generated with two-step PCR. The identities of individual clones were verified through double-strand plasmid sequencing. Cwc2 variants were overexpressed in the *Escherichia coli* strain BL21(DE3) at 18 °C using pET15b vectors with an N-terminal His_6 tag or pET21b vectors with a C-terminal His_6 tag. Cwc2 ZnF domain and RRM were cloned in pET15b and pBB75 vectors respectively and were co-expressed in *E. coli* strain BL21(DE3). The soluble fraction of the *E. coli* lysate was purified over a Ni-NTA (Ni^{2+}-nitrilotriacetate) column (Qiagen). After affinity purification, all proteins were further purified by cation-exchange chromatography (Source-15S; GE Healthcare) and size-exclusion chromatography (Superdex-200; GE Healthcare). The protein concentrations were determined by spectroscopic measurement at 280 nm. Crystals were grown at 18 °C using the hanging-drop vapour diffusion method. Rod-shaped crystals appeared overnight in the well buffer containing 21 % PEG [poly(ethylene glycol)] 3350, 200 mM ammonium citrate and 100 mM sodium citrate, pH 6.5, and grew to full size in 3 days.

Data collection, structure determination and refinement

The complex of Cwc2-(1–121) and Cwc2-(133–227) SAD (single-wavelength anomalous dispersion) and native data were collected at the SSRF (Shanghai Synchrotron Radiation Facility) beamline BL17U. Cwc2-(1–227) native data was collected on the Rigaku Saturn 944 + CCD (charge-coupled-device) configured with the Rigaku MicroMax-007HF generator. All data were integrated and scaled using the HKL2000 package. Further processing was carried out using programs from the CCP4 suite [23].

The zinc position in the Zn-SAD data of the complex of Cwc2-(1–121) and Cwc2-(133–227) was determined by the program SHELXD [24]. The identified single zinc atom was refined and the initial phases were generated in the program PHASER [25] with the SAD experimental phasing module. The real-space constraints were applied to the SAD electron density with density modification. A crude model was traced automatically using the program BUCCANEER [26] and was optimized further by RESOLVE [27] together with PHASER. Manual model building and refinement were performed iteratively with COOT [28] and PHENIX [29]. The Cwc2 model obtained from Zn-SAD data was used for molecular replacement with the program PHASER into the native data of the complex of Cwc2-(1–121) and Cwc2-(133–227) and then using this complex native structure as a molecular replacement model for Cwc2-(1–227) native structure determination. Both structures were refined with COOT and PHENIX iteratively. Data collection and refinement statistics are summarized in Table 1.

Limited proteolysis assay

The full-length Cwc2 protein was incubated with increasing concentrations of trypsin (Sigma) in 20 μl reaction buffer containing 20 mM Tris/HCl, pH 8.0, and 150 mM NaCl at room temperature (25 °C) for 10 min, followed by addition of 0.2 μl 100 mM PMSF. Half of each sample was separated on SDS/PAGE (15 % gel) and stained with Coomassie Blue R250. The stable band was excised and identified by MS analysis with Q-Star (ABI) after in-gel trypsin digestion.

EMSA (electrophoretic mobility-shift assay)

The yeast U6 snRNA was prepared and ^{32}P-labelled according to a previously published procedure [30]. For EMSA, a range of different concentrations of Cwc2 or mutant proteins were incubated with 20000 d.p.m. ^{32}P-labelled RNA probe for 30 min on ice in reaction buffer (20 mM Hepes/KOH, pH 7.9, 300 mM NaCl, 10 mM $ZnCl_2$, 10 % glycerol, 10 mM dithiothreitol, 0.45 mg/ml *E. coli* tRNA and 0.8 mg/ml BSA), followed by addition of 0.4 vol. glycerol loading dye (0.05 % Bromophenol Blue, 10 % glycerol) [9]. Reactions were then resolved on 15 cm native 5 % acrylamide gels (37.5:1 acrylamide/bisacrylamide) in 0.5× Tris-borate buffer containing 6 % glycerol at 20 V/cm for approximately 1.5 h. Dried gels were exposed to phosphorimager screens and analysed by a Typhoon 9400 variable scanner (Amersham Pharmacia).

RESULTS

Overall structure of the N-terminus of Cwc2

Cwc2 from *Saccharomyces cerevisiae* has been reported to contain a CCCH-type ZnF domain and a non-consensus RRM [9] (Figure 1A). The C-terminal sequences of Cwc2 are known to interact with Prp19 [10]. Sequence alignment revealed that the N-terminal two-thirds of Cwc2 are highly conserved in its orthologues from yeast to human, whereas their C-terminal sequences show little homology [9] (see Supplementary Figure S1 at http://www.BiochemJ.org/bj/441/bj4410591add.htm). To identify stable structural core domain(s) of Cwc2, the full-length protein (residues 1–337) was subjected to digestion by increasing amounts of trypsin. Figure 1(B) clearly shows that Cwc2 contains a trypsin-resistant core domain with an apparent molecular mass of approximately 25 kDa. The boundaries of this stable core domain were identified by MS to include residues 1–237 (results not shown).

We crystallized the structural core domain (residues 1–227) in the $P2_12_12_1$ space group (Table 1). The structure was determined at 1.9 Å resolution using a zinc-based SAD method (Table 1). The structural core domain of Cwc2 has a globular appearance, with a diameter of approximately 42 Å. As anticipated, the overall structure is composed of two domains (Figures 2A and 2B), a ZnF domain (61–120) and an RRM (136–227). The 15-residue linker sequence (121–135) between these two domains was named the RB loop (RNA-binding loop) owing to its essential role in RNA binding (described below). Five amino acids (127–131) in the middle of the RB loop had discontinuous electron density, suggesting a flexible conformation. Analysis of the surface electrostatic potential revealed a positively charged continuous strip which extends from the RB loop to the β-strands of RRM (Figure 2B), suggesting a potential role in RNA binding. This strip consists of at least eight positively charged residues: three lysine residues (Lys^{132}, Lys^{133} and Lys^{135}) from the C-terminal end of the RB loop, two arginine residues (Arg^{172} and Arg^{174}) from β2, Lys^{187}

Table 1 Data collection and refinement statistics

One crystal was used for each structure. Values in parentheses are for the highest resolution shell. $R_{merge} = \Sigma h \Sigma i |Ih,i - Ih| / \Sigma h \Sigma i Ih,i$, where Ih is the mean intensity of the i observations of symmetry-related reflections of h. $R = \Sigma |F_{obs} - F_{calc}| / \Sigma F_{obs}$, where F_{calc} is the calculated protein structure factor from the atomic model (R_{free} was calculated with 5 % of the reflections selected randomly).

Parameter	Zinc-SAD (1–121 + 133–227)	Cwc2-native (1–121 + 133–227) (PDB code 3U1L)	Cwc2-native (1–227) (PDB code 3U1M)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions			
a, b, c (Å)	44.91, 54.49, 112.01	44.85, 54.37, 111.77	44.72, 54.34, 111.38
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Wavelength (Å)	1.26770	0.96393	1.54178
Resolution (Å)	40∼2.65 (2.74∼2.65)	40∼1.64 (1.70∼1.64)	40∼1.90 (1.97∼1.90)
R_{merge} (%)	6.8 (14.9)	6.0 (33.9)	4.1 (11.7)
$I/\sigma I$	25.2 (15.4)	21.4 (3.4)	35.1 (10.2)
Completeness (%)	99.1 (99.9)	99.1 (95.5)	99.5 (95.6)
Redundancy	5.3	3.9	6.1
Resolution (Å)	40∼2.65	40∼1.64	40∼1.90
Number of reflections	8391	34029	22066
R_{work}/R_{free} (%)	17.12/18.93	15.89/19.37	
Number of atoms			
Protein	1787	1784	
Ligand/ion	1	1	
Water	303	326	
B-factors			
Protein	23.22	19.19	
Ligand/ion	11.01	11.84	
Water	35.47	29.63	
rmsd			
Bond lengths (Å)	0.006	0.011	
Bond angles (°)	0.977	1.177	
Ramachandran plot statistics (%)			
Most favoured	93.6	93.2	
Additional allowed	6.4	6.8	
Generously allowed	0.0	0.0	
Disallowed	0.0	0.0	

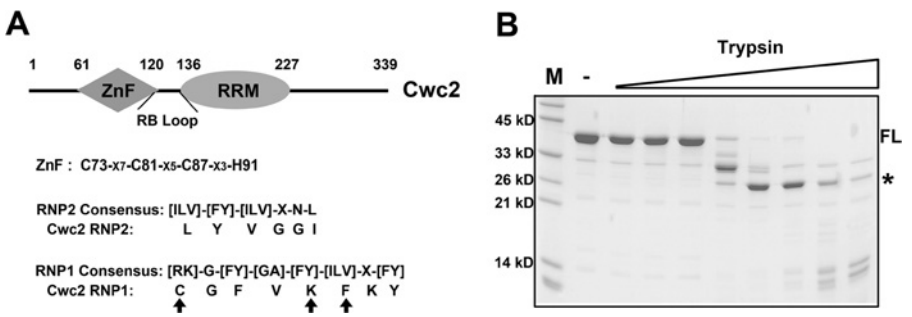

Figure 1 Identification of a structural core domain in Cwc2

(**A**) Domain structure and sequence features of Cwc2. The full-length Cwc2 protein contains an N-terminal ZnF domain, followed by a RRM and an extended C-terminal sequence. Assignment of the conserved sequences of ZnF and RRM based on the X-ray structure is shown below. The RB loop is the RNA-binding loop connecting the ZnF domain and the RRM. Arrows indicate the non-conserved residues in Cwc2. (**B**) Identification of a structural core domain in Cwc2 by trypsin digestion. The result, shown on a SDS/PAGE gel, led to the identification of residues 1–237 as the structural core domain (indicated by an asterisk, for more details see the Experimental section). FL, full length protein; M, protein markers with molecular mass on the left-hand side in kDa (kD).

from β3, Lys^{224} from β4 and Lys^{179} from the loop connecting β2 and β3 (Figure 2C and Supplementary Figure S1).

Unexpectedly, the ZnF domain is closely stacked against the RRM (Figure 3A), raising the possibility that these two domains might interact with each other independently without any covalent linker. To examine this scenario, we co-expressed these two domains (1–121 and 122–227, or 1–132 and 133–227). Interestingly, the untagged ZnF domain could be pulled down by Ni-NTA resin only when it was co-expressed with the His_6-tagged RRM (see the Supplementary Experimental section and Supplementary Figure S2 at http://www.BiochemJ.org/bj/441/bj4410591add.htm). Consistent with this result, 11 residues of the RB loops in the Cwc2 structural core domain (residues 1–227) could be removed by elevated concentrations of proteases. The protease-resistant core was also crystallized and the structure was determined at 1.7 Å resolution (see Supplementary Figure S3 at http://www.BiochemJ.org/bj/441/bj4410591add.htm and Table 1). The structure is almost identical to that of the intact Cwc2 core domain, with a rmsd (root mean squared deviation) of 0.142 Å for all aligned Cα atoms.

The interaction of RRM with ZnF is primarily mediated by the α10 helix. The interaction includes a network of extensive

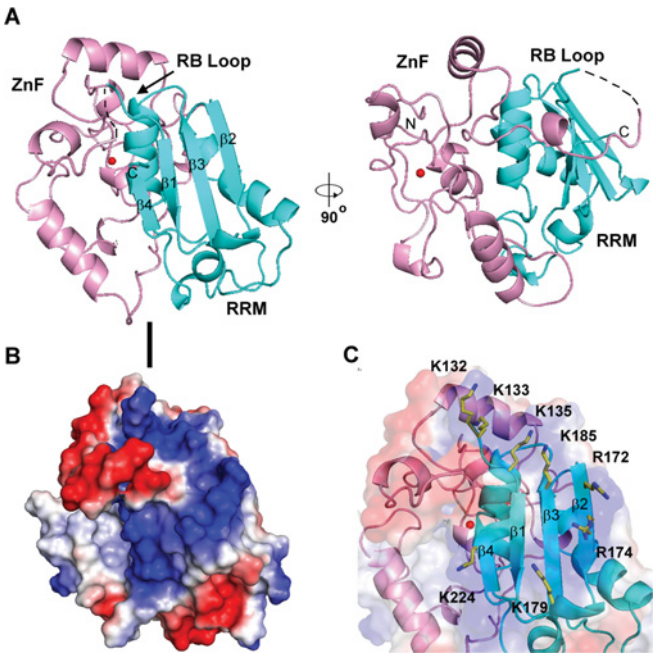

Figure 2 Overall structure of the structural core domain of Cwc2

(**A**) Overall structure of Cwc2 is shown in two perpendicular views. The ZnF domain and the RRM are coloured pink and cyan respectively. The zinc atom is shown as a red sphere. The five residues in the middle region of the RB loop are represented by a broken line because they are disordered in the crystals. (**B**) Surface electrostatic potential of Cwc2. There is a positively charged strip from the C-terminal end of the RB loop to the β-strands of RRM. (**C**) A close-up view of the eight constituent residues in the positively charged strip. The side chains are shown in stick representation. All structural Figures were prepared using PyMOL (http://www.pymol.org).

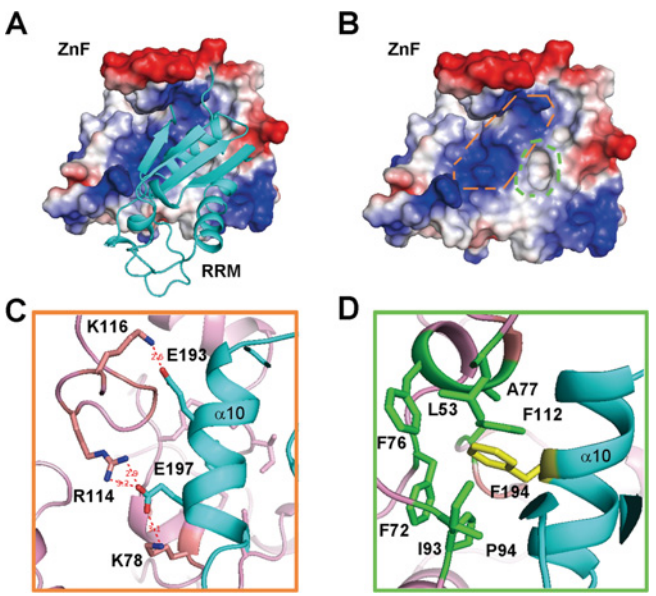

Figure 3 The ZnF domain and the RRM appear to form a single folding unit

(**A**) The ZnF domain is shown in surface electrostatic potential to highlight the interface between two domains. (**B**) The ZnF domain is shown in the same orientation as (**A**), but in the absence of the RRM. Note the positive charges and hydrophobic pocket on the surface of the ZnF domain. (**C**) A close-up view of the hydrogen bonds at the interface between the ZnF domain and the RRM. Hydrogen bonds are represented by broken red lines. Structural elements from ZnF and RRM are coloured pink and cyan respectively. (**D**) A close-up view of the van der Waals interactions at the interface between the ZnF domain and the RRM.

hydrophobic contacts, and four inter-domain hydrogen bonds. Notably, the hydrophobic contacts are mediated by the single aromatic residue Phe[194] on α10 of RRM, which is nestled in a hydrophobic cave formed by seven non-polar amino acids from ZnF: Leu[53], Phe[72], Phe[76], Ala[77], Ile[93], Pro[94] and Phe[112]. Reinforcing these hydrophobic contacts, the carboxylate side-chain of Glu[197] on α10 of RRM accepts three charge-stabilized hydrogen bonds from the side-chains of Arg[114] and Lys[78] on ZnF (Figures 3C and 3D). In addition, Glu[193] on α10 of RRM also accepts a hydrogen bond from Lys[116] of the ZnF motif (Figure 3C). Glu[193], Phe[194] and Glu[197], which make important contributions to the RRM–ZnF interaction, are highly conserved from yeast to human. Mutation of these residues led to abrogation of the interaction between RRM and ZnF, which was validated by pull-down assay (Supplementary Figure S2). These results suggested that the ZnF domain indeed interacts with the RRM to form a stable structure in the absence of a connecting loop.

The structure of the CCCH ZnF domain

The CCCH ZnF domain (residues 1–120) comprises seven short α-helices. The bound zinc ion is tetrahedrally co-ordinated by three cysteine residues (Cys[73], Cys[81] and Cys[87]) and one histidine residue (His[91]), which are arranged in a Cys-X_7-Cys-X_5-Cys-X_3-His sequence (Figure 4A and Supplementary Figure S1). This zinc-binding core (residues 72–93) is surrounded by helices α2 and α6 and the N-terminal loop sequences 1–14 and 18–30 (Figure 4A). The position of the zinc atom was determined by its anomalous signal, and the electron density of the entire metal-binding core is of excellent quality (see Supplementary Figure S4 at http://www.BiochemJ.org/bj/441/bj4410591add.htm). The zinc-binding core (residues 72–93) only contains a short helix α5 between the first two cysteine residues and is preceded by a loop and followed by another loop (Figure 4B). This structural fold comprises few secondary structural elements and is different from those of eight known ZnF fold groups [13]. A novel ZnF fold has been identified in the single-stranded RNA-binding protein TIS11d [15]. Structural alignment revealed that the ZnF domain from Cwc2 adopts a similar fold as the first or second ZnF domain of TIS11d (Figure 4C). We propose that the ZnF domains present in Cwc2 and TIS11d are classified as a new subtype of ZnF structure (Figure 4C).

TIS11d binds to RNA via hydrogen bonds, mediated by positively charged residues in TIS11d, and stacking interactions between two conserved aromatic residues and the RNA bases [15]. Mutation of either of the two aromatic residues in TIS11d resulted in abrogation of the RNA-binding activity [31]. Among the two aromatic residues of TIS11d, only one appears to be conserved in Cwc2. This residue, Tyr[89] in Cwc2, is exposed to solvent and thus might be involved in RNA binding. However, the missense mutation Y89A in Cwc2 only exhibited a moderate effect on RNA binding (see Supplementary Figure S5 at http://www.BiochemJ.org/bj/441/bj4410591add.htm), suggesting a limited role by the aromatic residues of the ZnF domain. In addition, the ZnF domain of Cwc2 does not contain the corresponding residues that mediate the critical hydrogen bonds in TIS11d–RNA interactions.

Because the isolated ZnF domain of Cwc2 remains completely insoluble, we are unable to assess its RNA-binding ability. The observation that several α-helices (α1, α2 and α6) and two loops (1–14 and 18–30) surround the ZnF domain raises the possibility that the ZnF domain could interact with RNA under conditions in which these α-helices and/or N-terminal loops undergo major conformational changes during dynamic spliceosome assembly.

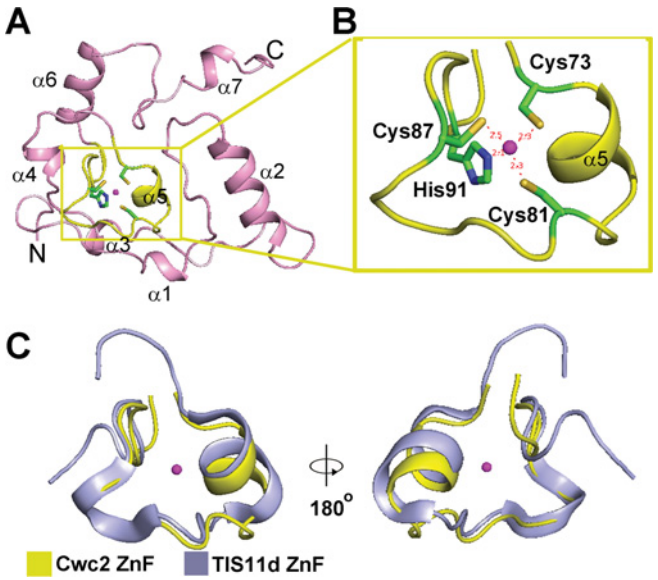

Figure 4 Structural features of the Cwc2 ZnF domain

(**A**) Structural features of the ZnF domain. The zinc-binding core (yellow inset) is surrounded by the N-terminal loops and helices α2 and α6. (**B**) A close-up view of the zinc-binding core structure (residues 72–94). The side chains of the four zinc-co-ordinating residues are shown. (**C**) Structural superposition of the Cwc2 ZnF (yellow) with the second ZnF of TIS11d (residues 187–220) (light blue) (PDB code 1RGO [15]). The zinc atom is shown as a magenta sphere.

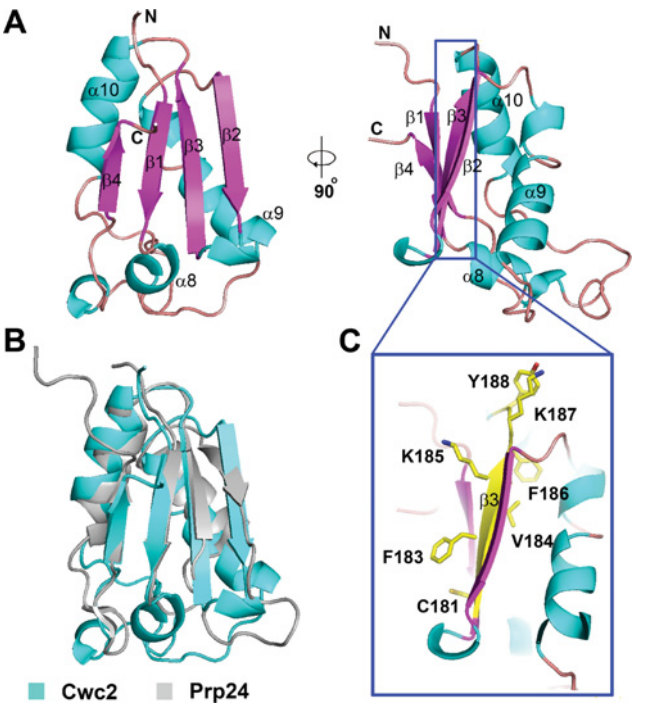

Figure 5 Structural features of the Cwc2 RRM

(**A**) Ribbon representation of the RRM in two perpendicular views. The α-helices are coloured cyan and the β-sheets are coloured magenta. Loops are shown in pink. The blue box indicates the position of the atypical RNP1 residues. (**B**) Superposition of the Cwc2 RRM (cyan) with the second RRM of Prp24 (grey, PDB code 2KH9 [22]), a component of the U6 snRNP. (**C**) A close-up view on the atypical RNP1 residues.

The structure of the RRM

The RRM of Cwc2 adopts a canonical $\beta\alpha\beta\beta\alpha\beta$ topology, with a four-stranded antiparallel β-sheet (β1, β2, β3 and β4) packed against two α-helices (α9 and α10) (Figure 5A). The overall structure is similar to that of other typical RRM, as exemplified by superposition with the second RRM of Prp24 [22] (Figure 5B). In addition, it contains a short helix, α8, between strand β1 and helix α9.

To examine the role of RRM in RNA binding, we used an RNA-binding assay with the full-length 112-nucleotide U6 snRNA that Cwc2 directly interacts with [9]. As anticipated, the full-length Cwc2 and the trypsin-resistant core domain exhibited robust RNA-binding activity and showed moderate binding affinity differences (Figure 6A, lanes 1–10, and Supplementary Figure S6 at http://www.BiochemJ.org/bj/441/bj4410591add.htm). In contrast, the C-terminal fragment of Cwc2 (residues 227–339) exhibited no detectable RNA-binding activity (Figure 6A, lanes 11–15). These findings demonstrate that Cwc2 binds to RNA via its structural core domain. Intriguingly, the RRM (122–227) alone failed to bind RNA (Figure 6B, lanes 1–5).

A representative RRM usually recognizes RNA using two conserved sequence elements, RNP1 in β3 and RNP2 in β1. Aromatic amino acids from these two elements stack with two nucleotides, whereas the first positively charged residue of RNP1 (lysine or arginine) donates a hydrogen bond to the phosphate group between the two nucleotides [16,17]. On the basis of sequence alignment (Figure 1A and Supplementary Figure S1), the corresponding aromatic residues were identified to be Tyr^{138} in RNP2 and Phe^{183} in RNP1 in Cwc2. In addition, a positively charged residue, Lys^{179}, is located in close proximity to RNP1. A triple mutation in the RRM of the core domain (Y138A, K179A and F183A) led to complete abolishment of RNA binding (Figure 6B, lanes 6–10), confirming an essential role for these residues in RNA binding. The loss of RNA binding was not due to structural disruption, because the triple mutant Cwc2 (Y138A/K179A/F183A) exhibited the same solution behaviour on gel filtration as the wild-type Cwc2 (see Supplementary Figure S7 at http://www.BiochemJ.org/bj/441/bj4410591add.htm). Together with the observation that RRM alone is unable to bind to RNA, these results suggest a requirement of additional structural elements in Cwc2 for RNA binding.

The RB loop plays an important role in RNA binding

The extended RB loop (residues 122–135) connects the ZnF domain and the RRM. The C-terminal portion of the RB loop contains a stretch of positively charged amino acids, RKKNK (residues 131–135) (Supplementary Figure S1). These residues appear to form a positively charged strip together with a number of other residues from several β-strands of the RRM (Figures 2B and 2C and Supplementary Figure S1), hinting at a potential RNA binding site. To examine this hypothesis, we individually mutated the basic amino acids (Arg^{131}–Lys^{135}) of the RB loop to alanine. These missense mutations failed to alter the RNA-binding activity of Cwc2 (results not shown). However, when all four positively charged amino acids were replaced by an alanine residue, the RNA-binding activity was nearly abolished for the quadruple mutant (R131A/K132A/K133A/K135A) (Figure 6C, lanes 1–5). Therefore, we named this connecting loop the RB loop.

To further examine the importance of the RB loop integrity in RNA binding, we co-expressed two discrete domains with the RB loop sequences linked to either ZnF or RRM (residues 1–135 and 136–227, or 1–121 and 122–227). These proteins were purified to homogeneity and tested for their RNA-binding activity *in vitro*.

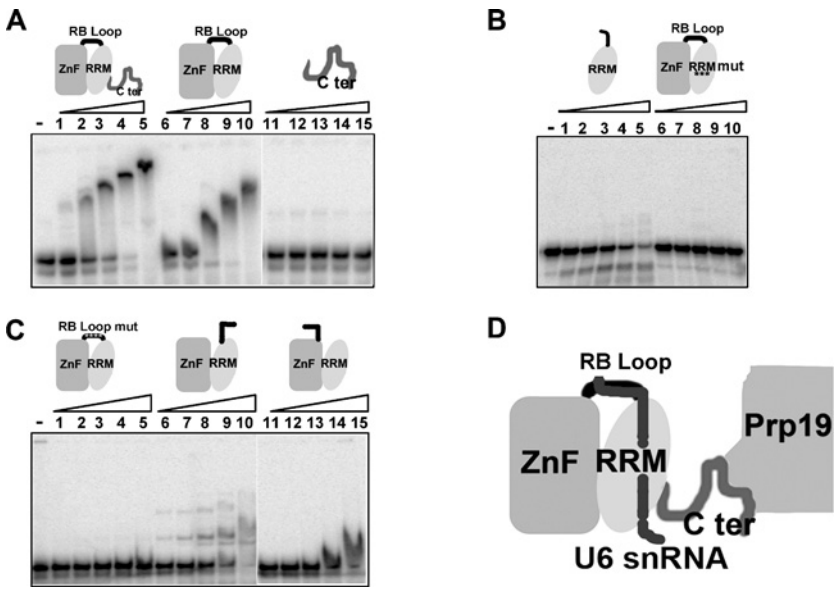

Figure 6 Characterization of RNA binding by Cwc2

(**A**) The structural core domain of Cwc2, but not the C-terminal flexible sequences, binds to U6 snRNA. A fixed amount of radiolabelled RNA probe was incubated with increasing concentrations of Cwc2; the mixture was subject to EMSA. The Cwc2 variants are represented as cartoons on top of the gel. The five Cwc2 concentrations used were 2, 3, 4.5, 6.7 and 10 μM. (**B**) The RRM (122–227) alone and the quadruple mutant of Cwc2 exhibited markedly decreased RNA-binding activity. Asterisks indicate positions of the mutations. Mut, mutant. (**C**) The integrity of the RB loop is essential for RNA binding. Protein concentrations are the same as in (**A**). (**D**) A proposed model of RNA–Cwc2 interactions. Our biochemical data suggest that RNA (shown as a line with chequerboard pattern) may bind to the extended basic strip on the surface of Cwc2. The C-terminal sequences of Cwc2 is thought to interact with Prp19.

Strikingly, the RNA-binding mode of ZnF and the RRM–RB loop complex (residues 1–121 and 122–227) was dramatically changed (Figure 6C, lanes 6–10). Additionally, there was little RNA-binding activity for ZnF–RB loop and RRM complex (residues 1–135 and 136–227) (Figure 6C, lanes 11–15). These results strongly suggest that the covalent linkage of the RB loop with the ZnF and RRM domains is important for U6 snRNA binding.

Previous structural studies on RRM revealed a critical role in RNA binding for both the β-sheet surface of the RRM and the loops connecting the β-sheets and α-helices [16,32–35]. In addition, the N-terminal extension of the 65 kDa C-terminal RRM facilitated RNA-binding activity indirectly via stabilization of the RRM structure [36]. In the present study, we provide the first piece of evidence that the linker sequences outside the RRM – those connecting the ZnF domain and the RRM – directly interact with RNA.

DISCUSSION

In the present study, we report the high-resolution crystal structure of the bulk of Cwc2, an essential protein for RNA splicing and the only known RNA-binding factor in the NTC. An unanticipated structural feature is the tight association between the ZnF domain and the RRM, and this feature appears to be essential for the RNA-binding activity of Cwc2. Guided by the structure, we performed RNA-binding studies and identified essential amino acids on the surface of the structure.

Surprisingly, these amino acids do not simply define a localized surface epitope. Rather, they collectively form an elongated strip that extends from the RB loop on one side of the globular Cwc2 protein to the β-strands of the RRM located on the opposite side. On the basis of these results and analyses, we propose a novel RNA-binding model for the Cwc2 protein (Figure 6D). Whereas the C-terminal portion of Cwc2 interacts with Prp19, the scaffold protein of the NTC, the RB loop and RRM bind to RNA through the positively charged strip (Figure 6D). Supporting this model, the critical aromatic residues in the RRM of Cwc2 are highly conserved from yeast to human [9] (see Supplementary Figure S1). Intriguingly, however, the positively charged RB loop residues required for RNA binding are only highly conserved in yeast but less so in higher organisms (Supplementary Figure S1). This could reflect evolutionary differences in the NTC. For example, the human Cwc2 orthologue RBM22 contains only one lysine residue in the C-terminal portion of the inter-domain connecting loop, suggesting a variation in RNA binding in humans.

One important unanswered question is whether Cwc2 binds to specific RNA sequences or whether it merely exhibits relatively non-specific RNA-binding activity to assist the RNA splicing function. The gradual shift of the U6 snRNA in the presence of increasing amounts of Cwc2 is consistent with relatively non-specific RNA-binding activity. But this could be due to the lack of a specific binding site in the U6 snRNA. Additionally, Cwc2 failed to bind to different RNA ligands comprising stem-loop, single-stranded or double-stranded RNA from U6 snRNA respectively (see Supplementary Figure S8 at http://www.BiochemJ.org/bj/441/bj4410591add.htm). We have begun to address this issue by attempting to identify potential RNA binding site(s) from a pool of random and degenerate RNA sequences. The preliminary results do not support stringent RNA sequence recognition by Cwc2.

Our biochemical assay indicated that the Cwc2 RRM alone is insufficient for binding to RNA. This might be due to its poorly conserved RNP1 sequence compared with the consensus RNP1 motif (Figure 1A). The first residue of RNP1 in Cwc2 is cysteine, instead of a positively charged amino acid in other RNP1 sequences, which directly binds to a phosphate group of RNA. In addition, a lysine residue (Lys^{185}) is located at the fifth position and its side chain contributes to the positively charged strip (Figures 2C and 5C). Importantly, the presence of Lys^{185} results in the register shift of one amino acid for Phe^{186}, which is now buried in the hydrophobic core of RRM and unable to interact

with the RNA base, as observed for the corresponding residue in the consensus RNP1 motif.

RNA splicing is a multi-step reaction requiring multiple protein–RNA complexes. In these steps, the NTC plays an obligate role in the regulation of spliceosome rearrangement and maintenance of splicing fidelity [1]. The present study represents an important step towards deciphering the mysteries of RNA binding by the NTC. The results are somewhat unexpected, which gives rise to an atypical model of RNA binding. Our proposed model of RNA binding by Cwc2, which is based on mutational analysis, remains to be experimentally verified through biochemical and structural investigations.

AUTHOR CONTRIBUTION

Peilong Lu, Guifeng Lu and Ping Yin performed all of the experiments and analysed the experimental data. Li Wang and Wenqi Li purified the recombinant proteins. Chuangye Yan determined the structure. Ping Yin supervised the project and prepared the manuscript.

ACKNOWLEDGEMENTS

We are grateful to members of Yigong Shi's laboratory for careful discussion. We thank the scientists J. He and S. Huang at the beamline BL17U of the Shanghai Synchrotron Radiation Facility.

FUNDING

This work was supported by the Special China Postdoctoral Science Foundation [grant number 201003126].

REFERENCES

1 Hogg, R., McGrail, J. C. and O'Keefe, R. T. (2010) The function of the NineTeen Complex (NTC) in regulating spliceosome conformations and fidelity during pre-mRNA splicing. Biochem. Soc. Trans. **38**, 1110–1115
2 Wahl, M. C., Will, C. L. and Luhrmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. Cell **136**, 701–718
3 Ohi, M. D., Vander Kooi, C. W., Rosenberg, J. A., Ren, L., Hirsch, J. P., Chazin, W. J., Walz, T. and Gould, K. L. (2005) Structural and functional analysis of essential pre-mRNA splicing factor Prp19p. Mol. Cell Biol. **25**, 451–460
4 Chan, S. P., Kao, D. I., Tsai, W. Y. and Cheng, S. C. (2003) The Prp19p-associated complex in spliceosome activation. Science **302**, 279–282
5 Ajuh, P., Kuster, B., Panov, K., Zomerdijk, J. C., Mann, M. and Lamond, A. I. (2000) Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry. EMBO J. **19**, 6569–6581
6 Villa, T. and Guthrie, C. (2005) The Isy1p component of the NineTeen complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. Genes Dev. **19**, 1894–1904
7 Ren, L., McLean, J. R., Hazbun, T. R., Fields, S., Vander Kooi, C., Ohi, M. D. and Gould, K. L. (2011) Systematic two-hybrid and comparative proteomic analyses reveal novel yeast pre-mRNA splicing factors connected to Prp19. PLoS One. **6**, e16719
8 Tarn, W. Y., Hsu, C. H., Huang, K. T., Chen, H. R., Kao, H. Y., Lee, K. R. and Cheng, S. C. (1994) Functional association of essential splicing factor(s) with PRP19 in a protein complex. EMBO J. **13**, 2421–2431
9 McGrail, J. C., Krause, A. and O'Keefe, R. T. (2009) The RNA binding protein Cwc2 interacts directly with the U6 snRNA to link the nineteen complex to the spliceosome during pre-mRNA splicing. Nucleic Acids Res. **37**, 4205–4217
10 Vander Kooi, C. W., Ren, L., Xu, P., Ohi, M. D., Gould, K. L. and Chazin, W. J. (2010) The Prp19 WD40 domain contains a conserved protein interaction region essential for its function. Structure **18**, 584–593
11 Hall, T. M. (2005) Multiple modes of RNA recognition by zinc finger proteins. Curr. Opin. Struct. Biol. **15**, 367–373
12 Brown, R. S. (2005) Zinc finger proteins: getting a grip on RNA. Curr. Opin. Struct. Biol. **15**, 94–98
13 Krishna, S. S., Majumdar, I. and Grishin, N. V. (2003) Structural classification of zinc fingers: survey and summary. Nucleic Acids Res. **31**, 532–550
14 Gamsjaeger, R., Liew, C. K., Loughlin, F. E., Crossley, M. and Mackay, J. P. (2007) Sticky fingers: zinc-fingers as protein-recognition motifs. Trends Biochem. Sci. **32**, 63–70
15 Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J. and Wright, P. E. (2004) Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. Nat. Struct. Mol. Biol. **11**, 257–264
16 Clery, A., Blatter, M. and Allain, F. H. (2008) RNA recognition motifs: boring? Not quite. Curr. Opin. Struct. Biol. **18**, 290–298
17 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
18 Teplova, M., Yuan, Y. R., Phan, A. T., Malinina, L., Ilin, S., Teplov, A. and Patel, D. J. (2006) Structural basis for recognition and sequestration of UUU(OH) 3′ temini of nascent RNA polymerase III transcripts by La, a rheumatic disease autoantigen. Mol. Cell. **21**, 75–85
19 Khoshnevis, S., Neumann, P. and Ficner, R. (2010) Crystal structure of the RNA recognition motif of yeast translation initiation factor eIF3b reveals differences to human eIF3b. PLoS ONE **5**, e12784
20 Rideau, A. P., Gooding, C., Simpson, P. J., Monie, T. P., Lorenz, M., Huttelmaier, S., Singer, R. H., Matthews, S., Curry, S. and Smith, C. W. (2006) A peptide motif in Raver1 mediates splicing repression by interaction with the PTB RRM2 domain. Nat. Struct. Mol. Biol. **13**, 839–848
21 Schellenberg, M. J., Edwards, R. A., Ritchie, D. B., Kent, O. A., Golas, M. M., Stark, H., Luhrmann, R., Glover, J. N. and MacMillan, A. M. (2006) Crystal structure of a core spliceosomal protein interface. Proc. Natl. Acad. Sci. U.S.A. **103**, 1266–1271
22 Bae, E., Reiter, N. J., Bingman, C. A., Kwan, S. S., Lee, D., Phillips, Jr, G.N., Butcher, S. E. and Brow, D. A. (2007) Structure and interactions of the first three RNA recognition motifs of splicing factor prp24. J. Mol. Biol. **367**, 1447–1458
23 Collaborative Computational Project, Number 4 (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. **50**, 760–763
24 Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with SHELXD. Acta Crystallogr. D Biol. Crystallogr. **58**, 1772–1779
25 McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. and Read, R. J. (2007) Phaser crystallographic software. J. Appl. Crystallogr. **40**, 658–674
26 Cowtan, K. (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. D Biol. Crystallogr. **62**, 1002–1011
27 Terwilliger, T. C. (2003) Automated main-chain model building by template matching and iterative fragment extension. Acta Crystallogr. D Biol. Crystallogr. **59**, 38–44
28 Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. **60**, 2126–2132
29 Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. and Terwilliger, T. C. (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr. D Biol. Crystallogr. **58**, 1948–1954
30 Licht, K., Medenbach, J., Luhrmann, R., Kambach, C. and Bindereif, A. (2008) 3′-cyclic phosphorylation of U6 snRNA leads to recruitment of recycling factor p110 through LSm proteins. RNA **14**, 1532–1538
31 Lai, W. S., Kennington, E. A. and Blackshear, P. J. (2002) Interactions of CCCH zinc finger proteins with mRNA: non-binding tristetraprolin mutants exert an inhibitory effect on degradation of AU-rich element-containing mRNAs. J. Biol. Chem. **277**, 9606–9613
32 Dominguez, C., Fisette, J. F., Chabot, B. and Allain, F. H. (2010) Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. Nat. Struct. Mol. Biol. **17**, 853–861
33 Skrisovska, L., Bourgeois, C. F., Stefl, R., Grellscheid, S. N., Kister, L., Wenter, P., Elliott, D. J., Stevenin, J. and Allain, F. H. (2007) The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. EMBO Rep. **8**, 372–379
34 Dominguez, C. and Allain, F. H. (2006) NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition. Nucleic Acids Res. **34**, 3634–3645
35 Auweter, S. D., Fasan, R., Reymond, L., Underwood, J. G., Black, D. L., Pitsch, S. and Allain, F. H. (2006) Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. EMBO J. **25**, 163–173
36 Netter, C., Weber, G., Benecke, H. and Wahl, M. C. (2009) Functional stabilization of an RNA recognition motif by a noncanonical N-terminal expansion. RNA **15**, 1305–1313

Received 29 July 2011/27 September 2011; accepted 30 September 2011
Published as BJ Immediate Publication 30 September 2011, doi:10.1042/BJ20111385

Biochem. J. (2012) **441**, 591–597 (Printed in Great Britain) doi:10.1042/BJ20111385

online data

SUPPLEMENTARY ONLINE DATA

Structure of the mRNA splicing complex component Cwc2: insights into RNA recognition

Peilong LU*†[1], Guifeng LU*‡[1], Chuangye YAN*†, Li WANG§, Wenqi LI*† and Ping YIN*†[2]

*Center of Structural Biology, Tsinghua University, Beijing 100084, China, †School of Life Sciences, Tsinghua University, Beijing 100084, China, ‡School of Medicine, Tsinghua University, Beijing 100084, China, and §School of Life Sciences, Peking University, Beijing 100084, China

EXPERIMENTAL

Pull-down assay

Cwc2 ZnF domain (1–121) and two constructs of RRM (122–227) (wild-type and triple mutant E193K/F194E/E197K) were cloned in pBB75 and pET15b vectors respectively. They were co-expressed in the *E. coli* strain BL21(DE3). As a control, RRM (with or without the triple mutation) alone was expressed in *E. coli* strain BL21(DE3). They were induced under the same condition and the soluble fraction of the *E. coli* lysate was purified over a Ni-NTA column (Qiagen). After affinity purification, all proteins were further purified by cation exchange chromatography (Source-15S, GE Healthcare). The same amount of the purified proteins was loaded on to SDS/PAGE (16 % gel) and run at 250V for 30 min.

[1] These authors contributed equally to this work.

[2] To whom correspondence should be addressed (email yinping@biomed.tsinghua.edu.cn).

The structure of the mRNA splicing complex component Cwc2 amino acids 1-121 + 133-227 and 1-121 alone have been deposited in the Protein Data Bank under accession numbers 3U1L and 3UIM respectively.

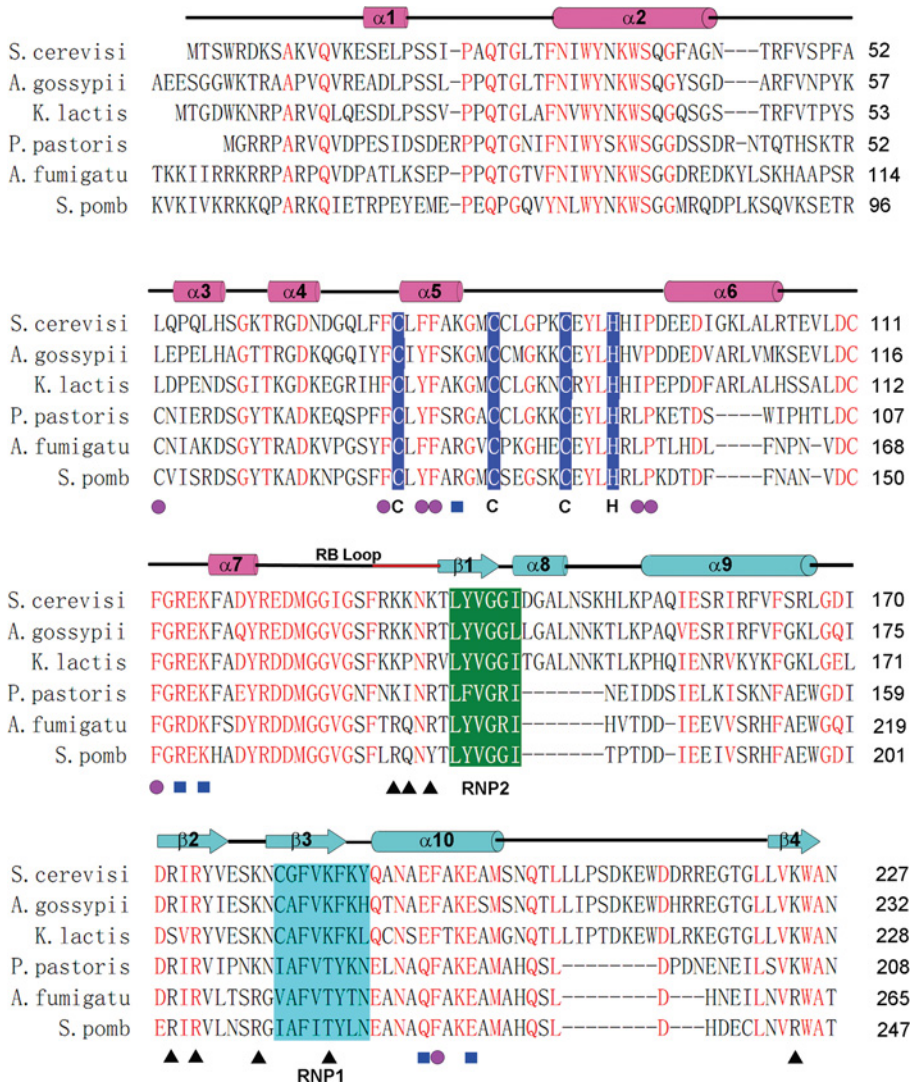

Figure S1 Sequence alignment of the Cwc2 core region

Sequence alignment of the Cwc2 core region with its orthologues from *Ashbya gossypii* (NCBI-GI: 302308872), *Kluyveromyces lactis* (NCBI-GI: 50305857), *Pichia pastoris* (NCBI-GI: 254569804), *Aspergillus fumigatus* (NCBI-GI: 70998470) and *Schizosaccharomyces pombe* (NCBI-GI: 19114249). RNP1 and RNP2 sequences are highlighted in cyan and green respectively. The four zinc-co-ordinating residues are highlighted in blue. Closed pink circles indicate the residues mediating inter-domain interactions via hydrophobic contacts. Closed blue squares indicate the residues mediating inter-domain interactions via hydrogen bonds. Closed black triangles indicate the residues constituting the positively charged strip.

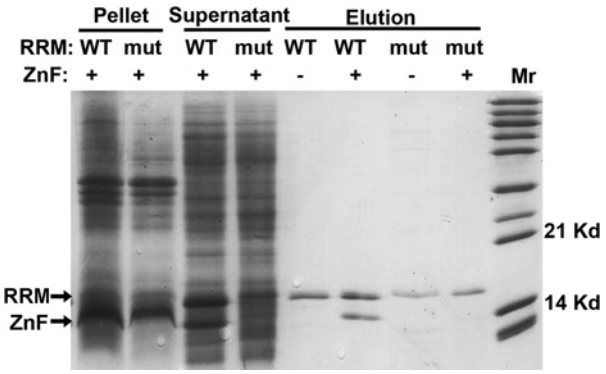

Figure S2 The ZnF domain interacts with the RRM motif

The interaction of the ZnF domain with the RRM motif was analysed by pull-down assay (see the Supplementary Experimental section). The SDS/PAGE gel shows that the His_6-tagged Cwc2 RRM motif (122–227) was able to interact with the untagged ZnF domain (1–121), but the RRM triple mutation was not. mut, triple mutation of RRM (E193K/F194E/E197K); WT, wild-type RRM motif.

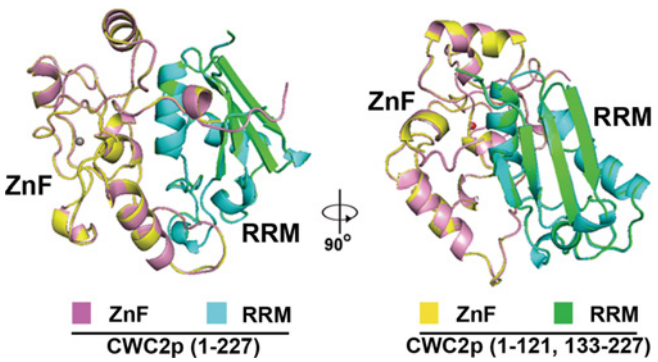

Figure S3 The binary complex of ZnF domain and RRM motif (1–121 + 133–227) represents a very similar structure to the Cwc structural core (1–227)

Structural comparison between Cwc2-(1–227) (PDB code 3U1M) and the binary complex of ZnF and RRM (1–121 + 133–227) (PDB code 3U1L) in two perpendicular views. The ZnF and RRM of Cwc2-(1-227) are colored pink and cyan respectively. Their counterparts in the binary complex are coloured yellow and green respectively.

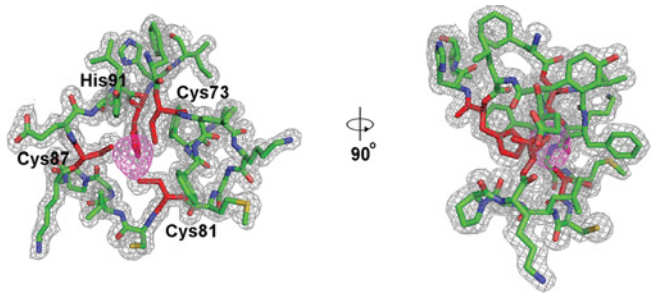

Figure S4 The electron density map of ZnF (72–93)

The electron density map shows part of an experimental map calculated using the final refined structure. The electron density map (grey wire-frame contoured at 1σ) closely matches the atomic model. The anomalous difference map was calculated using Zn-SAD data. The anomalous peaks are contoured at 4σ (magenta mesh). The side chains of the four zinc-co-ordinating residues are coloured red.

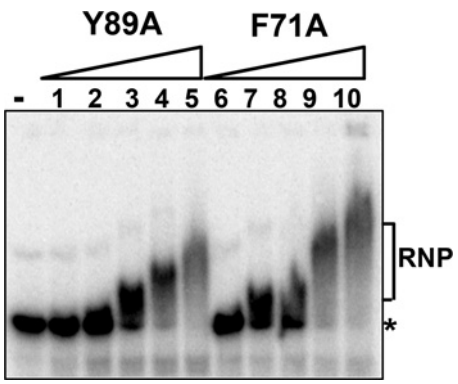

Figure S5 The missense mutations Y89A or F71A in the Cwc2 ZnF domain exhibited moderate effect on RNA binding by EMSA analysis

Lanes 1–5 show the Y89A mutant and lanes 6–10 indicate the F71A mutant (final concentrations of protein: 2, 3, 4.5, 6.7 and 10 μM). The asterisk indicates unbound RNA probe. RNP indicates the RNA–protein complex.

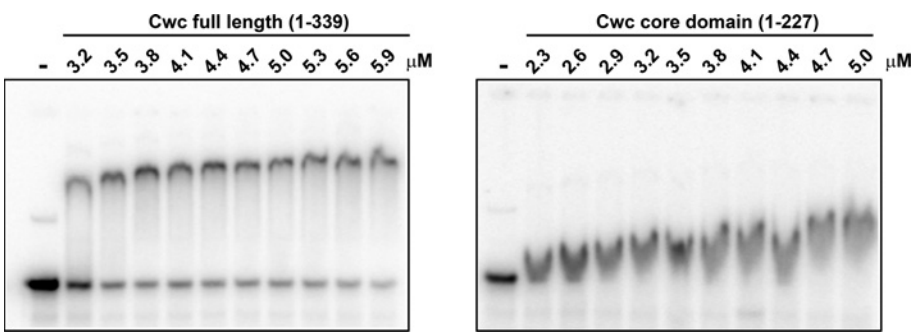

Figure S6 Binding of full-length Cwc2 protein or the structural core domain (1–227) to U6 snRNA

The indicated concentrations of full-length Cwc2 protein or structural core domain (1–227) proteins were incubated with radiolabelled U6 snRNA probe.

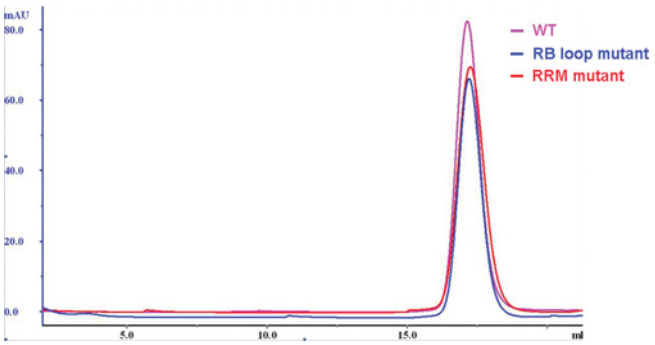

Figure S7 The similar behaviour of the Cwc structural core (1–227) and the mutant proteins used in EMSA assay

Size-exclusion chromatography analysis of proteins was performed with Superdex 200hr10/30 (GE healthcare) in lysis buffer (25 mM Tris/HCl, pH 8.0, and 150 mM NaCl). RB loop mutant, quadruple mutant in RB loop; RRM mutant, triple mutant in RRM motif (for more details, see the Results section of the main text); WT, wild-type Cwc-(1–227).

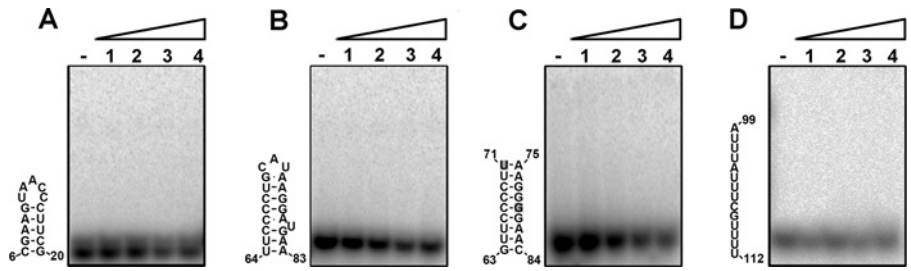

Figure S8 Cwc2 is unable to bind to typical secondary structural RNA elements of U6 snRNA

The full-length Cwc2 was incubated with different 5′-^{32}P-labelled RNA ligands. All RNA ligands was designed based on U6 snRNA secondary structure [1]. (**A**) U6 snRNA 5′ stem-loop (6–20); (**B**) U6 snRNA 3′ stem-loop containing U4/U6 stemII (64–83); (**C**) U6 snRNA 3′ stem as double-stranded RNA. The moderately changed ribonucleotides are shaded in grey. (**D**) U6 snRNA 3′ Lsm binding site as single-stranded RNA (99–112). The four Cwc2 concentrations used in each panel were 3, 4.5, 6.7 and 10 μM (lanes 1–4 respectively).

REFERENCE

1 Karaduman, R., Dube, P., Stark, H., Fabrizio, P., Kastner, B. and Lührmann, R. (2008) Structure of yeast U6 snRNPs: arrangement of Prp24p and the LSm complex as revealed by electron microscopy. RNA **14**, 2528–2537

Received 29 July 2011/27 September 2011; accepted 30 September 2011
Published as BJ Immediate Publication 30 September 2011, doi:10.1042/BJ20111385