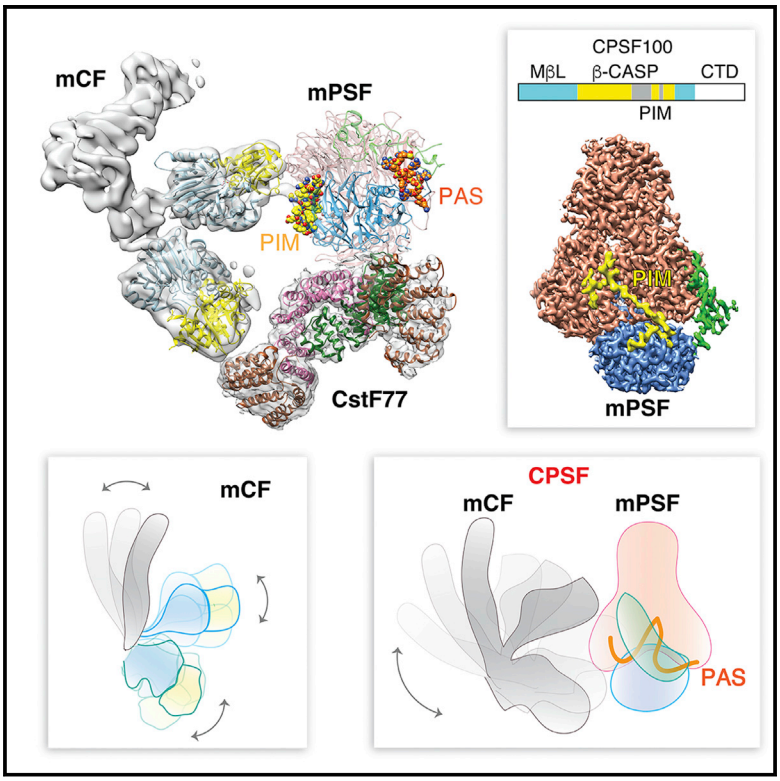
Article

Molecular Cell

Structural Insights into the Human Pre-mRNA 3′-End Processing Machinery

Graphical Abstract



Highlights

- An mPSF interaction motif (PIM) in CPSF100 tethers mCF to mPSF in CPSF
- The overall conformation of mCF and its location relative to mPSF are highly dynamic
- CstF is bound to mPSF through its CstF77 subunit
- CPSF160 and WDR33 of mPSF form the core of the pre-mRNA 3′-end-processing machinery

Authors

Yixiao Zhang, Yadong Sun, Yongsheng Shi, Thomas Walz, Liang Tong

Correspondence

twalz@rockefeller.edu (T.W.), ltong@columbia.edu (L.T.)

In Brief

Cryo-EM and biochemical studies on CPSF and CstF in the human pre-mRNA 3′-end processing machinery have offered the first insights into the three-dimensional organization of this machinery.

Zhang et al., 2020, Molecular Cell 77, 1–10
February 20, 2020
https://doi.org/10.1016/j.molcel.2019.11.005

CellPress

Please cite this article in press as: Zhang et al., Structural Insights into the Human Pre-mRNA 3′-End Processing Machinery, Molecular Cell (2019), https://doi.org/10.1016/j.molcel.2019.11.005

Molecular Cell
Article

Structural Insights into the Human Pre-mRNA 3′-End Processing Machinery

Yixiao Zhang,[1,4] Yadong Sun,[2,4] Yongsheng Shi,[3] Thomas Walz,[1,*] and Liang Tong[2,5,*]
[1]Laboratory of Molecular Electron Microscopy, Rockefeller University, New York, NY 10065, USA
[2]Department of Biological Sciences, Columbia University, New York, NY 10027, USA
[3]Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: twalz@rockefeller.edu (T.W.), ltong@columbia.edu (L.T.)
https://doi.org/10.1016/j.molcel.2019.11.005

SUMMARY

The mammalian pre-mRNA 3′-end-processing machinery consists of cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), and other proteins, but the overall architecture of this machinery remains unclear. CPSF contains two functionally distinct modules: a cleavage factor (mCF) and a polyadenylation specificity factor (mPSF). Here, we have produced recombinant human CPSF and CstF and examined these factors by electron microscopy (EM). We find that mPSF is the organizational core of the machinery, while the conformations of mCF and CstF and the position of mCF relative to mPSF are highly variable. We have identified by cryo-EM a segment in CPSF100 that tethers mCF to mPSF, and we have named it the PSF interaction motif (PIM). Mutations in the PIM can abolish CPSF formation, indicating that it is a crucial contact in CPSF. We have also obtained reconstructions of mCF and CstF77 by cryo-EM, assembled around the mPSF core.

INTRODUCTION

In eukaryotes, most messenger RNA precursors (pre-mRNAs) must undergo extensive processing before they can be exported to the cytoplasm and translated into proteins (Proudfoot, 2011; Xiang et al., 2014; Yang and Doublié, 2011). At the 3′ end, these pre-mRNAs are cleaved at a specific location followed by the addition of a poly(A) tail. Studies over the years have identified a large machinery of many protein factors that is required for this 3′ end processing (Mandel et al., 2008; Shi et al., 2009; Shi and Manley, 2015; Zhao et al., 1999), including the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), and poly(A) polymerase (PAP) in mammals.

Recent studies showed that CPSF consists of two sub-complexes, mPSF (mammalian polyadenylation specificity factor) and mCF (mammalian cleavage factor) (Chan et al., 2014; Schönemann et al., 2014). mPSF contains CPSF160, WDR33, CPSF30, and Fip1. WDR33 and CPSF30 recognize the AAUAAA polyadenylation signal (PAS) to define the cleavage site (Chan et al., 2014; Schönemann et al., 2014), while Fip1 recruits PAP to catalyze the polyadenylation (Helmling et al., 2001; Kaufmann et al., 2004; Meinke et al., 2008). mCF contains CPSF73, CPSF100, and symplekin, with CPSF73 being the endonuclease that catalyzes the cleavage of the pre-mRNA (Mandel et al., 2006b). mCF is also required for replication-dependent histone pre-mRNA 3′ end processing (Sullivan et al., 2009).

CPSF160 contains three β-propellers (BPA, BPB, and BPC) and a C-terminal domain (CTD, residues 1351–1443) (Clerici et al., 2017, 2018; Sun et al., 2018) (Figure 1A). WDR33 contains a WD40 domain near the N terminus and an extended C-terminal segment with unknown function. CPSF30 contains five zinc fingers (ZF1–ZF5) and a zinc knuckle. CPSF73 contains a metallo-β-lactamase domain, a β-CASP domain and a CTD. CPSF100 is a weak sequence homolog of CPSF73, and the β-CASP domain of CPSF100 contains a highly hydrophilic and generally poorly conserved segment (Figure 1A) (Kolev et al., 2008; Mandel et al., 2006a, 2006b). Symplekin contains an N-terminal domain (NTD) that interacts with the protein phosphatase Ssu72 (Xiang et al., 2010, 2012), a middle region that interacts with CstF64 (Ruepp et al., 2011), and a CTD, the first part of which interacts with CPSF73 (Ghazy et al., 2009).

CstF contains three subunits, CstF50, CstF64, and CstF77. CstF50 has a WD40 domain (Yang et al., 2018) (Figure 1A). CstF64 contains an RNA recognition module (RRM) at the N terminus that binds the G/U-rich downstream element (DSE) of the pre-mRNA (Pérez Cañadillas and Varani, 2003). The HAT (half a tetratricopeptide repeat) domain of CstF77 consists of two subdomains, HAT-N and HAT-C, and forms a bow-shaped dimer (Bai et al., 2007; Legrand et al., 2007).

We and others reported recently the structure of human mPSF in complex with the AAUAAA polyadenylation signal, revealing the molecular basis for this crucial event in 3′ end processing (Clerici et al., 2018; Sun et al., 2018). WDR33 and especially ZF2 and ZF3 of CPSF30 directly contact the RNA, which also contains a Hoogsteen base pair between the U at the third position and the A at the sixth position. CPSF160 is a scaffold that pre-organizes WDR33 and CPSF30 for this recognition. The yeast homologs of the three mammalian proteins have a similar organization (Casañal et al., 2017), although how they recognize RNA is currently not known. Further studies with the yeast

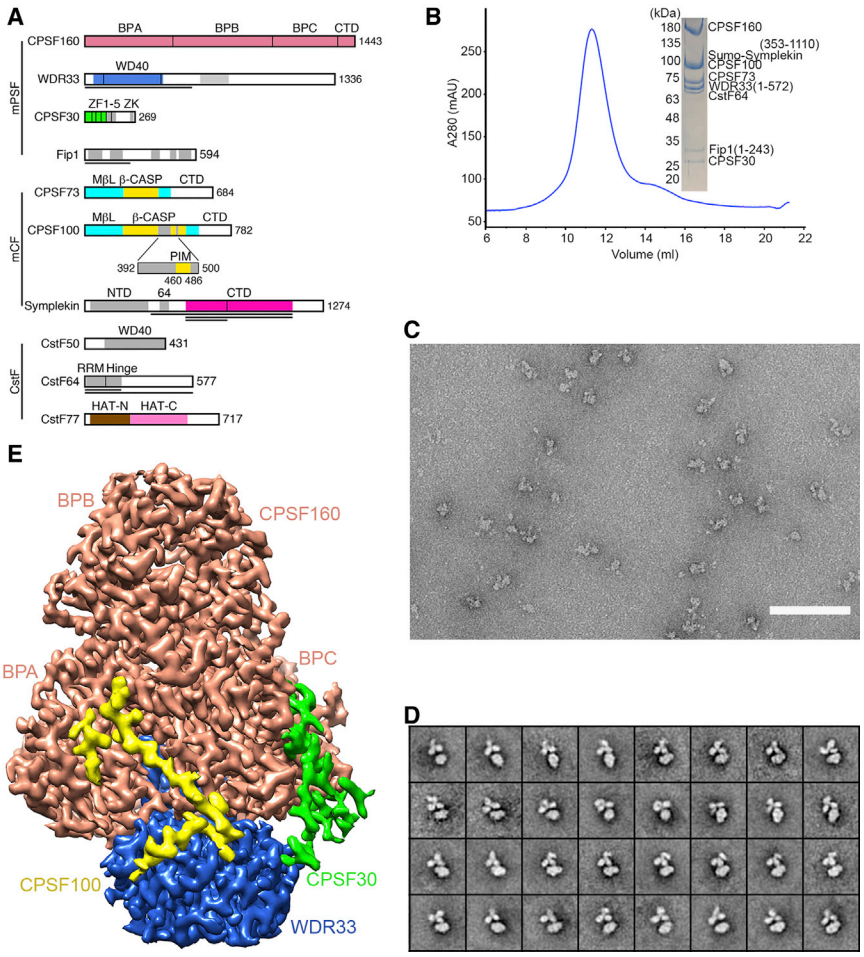


Figure 1. Structural Studies of Human CPSF

(A) Domain organizations of selected subunits of the 3′-end-processing machinery studied here. The collagen-like segment in WDR33 is in gray. The highly hydrophilic segment in the β-CASP domain of CPSF100 is in gray, and the PSF interaction motif (PIM) in this segment is in yellow. An expanded view of this hydrophilic segment is also shown. The vertical bar in the symplekin CTD marks the end of its N-terminal segment that interacts with CPSF73. The truncated constructs of the proteins are indicated with the dark lines, while full-length proteins are used for the others. Domains in color represent those with structural information from this study. ZK, zinc knuckle; MβL, metallo-β-lactamase.

(B) Gel-filtration profile of the human CPSF sample, containing full-length CPSF160, CPSF73, CPSF100, CPSF30, residues 1–572 of WDR33, residues 1–243 of Fip1, and residues 353–1110 of symplekin as a SUMO fusion protein. Full-length human CstF64 is also included, to interact with symplekin. The PAS RNA was not included in this sample. Inset: SDS-PAGE gel of the complex. The bands for SUMO-symplekin (96 kDa) and CPSF100 (88 kDa) are overlapped.

(C) Area of a negative-stain EM image of the CPSF sample. Scale bar: 100 nm.

(D) Selected 2D class averages of negatively stained CPSF, showing that the particles consist of a well-ordered core and a flexible trilobal structure. Residues 538–762 of symplekin are present in this sample. Side length of individual averages: 46 nm. See Figure S2 for all the 2D class averages.

(E) Cryo-EM 3D reconstruction of the well-ordered core of the CPSF sample. The densities for CPSF160, WDR33, and CPSF30 are shown in salmon, light blue, and green, respectively. The additional density that does not belong to mPSF is shown in yellow, and it belongs to CPSF100. Produced with Chimera (Pettersen et al., 2004).

See also Figures S1–S4; Video S1.

machinery successfully reconstituted the cleavage and polyadenylation activity *in vitro* (Hill et al., 2019), which also showed that the structure of the machinery is highly dynamic.

Here, we have produced recombinant human CPSF, alone and in complex with CstF64, as well as mPSF in complex with CstF, and examined these complexes by electron microscopy (EM). We find that mPSF forms a well-ordered core of CPSF and the processing machinery, while the conformations of mCF and CstF and the position of mCF relative to mPSF are highly flexible. Nonetheless, we have identified a segment in CPSF100 that tethers mCF to mPSF, and we have named this segment the PSF interaction motif (PIM). Mutations in the PIM can abolish CPSF formation, indicating that it is a crucial contact for CPSF. We have also obtained a reconstruction of mCF at 7.4 Å resolution by cryoelectron microscopy (cryo-EM), in the context of its complex with mPSF, as well as a structure of CstF bound to mPSF at 3.6 Å resolution. These studies have provided molecular insights into the organization of CPSF and CstF in the human pre-mRNA 3′-end-processing machinery.

RESULTS

Structure Determination

To produce samples of human CPSF for structural studies, we first expressed its two sub-complexes, mPSF and mCF, separately in baculovirus-infected insect cells. The expression and purification of mPSF followed the same protocol as we described earlier, which produced a well-behaved sample for EM studies (Sun et al., 2018). The mCF sample contained full-length CPSF73, full-length CPSF100, and a segment of symplekin (Figure 1A). Various segments of symplekin were examined, including one containing residues 353–1110 (missing the NTD that interacts with Ssu72 (Xiang et al., 2010) and the poorly conserved C-terminal region) and one containing residues 538–1110 (Figure S1). Because the 353–1110 segment also contains the region (residues 391–465) that interacts with CstF64 (Ruepp et al., 2011) (Figure 1A), we included full-length human CstF64 in this sample as well. We examined the purified mCF samples by negative-stain EM and found them to be highly flexible conformationally (Figure S1). mCF on its own has a trilobal

Table 1. Cryo-EM Data Collection, Refinement, and Validation Statistics

	mPSF-CPSF100 PIM (CPSF160-WDR33-CPSF30-CPSF100 PIM) (EMDB-20860) (PDB 6URG)	mCF[a] (CPSF73-CPSF100-symplekin) (EMDB-20859)	mPSF-CstF (CPSF160-WDR33-CPSF30-PAS RNA-CstF77) (EMDB-20861) (PDB 6URO)
Data Collection and Processing			
Magnification	22,500	22,500	22,500
Voltage (kV)	300	300	300
Electron exposure (e^{-}/Å^2)	70	71	70
Defocus range (μm)	1.2∼2.5	1.2∼2.5	1.2∼2.5
Pixel size (Å)	1.06	1.07	1.07
Symmetry imposed	C1	C1	C1
Image stacks (no.)	4,819	7,608	4,095
Initial particle images (no.)	1,539,569	6,870,618	1,723,794
Final particle images (no.)	859,796	35,040	50,092
Map resolution (Å)	3.0	7.4	3.6
Fourier shell correlation (FSC) threshold	0.143	0.143	0.143
Map sharpening B factor (Å^2)	–110	–323	–81
Refinement			
Number of protein residues	1,661		2,750
Number of RNA nucleotides	0		8
Number of metal ions	1		3
Rmsds			
Bond lengths (Å)	0.01		0.01
Bond angles (°)	0.88		0.99
PDB validation			
Clash score	8		9
Poor rotamers (%)	1		1
Ramachandran plot			
Favored (%)	91.86		92.30
Allowed (%)	8.08		7.58
Disallowed (%)	0.06		0.11

[a]Images for the mPSF-CPSF100 PIM reconstruction are also used for the mCF reconstruction.

structure overall, but the relative positions of the three lobes show substantial variations among the particles (Figure S1). The inclusion of CstF64 with the longer segment of symplekin did not improve the sample (data not shown). This conformational flexibility precluded us from obtaining a high-resolution cryo-EM reconstruction of mCF alone.

We mixed purified mPSF and mCF and successfully purified recombinant human CPSF (Figure 1B). Negative-stain EM studies showed that the sample still had substantial structural variability (Figure 1C), irrespective of whether the PAS RNA was present or not, although some of the particles displayed a trilobal structure attached to a shape consistent with mPSF (Figure 1D; Figure S2). To obtain a cryo-EM structure, we collected a large number of images, on samples with and without PAS RNA, and selected nearly 1.2 million particles for analysis. We were able to obtain a cryo-EM reconstruction for the well-ordered core of CPSF (which is primarily mPSF) at 3.0 Å resolution (Figures S3 and S4; Table 1). One of the 3D classes showed additional density representing a part of mCF connected to the core (Figure S3). After careful analyses, we were able to obtain a reconstruction of mCF in this complex at 7.4 Å resolution (Figures S4 and S5).

A Segment of CPSF100 Tethers mCF to mPSF

The overall shape of the cryo-EM reconstruction for the well-ordered core of CPSF is similar to that of mPSF that was reported earlier (Clerici et al., 2018; Sun et al., 2018), with the exception that ZF2 and ZF3 of CPSF30 had essentially no density as the PAS RNA was not included in this sample (Figure 1E). An atomic model for CPSF160, WDR33, and CPSF30 (N-terminal segment plus ZF1) could be built based on the density (Figure 2A), which had an overall root-mean-square (rms) distance between equivalent Cα atoms of 0.4 Å to the structure reported earlier, indicating that the two structures are essentially identical overall.

The detailed analysis also reveals that there is extra density in this reconstruction that does not belong to mPSF (Figure 1E) and is not present in the earlier reconstruction for mPSF alone. The density contacts both CPSF160 and WDR33, with an extended segment in the middle. After examining many possibilities, the

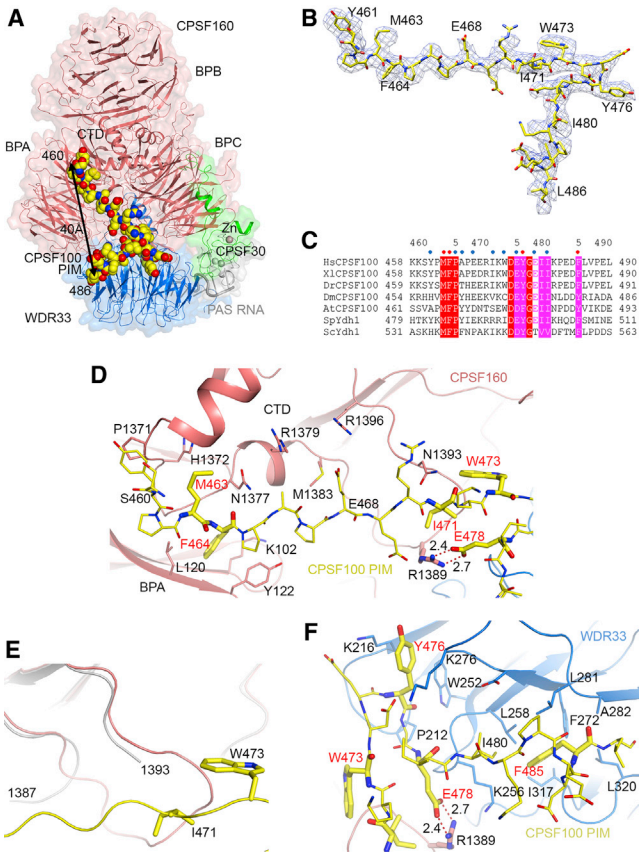


Figure 2. The PSF Interaction Motif (PIM) of CPSF100 Tethers mCF to mPSF in CPSF

(A) Overall structure of the well-ordered core of CPSF, colored as in Figure 1A. The molecular surface of mPSF is shown as a transparent surface. The PIM of CPSF100 is shown as ball-and-stick models (yellow). The ZF2 and ZF3 and the PAS RNA are shown for reference (gray), but no density was observed for them in this reconstruction as the PAS RNA was not included in this sample.

(B) Cryo-EM density for the CPSF100 PIM (residues 460–486).

(C) Sequence conservation of the CPSF100 PIM. An alignment of PIM residues in human (Hs), *X. laevis* (Xl), *D. rerio* (Dr), *D. melanogaster* (Dm), *A. thaliana* (At), *S. pombe* (Sp), and *S. cerevisiae* (Sc) CPSF100 homologs are shown. Strictly conserved residues are highlighted in red, and well-conserved residues are in magenta. Red and blue dots above the alignment indicate residues with >100 and 50–100 Å^2 buried surface areas in the complex, respectively. The residue numbers above the sequence are for human CPSF100.

(D) Detailed interactions between the N-terminal and middle regions of the CPSF100 PIM (yellow) with the BPA and CTD of CPSF160 (salmon). Residues in the CPSF100 PIM having extensive interactions with mPSF are shown as sticks and labeled in red.

(E) Conformational change for a loop in the CTD of CPSF160 (salmon) in the complex with CPSF100 PIM (yellow) compared to the structure of mPSF alone (gray).

(F) Detailed interactions between the C-terminal region of the CPSF100 PIM (yellow) with WDR33 (light blue). Produced with PyMOL (https://pymol.org/2/). See also Figures S3 and S4.

majority of this density was interpreted as belonging to residues 460–486 of CPSF100 (Figures 2A and 2B). These residues are located in the highly hydrophilic segment of CPSF100 (residues 392–500, Figure 1A), and this segment is removed by proteolysis during the crystallization of its yeast homolog Ydh1 (Mandel et al., 2006a, 2006b). Residues in this segment are generally poorly conserved among CPSF100 homologs. However, residues 462–485 are highly conserved, even in the plant and yeast homologs (Figure 2C); in fact, it is the only conserved region in this segment of CPSF100, supporting the important role of these residues in tethering mCF to mPSF. We have named this region of CPSF100 the PSF interaction motif (PIM). This hydrophilic segment does not exist in CPSF73 homologs (Figure 1A), and they cannot interact with mPSF in a similar fashion.

The reconstruction contains another, smaller piece of density, located next to residue 462 of CPSF100 (Figure 1E). The residues in this density could also be in the hydrophilic segment. However, the exact assignment of this density is not certain, due mostly to its short length (about 5 residues), and it will not be described further here.

Interactions between CPSF100 and mPSF

The PIM of CPSF100 makes extensive contacts with CPSF160 and WDR33, burying ~1,550 Å^2 of its surface area in the interface. The buried surface area in the interface with CPSF160 is ~850 Å^2, involving the N-terminal region of the PIM (residues 461–466) and the extended middle region (residues 467–473). The buried surface area in the interface with WDR33 is ~700 Å^2, involving the C-terminal region of the PIM (residues 474–486), which includes a β reverse turn followed by a segment that is oriented at nearly a 90° angle relative to the middle region of the PIM (Figure 2A). The distance between the N and C termini of the PIM is ~40 Å, which allows it to contact both CPSF160 and WDR33.

Residues in the CPSF100 PIM making large contributions to the interface with mPSF are highly conserved among CPSF100 homologs (Figure 2C). The interactions between the PIM and CPSF160 and WDR33 are primarily hydrophobic and van der Waals in nature. In the N-terminal region of the PIM, Met463 and Phe464 make the largest contributions, with the side chain of Phe464 almost entirely buried (Figures 2C and 2D). This region of the PIM contacts the BPA and CTD of CPSF160. In the middle region, Ile471 and Trp473 of the PIM have interactions with a loop in the CTD of CPSF160. This loop is mostly disordered and assumes a different conformation in the structure of mPSF alone (Sun et al., 2018) and becomes ordered in the complex with mCF, likely stabilized by its contacts with the PIM (Figure 2E).

In the interface with WDR33, residues Tyr476 (in the β reverse turn) and Phe485 in the C-terminal region of the PIM make the largest contributions (Figure 2F). This binding site is formed by the bottom face of three consecutive blades of the WD40 propeller of WDR33. Tyr476 is located between the first two blades, while Phe485 is in a hydrophobic pocket between the second two blades. There is also an ion pair between Glu478 in this C-terminal region of the PIM and Arg1389 of CPSF160 (Figures 2D and 2F), which may allow the main chain of Glu478 to make an ~90° turn so that the C-terminal region can maintain contacts with WDR33.

Mutations in CPSF100 PIM Block CPSF Formation

To obtain biochemical evidence for the interactions between CPSF100 PIM and mPSF, we introduced mutations in the PIM

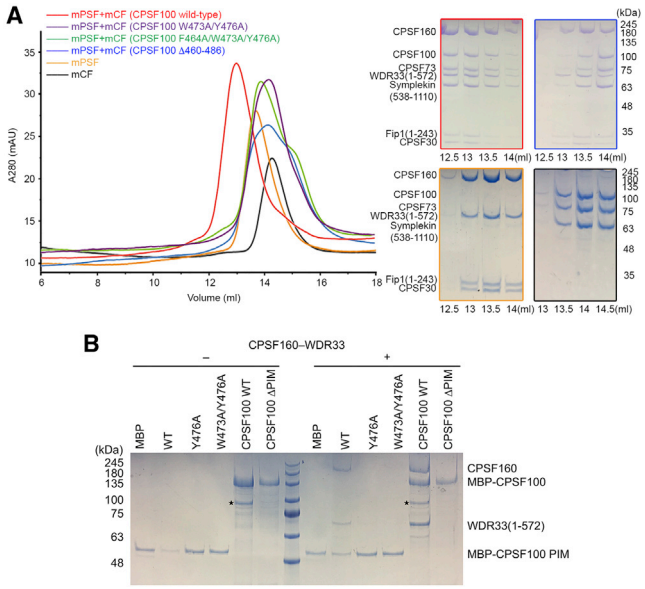


Figure 3. Mutations in the CPSF100 PIM Disrupt CPSF Formation

(A) Gel-filtration profiles for mixtures of mPSF with wild-type and mutant mCF, as well as those of wild-type mPSF and mCF alone. The SDS gels in the red, blue, gold, and black boxes correspond to the red, blue, gold, and black UV traces, respectively. Site-specific and deletion mutations of the CPSF100 PIM abolish the formation of CPSF. For example, the mPSF and mCF complexes no longer co-migrate when the PIM is deleted (blue box). The mCF mutants alone migrate at the same position as that of the wild-type mCF.

(B) Mutations in the CPSF100 PIM block pull-down of CPSF160-WDR33. Full-length CPSF100, CPSF100 ΔPIM, and PIM alone were expressed as fusion proteins with MBP, which were used to pull down the CPSF160-WDR33 complex. Wild-type (WT) PIM and CPSF100 can pull down CPSF160-WDR33, while mutations in the PIM and its deletion abolished the pull-down. MBP alone was included as a control and could not pull down CPSF160-WDR33. An impurity band in the MBP-CPSF100 fusion protein is indicated with the asterisk.

and assessed their effects on the formation of CPSF. We purified wild-type (WT) and mutant mCF carrying site-specific mutations W473A/Y476A and F464A/W473A/Y476A or a deletion of the entire PIM (residues 460–486) in CPSF100, and examined their mixtures with mPSF by gel filtration. The mixture of wild-type mPSF and mCF produced a clear shift in the migration compared to that of the two complexes alone, indicating the formation of CPSF (Figure 3A). In comparison, all three mutations essentially blocked the formation of CPSF, confirming the importance of the CPSF100 PIM.

We also expressed the CPSF100 PIM (residues 460–486) as a fusion protein with maltose binding protein (MBP) in bacteria, and examined the ability of purified WT and mutant PIMs to pull down purified CPSF160-WDR33 complex. The Y476A single-site and the W473A/Y476A double mutations blocked the interaction (Figure 3B). We also produced full-length CPSF100 as an MBP fusion protein. While wild-type CPSF100 can pull down CPSF160-WDR33, the mutant lacking the PIM cannot. These data demonstrate that the CPSF100 PIM alone is sufficient for interaction with CPSF160-WDR33.

Overall Structure of mCF

While a good quality reconstruction was obtained for the core region of CPSF, the quality of the density for mCF was much poorer. In about 100,000 particles (~9% of the total particles examined) that were used for 3D classification, some density for mCF was observed (Figures 4A and S3). To improve the quality of this reconstruction, we masked out the density for mPSF from the particles as well as identified additional particles using the mCF density as a template (Figure S5). We then calculated a reconstruction for mCF using only 2D classes that clearly showed the mCF density (Figure 4B). This led to a density map at 7.4 Å resolution, using 35,000 particles (Figures S4 and S5).

Consistent with the negative-stain EM images of mCF alone (Figure S1), the structure of mCF in this complex with mPSF is also trilobal (Figures 4C and 4D). While we were not able to achieve a resolution that was sufficient for building atomic models, the reconstructed density contained sufficient features such that it was recognizable that two of the lobes corresponded to the metallo-β-lactamase and β-CASP domains of CPSF73 and CPSF100. The third lobe therefore was assigned to symplekin. This lobe has an elongated shape and contains density that is indicative of helices (Figures 4C and 4D), consistent with the all helical structure of this domain based on secondary structure predictions.

The overall structures of the metallo-β-lactamase and β-CASP domains of CPSF73 and CPSF100 are expected to be similar to each other at this resolution, and therefore we were not able to make a definitive assignment as to which subunit is located in which lobe of the reconstruction. In one assignment, the β-CASP domain of CPSF100 is located close to the PIM that is bound to CPSF160-WDR33 (Figure 4C), and, in the other assignment, the β-CASP domain of CPSF73 would be close to the PIM (Figure 4D). While the first assignment could more likely be correct because it would place CPSF100 close to the PIM, the second assignment could not be ruled out because the linkers from the PIM to the rest of the β-CASP domain in CPSF100 are quite long (Figure 1A) and because there appears to be some density in this assignment for the long β-hairpin observed in the yeast CPSF100 homolog Ydh1 that brackets the highly hydrophilic segment in the β-CASP domain (Mandel et al., 2006b) (Figure S6). Further studies at higher resolution are needed to definitively resolve this ambiguity, although it does not affect the conclusions from this study.

The central core of this trilobal structure would be composed of the CTDs of CPSF73 and CPSF100, and a segment of symplekin. The C termini of the metallo-β-lactamase domains of CPSF73 and CPSF100 are both located directly next to this density (Figures 4C and 4D). Currently, there is no atomic structure information on the CTDs of CPSF73 and CPSF100, and it is not known whether they have similarity to the CTDs of their homologs IntS9 and IntS11 (Wu et al., 2017). The reconstruction for this part of mCF does not contain sufficient features to make a clear assessment. Earlier studies have shown that the CTDs of CPSF73 and CPSF100 interact with each other (Dominski et al., 2005), and their homologs in yeast interact with each other and with the symplekin homolog Pta1 (Ghazy et al., 2009; Hill et al., 2019; Lidschreiber et al., 2018).

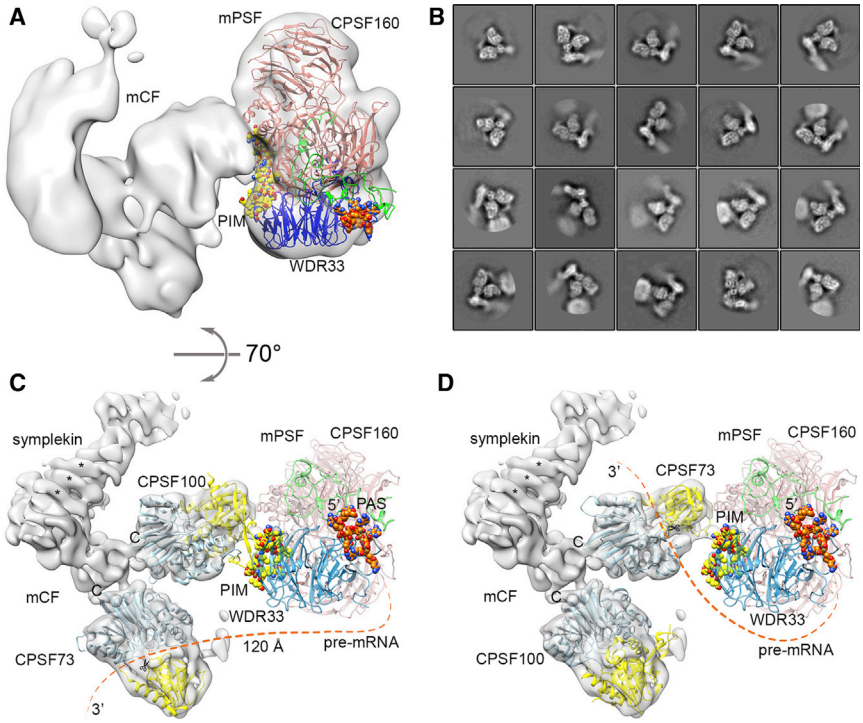


Figure 4. 3D Reconstruction for mCF and Overall Structure of CPSF

(A) After an initial 3D classification, only one class of particles showed density for mCF. The atomic models for mPSF and the CPSF100 PIM are also shown. The PAS RNA and the PIM are in ball-and-stick representation. The view is related to that of Figure 2A by a 90° rotation around the vertical axis.

(B) 2D classification of mCF based on cryo-EM images, obtained after subtraction of mPSF or using mCF as template. Side length of individual averages: 27 nm.

(C) Final 3D reconstruction of mCF at ~7 Å resolution (gray surface), fitted with the atomic structures of the metallo-β-lactamase (cyan) and β-CASP domains (yellow) of CPSF73 and CPSF100. The position of mPSF is shown for reference, as semitransparent ribbons. The AAUAAA PAS RNA and the CPSF100 PIM are shown as ball-and-stick models in orange and yellow, respectively. A possible path of the pre-mRNA from the PAS to the CPSF73 active site (scissors) is indicated with the dashed line, mostly to indicate the distance between the two. Several of the features in the density for symplekin that are consistent with helices are indicated with the asterisks. The view is related to that of panel (A) by a 70° rotation around the horizontal axis.

(D) Same as (C) except that the positions of CPSF73 and CPSF100 have been swapped. This possibility cannot be excluded based solely on the current reconstruction.

See also Figures S4–S6.

Overall Structure of CPSF

Our studies have provided a model for how the subunits of CPSF are arranged relative to each other. mPSF forms the core of the factor, and mCF is arranged to its side, tethered to mPSF by the CPSF100 PIM. Because of the long, disordered linkers between the PIM and the rest of CPSF100, the position of mCF relative to mPSF is highly flexible (Figures 1D and S2). There may be additional contacts between the β-CASP domain of CPSF100 (Figure 4C), or CPSF73 (Figure 4D), with the CTD of CPSF160 in this organization. Nonetheless, the small footprint of this contact and the trilobal structure of mCF produce a much larger overall size for this structure of CPSF (longest dimension of ~170 Å) compared to mPSF alone (40 × 80 × 110 Å^3).

In this structure of CPSF, the distance between the active site of CPSF73 and the AAUAAA PAS is likely more than 120 Å (Figures 4C and 4D). Since the cleavage site of the pre-mRNA is usually 20 nucleotides downstream of the PAS (Tian et al., 2005), the CPSF73 active site is too far from the PAS binding site in the current state of CPSF. Therefore, a change in the position of mCF would be necessary to initiate cleavage of the pre-mRNA. In fact, we do observe many other arrangements of mCF relative to mPSF in the EM images (Video S1), but these other states are not populated sufficiently for us to achieve a stable reconstruction.

Structure of mPSF in Complex with CstF

We expressed and purified human CstF in insect cells, which contained full-length CstF50 and CstF77 and residues 1–195 of CstF64 (covering the RRM and hinge domains) (Figure 1A). Purified human mPSF and CstF formed a stable complex (Figure 5A), and we have produced a cryo-EM reconstruction of this complex at 3.6 Å resolution (Figure 5B; Figures S4 and S7; Table 1). The HAT-N domain of CstF77 has weaker density, suggesting that it is somewhat mobile. We could readily dock the structures of mPSF (Sun et al., 2018) and the HAT domain dimer of CstF77 (Bai et al., 2007) into the density (Figure 5C). The rest of CstF, including CstF50, CstF64, and the C-terminal segment of CstF77, is not observed in this reconstruction, likely due to their flexibility in the different particles. In fact, our negative-stain EM studies of CstF alone showed the presence of CstF50 and CstF64, mostly associated with the HAT-N domain and possibly the C-terminal segment of CstF77 (Figure S2). However, the positions of CstF50 and CstF64 relative to the HAT domain dimer are highly variable.

The HAT domain dimer of CstF77 is bound to the side of mPSF, making contacts to both CPSF160 and WDR33 (Figure 5C). In contrast to mCF, the position of CstF77 relative to mPSF appears to be stable among the complexes. The HAT-C domains of both CstF77 molecules have direct contacts with mPSF, and the interactions do not follow the 2-fold symmetry of this dimer. The HAT-N domains do not appear to be involved in the interactions with mPSF in the current structure.

Roughly one-quarter of the convex surface of the HAT-C dimer is in contact with mPSF, burying approximately 1,000 Å^2 of its surface area. For one HAT-C domain, the loops connecting the two helices in its first five HAT repeats (6–10) are in the interface

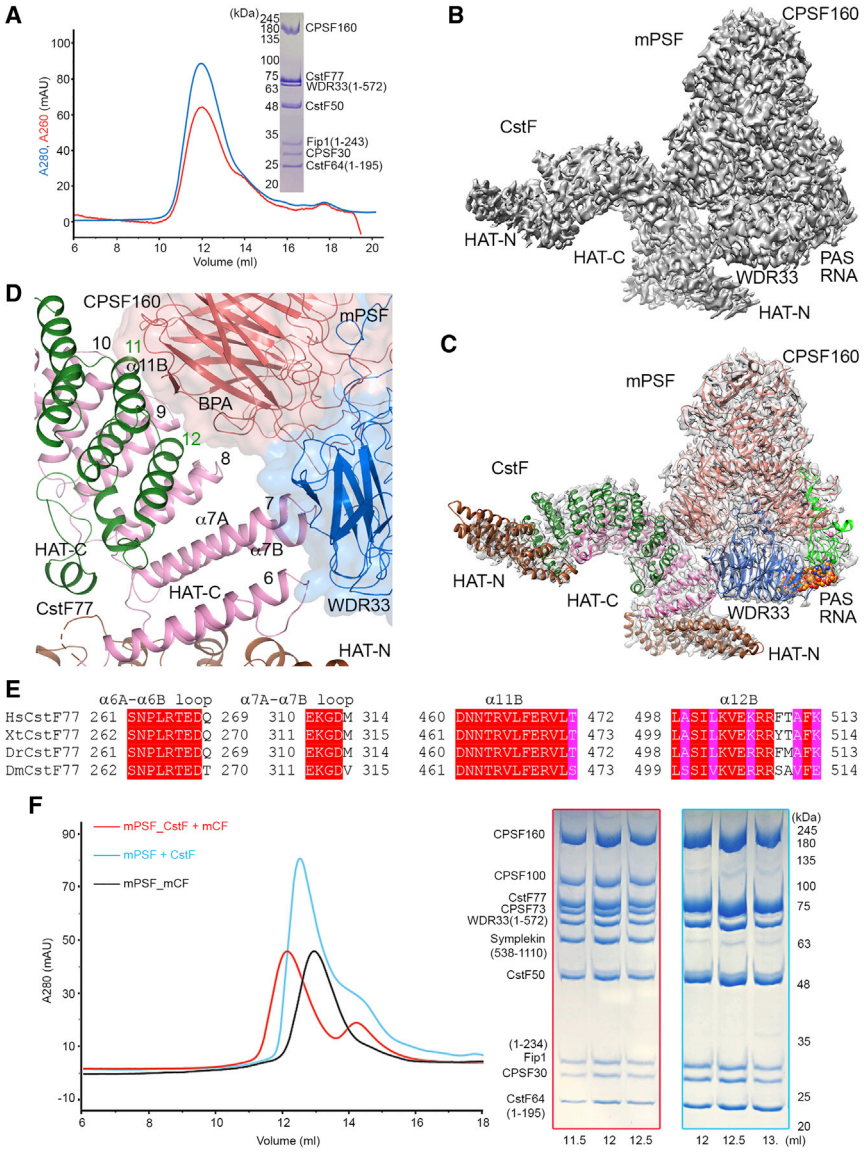


Figure 5. 3D Reconstruction for the Human mPSF-CstF Complex

(A) Gel-filtration profile of purified human mPSF-CstF complex. The absorbance at 280 nm (blue) and 260 nm (red) are shown. The observed A260/280 ratio for the sample is 0.86, and the calculated ratio is 0.94. The SDS-PAGE gel is for the fraction at the peak, 11.5–12 mL.

(B) 3D reconstruction of the mPSF-CstF complex at 3.6 Å resolution (gray). The map is displayed at a lower threshold to show the weak density of the HAT-N domain, and thus some noise is visible on the map surface.

(C) 3D reconstruction of the mPSF-CstF complex (gray surface) fitted with the atomic structures of mPSF and the HAT domain dimer of CstF77 (Bai et al., 2007). The HAT-C domains are colored in pink and dark green, and the HAT-N domains are in brown.

(D) Close-up of the interface between the HAT domain of CstF77 and mPSF.

(E) Sequence conservation of CstF77 residues in the interface with mPSF.

(F) Gel-filtration profile (red curve) for a mixture of mPSF-CstF complex with mCF, showing the formation of a ternary complex of the three factors. The ternary complex (first peak) migrates at an earlier position compared to the mPSF-mCF complex (black curve) or a mixture of mPSF and CstF (cyan curve). The second peak of the red curve contains excess mCF, and that of the cyan curve contains excess CstF. The SDS gels correspond to the red and cyan curves.

See also: Figures S2 and S7.

(Figure 5D). The first two loops contact WDR33, and the following three loops contact BPA of CPSF160. For the other HAT-C domain, the last two repeats (11 and 12) contact the BPA of CPSF160, involving the exposed surface of the second helix in each repeat. Residues at the interface between mPSF and CstF are mostly hydrophilic and ionic in nature, and are highly conserved among animal CstF77 homologs (Figure 5E). However, not sufficient density was observed for the side chains of residues in the interface, due to the limited resolution of this reconstruction, to allow a detailed analysis of their interactions.

Overall Structure of the Human pre-mRNA 3′-End-Processing Machinery

The binding of mCF and CstF to mPSF does not appear to be mutually exclusive, as there are no steric hindrances among their bound positions. This is confirmed by our biochemical observations that purified mPSF, mCF, and CstF can form a ternary complex (Figure 5F). By combining the two separate structures of the mPSF-mCF and mPSF-CstF complexes, we have produced a three-dimensional model for how mPSF, mCF, and CstF are organized in the human pre-mRNA 3′-end-processing machinery (Figures 6A–6C), which account for a large portion of this machinery. Symplekin is located furthest from mPSF and CstF, consistent with experimental data showing that it does not stably associate with mPSF (Schönemann et al., 2014). Besides participating in the core of mCF, the rest of the symplekin CTD may contact the other factors in the machinery.

As described earlier for CPSF, a change in the position of mCF relative to mPSF would be necessary to bring CPSF73 to the cleavage site in the pre-mRNA. Such a rearrangement is likely to happen, given the large movement of mCF relative to mPSF that we observe in the EM images (Video S1).

DISCUSSION

Our studies have provided insights into the organization of the CPSF and CstF components in the human pre-mRNA 3′-end-processing machinery (Figure 6D). The CPSF160-WDR33 complex in mPSF constitutes the core of this machinery, and it likely

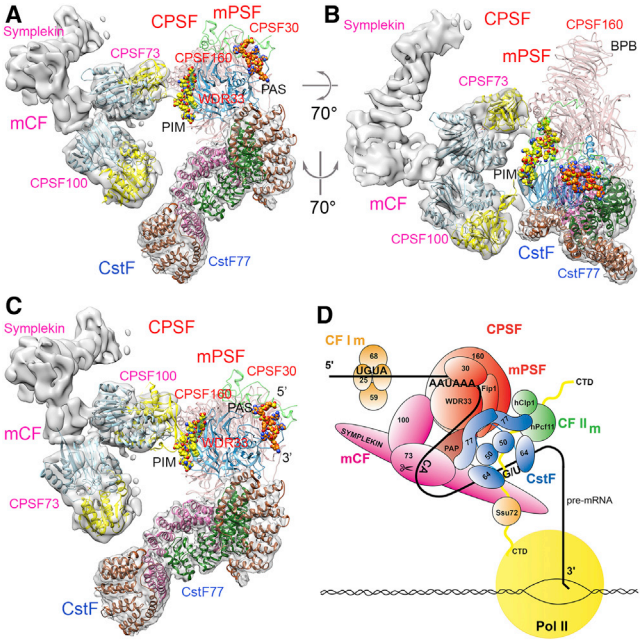


Figure 6. A Model for the CPSF-CstF Components of the Human pre-mRNA 3′-End-Processing Machinery

(A) A model for CPSF-CstF composed by overlaying the structures of the mPSF-mCF and mPSF-CstF complexes. The various proteins and sub-complexes are labeled. The view is similar to that of Figure 4D.

(B) Another view of the CPSF-CstF complex. The BPB domain of CPSF160 is labeled and is not involved in any contacts in this complex.

(C) An alternative model for CPSF-CstF, where the positions of CPSF73 and CPSF100 are swapped compared to those in (A).

(D) A schematic drawing of the mammalian canonical 3′-end-processing machinery.

undergoes only local (if any) conformational changes upon binding other protein factors. CPSF30-PAS RNA, mCF, and CstF are bound to different regions of this core. In fact, they are separated by about 120° around the surface of CPSF160-WDR33 (Figure 6A). For CPSF160, its BPA, BPC, and CTD domains are involved in interactions with these other protein factors, while its BPB domain is free (Figure 6B). The BPB domain could also recruit other proteins, as observed for the homologous DDB1-DDB2 complex, which serves as the core for recruiting many factors for DNA damage repair (Angers et al., 2006; Fischer et al., 2011). Therefore, it remains to be seen which protein(s) in the machinery contact the BPB of CPSF160.

A common theme that emerges from our observations with mCF, CstF, and their complexes with mPSF is the highly dynamic nature of the human pre-mRNA 3′-end-processing machinery, which was also reported for the yeast machinery (Hill et al., 2019). It might be possible that inclusion of other protein factors in the mammalian machinery could lead to its stabilization and reduce the flexibility. It might also be possible that this dynamic behavior is a natural property of this machinery. Moreover, the distance between the active site of the endonuclease CPSF73 and the PAS RNA is too long in the current model of CPSF (Figures 4C and 4D), suggesting that it represents an inactive state of the machinery. The structural flexibility in the machinery could allow it to recognize the sequence elements in the pre-mRNA appropriately and help to bring CPSF73 close to the cleavage site in the substrate. After the cleavage and polyadenylation, the machinery needs to be disassembled and reassembled on a different pre-mRNA.

The structure of the metallo-β-lactamase and β-CASP domains of CPSF73 alone is in a closed state and does not have sufficient space to accommodate the pre-mRNA in the active site (Mandel et al., 2006b), and a recent structure of the yeast homolog Ysh1 is also in a closed state (Hill et al., 2019). Besides positioning CPSF73 near the cleavage site through the dynamic rearrangement of the subunits in the machinery, an activation of CPSF73 itself appears necessary to allow pre-mRNA binding and cleavage. The molecular mechanism of this activation will require further studies.

PAP is recruited to the machinery by Fip1 (Helmling et al., 2001; Kaufmann et al., 2004; Meinke et al., 2008), which in turn interacts with ZF4 and ZF5 of CPSF30 (Barabino et al., 2000). These two zinc fingers are disordered in the earlier reconstruction of mPSF in complex with PAS RNA (Sun et al., 2018). Therefore, the complex between PAP and the rest of the machinery may also be highly dynamic. This may help PAP to be positioned at the newly released 3′ end of the RNA for polyadenylation after the cleavage reaction, and it may also help to release the inhibition of PAP activity by Fip1 (Helmling et al., 2001).

CPSF100 is a weak sequence homolog of CPSF73, and its functional role in the pre-mRNA 3′-end-processing machinery has been a mystery. Our studies show that one function of CPSF100 is to tether mCF to mPSF, which is required for the formation of CPSF. In addition, the interactions between the CTDs of CPSF100 and CPSF73 may help to ensure that the endonuclease and symplekin are recruited to the pre-mRNA. Therefore, our studies have revealed crucial roles for CPSF100 in this machinery.

The HAT domain of CstF77 contacts mPSF, while the C-terminal region of CstF77 recruits CstF50 (Yang et al., 2018) and CstF64 (Moreno-Morcillo et al., 2011; Paulson and Tong, 2012). The dimeric CstF may allow the RRMs from the two CstF64 proteins to recognize the DSE optimally (Figure 6D) (Bai et al., 2007; Yang et al., 2018). On the other hand, the binding site for mPSF on CstF77 does not obey its 2-fold symmetry, suggesting that there may be a second, equivalent binding site for mPSF on the CstF77 dimer. Using the binding mode of mPSF observed here, a second copy of mPSF could be modeled based on the 2-fold symmetry of the dimer. There are no steric clashes between the two copies of mPSF in this model. However, our gel-filtration data show that only one copy of mPSF is bound to CstF, even with mPSF in 2-fold molar excess. Further studies will be necessary to illuminate how the stoichiometry between mPSF (and thereby CPSF) and CstF is determined.

Our observations on the overall organization of the human pre-mRNA 3′-end-processing machinery are likely to be relevant to the machineries in other organisms as well. Especially, the PIM is highly conserved from yeast to humans, suggesting that it may have a similar role in tethering components of the cleavage and polyadenylation factor (CPF) in the yeast machinery. In addition, mCF together with CstF64 constitute the histone cleavage complex (HCC), responsible for 3′-end-processing of metazoan

replication-dependent histone pre-mRNAs (Marzluff and Koreski, 2017). The structural information on mCF could also help to understand the molecular mechanism of that machinery.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Protein expression and purification
  - CPSF-CstF64 complex formation
  - mPSF-CstF complex formation
  - mPSF-mCF-CstF complex formation
  - *In vitro* pulldown assay
  - EM sample preparation and data collection
  - Image processing
  - Model building and refinement
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2019.11.005.

ACKNOWLEDGMENTS

We thank Ed Eng, Bill Rice, Mykhailo Kopylov, and Bob Grassucci for help with data collection at the New York Structural Biology Center and Mark Ebrahim and Johanna Sotiris for help with grids screening at the Evelyn Gruss Lipper Cryo-Electron Microscopy Resource Center at The Rockefeller University. This research is supported by NIH grants R35GM118093 (to L.T.) and R01GM090056 (to Y. Shi). Some of this work was performed at the Simons Electron Microscopy Center and National Resource for Automated Molecular Microscopy located at the New York Structural Biology Center, supported by grants from the Simons Foundation (349247), NYSTAR, and the NIH National Institute of General Medical Sciences (GM103310) with additional support from the Agouron Institute (F00316) and NIH S10 OD019994.

AUTHOR CONTRIBUTIONS

Y.Z. carried out EM data collection and analysis, EM reconstruction, model building, and refinement. Y. Sun prepared all the samples for the EM analysis, carried out the mutagenesis experiments on the PIM, and performed model building and structure refinement. L.T., T.W., and Y. Shi supervised the research and analyzed the data. L.T., Y. Sun, Y.Z., and T.W. wrote the paper, and all authors commented on the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 20, 2019
Revised: September 16, 2019
Accepted: October 29, 2019
Published: December 3, 2019

REFERENCES

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.

Angers, S., Li, T., Yi, X., MacCoss, M.J., Moon, R.T., and Zheng, N. (2006). Molecular architecture and assembly of the DDB1-CUL4A ubiquitin ligase machinery. Nature *443*, 590–593.

Bai, Y., Auperin, T.C., Chou, C.-Y., Chang, G.-G., Manley, J.L., and Tong, L. (2007). Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. Mol. Cell *25*, 863–875.

Barabino, S.M.L., Ohnacker, M., and Keller, W. (2000). Distinct roles of two Yth1p domains in 3′-end cleavage and polyadenylation of yeast pre-mRNAs. EMBO J. *19*, 3778–3787.

Casañal, A., Kumar, A., Hill, C.H., Easter, A.D., Emsley, P., Degliesposti, G., Gordiyenko, Y., Santhanam, B., Wolf, J., Wiederhold, K., et al. (2017). Architecture of eukaryotic mRNA 3′-end processing machinery. Science *358*, 1056–1059.

Chan, S.L., Huppertz, I., Yao, C., Weng, L., Moresco, J.J., Yates, J.R., 3rd, Ule, J., Manley, J.L., and Shi, Y. (2014). CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. Genes Dev. *28*, 2370–2380.

Clerici, M., Faini, M., Aebersold, R., and Jinek, M. (2017). Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. eLife *6*, e33111.

Clerici, M., Faini, M., Muckenfuss, L.M., Aebersold, R., and Jinek, M. (2018). Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. Nat. Struct. Mol. Biol. *25*, 135–138.

Dominski, Z., Yang, X.-C., Purdy, M., Wagner, E.J., and Marzluff, W.F. (2005). A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. Mol. Cell. Biol. *25*, 1489–1500.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Fischer, E.S., Scrima, A., Böhm, K., Matsumoto, S., Lingaraju, G.M., Faty, M., Yasuda, T., Cavadini, S., Wakasugi, M., Hanaoka, F., et al. (2011). The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation. Cell *147*, 1024–1039.

Ghazy, M.A., He, X., Singh, B.N., Hampsey, M., and Moore, C. (2009). The essential N terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3′-end processing. Mol. Cell. Biol. *29*, 2296–2307.

Helmling, S., Zhelkovsky, A., and Moore, C.L. (2001). Fip1 regulates the activity of Poly(A) polymerase through multiple interactions. Mol. Cell. Biol. *21*, 2026–2037.

Hill, C.H., Boreikaitė, V., Kumar, A., Casañal, A., Kubík, P., Degliesposti, G., Maslen, S., Mariani, A., von Loeffelholz, O., Girbig, M., et al. (2019). Activation of the endonuclease that defines mRNA 3′ ends requires incorporation into an 8-subunit core cleavage and polyadenylation factor complex. Mol. Cell *73*, 1217–1231.e11.

Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. EMBO J. *23*, 616–626.

Kolev, N.G., Yario, T.A., Benson, E., and Steitz, J.A. (2008). Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′-end maturation. EMBO Rep. *9*, 1013–1018.

Legrand, P., Pinaud, N., Minvielle-Sébastia, L., and Fribourg, S. (2007). The structure of the CstF-77 homodimer provides insights into CstF assembly. Nucleic Acids Res. *35*, 4515–4522.

Lidschreiber, M., Easter, A.D., Battaglia, S., Rodríguez-Molina, J.B., Casañal, A., Carminati, M., Baejen, C., Grzechnik, P., Maier, K.C., Cramer, P., and

Passmore, L.A. (2018). The APT complex is involved in non-coding RNA transcription and is distinct from CPF. Nucleic Acids Res. *46*, 11528–11538.

Mandel, C.R., Gebauer, D., Zhang, H., and Tong, L. (2006a). A serendipitous discovery that in situ proteolysis is essential for the crystallization of yeast CPSF-100 (Ydh1p). Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *62*, 1041–1045.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006b). Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. Nature *444*, 953–956.

Mandel, C.R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3′-end processing. Cell. Mol. Life Sci. *65*, 1099–1122.

Marzluff, W.F., and Koreski, K.P. (2017). Birth and death of histone mRNAs. Trends Genet. *33*, 745–759.

Meinke, G., Ezeokonkwo, C., Balbo, P., Stafford, W., Moore, C., and Bohm, A. (2008). Structure of yeast poly(A) polymerase in complex with a peptide from Fip1, an intrinsically disordered protein. Biochemistry *47*, 6859–6869.

Moreno-Morcillo, M., Minvielle-Sébastia, L., Fribourg, S., and Mackereth, C.D. (2011). Locked tether formation by cooperative folding of Rna14p monkeytail and Rna15p hinge domains in the yeast CF IA complex. Structure *19*, 534–545.

Ohi, M., Li, Y., Cheng, Y., and Walz, T. (2004). Negative staining and image classification - powerful tools in modern electron microscopy. Biol. Proced. Online *6*, 23–34.

Paulson, A.R., and Tong, L. (2012). Crystal structure of the Rna14-Rna15 complex. RNA *18*, 1154–1162.

Pérez Cañadillas, J.M., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. EMBO J. *22*, 2821–2830.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev. *25*, 1770–1782.

Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods *14*, 290–296.

Rohou, A., and Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J. Struct. Biol. *192*, 216–221.

Ruepp, M.D., Schweingruber, C., Kleinschmidt, N., and Schümperli, D. (2011). Interactions of CstF-64, CstF-77, and symplekin: implications on localisation and function. Mol. Biol. Cell *22*, 91–104.

Sari, D., Gupta, K., Thimiri Govinda Raj, D.B., Aubert, A., Drncová, P., Garzoni, F., Fitzgerald, D., and Berger, I. (2016). The MultiBac baculovirus/insect cell expression vector system for producing complex protein biologics. Adv. Exp. Med. Biol. *896*, 199–215.

Schönemann, L., Kühn, U., Martin, G., Schäfer, P., Gruber, A.R., Keller, W., Zavolan, M., and Wahle, E. (2014). Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. Genes Dev. *28*, 2381–2393.

Shi, Y., and Manley, J.L. (2015). The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. Genes Dev. *29*, 889–897.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J., and Manley, J.L. (2009). Molecular architecture of the human pre-mRNA 3′ processing complex. Mol. Cell *33*, 365–376.

Sullivan, K.D., Steiniger, M., and Marzluff, W.F. (2009). A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. Mol. Cell *34*, 322–332.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., and Carragher, B. (2005). Automated molecular microscopy: the new Leginon system. J. Struct. Biol. *151*, 41–60.

Sun, Y., Zhang, Y., Hamilton, K., Manley, J.L., Shi, Y., Walz, T., and Tong, L. (2018). Molecular basis for the recognition of the human AAUAAA polyadenylation signal. Proc. Natl. Acad. Sci. USA *115*, E1419–E1428.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., and Ludtke, S.J. (2007). EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol. *157*, 38–46.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. *33*, 201–212.

Wu, Y., Albrecht, T.R., Baillat, D., Wagner, E.J., and Tong, L. (2017). Molecular basis for the interaction between Integrator subunits IntS9 and IntS11 and its functional importance. Proc. Natl. Acad. Sci. USA *114*, 4394–4399.

Xiang, K., Nagaike, T., Xiang, S., Kilic, T., Beh, M.M., Manley, J.L., and Tong, L. (2010). Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature *467*, 729–733.

Xiang, K., Manley, J.L., and Tong, L. (2012). An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72. Genes Dev. *26*, 2265–2270.

Xiang, K., Tong, L., and Manley, J.L. (2014). Delineating the structural blueprint of the pre-mRNA 3′-end processing machinery. Mol. Cell. Biol. *34*, 1894–1910.

Yang, Q., and Doublié, S. (2011). Structural biology of poly(A) site definition. Wiley Interdiscip. Rev. RNA *2*, 732–747.

Yang, Z., Fang, J., Chittuluru, J., Asturias, F.J., and Penczek, P.A. (2012). Iterative stable alignment and clustering of 2D transmission electron microscope images. Structure *20*, 237–247.

Yang, W., Hsu, P.L., Yang, F., Song, J.E., and Varani, G. (2018). Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3′-end processing signals. Nucleic Acids Res. *46*, 493–503.

Zhao, J., Hyman, L., and Moore, C. (1999). Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol. Mol. Biol. Rev. *63*, 405–445.

Zheng, S.Q., Palovcak, E., Armache, J.P., Verba, K.A., Cheng, Y., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods *14*, 331–332.

Zivanov, J., Nakane, T., Forsberg, B., Kimanius, D., Hagen, W.J.H., Lindahl, E., and Scheres, S.H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife *7*, e42166. Published online November 9, 2018. https://doi.org/10.7554/eLife.42166.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E.coli DH5α	Thermo Fisher Scientific	Cat#18265017
E.coli DH10 EMBacY	Geneva Biotech	
E.coli BL21 star	Thermo Fisher Scientific	Cat#C6010-03
E.coli BW23473	Qualityard	Cat#QYV0579
Chemicals, Peptides, and Recombinant Proteins		
ESF 921 Insect Cell Culture Medium	Expression Systems	Cat#96-001-01
Protease Inhibitor Cocktail	Sigma-Aldrich	Cat#11836170001
Imidazole	Sigma-Aldrich	Cat#56750
TEMED	Sigma-Aldrich	CAS 110-18-9
Ammonium persulfate (APS)	Sigma-Aldrich	GE17-1311-01
30% Acrylamide/Bis Solution, 29:1	BIO-Rad	Cat#1610156
Ni-NTA Agarose	QIAGEN	Cat#30230
Amylose Resin	New England Biolabs	Cat#E8021S
Cre Recombinase	New England Biolabs	Cat#M0298S
DpnI	New England Biolabs	Cat#R0176L
Phusion high-fidelity DNA polymerase	New England Biolabs	Cat#M0530S
Cellfectin II Reagent	Thermo Fisher Scientific	Cat#10362100
Transfection Medium	Expression Systems	Cat# 95-020-100
Recombinant protein complex: human “mCF” CPSF73-CPSF100-Symplekin (538-1110)	This work	N/A
Recombinant protein complex: human “mCF” CPSF73-CPSF100-Symplekin (538-762)	This work	N/A
Recombinant protein complex: human “mCF+CstF64” CPSF73-CPSF100-Symplekin (sumo-353-1110)-CstF64	This work	N/A
Recombinant protein complex: human “mCF” CPSF73-CPSF100 W473A/Y476A-Symplekin (538-1110)	This work	N/A
Recombinant protein complex: human “mCF” CPSF73-CPSF100 F464A/W473A/Y476A -Symplekin (538-1110)	This work	N/A
Recombinant protein complex: human “mCF” CPSF73-CPSF100 del(460-486)-Symplekin (538-1110)	This work	N/A
Recombinant protein complex: human CPSF160-WDR33 (1-572)	This work	N/A
Recombinant protein complex: human CPSF30-Fip1(1-243)	This work	N/A
Recombinant protein complex: human CstF77-CstF50-CstF64(1-195)	This work	N/A
Recombinant protein complex: “CPSF+CstF64” CPSF160-WDR33(1-572)-CPSF30-Fip1(1-243)-CPSF73-CPSF100-Symplekin(sumo-353-1110)-CstF64	This work	N/A
Recombinant protein complex: “mPSF+CstF” CPSF160-WDR33(1-572)-CPSF30-Fip1(1-243)-CstF77-CstF50-CstF64(1-195)	This work	N/A
Recombinant protein complex: “mPSF+mCF+CstF” CPSF160-WDR33(1-572)-CPSF30-Fip1(1-243)-CPSF73-CPSF100-Symplekin(538-1110)-CstF77-CstF50-CstF64(1-195)	This work	N/A
Recombinant protein: MBP	This work	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Recombinant protein: MBP-CPSF100 (460-486)	This work	N/A
Recombinant protein: MBP-CPSF100 (460-Y476A-486)	This work	N/A
Recombinant protein: MBP-CPSF100 (460-W473A/ Y476A-486)	This work	N/A
Recombinant protein: MBP-CPSF100	This work	N/A
Recombinant protein: MBP-CPSF100 delete (460-486)	This work	N/A
Deposited Data		
Mendeley raw data (uncropped gels)	This work	https://dx.doi.org/10.17632/f3v9w58m3j.1
CPSF160-WDR33-CPSF30-CPSF100 cryo-EM structure	This work	PDB: 6URG
CPSF160-WDR33-CPSF30-CPSF100 cryo-EM map	This work	EMD: 20860
CPSF160-WDR33-CPSF30-PAS RNA-CstF77 cryo-EM structure	This work	PDB: 6URO
CPSF160-WDR33-CPSF30-PAS RNA-CstF77 cryo-EM map	This work	EMD: 20861
CPSF73-CPSF100-Symplekin cryo-EM map	This work	EMD: 20859
Crystal structure of human CPSF73 (used for Figure 4 and Figure 6)	(Mandel et al., 2006b)	PDB: 2I7V
Crystal structure of yeast CPSF100 (used for Figure 4 and Figure 6)	(Mandel et al., 2006b)	PDB: 2I7X
Crystal structure of mouse CstF77 (used for Figure 5 and Figure 6)	(Bai et al., 2007)	PDB: 2OOE
Experimental Models: Cell Lines		
Sf9	Thermo Fisher Scientific	Cat#11496-015
High5	Thermo Fisher Scientific	Cat#B855-02
Oligonucleotides		
RNA (used in mPSF+CstF complex) 5′-UUCACAAA UAAAGCAUUUUUUUCACUGCAUUCUAGUUGUG GUUUGUCC-FAM-3′	This work	IDT fluorescent oligo
Primer sequences	This work	See Table S1
Recombinant DNA		
CPSF73-CPSF100 in the pSPL vector	This work	N/A
Symplekin (6xHis-538-1110) in the pFL vector	This work	N/A
Fusion CPSF73-CPSF100-Symplekin(6xHis-538-1110) (baculovirus expression)		N/A
CstF64 in the pSPL vector	This work	N/A
Symplekin(6xHis-Sumo-353-1110) in the pFL vector	This work	N/A
Fusion CstF64-Symplekin(6xHis-Sumo-353-1110) (baculovirus expression)	This work	N/A
CPSF73-CPSF100 in the pFL vector (baculovirus expression)	This work	N/A
CPSF160 in the pKL vector (baculovirus expression)	This work	N/A
WDR33(6xHis-1-572) in the pFL vector (baculovirus expression)	This work	N/A
CPSF30_6xHis in the pFL vector	This work	N/A
Fip1(1-243) in the pUCDM vector	This work	N/A
Fusion CPSF30_6xHis-Fip1(1-243) (baculovirus expression)	This work	N/A
CPSF73 in the pSPL vector	This work	N/A
Symplekin(6xHis-538-1110) in the pFL vector	This work	N/A
Fusion CPSF73-Symplekin(6xHis-538-1110) (baculovirus expression)	This work	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
CPSF100 mutants in the pFL vector (baculovirus expression)	This work	N/A
CstF77-CstF50 in the pFL vector (baculovirus expression)	This work	N/A
CstF64 (6xHis-1-195) in the pFL vector (baculovirus expression)	This work	N/A
6xHis-MBP-TEV in the pRSFDuet vector	This work	N/A
6xHis-MBP-TEV-CPSF100 in the pRSFDuet vector	This work	N/A
6xHis-MBP-TEV-CPSF100 delete PIM in the pRSFDuet vector	This work	N/A
6xHis-MBP-TEV-CPSF100 PIM in the pRSFDuet vector	This work	N/A
6xHis-MBP-TEV-CPSF100 PIM Y476A in the pRSFDuet vector	This work	N/A
6xHis-MBP-TEV-CPSF100 PIM W473A/Y476A in the pRSFDuet vector	This work	N/A
Software and Algorithms		
COOT	(Emsley and Cowtan, 2004)	https://www2.mrc-lmb.cam.ac.uk/Personal/pemsley/coot/
Phenix.refine	(Adams et al., 2002)	https://www.phenix-online.org/
RELION3.0	(Zivanov et al., 2018)	https://www3.mrc-lmb.cam.ac.uk/relion/index.php?title=Main_Page
UCSF Chimera	(Pettersen et al., 2004)	http://www.cgl.ucsf.edu/chimera
PyMOL	Schrodinger	https://www.pymol.org/2/
MotionCor2	(Zheng et al., 2017)	http://msg.ucsf.edu
Leginon	(Suloway et al., 2005)	https://emg.nysbc.org/redmine/projects/leginon/wiki/Leginon_Homepage
CTFFIND4	(Rohou and Grigorieff, 2015)	http://grigoriefflab.janelia.org/ctffind4
Gautomatch	Kai Zhang	https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/
CryoSPARC	(Punjani et al., 2017)	https://cryosparc.com/
Other		
R 1.2/1.3 300 mesh gold grids	Quantifoil	Cat# Q3100AR1.3

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Liang Tong (ltong@columbia.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All gene cloning, manipulation and plasmid propagation steps involving pFL, pKL and pRSF vectors were carried out in bacterial *E. coli* DH5α cells grown in LB media supplemented with appropriate antibiotics. *E. coli* BW23473 cells were used for constructs in pSPL and pUCDM vectors. *E. coli* DH10 EMBacY cells were used for bacmid isolation. BL21 Star (DE3) cells were used for MBP-CPSF100 fusion protein expression in LB media. The cells were induced with 0.5 mM IPTG and grown at 16°C for 16-20 h. For all other recombinant protein complexes, Sf9 cell line was used for baculovirus formation and amplification. High5 cell line was used for baculovirus-driven overexpression as described below.

METHOD DETAILS

Protein expression and purification

All the complexes were expressed in insect cells using Multibac technology (Sari et al., 2016) (Geneva Biotech). For the human CPSF73-CPSF100-symplekin complex, CPSF73 and CPSF100 were cloned into the pSPL donor vector. Symplekin (residues 538-1110) was cloned into the pFL acceptor vector and a 6 × His tag was added to the N terminus of symplekin. The donor was fused to the acceptor by Cre recombinase.

For the human CPSF73-CPSF100-symplekin-CstF64 complex, CstF64 and symplekin (residues 353-1110) were cloned into the pSPL and pFL vector, respectively, and a 6 × His-SUMO tag was added to the N terminus of symplekin. These two vectors were fused together by Cre recombinase. CPSF73 and CPSF100 were cloned into the pFL vector. High5 cells were co-infected with 12 mL CstF64-symplekin P2 virus and 12 mL CPSF73-CPSF100 P2 virus.

For the human CPSF160-WDR33 complex, CPSF160 and WDR33 (residues 1-572) were cloned into the pKL and pFL vectors, respectively, and a 6 × His tag was added to the N terminus of WDR33. High5 cells were co-infected with 15 mL CPSF160 P2 virus and 10 mL WDR33 P2 virus.

For the human CPSF30-Fip1 complex, CPSF30 and Fip1 (residues 1-243) were cloned into the pFL acceptor and the pUCDM donor vector, respectively, and a 6 × His tag was added to the C terminus of CPSF30. These two vectors were fused together by Cre recombinase. We used isoform 2 of CPSF30, in which residues 191-215 are absent.

For the human CPSF73-CPSF100 mutant-symplekin complex, CPSF73 and symplekin (residues 538-1110) were cloned into the pSPL donor and the pFL acceptor vectors, and a 6 × His tag was added to the N terminus of symplekin. These two vectors were fused together by Cre recombinase. CPSF100 with double mutant W473A/Y476A, triple mutant F464A/W473A/Y476A, and internal deletion of residues 460-486 were cloned into the pFL vector, respectively. High5 cells were co-infected with 10 mL CPSF73-symplekin P2 virus and 10 mL CPSF100 mutant P2 virus.

For human CstF, full-length CstF77 and CstF50 were cloned into the same pFL vector. CstF64 (residues 1-195) was cloned into another pFL vector and a 6 × His tag was added to its N terminus. High5 cells were co-infected with 10 mL CstF77-CstF50 P2 virus and 10 mL CstF64 P2 virus.

Bacmids for all the complexes were generated in DH10EMBacY competent cells (Geneva Biotech) by transformation. Baculoviruses were generated by transfecting bacmids into Sf9 cells using Cellfectin II (Thermo Fisher Scientific). P1 viruses were cultured at 27°C for 5 days, and P2 viruses for large-scale infection were amplified from P1 viruses in 50 mL Sf9 cells at 27°C for 3 days. One liter of High5 cells (1.8×10^6 cells ml^{-1}) cultured in ESF 921 medium (Expression Systems) was infected with a suitable amount of P2 virus at 27°C with constant shaking. Cells were harvested after 48 h by centrifugation at 2000 rpm for 13 min.

For purification, the cell pellet was re-suspended and lysed by sonication in 100 mL buffer containing 25 mM Tris (pH 8.0), 300 mM NaCl and one protease inhibitor cocktail tablet (Sigma). The cell lysate was then centrifuged at 13,000 rpm for 45 min at 4°C. The supernatant was incubated with nickel beads for 1 h at 4°C. The beads were then washed 4 times with 50 bed volumes of wash buffer (25 mM Tris (pH 8.0), 150 mM NaCl and 20 mM imidazole) and eluted with 25 mM Tris (pH 8.0), 150 mM NaCl and 250 mM imidazole. The protein was further purified by chromatography using a HiTrap Q column (GE Healthcare) and a Superose 6 10/300 GL column (GE Healthcare).

The peak fractions of the CPSF73-CPSF100-symplekin complex were used for EM studies in a buffer containing 25 mM Tris (pH 8.0), 150 mM NaCl and 5 mM DTT. For other complexes, fractions of interest were concentrated to 1∼3 mg/ml, and stored at –80°C.

CPSF-CstF64 complex formation

Purified CPSF160-WDR33 complex, CPSF30-Fip1 complex and CPSF73-CPSF100-symplekin-CstF64 complex were mixed at a molar ratio of 1:1.2:1. The reaction mixture was incubated on ice for 1 h, with or without PAS RNA, and then purified by gel filtration using a Superose 6 10/300 GL column (GE Healthcare), in a running buffer containing 25 mM Tris (pH 8.0), 150 mM NaCl and 5 mM DTT. Fractions of interest were used for EM studies. mPSF and mCF mutants mixing assays used the same method. The mixtures did not contain PAS RNA.

mPSF-CstF complex formation

Purified CPSF160-WDR33 complex, CPSF30-Fip1 complex, and CstF77-CstF64-CstF50 complex were mixed at a molar ratio of 1:1.3:1 in the presence of the 48-mer RNA oligonucleotide UUCACAAAUAAAGCAUUUUUUU-CACUGCAUUCUAGUUGUGGUUU GUCC (with 3′ end 6-FAM label; IDT). The mixture was incubated on ice for 1 h and then purified by gel filtration using a Superose 6 10/300 GL column (GE Healthcare), in a running buffer containing 20 mM HEPES (pH 7.5), 100 mM NaCl, 5 mM EDTA and 5 mM DTT. Fractions of interest were used for EM studies.

mPSF-mCF-CstF complex formation

Purified CPSF73-CPSF100-symplekin(538-1110) (mCF) was mixed with mPSF-CstF complex at a molar ratio of 2:1, and RNA was not included in the reaction. The mixture was incubated on ice for 1 h and then purified by gel filtration using a Superose 6 10/300 GL column (GE Healthcare), in a running buffer containing 25 mM Tris (pH 8.0), 150 mM NaCl and 5 mM DTT.

In vitro pulldown assay

Human full-length CPSF100, CPSF100 ΔPIM, and PIM alone (wild-type, single mutant Y476A and double mutant W473A/Y476A) were cloned into the pRSFDuet vector (Novagen) and overexpressed in *E. coli* BL21 (DE3) Star cells. A 6 × His tag followed by maltose binding protein (MBP) was added to the N terminus of CPSF100, separated by a TEV protease cleavage site. Cell cultures were grown at 37°C in LB (Sigma) containing 35 μg ml^{-1} kanamycin. When 0.5 mL cell cultures reached an OD_{600} of 0.6∼0.7, protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at room temperature overnight. The supernatant

after sonication and centrifugation was incubated with 30 μL amylose resin for 1 h at 4°C. The resin was washed twice with 1 mL buffer containing 25 mM Tris (pH 8.0), 150 mM NaCl. To test the interaction between the CPSF100 segment and the CPSF160-WDR33 complex, an equal molar amount of CPSF160-WDR33 complex was loaded onto MBP-CPSF100 bound amylose resin. The mixture was incubated on ice for 30 min. The resin was washed three times with 1 mL buffer. The resin was loaded onto 4%–20% gradient SDS-PAGE gel and detected by Coomassie blue staining.

EM sample preparation and data collection

The homogeneity of samples was first assessed by negative-stain EM with 0.7% (w/v) uranyl formate as described (Ohi et al., 2004). Using a Philips CM10 electron microscope operated at 100 kV, 67 images were collected for the mCF complex, 68 images for the CPSF complex, and 50 images for the CstF complex. The images were recorded at a defocus of –1.5 μm on an XR16L-ActiveVu charge-coupled device camera (AMT) at a nominal magnification of 52,000 × (calibrated pixel size of 2.4 Å at the specimen level).

Freshly purified CPSF samples were centrifuged at 13,000 × *g* for 2 min to remove protein aggregates. The protein concentration was measured with a NanoDrop spectrophotometer (Thermo Fisher Scientific) and adjusted to 0.25 mg·ml^{-1}. All cryo-EM specimens were prepared with a Vitrobot Mark VI (FEI) set at 4°C and 100% humidity. A 4 μL aliquot was applied to a glow-discharged Quantifoil 300 mesh 1.2/1.3 gold grid (Quantifoil). The grid was blotted for 4 s at a blot force setting of −2 and plunged into liquid ethane cooled by liquid nitrogen.

Three datasets (CPSF-Krios1, CPSF-Krios3 and mPSF-CstF-Krios1) were collected on two Titan Krios electron microscopes (Krios1 and Krios3) in the Simons Electron Microscopy Center at the New York Structural Biology Center using Leginon (Suloway et al., 2005). The images were recorded with a K2 Summit camera in counting mode at a nominal magnification of 22,500 × (calibrated pixel sizes on the specimen level of 1.07 Å for CPSF-Krios1 and mPSF-CstF-Krios1, and 1.06 Å for CPSF-Krios3) and a defocus range from −1.2 to −2.5 μm. Exposures of 10 s were dose-fractionated into 40 frames (250 ms per frame), with a dose rate of 8 electrons·pixel^{-1}·s^{-1}, resulting in a total dose of 70 electrons·Å^{-2} for CPSF-Krios1 and mPSF-CstF-Krios1, and 71 electrons·Å^{-2} for CPSF-Krios3.

Image processing

EMAN2 (Tang et al., 2007) was used to pick 14,883 particles of the negatively stained mCF that were extracted into 100 × 100-pixel images, 21,315 particles for the negatively stained CPSF that were extracted into 192 × 192-pixel images, and 11,550 particles for the negatively stained CstF that were extracted into 112 × 112-pixel images. All particle images were resized to 64 × 64 pixels. After centering, the particles were subjected to classification with the iterative stable alignment and clustering algorithm (Yang et al., 2012), specifying 100 images per group and a pixel error threshold of 0.7. Six generations produced 97 averages for mCF, 116 averages for CPSF and 146 averages for CstF.

For the cryo-EM data, the image stacks were motion-corrected and dose-weighted in MotionCor2 (Zheng et al., 2017). The CTF parameters were determined with CTFFIND4 (Rohou and Grigorieff, 2015). For the CPSF-Krios3 dataset, 1,539,569 particles were automatically picked with Gautomatch (https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) from 4,095 micrographs. The particles were windowed into 192 × 192-pixel images and subjected to 2D classification in RELION-3, which was also used for all further image processing (Zivanov et al., 2018). Particles in classes that generated averages showing clear structural features were combined (1,177,628 particles) and subjected to 3D classification into seven classes using as initial reference the density map of the CPSF160-WDR33 complex we determined before (Sun et al., 2018) filtered to a resolution of 30 Å. Four classes produced maps with clear fine structural features, and the particles of these classes were combined. Refinement yielded a density map at 3.5 Å resolution. To further improve the map, the particles were re-centered, re-extracted into bigger images of 256 × 256 pixels, and subjected to CTF refinement and Bayesian polishing, resulting in a map at 3.0 Å resolution (Figure S3). The core of CPSF together with two fragments of CPSF100 were well resolved, but most density for mCF was missing from the map. During the initial 3D classification step, one class showed additional density extending laterally from the core. The particles in this class were re-centered, re-extracted into bigger images of 360 × 360 pixels, and subjected to 3D classification into five classes. One class showed a more complete mCF complex, but refinement of this class showed that the map suffered from particles assuming strongly preferred orientations. Despite this problem, the map showed that mCF adopts a quite well-defined position relative to the CPSF core (Figure S3).

The CPSF-Krios1 dataset (7,608 movie stacks) was recorded from grid areas with a thick ice layer. To combine it with the CPSF-Krios3 dataset, the pixel size from CPSF-Krios3 was rescaled from 1.06 Å to 1.07 Å. Processing of the combined Krios datasets did not improve the quality of the map for the core structure. To generate a better map for the mCF complex, two approaches were used to identify particles containing the mCF subunit. Approach 1: Particles were auto-picked with Gautomatch using averages of the core as templates. The 5,104,366 auto-picked particles were extracted into 192 × 192-pixel images and subjected to 2D classification into 150 classes. The classes that generated averages showing clear secondary structures were combined (3,683,475 particles) and refined using the core structure as reference. After refinement, the particles were re-extracted into 256 × 256-pixel images, applying a shift so that the putative mCF was in the center of the extracted images. The re-extracted particles were subjected to another round of refinement, after which the signal for the core was subtracted from the images. The signal-subtracted images were subjected to 2D classification. The classes showing the mCF complex were selected for further processing. Approach 2: Particles were auto-picked with Gautomatch using averages of mCF generated with Approach 1 as templates. The 5,331,049 auto-picked particles were extracted into 180 × 180-pixel images and subjected to 2D classification into 150 classes. Classes that produced averages showing

features of mCF were combined (589,047 particles), re-extracted into 256 × 256-pixel images, and combined with the 182,521 particles selected with Approach 1. The combined particles were subjected to 3D classification using as reference the initial model for mCF obtained with cryoSPARC (Punjani et al., 2017). Four classes that showed fine structural features were combined (162,939 particles) and duplicate particles, i.e., particles that were less than 80 Å apart, were removed by the subset selection option in RELION-3. The remaining 144,906 particles were subjected to another round of 3D classification. The class showing the best structural features was refined to yield a map at 7.4 Å resolution. The same class was also refined using different masks to improve the resolution in different map regions (Figure S5) Fourier shell correlation curves, local resolution maps, and resolution-filtered maps were calculated in RELION-3 (Figure S4).

For the mPSF-CstF-Krios1 dataset, 1,723,794 particles were automatically picked from 4819 micrographs with Gautomatch. The particles were windowed into 192 × 192-pixel images and subjected to 2D classification into 150 classes. The particles from the 2D classes that showed clear secondary structures were selected (869,765 particles) and subjected to 3D classification into 8 classes using the CPSF core structure as initial model. The 58,111 particles in the class that showed density for CstF77 were selected, re-extracted into 280 × 280-pixel images, and subjected to 3D classification into four classes. The particles in the class showing clear fine structural features were combined (50,092) and refined, yielding a density map at 3.7 Å resolution. CTF refinement and Bayesian polishing improved the map to 3.6 Å resolution. Fourier shell correlation curves, local resolution maps, and resolution-filtered maps were calculated in RELION-3 (Figure S7).

Model building and refinement

We used the cryo-EM structure of CPSF160-WDR33-CPSF30-PAS RNA complex (PDB ID 6DNH) (Sun et al., 2018), the crystal structures of CPSF73 (PDB ID 2I7V) and Ydh1 (PDB ID 2I7X) (Mandel et al., 2006b), and the crystal structure of CstF77 (PDB ID 2OOE) (Bai et al., 2007) as the starting models, and fitted them into the cryo-EM density map with Chimera (Pettersen et al., 2004). All manual model building was performed with Coot (Emsley and Cowtan, 2004). A procedure similar to that used for tracing electron density maps in crystallography was used for identifying the PIM. A tentative sequence for the segment was assigned based on the sizes of the side-chain density. This tentative sequence was then compared, in both directions, against all sequence segments from all the proteins in the sample. A score was calculated for each comparison based on the sizes of the side chains being compared, using a locally produced program. The segment in CPSF100 produced a score that was 1.1σ above the others. The atomic models were refined using phenix.real_space_refine (Adams et al., 2002). The statistics from the structure determination is summarized in Table 1.

Replication

N/A

Strategy for randomization

N/A

Blinding

N/A

Sample size estimation and statistical method of computation

N/A

Inclusion and exclusion criteria

N/A

QUANTIFICATION AND STATISTICAL ANALYSIS

N/A

DATA AND CODE AVAILABILITY

The PDB accession number for the coordinates of mPSF in complex with CPSF100 PIM is 6URG, and that mPSF in complex with CstF77 is 6URO. The EMDB accession number for the mCF reconstruction is 20859. The raw SDS gel images are available at https://dx.doi.org/10.17632/f3v9w58m3j.1.