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*†‡

*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 CX₄CX₃CX₂H (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 RRM [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 motif. 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, Ni²⁺-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).
3 The structure of the mRNP binding 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 3U1M 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 \( \beta \)-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 Zn\(_{1}\) 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\(_2\)+-nitrotetracyclate) 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 \( \mu \)l 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 \( \mu \)l 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% native 5% acrylamide gels (37.5:1 acrylamide/bisacrylamide) in 0.5× Tris-borate buffer containing 6% glycerol at 20 °C/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.html). 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\(_2\)\(_1\)2\(_2\) 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 (RNA-binding loop) 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.

We crystallized the structural core domain (residues 1–227) in the P2\(_2\)\(_1\)2\(_2\) 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 (RNA-binding loop) 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.
Table 1  Data collection and refinement statistics

One crystal was used for each structure. Values in parentheses are for the highest resolution shell. 

\[ R_{merge} = \frac{\sum h i |I_h,i| - \langle I_h,i \rangle}{\sum h i \langle I_h,i \rangle} \]

where \( I_h,i \) is the mean intensity of the \( i \) observations of symmetry-related reflections of \( h \). 

\[ R = \frac{\sum |F_{obs} - F_{calc}|}{\sum F_{obs}} \]

\( R_{free} \) was calculated with 5% of the reflections selected randomly.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Zinc-SAD (1–121 + 133–227)</th>
<th>Cwc2-native (1–121 + 133–227) (PDB code 3U1L)</th>
<th>Cwc2-native (1–227) (PDB code 3U1M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P2₁₂₁₂₁</td>
<td>P2₁₂₁₂₁</td>
<td>P2₁₂₁₂₁</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>44.91, 54.49, 112.01</td>
<td>44.85, 54.37, 111.77</td>
<td>44.72, 54.34, 111.38</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.00, 90.00, 90.00</td>
<td>90.00, 90.00, 90.00</td>
<td>90.00, 90.00, 90.00</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.26770</td>
<td>0.96393</td>
<td>1.54178</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>40<del>2.65 (2.74</del>2.65)</td>
<td>40<del>1.64 (1.70</del>1.64)</td>
<td>40<del>1.90 (1.97</del>1.90)</td>
</tr>
<tr>
<td>Rmerge (%)</td>
<td>6.8 (14.9)</td>
<td>6.0 (33.9)</td>
<td>4.1 (11.7)</td>
</tr>
<tr>
<td>I/σ</td>
<td>25.2 (15.4)</td>
<td>21.4 (3.4)</td>
<td>35.1 (10.2)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.1 (99.9)</td>
<td>99.1 (95.5)</td>
<td>99.5 (95.6)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>5.3</td>
<td>3.9</td>
<td>6.1</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>40~2.65</td>
<td>40~1.64</td>
<td>40~1.90</td>
</tr>
<tr>
<td>Number of reflections</td>
<td>8391</td>
<td>34029</td>
<td>22066</td>
</tr>
<tr>
<td>Rwork/Rfree (%)</td>
<td>17.12/18.93</td>
<td>15.89/19.37</td>
<td></td>
</tr>
<tr>
<td>Number of atoms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>1787</td>
<td>1784</td>
<td></td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>303</td>
<td>326</td>
<td></td>
</tr>
<tr>
<td>B-factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>23.22</td>
<td>19.19</td>
<td></td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>11.01</td>
<td>11.84</td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>35.47</td>
<td>29.63</td>
<td></td>
</tr>
<tr>
<td>rmsd</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.006</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.977</td>
<td>1.177</td>
<td></td>
</tr>
<tr>
<td>Ramachandran plot statistics (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most favoured</td>
<td>93.6</td>
<td>93.2</td>
<td></td>
</tr>
<tr>
<td>Additional allowed</td>
<td>6.4</td>
<td>6.8</td>
<td></td>
</tr>
<tr>
<td>Generously allowed</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Disallowed</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

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, Lys²²⁴ from β4 and Lys¹⁷⁹ 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 His₆-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.
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

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

Copyright © 2012 Biochemical Society.
addition, a positively charged residue, Lys 179, is located in a four-stranded antiparallel β sheet (residues 187–220) (light blue) (PDB code 1RGO [15]). The zinc atom is shown as a magenta sphere. The RRM of Cwc2 adopts a canonical βαββαββ 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, RNPI in β3 and RNPII in β1. Aromatic amino acids from these two elements stack with two nucleotides, whereas the first positively charged residue of RNPI (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 Tyr138 in RNPII and Phe160 in RNPI in Cwc2. In addition, a positively charged residue, Lys179, is located in close proximity to RNPI. 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 (Arg131–Lys135) 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.
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 – directly interacting with RNA. Strikingly, 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).

**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). 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 (Lys185) is located at the fifth position and its side chain contributes to the positively charged strip (Figures 2C and 5C). Importantly, the presence of Lys185 results in the register shift of one amino acid for Phe186, which is now buried in the hydrophobic core of RRM and unable to interact with RNA.

This work was supported by the National Natural Science Foundation, China (30825021, 30130010), the Key Project of Shanghai, China (01dz22607), and the Chinese Academy of Sciences (KGCX2-YW-257). We thank Drs. W. Zhang and Y. Niu for their assistance with electron microscopy and Dr. H. Liu for discussions about the C-terminal sequences of Cwc2.
with the RNA base, as observed for the corresponding residue in the consensus RNPI 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
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: 302306872), Kluyveromyces lactis (NCBI-GI: 50935857), 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.
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 His6-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.

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.

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.

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.

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′-32P-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


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