Supplementary Materials for

Structure of an active human histone pre-mRNA 3′-end processing machinery

Yadong Sun,* Yixiao Zhang,* Wei Shen Aik, Xiao-Cui Yang, William F. Marzluff, Thomas Walz,#
Zbigniew Dominski# & Liang Tong#

*These authors contributed equally to this work.
#correspondence to: ltong@columbia.edu, zbigniew_dominski@med.unc.edu,
twalz@mail.rockefeller.edu

This PDF file includes:

Other Supplementary Materials for this manuscript includes the following:

Movies S1 to S2

Materials and Methods

Expression and purification of Histone Cleavage Complex (HCC)

Human CPSF73, CPSF100, Symplekin (residues 353-1110, 30-1160 or 30-1101) and CstF64 were co-expressed in insect cells using Multibac technology (*31*) (Geneva Biotech). CPSF73 and CPSF100 were cloned into the pFL acceptor vector. CstF64 was cloned into the pSPL donor vector. These two vectors were fused together by Cre recombinase. N-terminal 6×His-SUMO-tagged symplekin was cloned into another pFL vector. One liter of High5 cells (1.8×10^6 cells/ml) cultured in ESF 921 medium (Expression Systems) was co-infected with 12 ml CPSF73-CPSF100-CstF64 P2 virus and 12 ml symplekin P2 virus at 27 °C with constant shaking. Cells were harvested after 48 h by centrifugation at 2,000 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). Fractions of interest were concentrated to 1 mg/ml, and used for binding assay with U7 snRNP-FLASH-SLBP-H2a* RNA complex.

Expression and purification of Lsm11-Lsm10 complex

Human Lsm11 (with an internal deletion of residues 211-322) and Lsm10 were cloned into the same pFL vector and co-expressed in insect cells. A 6×His tag was added to the N terminus of Lsm11 and a maltose binding protein (MBP) was added to the N terminus of Lsm10, separated by a TEV protease cleavage site. One liter of High5 cells (1.8×10^6 cells/ml) was infected with 25 ml Lsm11-Lsm10 P2 virus.

For purification, the cell pellet was re-suspended and lysed by sonication in 100 ml buffer containing 20 mM Tris (pH 7.5), 500 mM NaCl, 5% (v/v) glycerol 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 (20 mM Tris (pH 7.5), 500 mM NaCl and 40 mM imidazole) and eluted with 20 mM Tris (pH 7.5), 500 mM NaCl, 500 mM imidazole and 5% (v/v) glycerol. The eluate was diluted 2.5 times with buffer containing 20 mM Hepes (pH 7.5) and 5 mM DTT. The protein was further purified by chromatography using a HiTrap Heparin column (GE Healthcare). Fractions of interest were concentrated to 3.4 mg/ml, and stored at –80 °C.

Expression and purification of SmD3-SmB complex

Human SmD3 and SmB (residues 1-95) were cloned into the pET28a (Novagen) and pCDFDuet vector, respectively. A 6×His tag was added to the N terminus of SmD3. Both genes were over-expressed in *E. coli* BL21 (DE3) Star strain (Novagen).

For purification, the cell pellet was re-suspended and lysed by sonication in 100 ml buffer containing 20 mM Tris (pH 7.5), 500 mM NaCl, 10 mM imidazole, 5% (v/v) glycerol, 17.8 μg/mL PMSF, 10 mM β-mercaptoethanol. The protein was purified using nickel affinity and heparin affinity, as described above. The gradient is starting with 40% buffer B to 100% buffer B (20 mM Hepes (pH 7.5), 1 M NaCl, 5 mM DTT). Buffer A contained 20 mM Hepes (pH 7.5) and 5 mM DTT. Fractions of interest were concentrated to 4.4 mg/ml, and stored at –80 °C.

Expression and purification of SmF-SmE-SmG complex

SmE and SmF were cloned into pCDFDuet MCS1 and MCS2, respectively. SmG was cloned into pET26b. A 6×His tag was added to the C terminus of SmG. All three proteins were co-expressed in *E. coli* BL21 Star (DE3) strain (Novagen) and purified as described above for the SmD3-SmB complex. Fractions of interest were concentrated to 9.5 mg/ml, and stored at –80 °C.

Expression and purification of SLBP

SLBP (residues 125-270) was cloned into the pFL vector. A 6×His tag was added to the N terminus. The construct was expressed in Hi5 cell and purified using nickel affinity and anion exchange, as described above. Fractions of interest were concentrated to 1.8 mg/ml, and stored at –80 °C.

Expression and purification of FLASH, symplekin NTD and Ssu72

FLASH (residues 51-137) double cysteine mutant C54S/C83A was expressed and purified as described previously (*14*). These two cysteine residues are located near the interface of the FLASH coiled-coil dimer. They were mutated to avoid the fortuitous formation of disulfide bonds through oxidation, which could introduce artifacts in the assays. Symplekin NTD (residues 30-360) and Ssu72 were expressed and purified following published protocols (*16*).

RNAs

Human U7 snRNA (nucleotides 2-61) and mouse H2a pre-mRNA (52 nucleotides, modified so that it can form 15 consecutive base pairs with the U7 snRNA) were purchased from IDT. The sequences of the RNAs are shown in fig. S1. The synthetic pre-mRNA (Dharmacon) used in cleavage assays also carries a biotin label at the 3′ end.

In vitro reconstitution of histone pre-mRNA processing machinery for EM studies

An equimolar mixture of human U7 snRNA and modified mouse H2a pre-mRNA (H2a*) was annealed in 100 μl reconstitution buffer A containing 20 mM Hepes (pH 7.5), 500 mM NaCl, 5 mM EDTA, and 5 mM DTT by heating to 90 °C for 5 min and snap-cooling on ice for 10 min. The annealed RNA was added to a ~700 μl mixture in reconstitution buffer A containing equimolar Lsm11-Lsm10, SmD3-SmB, SmG-SmE-SmF complex and 2 molar equivalents of FLASH. The mixture was incubated at 30 °C for 30 min, followed by 37 °C for 15 min (*32*) and then cooled on ice for 10 min. Purified SLBP was then added to the mixture and incubated on ice for another 5 min. After U7 snRNP-FLASH-SLBP-H2a* complex formation, the MBP tag on Lsm10 was removed by TEV protease at a ratio of 1:3 (w/w). The reaction mixture was incubated overnight at 4 °C and then purified by gel filtration using a Superose 6 10/300 GL column (GE Healthcare), in reconstitution buffer B containing 20 mM Hepes (pH 7.5), 100 mM NaCl, 10 mM EDTA, and 5 mM DTT. Fractions of interest were concentrated to 2.7 mg/ml, and store at –80 °C. The presence of RNA in the complex was confirmed with an A260/A280 ratio of ~1.7.

The histone pre-mRNA processing complex was formed by mixing purified HCC and U7 snRNP-FLASH-SLBP-H2a* complex at a molar ratio of 1:1.3. The reaction 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 reconstitution buffer B. Fractions of interest were used for EM studies.

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 (*33*). Before preparing grids for cryo-EM, the freshly purified protein sample was

centrifuged at 13,000 g for 2 min to remove potential protein aggregates, and the protein concentration was measured with a NanoDrop spectrophotometer (Thermos Fisher Scientific). The protein sample was kept on ice before grid preparation. A 3.5 μL aliquot at a concentration of 0.3 mg/ml was applied to a glow-discharged Quantifoil 400 mesh 1.2/1.3 gold grid (Quantifoil) that had been glow-discharged for 20 s at 40 mA with EMS 100X glow discharger. After 5 s, the grid was blotted for 4 s with 55/20mm filter paper (TED PELLA) at a blot force setting of −2 and plunged into cooled liquid ethane with a Vitrobot Mark VI (FEI) set at 22 °C and 100% humidity.

Two datasets (Data-Krios1 and Data-Krio3) 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 (*34*). 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 Data-Krios1, and 1.045 Å for Data-Krios3) and a defocus range from −1.2 to −2.5 μm. Exposures of 10 s were dose-fractionated into 50 frames (200 ms per frame), with an exposure rate of 8 electrons·pixel^{-1}·s^{-1}, resulting in a total exposure of 70 electrons·Å^{-2} for Data-Krios1, and 73 electrons·Å^{-2} for Data-Krios3.

Image processing

The image stacks were motion-corrected and dose-weighted in MotionCor2 (*35*). The CTF parameters were determined with CTFFIND4 (*36*). 1,517,598 particles were automatically picked from 3,363 micrographs for Data-Krios1 and 1,641,327 particles were automatically picked from 3,740 micrographs for Data-Krios3 with Gautomatch (www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/). The Data-Krios3 particles were windowed into 260×260-pixel images, rescaled to 130×130-pixel images, and subjected to 2D classification in RELION-3 (*37*). Particles in classes that generated averages showing very clear structural features were selected (130,245 particles) and subjected to 3D initial model building in RELION-3.

To combine the two datasets for 3D classification, Data-Krios1 particles were extracted into 334×334-pixel images, and Data-Krios3 particles were extracted into 342×342-pixel images and rescaled to 334×334-pixel images (342*1.045 Å / 334=1.07003 Å) within RELION particle extraction. 2D classification was skipped for the combined dataset and the initial model from Data-Krios3 was used as reference in 3D classification for 10 classes with the bin 2 particles. The 399,507 particles in the class that showed clear structural features were selected, re-centered, refined, and subjected to CtfRefinement and BayesianPolishing. The polished particles were subjected to another round of 3D classification into eight classes. The particles in the classes showing clear fine structural features for the core were combined (325,282 particles) and refined, yielding a density map at 3.2 Å resolution.

The particles in the class showing density for symplekin CTD and FLASH were selected (31,312 particles) and subjected to a masked classification focusing on symplekin. The particles in the class showing better symplekin density were selected (19,315 particles) and refined, yielding a density map at an overall resolution of 4.1 Å resolution. The masked refinement for CPSF73-CPSF100-N-terminal segment of symplekin CTD and symplekin CTD alone yielded maps at 4.1 Å and 4.7 Å resolution, respectively (fig. S4). Fourier shell correlation curves, local resolution maps, and resolution-filtered maps were calculated in RELION-3 (fig. S5).

Model building and refinement

For the core of the machinery, we used the crystal structures of human CPSF73 (PDB ID 2I7V) (*6*), yeast CPSF100 (PDB ID 2I7X) (*6*), spliceosomal U4 snRNP core domain (PDB ID 4WZJ) (*18*) and the N-terminal domain of human symplekin (PDB ID 3O2Q) (*16*) as starting models and fitted them into the

cryo-EM density map with Chimera using the "fit in map" tool (*30*). We then calculated predicted structures of human CPSF100, Lsm10 and Lsm11 from I-TASSER (*38*), and superimposed them onto yeast CPSF100, and SmD1 and SmD2 in U4 snRNP with Chimera. The metallo-β-lactamase domain and β-CASP domain of CPSF100 and the Sm cores of Lsm10 and Lsm11 fitted well into the EM density overall. The first CTDs of CPSF73 and CPSF100, and the N- and C-terminal extensions of Lsm10 and Lsm11 were manually built. The entire model was examined and adjusted to fit the EM density manually. All manual model building was performed with Coot (*39*), including torsion restraints, planar peptide restraints, trans peptide restraints and Ramachandran restraints. The atomic model for the core was refined using phenix.real_space_refine global minimization (default), morphing and simulated annealing rama potential (*40*).

In the active site of CPSF73, EM density for the two zinc ions was observed in the reconstruction. The crystal structure of human CPSF73 alone (PDB ID 2I7V) (*6*) was used to guide the model building for the active site region. In comparison, no EM density for metal ions was observed in the equivalent region of CPSF100.

For the entire machinery, the symplekin CTD was built as a collection of poly-alanine helices (with no sequence assignment due to the lack of sufficient side chain density) by using the 4.7 Å symplekin CTD masked refinement map in Coot. The core of HCC was modeled using the 4.1 Å HCC core masked refinement map, which suggested that the overall structure of this region was similar to that of their homologs IntS11 and IntS9 in the integrator complex. We fitted the IntS11-IntS9 CTD complex (PDB ID 5V8W) (*28*) into the EM density as CTD2 complex of CPSF73-CPSF100, without any further manual adjustment or refinement. FLASH (PDB ID 6AOZ) (*28*) and SLBP (PDB ID 4L8R) (*28*) were fitted into entire machinery map filtered to local resolution. EM density for the two helices in the FLASH dimer was clearly recognizable (Fig. 1E, fig. S5H). The directionality of the helices was assigned based on the biochemical observation that the LDLY motif just prior to the N-terminal end of the coiled-coil domain meditates HCC recruitment. This assignment is supported by mutagenesis data showing that removing the LDLY motif of FLASH or the second half of the symplekin CTD abolished HCC recruitment (fig. S13A).

Then we combined the models of the machinery core, CPSF73 and CPSF100 CTD2, symplekin CTD, FLASH, and SLBP by fitting all these maps together. The statistics from the structure determination is summarized in Tables S2 and S3.

Sequence alignment

Multiple alignment of amino acid sequences was produced with Clustal (*41*) and rendered with ESPript (*42*). Additional annotations were introduced manually.

In vitro assembly of U7 snRNP for functional studies

U7 snRNP used to test processing was assembled in a high salt buffer (15 mM Hepes (pH 7.9), 600 mM KCl, 15% (v/v) glycerol, 0.25 μg/μl yeast tRNA and 20 mM EDTA (pH 8)) and purified by gel filtration chromatography using Superose 6 Increase 3.2/300 column (GE Healthcare) and a running buffer containing 15 mM Hepes (pH 7.9), 75 mM KCl, 15% (v/v) glycerol and 20 mM EDTA (pH 8), as described (*13*).

In vitro processing

Processing reactions were reconstituted in a final volume of 10 μl containing 15 mM Hepes (pH 7.9), 75 mM KCl, 15% (v/v) glycerol, 20 mM EDTA (pH 8), yeast tRNA at 0.1 μg/μl and the following recombinant components: purified U7 snRNP, C54S/C83A FLASH mutant (amino acid residues 51-

137), HCC, SLBP (amino acid residues 125-223), and the N-terminal domain (NTD) of symplekin, when indicated. With the exception of FLASH, which was used at the final concentration of 0.5 μM, the remaining components were at 0.25 μM, unless otherwise indicated. Each processing reaction contained 0.5 ng of histone pre-mRNA (final concentration 2.5 nM), labeled at the 5′ end with ^{32}P and with a biotin at the 3′ end (*13*), and was incubated for 60 min at 32 °C. Radioactive RNA was recovered and analyzed by gel electrophoresis and autoradiography, as described (*43*).

Pull-down of U7 snRNP on streptavidin beads

U7 snRNP was assembled on U7 snRNA containing biotin at the 5′ end, and bound to C54S/C83A FLASH mutant (amino acid residues 51-137), as described (*13*). The U7 snRNP-FLASH complex (~100 pmol) was diluted in 1 ml of a buffer containing 15 mM Hepes (pH 7.9), 75 mM KCl, 15% (v/v) glycerol and 20 mM EDTA (pH 8), spun down in a microcentrifuge and the supernatant rotated with 25 μl of streptavidin beads for 60 min at 4 °C. The beads with the bound complex were exhaustively washed, resuspended in 1 ml of the same buffer containing 25 pmol of the HCC and 150 pmol of symplekin NTD and rotated for 60 min at 4 °C. The beads were exhaustively washed and the bound proteins separated in a 4-12% SDS/polyacrylamide gel and analyzed by silver staining and Western blotting.

Supplementary Text

Histone pre-mRNA 3′-end processing

The U7 snRNP plays a critical role in this processing (*44, 45*) (Fig. 1A). The heptameric Sm ring of the U7 snRNP contains two unique subunits, Lsm10 and Lsm11, replacing the SmD1 and SmD2 subunits of the spliceosomal Sm rings (*46, 47*). In vertebrates, Lsm11 contains a 150-residue N-terminal extension (Figs. 1A,B), and residues 20-65 interact with a coiled-coil dimer of the N-terminal region of FLASH (*9-11, 14*), and this Lsm11-FLASH complex recruits the HCC to the machinery (*9-12*). The symplekin component of HCC was previously identified as a heat labile factor required for histone pre-mRNA processing (*48*).

Lsm10 has small extensions at both termini (Fig. 1B), but it is not known whether they have any specific functions in processing. The RNA motif (Sm site) recognized by the U7 Sm ring is distinct from those in spliceosomal snRNAs, with a CUAG sequence at the 3′ end (fig. S1), which is also important for U7 snRNP assembly (*49, 50*). The stem-loop binding protein (SLBP) interacts with the 5′ arm of the stem-loop in the pre-mRNA (*15, 51*). It promotes the stable binding of the U7 snRNP to the pre-mRNA and it may also contact Lsm11-FLASH (*52*).

Cleavage activity requires the symplekin NTD

Our initial HCC sample contained full-length CPSF73, CPSF100 and CstF64, and a segment of symplekin (residues 353-1110) lacking its NTD, which interacts with the RNA polymerase II CTD phosphatase Ssu72 (*16, 53*). We refer to this complex as HCC(ΔNTD). The symplekin construct containing the NTD (residues 30-1160) was expressed at a much lower level. These constructs lacked the C-terminal segment of symplekin (up to residue 1274), which is poorly conserved and expected to be disordered. Earlier studies showed that optimized pre-mRNAs such as H2a* are efficiently processed without SLBP in mammalian nuclear extracts (*54, 55*).

However, we did not observe any cleavage activity when we incubated H2a* with reconstituted U7 snRNP, HCC(ΔNTD), FLASH (residues 51-137, C54S/C83A double mutant) (*14*), and SLBP (residues 125-223) (fig. S2A). Noting that this reaction was missing the NTD of symplekin, we then included purified NTD in *trans*, and remarkably observed robust cleavage activity (fig. S2A). The cleavage product had the precise size (26 nucleotides) as that generated by a nuclear extract (fig. S2B). Processing was inhibited by an anti-U7 oligonucleotide which binds the 5′ end of U7 snRNA. Besides the NTD, processing activity required U7 snRNP and the HCC, and was reduced in the absence of SLBP, reduced to a greater extent in the absence of FLASH, and essentially abolished in the absence of both SLBP and FLASH.

The NTD alone does not stably associate with the processing machinery, based on our pull-down (fig. S2C) and gel filtration (fig. S2D) experiments. Therefore, the NTD only transiently associates with the machinery when it is supplied in *trans*, but it is sufficient to activate processing.

To include the symplekin NTD in *cis* in the reaction, we examined additional expression constructs of symplekin, and found that the one encoding residues 30-1101 produced an HCC with sufficient yield. We observed cleavage activity that was no longer dependent on the NTD in *trans* using this HCC (fig. S3A). A mutation in the active site of CPSF73 (D75N/H76A, removing two of the zinc ligands (*6*)) abolished the activity (figs. S2E,S3A), consistent with CPSF73 being the endoribonuclease for the cleavage reaction. This reconstituted machinery is also active toward the wild-type H2a substrate (fig. S3B).

We next showed that Ssu72 potently inhibited the cleavage activity when the NTD was present in *trans* (fig. S3C). The K185A mutant of the NTD, which does not interact with Ssu72 (*16*), supported cleavage but was insensitive to Ssu72 inhibition (fig. S3C). The inhibitory activity of Ssu72 on reactions

where the NTD was present in *cis*, either in purified HCC (fig. S3D) or in a nuclear extract (fig. S3E), was substantially weaker, consistent with the low affinity of the NTD alone for the machinery. The Ssu72 triple mutant (T190A/V191A/F193A) that cannot interact with the NTD (*16*) had essentially no effect on the cleavage activity (fig. S3D). Overall, our biochemical studies demonstrate that we have successfully reconstituted an active histone pre-mRNA 3′-end processing machinery, and that the symplekin NTD is required for cleavage.

Roles of Ssu72 in histone and canonical pre-mRNA processing

Earlier data suggested that Ssu72 has a negative role in histone pre-mRNA processing in chicken DT40 cells (*56*). Removal of the NTD from symplekin also substantially increases misprocessing of histone pre-mRNAs in *Drosophila* tissue culture cells (*27*). On the other hand, the NTD may function in a different manner in canonical pre-mRNA processing, as Ssu72 has a positive role there (*56, 57*). The NTD of the symplekin homolog Pta1 in *trans* inhibits cleavage in yeast nuclear extracts (*53*) and the symplekin NTD inhibits transcription-coupled polyadenylation in human nuclear extracts (*16*). Pta1 was not present in the reconstituted active yeast canonical processing machinery (*24*).

Therefore, there is an interesting difference between the canonical and histone processing machineries in terms of the functional roles of the symplekin NTD and its binding partner Ssu72. Further studies are required to understand whether the opposite functions of these proteins in the two machineries can differentially affect expression of canonical and replication-dependent histone mRNAs and coupling between transcription, termination and processing.

The active site of Ssu72 is far away from the interface with symplekin NTD (fig. S6D) (*16*). Its inhibitory effect on histone processing is due to steric hindrance with the HDE-U7 duplex and CPSF73, and is unlikely dependent on the catalytic activity of Ssu72. Interestingly, the catalytic activity of Ssu72 is not required for canonical pre-mRNA processing in yeast (*58*).

CstF64 is not required for histone pre-mRNA 3′-end processing

We did not observe any density for the CstF64 subunit of HCC in our cryo-EM reconstructions, likely due to disorder. The symplekin segment between the NTD and the CTD, where the CstF64 binding site is located (*59*) (Fig. 1B), is disordered as well. Our biochemical data showed that leaving out CstF64 in the reaction had minimal effect, if any, on the cleavage (fig. S2F), demonstrating that CstF64 is not required for cleavage under the condition tested.

Recruitment of HCC by FLASH and Lsm11

Residues in FLASH containing the LDLY motif (*10*) contact the second half of the symplekin CTD (Fig. 1E). Strikingly, this essential interaction for HCC recruitment occurs at the far periphery of the machinery, more than 120 Å away from the active site of CPSF73. Residues 107-118 of Lsm11 tether CPSF73 of the HCC (Fig. 1D, figs. S11D,E). Overall, these observations explain earlier data showing the importance of the LDLY motif in FLASH and residues 65-130 in Lsm11 for HCC recruitment (*10, 11, 29*).

Two other highly conserved segments in the N-terminal extension of Lsm11 (residues 20-65 and 139-150, fig. S10B) also mediate important contacts in the machinery (Fig. 1E, fig. S11F). Residues 20-65 bind tightly to FLASH (*9*), although we did not observe these residues due to the limited resolution in that region of the machinery. The structure suggests the presence of an Lsm11(20-65)-FLASH-SLBP-SL quaternary complex (Fig. 1E), and the 2:1 FLASH:Lsm11 stoichiometry in the machinery is consistent with our previous biophysical data (*14*).

Implications for histone pre-mRNA processing

The structure shows that the segment of the pre-mRNA from the stem-loop to the cleavage site is mostly disordered and the stem-loop has no direct contact with the core of the machinery. In contrast, the backbone for the segment from the cleavage site to the HDE-U7 duplex is well ordered, explaining earlier biochemical data that the cleavage site is defined by the distance from the HDE-U7 duplex rather than the stem-loop (*60*). Moreover, the binding site for the duplex can accommodate longer duplexes because it is open at the top (Fig. 1D), explaining the 'length suppression' effect observed earlier (*13, 61*).

CPSF100 is a weak sequence homolog of CPSF73, and its role in pre-mRNA 3′-end processing has been a mystery (*62*). The structure shows that it serves several important functions in the histone processing machinery. Its CTD interacts with the CTDs of both CPSF73 and symplekin, ensuring the formation of the HCC (Fig. 4B). In addition, its β-CASP domain contacts the RNA duplex (Fig. 2A) and symplekin NTD, and its metallo-β-lactamase and β-CASP domains have extensive contacts with the equivalent domains in CPSF73, forming a pseudo dimer (figs. S11B,C).

Putative histone pre-mRNA processing cycle

The putative histone pre-mRNA processing cycle (Fig. 4C) is consistent with the structural information and biochemical data, although currently there is no direct experimental evidence for all the steps. A large fraction of U7 snRNP is in complex with FLASH and the HCC (without the pre-mRNA, state III in Fig. 4C) in human and *Drosophila* nuclear extracts, suggesting that this complex is stable and that it may be the basal form of the machinery in cells during S phase. In addition, in the presence of FLASH, HCC in nuclear extracts is recruited to U7 snRNP spontaneously (*13*), and FLASH is also required for stable association of the U7 snRNP to histone pre-mRNA by SLBP (*12*). Therefore, FLASH and HCC binding to U7 snRNP may precede pre-mRNA binding to the machinery, although we cannot exclude the possibility that these events happen concomitantly *in vivo*. Symplekin is present in the histone locus body (HLB) in *Drosophila* cells in S phase but not G1 phase, supporting the idea that the HCC is bound to U7 snRNP at the site of histone gene transcription (*63*). In comparison, U7 snRNP and FLASH are both present in the HLB at all cell-cycle stages. U7 snRNP is in an inactive form in mammalian cells not expressing histone mRNA due to the loss of a heat-labile factor (*64*), which was subsequently identified as symplekin (*48*).

After the binding of the pre-mRNA, it is likely that the recognition of the HDE-U7 duplex by symplekin NTD and CPSF100 leads to the activation of CPSF73, also aided by Lsm10 and Lsm11, and the pre-mRNA substrate is then presented to its active site and be cleaved. In comparison, the U7 snRNP-HCC complex (state III) is not by itself sufficient to activate CPSF73. Therefore, the machinery can only be activated by authentic pre-mRNAs, thus preventing unwanted cleavage activity on other RNAs. Such a mechanism also explains why the symplekin NTD is required for cleavage, as it is a major binding site for the HDE-U7 duplex. It is likely that the symplekin NTD stabilizes the bound position of the HDE-U7 duplex and the active conformation of HCC, facilitating the contacts between CPSF73 and CPSF100.

After pre-mRNA cleavage, the downstream cleavage product (DCP) of the pre-mRNA is degraded in the 5′-3′ direction, releasing the U7 snRNP for the next round of cleavage reaction (*5, 13*). While SLBP leaves the machinery with the mature mRNA, the rest of the machinery likely remains intact (state V), although some rearrangements of its subunits may occur as a result of cleavage. CPSF73 is required for this degradation, and the binding mode of the RNA substrate in RNase J (fig. S9E) may be consistent with a 5′-3′ exonuclease activity. The first four nucleotides of the DCP (27-30 in the current structure) are bound by the active site of CPSF73. The unwinding of the HDE-U7 duplex would allow

the DCP to move across the active site and be degraded, although the exact molecular mechanism of this activity will require further studies. The 5′-3′ exonuclease Xrn2 is not required for transcription termination of histone genes (*65, 66*), which occurs very close to the processing site both in mammals (*66*) and flies (*67*), and the site of termination is not affected by removal of Xrn2. Therefore, Xrn2 is unlikely involved in DCP degradation.

After degradation of the DCP, it is likely that the machinery is directly recycled for the next cycle of processing (solid arrow in Fig. 4C), which is supported by the presence of U7 snRNP-HCC complex in nuclear extracts. We cannot completely exclude the possibility that the machinery is disassembled and then re-assembled (dashed arrows in Fig. 4C). No exchange factors have been identified for the disassembly of the machinery.

The histone processing machinery is highly dynamic, due to the flexible linkers in the N-terminal extension of Lsm11 (Fig. 4C, state II) as well as the flexibility within the HCC/mCF itself (Fig. 4A, state III). Flexibility is also observed for the human (*25*) and yeast (*24*) canonical machineries, indicating that this may be a common feature for pre-mRNA processing. In the presence of the pre-mRNA, the recognition of the HDE-U7 duplex (by symplekin NTD, CPSF100 and CPSF73) and the SL-SLBP complex (by FLASH and Lsm11) helps to bring about rearrangements in the machinery core, so that it can catalyze the cleavage reaction. At the same time, even in this active state (IV), there is still flexibility for the periphery of the machinery, as indicated by the low-resolution nature of the reconstruction for symplekin CTD, FLASH, SL and SLBP (Fig. 1E, fig. S5), and the disorder of the symplekin NTD-CTD linker and its binding partner CstF64.

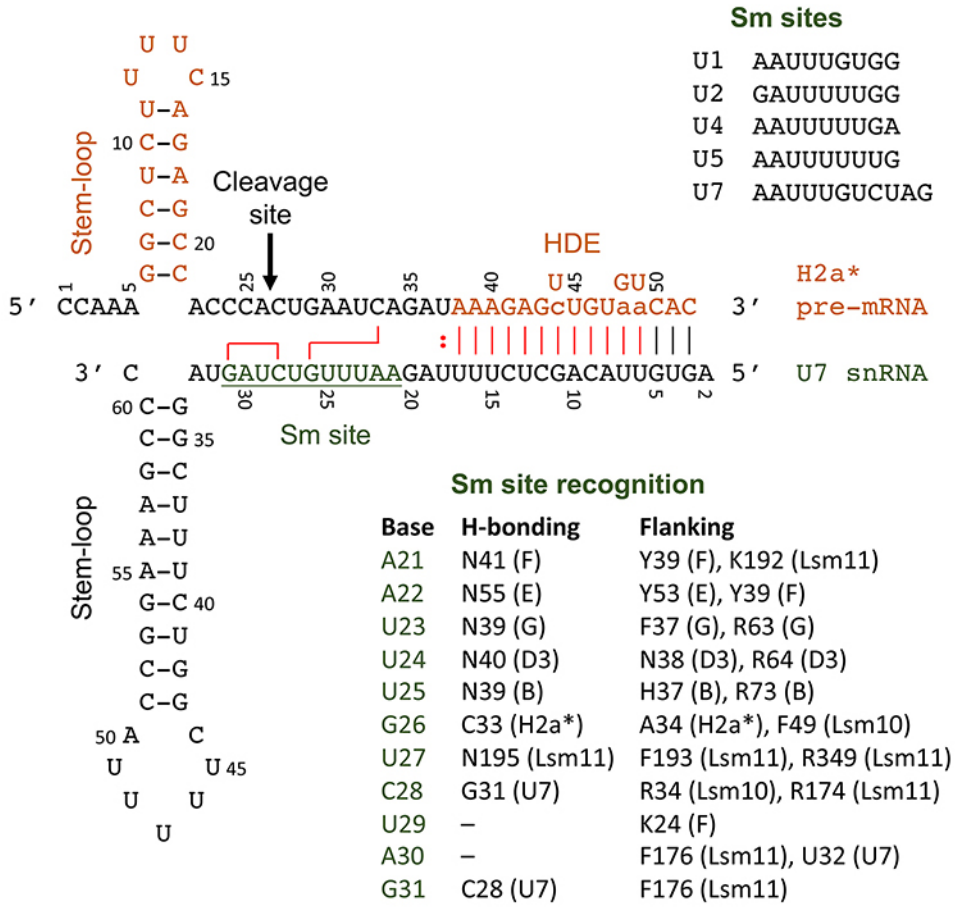

Fig. S1. Histone pre-mRNA 3′-end processing machinery.

Nucleotide sequences of the U7 snRNA and the mouse H2a pre-mRNA are shown. The HDE of the pre-mRNA was mutated so that it can form 15 consecutive Watson-Crick base pairs with the U7 snRNA, and this modified pre-mRNA is referred to as H2a*. The stem-loop and HDE are shown in orange, and the mutated nucleotides in the HDE are shown in lower case. The corresponding nucleotides in wild-type H2a are shown in upper case. The Watson-Crick base pairs are indicated with the lines, and the U-U base pair by a colon. The base pairs observed in the structure are shown in red. The Sm site is shown in green and underlined. The numbering schemes for the nucleotides in the two RNAs are indicated. Also shown are Sm site sequences of U1, U2, U4, U5 and U7 snRNAs, as well as the recognition of the U7 Sm site by subunits of the U7 Sm ring. Residues that hydrogen bond or flank the bases are indicated.

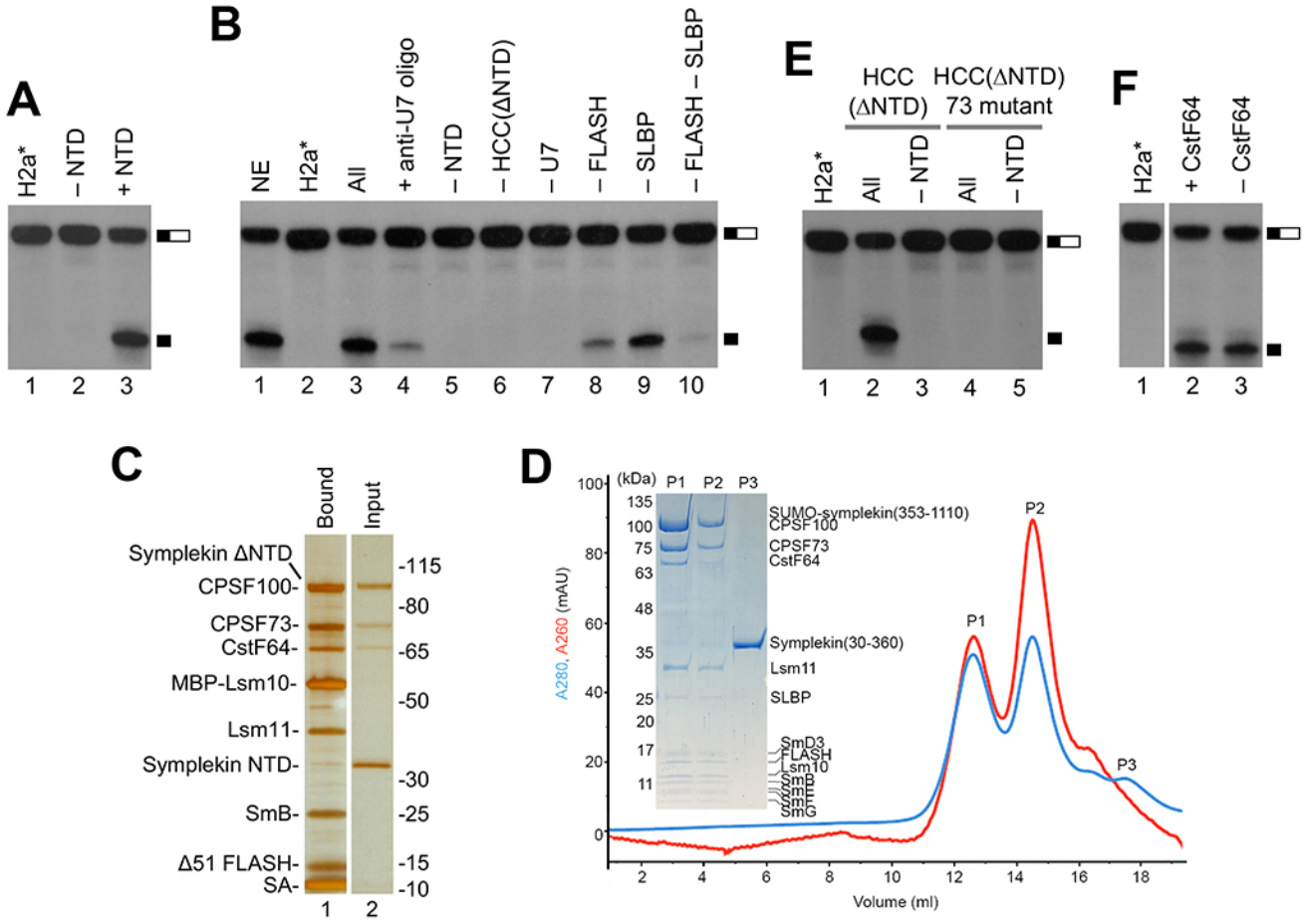

Fig. S2. Symplekin NTD is required for histone pre-mRNA 3′-end processing.

(**A**) The H2a* substrate has 64 nucleotides and is indicated with the black and white bar. The 5′ authentic cleavage product has 26 nucleotides and is indicated with the black bar. The HCC(ΔNTD) sample was used in these reactions, and the NTD was supplied in *trans*. (**B**) The reconstituted machinery generates a product of the same size as that from a nuclear extract (NE) (lane 1). The reaction in lane 3 contained all the recombinant factors. The HCC(ΔNTD) sample was used in these reactions, and the NTD was supplied in *trans*. (**C**) Symplekin NTD is not pulled down with the machinery using a biotin tag at the 3′ end of the U7 snRNA (lane 1). The NTD is at five-fold molar ratio to the HCC in the incubation mixture (lane 2). (**D**) Symplekin NTD (residues 30-360) does not co-migrate with the machinery containing HCC(ΔNTD) on a gel filtration column. Peak P2 lacks CstF64 and has lower content of the other subunits of the HCC. (**E**) The D75N/H76A double mutant in the active site of CPSF73 abolishes the processing activity when symplekin NTD is supplied in *trans*. (**F**) CstF64 is not required for cleavage activity. The HCC sample with CstF64 contained symplekin residues 353-1110, while that without CstF64 contained symplekin residues 538-1110. Symplekin NTD was supplied in *trans*.

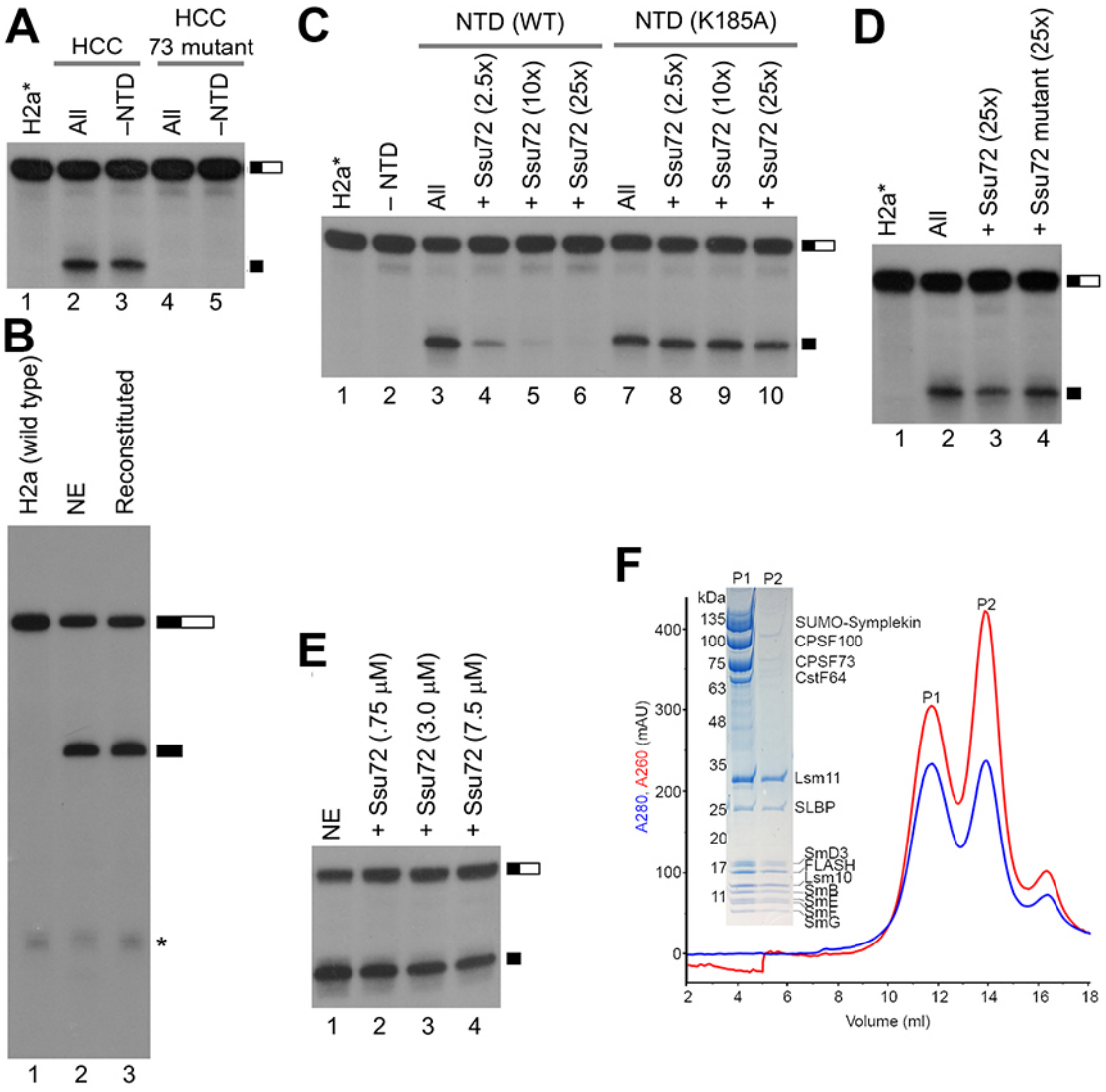

Fig. S3. Reconstitution of an active human histone pre-mRNA 3′-end processing machinery.

(**A**) The cleavage activity with an HCC containing symplekin NTD (residues 30-1101) is independent of the NTD supplied in *trans*, and is abolished by the D75N/H76A mutation in CPSF73. Unless indicated (–NTD), the NTD was also present *in trans*. (**B**) The reconstituted machinery produced only a single labeled product with wild-type H2a pre-mRNA, the same one as that generated by a nuclear extract (NE), indicating that full base pairing in the HDE-U7 duplex is not required for activity by the reconstituted machinery either. The symplekin NTD was present in *cis*. An impurity in the pre-mRNA is indicated with the asterisk. This substrate has 107 nucleotides, and the labeled product has 47 nucleotides. (**C**) Ssu72 is a potent inhibitor of the cleavage activity when the NTD is supplied in *trans*. The concentration of NTD in the assays was 0.3 µM. The K185A NTD mutant that does not interact with Ssu72 (*16*) is essentially insensitive to its inhibition, but supports normal cleavage activity. (**D**) Ssu72 shows weaker inhibition of H2a* processing when symplekin NTD is present in *cis*. The symplekin protein contains residues 30-1101. The triple mutant of Ssu72 that cannot interact with the NTD (*16*) shows no inhibition. Ssu72 was added after the formation of the HCC-U7 snRNP complex. It

could not compete as efficiently when the symplekin NTD is in *cis*, consistent with the lower affinity of the NTD in *trans* for the machinery. (**E**) Ssu72 shows weak inhibition of H2a* processing by a nuclear extract (NE) that contains full-length symplekin. (**F**) Purification of the reconstituted human histone pre-mRNA processing machinery by gel filtration. The stronger signal at 260 nm is due to the presence of the RNAs. Inset: SDS gel of the two peaks. The second peak lacks the HCC.

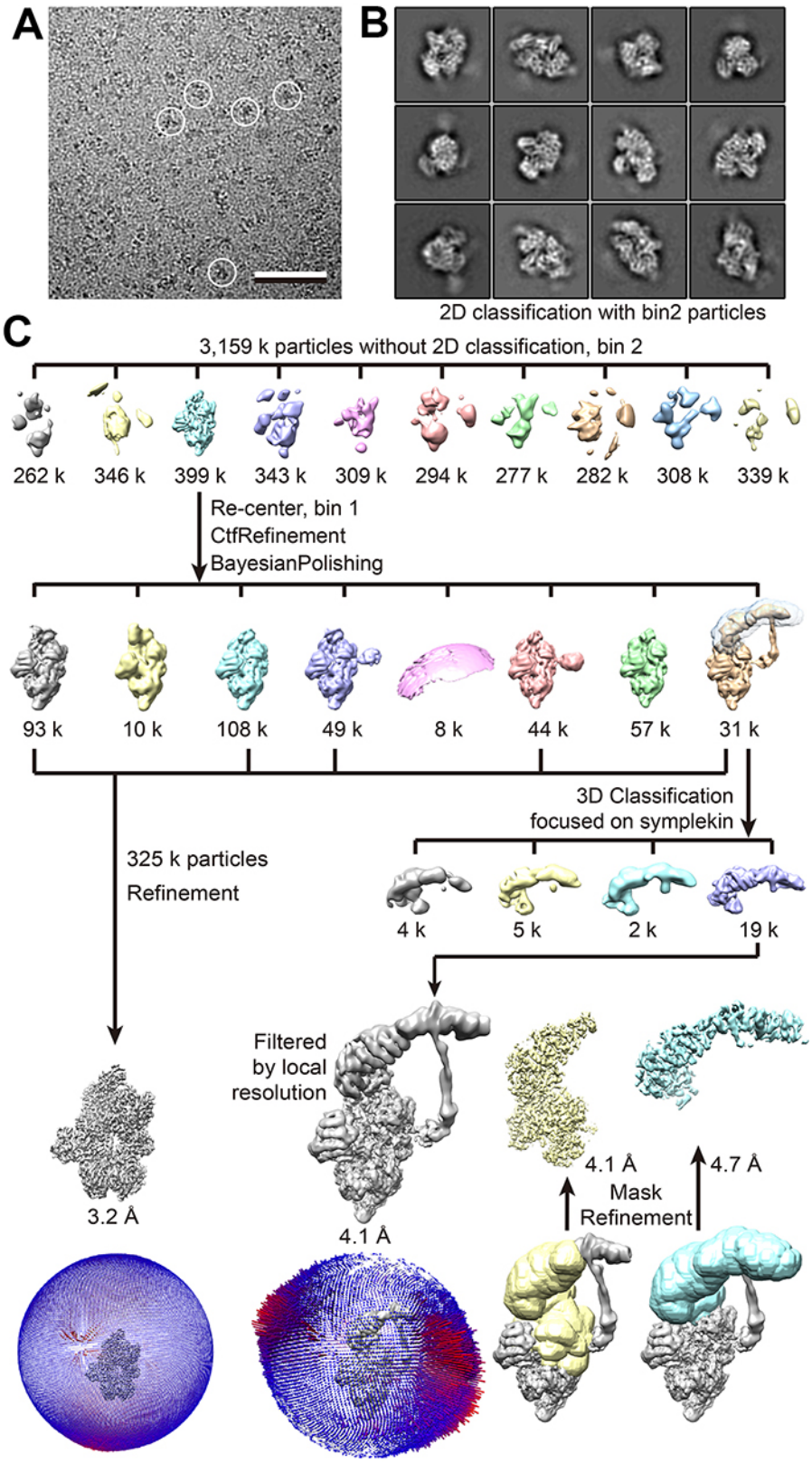
A
B
2D classification with bin2 particles
C
3,159 k particles without 2D classification, bin 2
262 k
346 k
399 k
343 k
309 k
294 k
277 k
282 k
308 k
339 k
Re-center, bin 1
CtfRefinement
BayesianPolishing
93 k
10 k
108 k
49 k
8 k
44 k
57 k
31 k
3D Classification
focused on symplekin
325 k particles
Refinement
4 k
5 k
2 k
19 k
Filtered
by local
resolution
4.1 Å
4.7 Å
Mask
Refinement
3.2 Å
4.1 Å

Fig. S4. Single-particle cryo-EM analysis of the histone pre-mRNA processing machinery.
(**A**) Area of a cryo-EM image of the vitrified machinery. Some particles are circled. Scale bar: 50 nm. (**B**) Selected 2D class averages obtained with RELION-3. The smeared "appendages" are mostly due to SLBP, which is highly flexible among the particles. Side length of individual averages: 27 nm. (**C**) Image-processing workflow for 3D classification and refinement in RELION-3. For the entire machinery, masks were applied during refinement to improve the local resolution for CPSF73-CPSF100-N-terminal segment of symplekin CTD (yellow) and symplekin CTD alone (cyan). Angular distribution of the particles is shown for the core and entire machinery. See Methods for details.

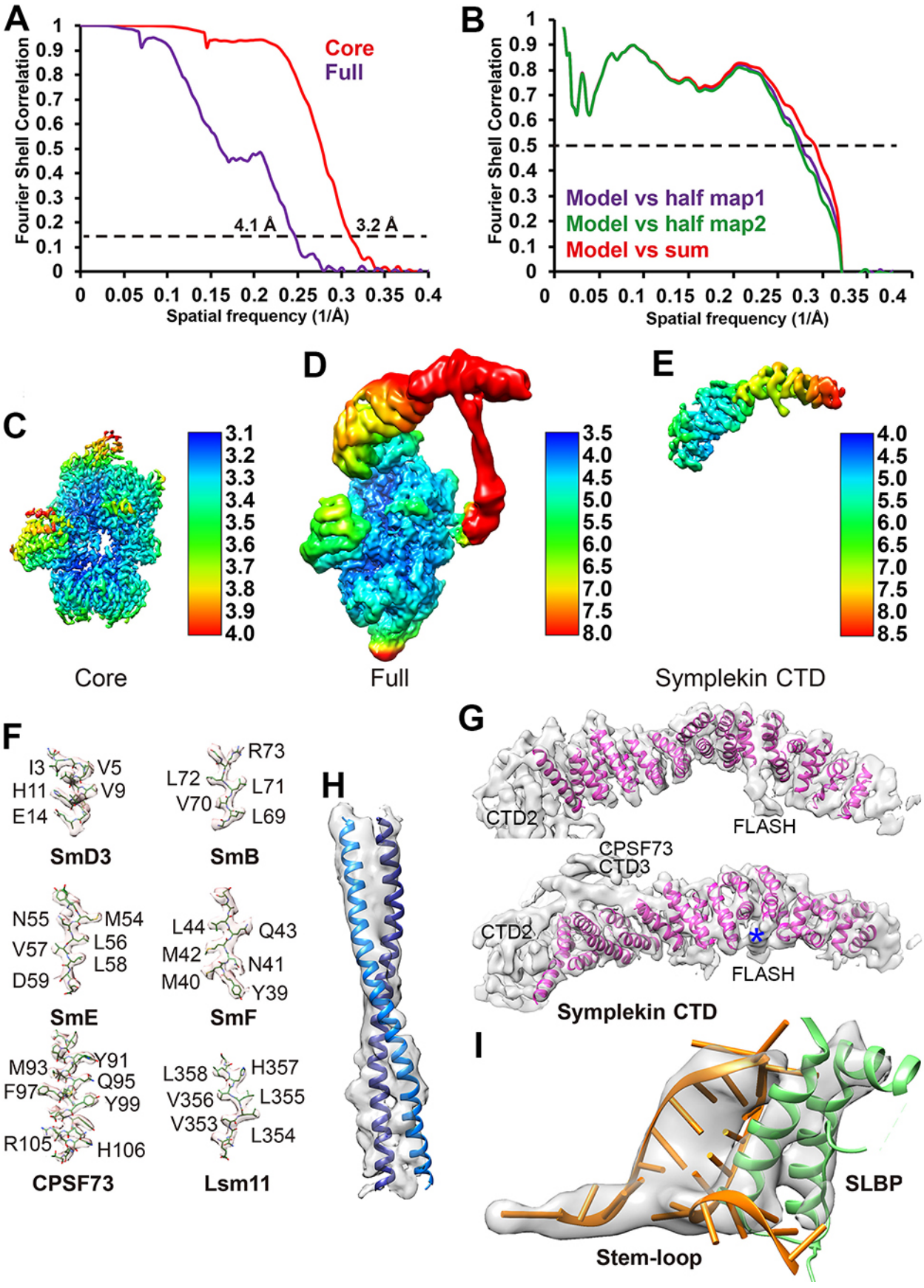
A
Fourier Shell Correlation
Core
Full
4.1 Å
3.2 Å
Spatial frequency (1/Å)
B
Fourier Shell Correlation
Model vs half map1
Model vs half map2
Model vs sum
Spatial frequency (1/Å)
C
Core
D
Full
E
Symplekin CTD
F
I3
V5
H11
V9
E14
SmD3
R73
L72
L71
V70
L69
SmB
N55
M54
V57
L56
L58
D59
SmE
L44
Q43
M42
N41
M40
Y39
SmF
M93
Y91
Q95
F97
Y99
R105
H106
CPSF73
L358
H357
V356
L355
V353
L354
Lsm11
H
G
CTD2
FLASH
CPSF73
CTD3
CTD2
FLASH
Symplekin CTD
I
SLBP
Stem-loop

Fig. S5. Quality assessment of the EM density maps.

(**A**) Gold-standard FSC (Fourier shell correlation) curves calculated between independently refined half maps for the density maps of the core of the machinery (red) and the full machinery (purple). (**B**) Cross-validation FSC curves for the core machinery; purple curve: refined model *versus* half map 1 used for refinement (work map); green curve: refined model *versus* half map 2 not used for refinement (free map); red curve: refined model *versus* the combined final map. The similarity of the "work" and "free" curves suggests no substantial over-fitting. (**C**,**D**,**E**) Local resolution maps for the core (C), the full machinery (the map was low-pass filtered by local resolution and not sharpened) (D), and a periphery of the machinery (E). (**F**) Representative cryo-EM densities from the core machinery map. (**G**) Cryo-EM density for symplekin CTD, in a periphery of the machinery corresponding mostly to panel E. Poly-alanine helices were built into the density, although the lack of side-chain density prevents building a complete model. The two views are related by a 90° rotation around the horizontal axis. Extra density for the CTD2 of CPSF73 and CPSF100, CTD3 of CPSF73, and FLASH is indicated. Density for an extra helix next to the symplekin CTD, possibly containing the LDLY motif of FLASH, is indicated with the star (blue). (**H**) Another view of the cryo-EM density for FLASH dimer, related to that in Fig. 1E by a 90° rotation around the vertical axis. (**I**) Cryo-EM density for SLBP and H2a* stem-loop. The crystal structure of the complex was docked into the EM density.

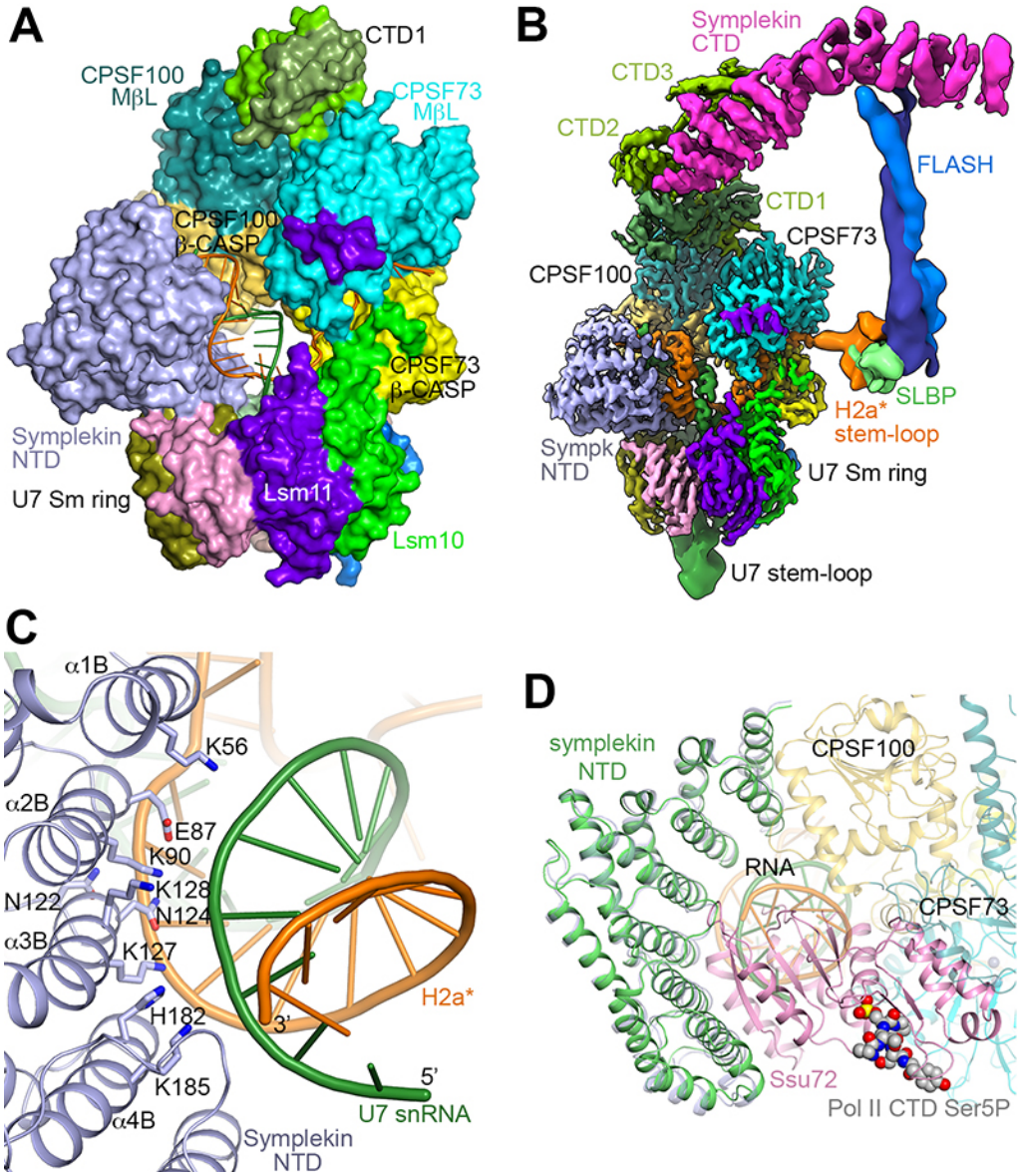

Fig. S6. Recognition of the duplex between U7 snRNA and H2a* HDE.

(**A**) Molecular surface of the proteins in the core of the machinery, colored as in Fig. 1B. The RNA molecules are shown as cartoons, with the U7 snRNA in dark green and the H2a* in orange. CPSF73 and symplekin NTD make direct contacts with the N- and C-terminal extensions of the Sm proteins. (**B**) Composite cryo-EM density map for the overall machinery. The view is the same as Fig. 1E. Produced with ChimeraX (*68*). (**C**) Interactions between the symplekin NTD and the RNA duplex. Side chains of NTD residues in the interface are shown as sticks. (**D**) Overlay of the structure of the machinery (in faded colors) with that of symplekin NTD (green) in complex with Ssu72 (pink) and the Pol II CTD peptide (gray, sphere model) (PDB entry 3O2Q). The NTD was used for the overlay, and Ssu72 clashes with the duplex and CPSF73.

Fig. S7. U-U base pairs may be common in HDE-U7 duplexes.

Alignment of human H2a and H2b histone pre-mRNA 3′-end sequences, showing the stem-loop and the HDE. The H2AA* sequence contains mutations in the HDE for complete Watson-Crick base pairing with the U7 snRNA. The cleavage site is indicated with the arrow. The U7 snRNA 5′-end sequence is shown at the bottom (5′ end on the right). Red letter indicates a possible U-U base pair between the HDE and U7. Cyan letter indicates a G-U wobble base pair. Uppercase letter indicates a Watson-Crick base pair, and lowercase letter indicates lack of a base pair.

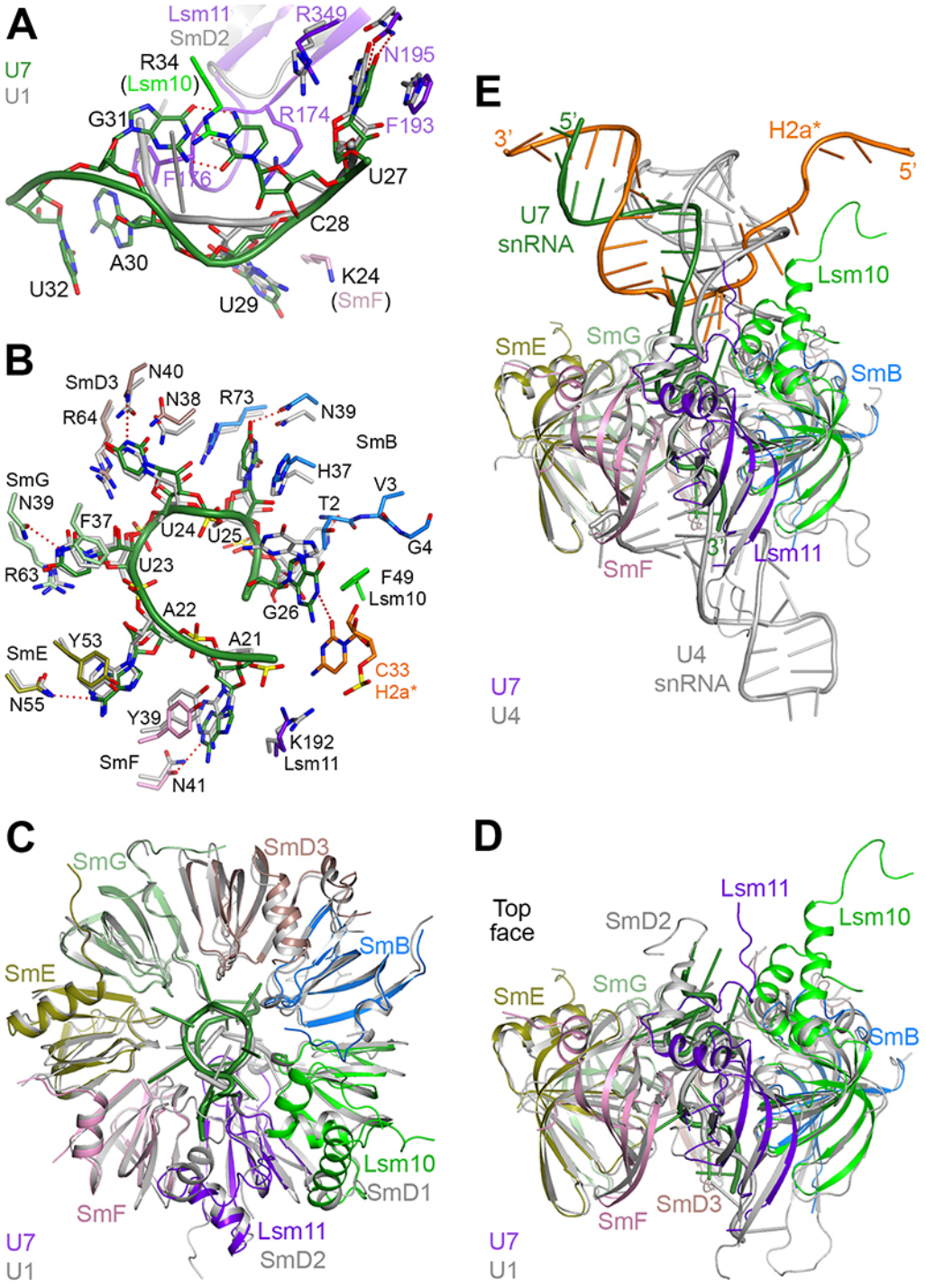

Fig. S8. Comparison of the U7 Sm ring with other Sm rings.

(**A**) Recognition of the C-G base pair in the CUAG sequence of the U7 Sm site. The base pair is flanked on one side by Arg34 of Lsm10, and on the other by Arg174 of Lsm11. C28 and G31 are conserved among most animal U7 snRNAs (*69*). Mutation of C28 to G reduced but did not abolish processing (*13*), and this base pair may also be important for the assembly of the U7 snRNP. Also shown is the

recognition of the U27 base by Lsm11. The binding mode of the Sm site in U1 snRNP is shown in gray (PDB entry 4PJO). The backbone conformation of the CUAG nucleotides is substantially different from those in U1 and U4 snRNAs, because of a longer loop in the Sm core of Lsm11 (residues 174-180) compared to SmD2 (residues 47-49). As a result, Lsm11 contacts 5 nucleotides of the extended 11-nucleotide Sm site in U7 snRNA, as compared to only two for SmD2. (**B**) Comparison of the binding modes of nucleotides 21-26 of the Sm site in U7 snRNP (in color) with those in U1 snRNP (in gray). G26 of U7 snRNA is flanked by A34 of the pre-mRNA and Phe49 of Lsm10 on its faces. The nucleotide equivalent to G26 in U1 snRNA (a G) has a different conformation and is not involved in any interactions. The equivalent nucleotide in U4 snRNA (a U) is shown as a thin stick model (in gray). It clashes with the N-terminal extension of SmB in U7 snRNP. (**C**) Overlay of the structure of U7 Sm ring reported here (in color) with that of the U1 Sm ring. The subunits of the U7 Sm ring are given different colors. The Sm site is located in the center of the ring and passes through the ring. (**D**) Same as panel C, viewed after a 90° rotation around the horizontal axis. The N- and C-terminal extensions are visible at the top face. Residues 1-138 of Lsm11 are not shown. (**E**) Overlay of the structure of U7 snRNP in complex with H2a* reported here (in color) with that of U4 snRNP (gray, PDB entry 4WZJ).

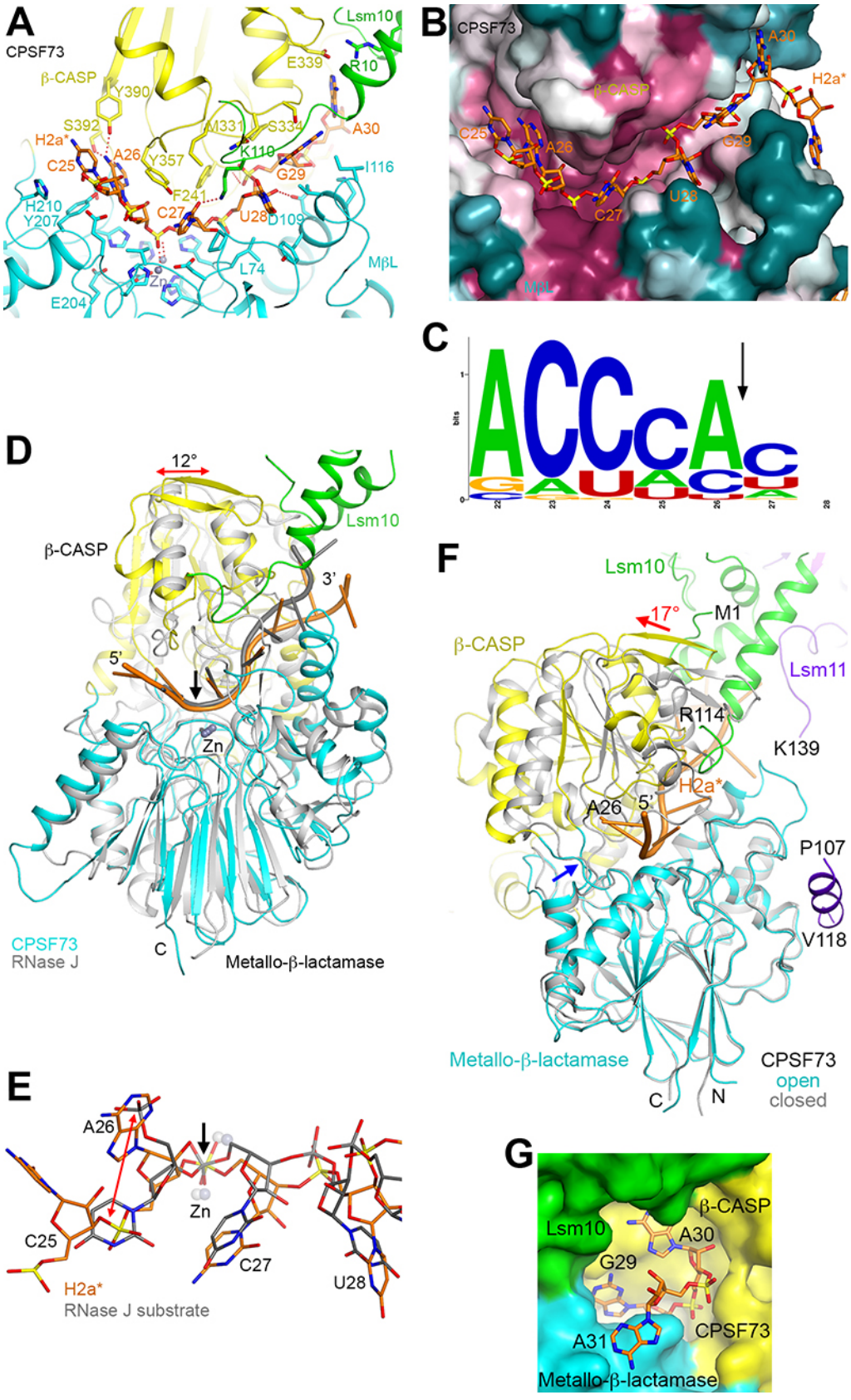
A
CPSF73
β-CASP
Lsm10
E339
R10
Y390
S392
M331
S334
A30
H2a*
A26
Y357
K110
I116
C25
G29
F241
H210
Y207
C27
U28
D109
MβL
L74
Zn
E204
B
CPSF73
A30
H2a*
β-CASP
C25
A26
G29
U23
C27
MβL
C
bits
D
12°
Lsm10
β-CASP
3'
5'
Zn
CPSF73
RNase J
C
Metallo-β-lactamase
F
Lsm10
17°
M1
β-CASP
Lsm11
R114
K139
H2a*
A26
5'
P107
V118
Metallo-β-lactamase
CPSF73
open
closed
C
N
E
A26
Zn
C25
C27
U28
H2a*
RNase J substrate
G
β-CASP
Lsm10
A30
G29
A31
CPSF73
Metallo-β-lactamase

Fig. S9. Recognition of the pre-mRNA substrate by CPSF73.

(**A**) Detailed interactions between CPSF73 and the H2a* pre-mRNA bound at the interface between the metallo-β-lactamase domain (MβL, cyan) and β-CASP domain (yellow). H2a* is shown in orange. The C-terminal extension of Lsm10 is in green. The phosphate groups for G29 and A30 have hydrogen-bonding interactions with backbone amides of CPSF73. The 2′ hydroxyl group of U28 is recognized by Asp109, but the other 2′ hydroxyl groups are not recognized. The bases are flanked on their sides, but generally show no hydrogen-bonding interactions, except for A26, which has hydrogen-bonding interactions to its N1 and N6 atoms from Ser392 and Tyr390, respectively. (**B**) The active site region of CPSF73 is highly conserved among homologs. Dark purple: highly conserved. Dark cyan: highly variable. Prepared with ConSurf (*70*). (**C**) Conservation of sequences near the cleavage site of 59 mammalian histone pre-mRNAs. The cleavage site is indicated with the arrow, and nucleotide numbers are the same as in fig. S1. There is no sequence preference at position 28, and hence no logo is shown there. Prepared with WebLogo (*71*) based on data in (*72*). (**D**) Overlay of the structure of CPSF73 reported here (in color) with that of RNase J (gray, PDB entry 4XWW). The cleavage site is indicated with the black arrow. The conformational difference for the β-CASP domain is indicated with the red arrow, which affects the binding mode for the 3′ part of the substrate. (**E**) Detailed comparison of the binding modes of the RNA in CPSF73 (in color) with that in RNase J (gray, PDB entry 4XWW). The swap of the base and phosphate group in the nucleotide just prior to the cleavage site is indicated by the red arrow. (**F**) Overlay of the structure of CPSF73 in the active state with that of CPSF73 in the inactive state, in complex with sulfate. The rms distances among equivalent Cα atoms are 0.63 and 0.98 Å for the metallo-β-lactamase and β-CASP domains when they are superimposed separately to the earlier structure, consistent with a rigid body movement of one domain relative to the other. The pivot points for this change are located in the linkers between the two domains, around residues His212 and Ala395 (blue arrow). The view is related to that of Fig. 3C by a 55° rotation around the vertical axis. (**G**) Nucleotides 29-31 exit a small tunnel in the surface of CPSF73. Nucleotide A31 is on the surface, and has weak density for its base (Fig. 3A).

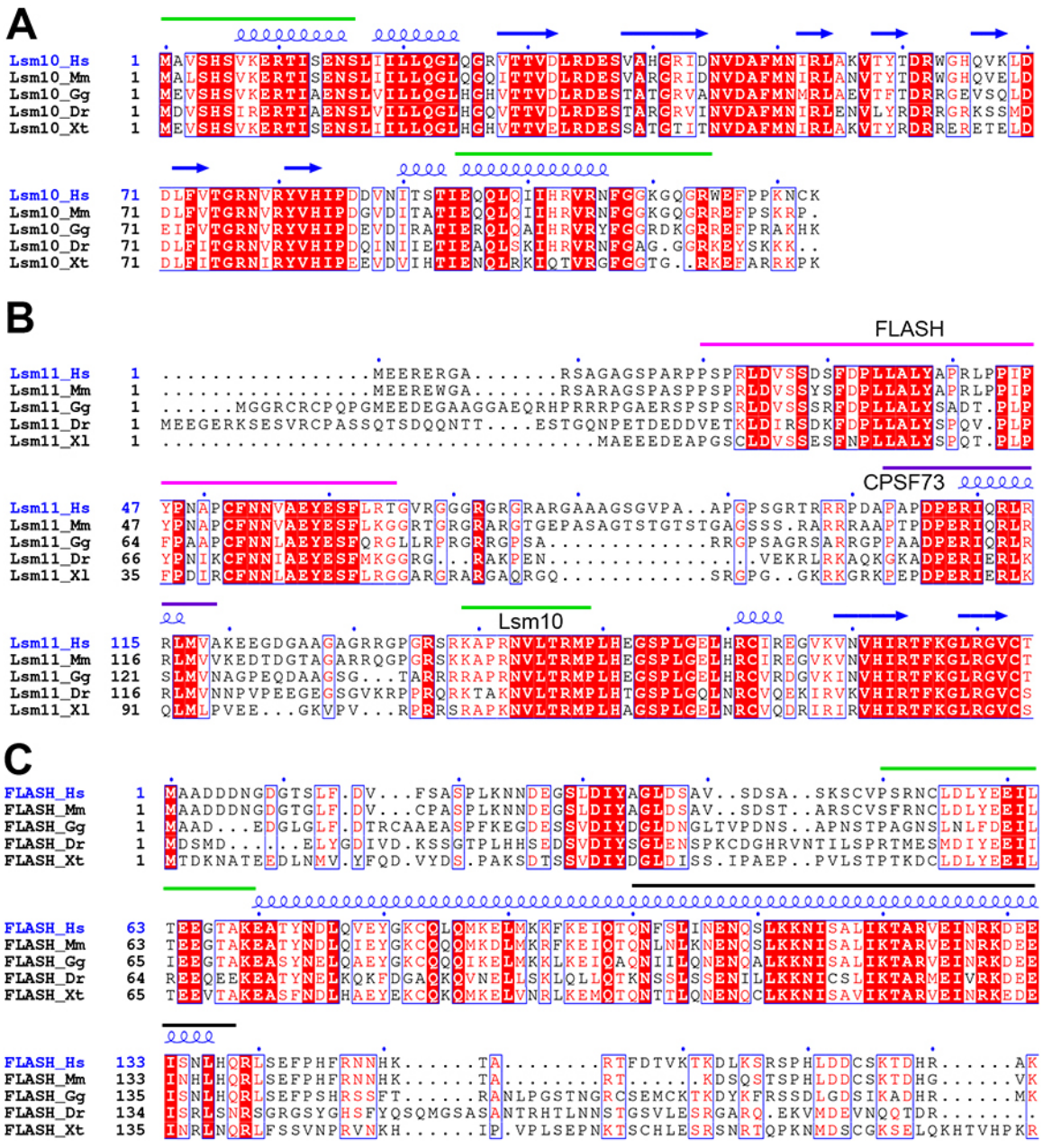

Fig. S10. Sequence conservation of Lsm10, Lsm11 and FLASH.

(**A**) Sequence alignment of selected vertebrate Lsm10 homologs. The N- and C-terminal extensions are indicated by the green lines. (**B**) Sequence alignment of selected vertebrate Lsm11 homologs. The regions that interact with FLASH, CPSF73 and Lsm10 are indicated. The insertion in the Sm core is not shown. (**C**) Sequence alignment of the N-terminal segment of selected vertebrate FLASH homologs. The region that interacts with HCC is indicated in green, and that with Lsm11 in black. The coiled-coil region is also shown. Hs: human; Mm: mouse; Gg: chicken; Dr: zebra fish; Xt: frog.

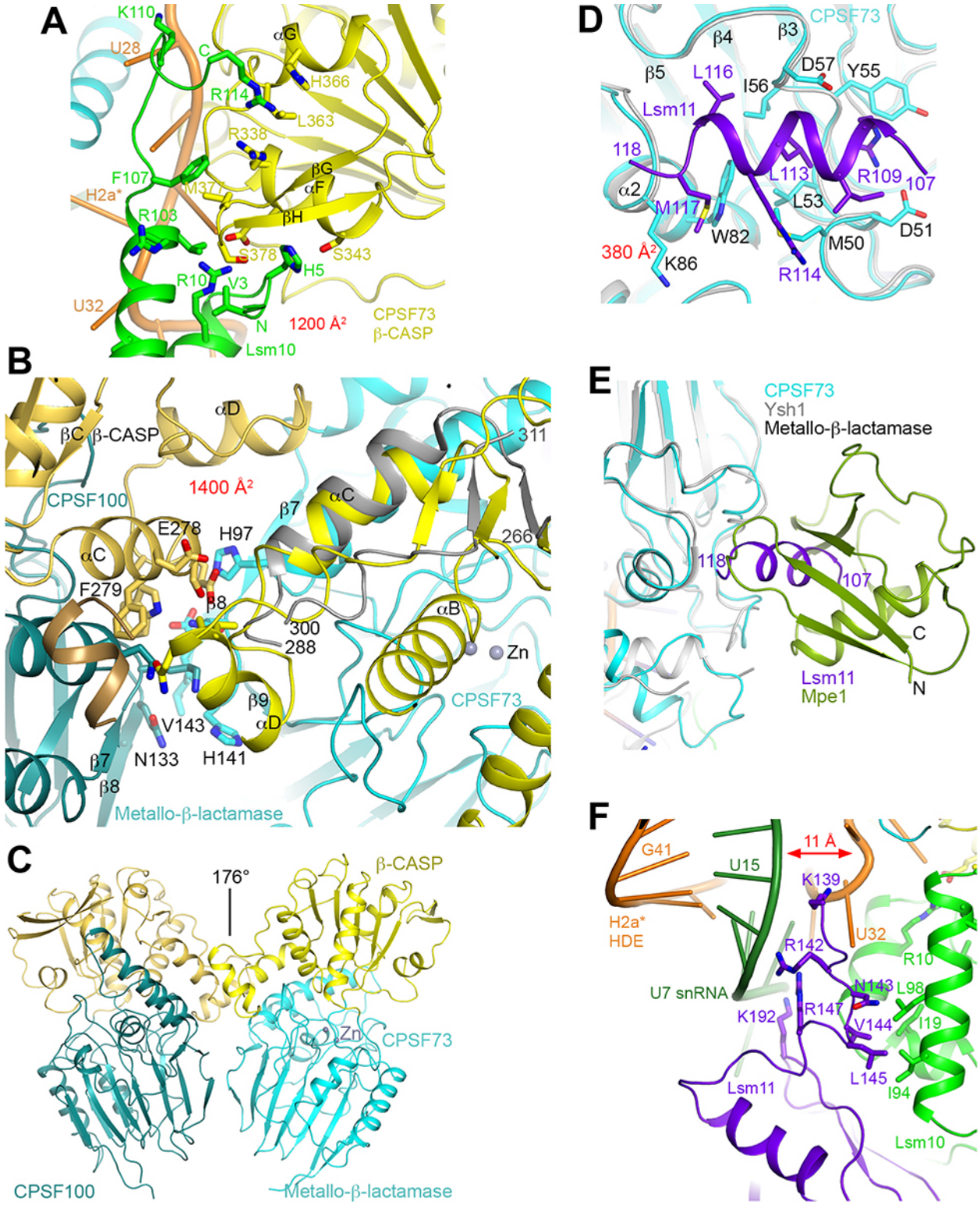
A
K110
U28
C
αG
H366
R114
L363
R338
F107
H2a*
M377
βG
αF
R103
βH
S378
S343
H5
R10
V3
U32
N
1200 Å²
CPSF73
β-CASP
Lsm10
B
αD
βC β-CASP
311
CPSF100
1400 Å²
β7
αC
E278
H97
266
αC
F279
β8
αB
300
288
Zn
β9
αD
CPSF73
V143
β7
N133
H141
β8
Metallo-β-lactamase
C
176°
β-CASP
Zn
CPSF73
CPSF100
Metallo-β-lactamase
D
CPSF73
β4
β3
β5
L116
D57
Y55
I56
Lsm11
118
L113
R109
107
α2
M117
L53
D51
W82
M50
380 Å²
K86
R114
E
CPSF73
Ysh1
Metallo-β-lactamase
118
107
C
Lsm11
Mpe1
N
F
G41
U15
11 Å
K139
H2a*
HDE
U32
R142
R10
U7 snRNA
N143
L98
K192
R147
V144
I19
L145
Lsm11
I94
Lsm10

Fig. S11. Interactions between selected subunits in the machinery.

(**A**) Interface between CPSF73 β-CASP domain (yellow) and the N- and C-terminal extensions of Lsm10 (green). The H2a* pre-mRNA is in orange. The N-terminal extension interacts with the loop connecting the βG-βH hairpin and helix αF of CPSF73. The C-terminal extension contacts helices αF and αG and the loop connecting the βG-βH hairpin. This extension is also located near the rim of the canyon and helps form the pre-mRNA binding site. (**B**) Interface between CPSF73 and CPSF100 metallo-β-lactamase and β-CASP domains. Residues 266-311 (with 288-300 being disordered) in the structure of CPSF73 alone are shown in gray. For the metallo-β-lactamase domain, the loop connecting strands β7 and β8 at the edge of the β-sandwich in CPSF73 contacts the equivalent loop in CPSF100. In the β-CASP domain, residues 288-300 of CPSF73 (containing helix αD) are disordered in the structure of CPSF73 alone. These residues are ordered in the machinery and have interactions with helix αC of the β-CASP domain of CPSF100. EM density was also observed for a short helix in this region (dark yellow), likely from the highly hydrophilic segment of CPSF100 since this helix is located close to the ends of that segment. However, we are not able to conclusively assign the residues that are in this helix due to its small size. (**C**) A pseudo-dimer of CPSF73 and CPSF100 in the HCC. The pseudo two-fold symmetry axis is indicated with the black line. The rotation overlays the metallo-β-lactamase and β-CASP domains, but the CTDs (not shown) do not obey this symmetry. (**D**) Interface between residues 107-118 in the N-terminal extension of Lsm11 (purple) and the metallo-β-lactamase domain of CPSF73. The structure of CPSF73 alone is shown in gray. These residues assume primarily a helical conformation and are bound to the surface of the metallo-β-lactamase domain of CPSF73. This helix contacts residues in the β4-β5 loop and helix α2 of CPSF73. However, the binding of this helix does not cause any conformational changes in the structure of CPSF73 compared to the closed, inactive form (*6*). (**E**) Overlay of the structure of CPSF73 metallo-β-lactamase domain (cyan) in complex with Lsm11 residues 107-118 (purple) with that of Ysh1 (gray) in complex with Mpe1 (green) (PDB entry 6I1D) (*24*). The binding site for Lsm11 overlaps with that for Mpe1. (**F**) Interface between Lsm11 and Lsm10, as well as between Lsm11 and the RNAs. Residues 139-150 in the N-terminal extension of Lsm11 are placed directly next to the helix in the C-terminal extension of Lsm10. The primarily hydrophobic contacts in this interface may provide stabilization of the C-terminal helix of Lsm10 and its interaction with CPSF73. The N-terminal extension of Lsm11 also contain three basic residues, which interact with the phosphodiester backbones of both U7 snRNA and pre-mRNA, stabilizing a close approach (within ~11 Å) between the duplex and the linker in the pre-mRNA from the cleavage site to the HDE. U32 of H2a* is also bound in this interface between Lsm10 and Lsm11.

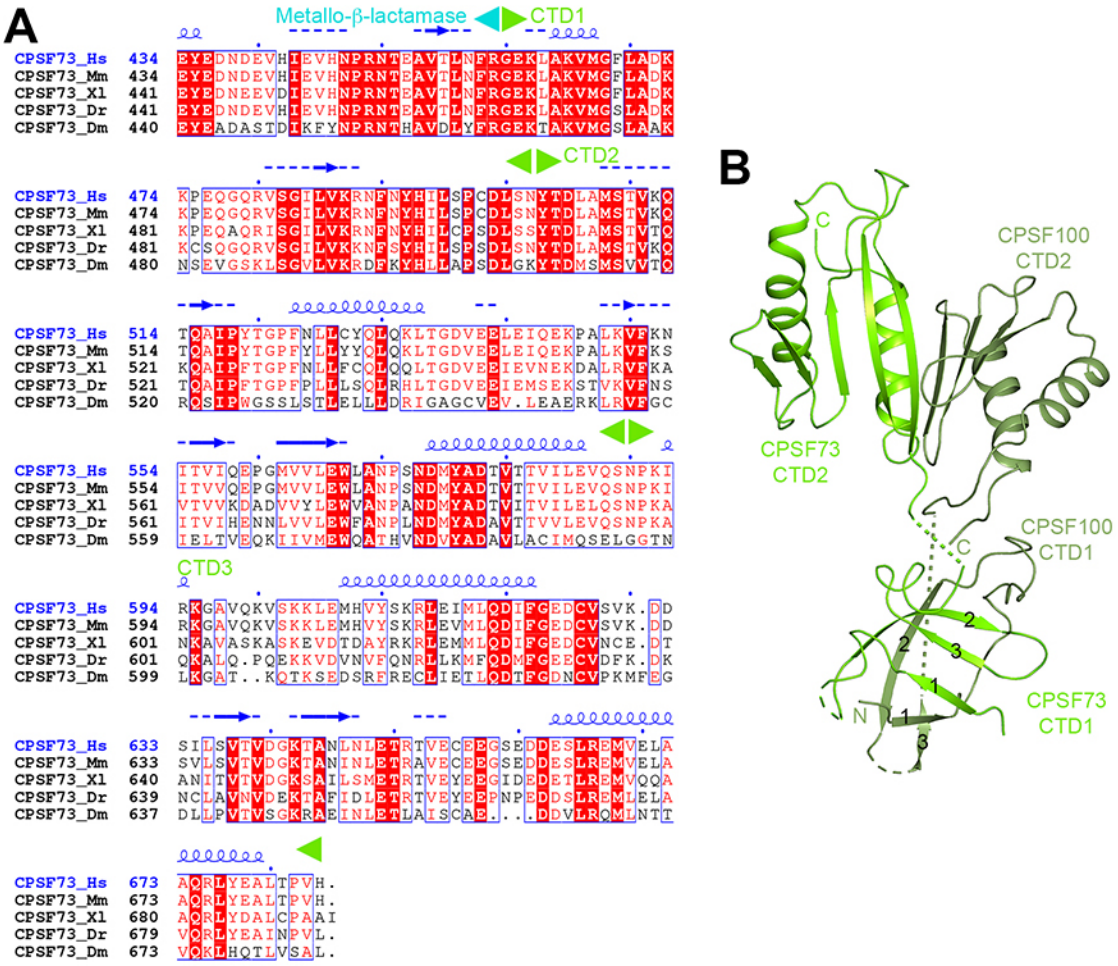

Fig. S12. The CTDs of CPSF73 and CPSF100.

(**A**) Sequence alignment of the CTDs of animal CPSF73 homologs. The domain boundaries are indicated. Hs: human, Mm: mouse, Xl: *Xenopus laevis*, Dr: *Danio rerio*, Dm: *Drosophila melanogaster*. (**B**) The structure of the CTD1 and CTD2 complexes of CPSF73 (light green) and CPSF100 (dark green). The structure of the CTD2 complex is based on that of the IntS11-IntS9 CTD complex (PDB entry 5V8W) (*28*). The connections between CTD1 and CTD2 are disordered and indicated with the dotted lines. The three strands in CTD1 are labeled.

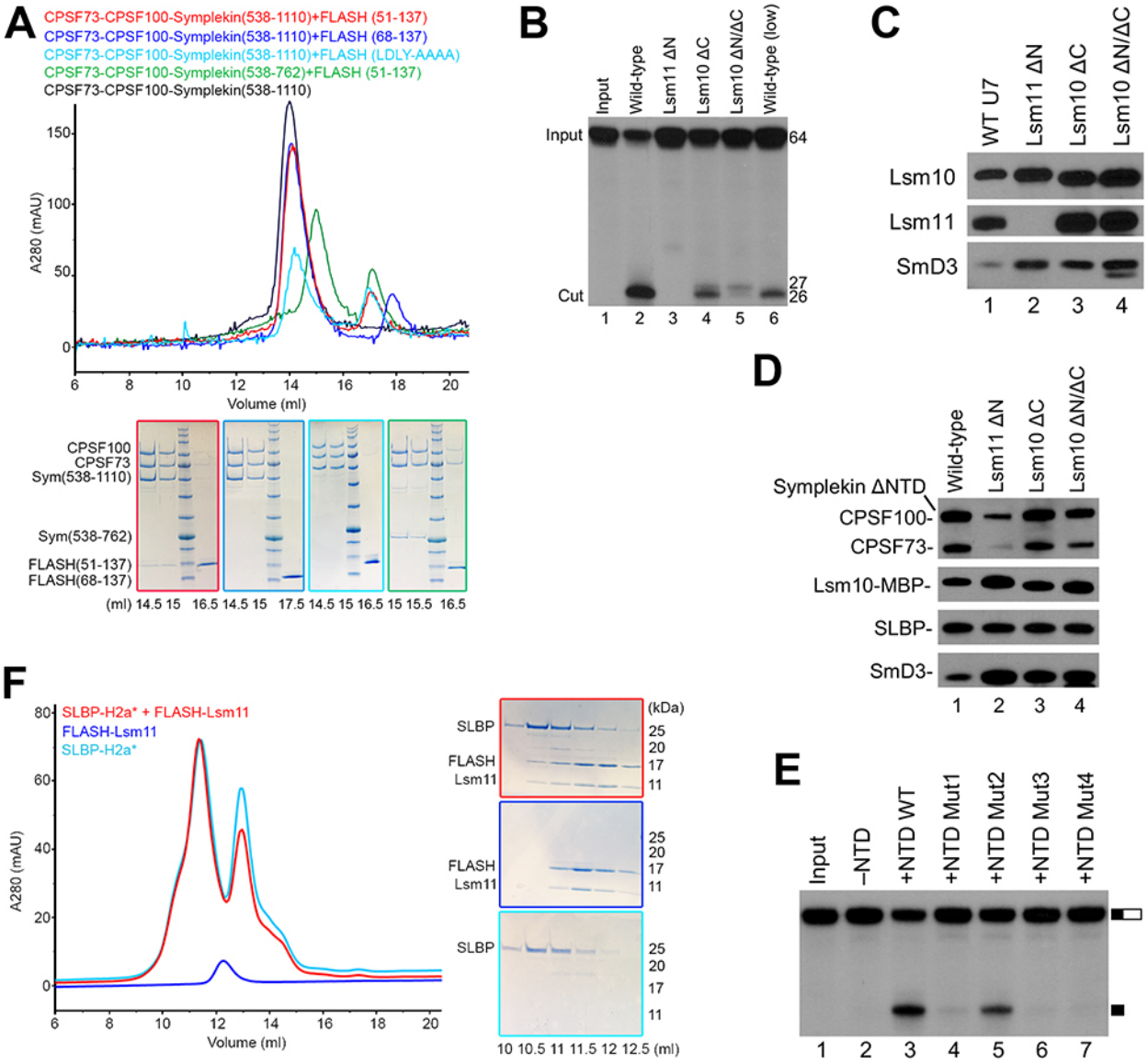

Fig. S13. Biochemical studies support the structural observations.

(**A**) HCC recruitment was abolished by deleting the LDLY motif of FLASH, mutating it to AAAA, or deleting the second half of the symplekin CTD. Gel filtration profiles (top) and SDS gels for the peak fractions (bottom) of mixtures of HCC and FLASH. The color of the box around each gel corresponds to the UV trace of the same color. The first half of the symplekin CTD (residues 538-762, green trace), or a FLASH mutant lacking residues 51-67 (blue trace, which include the LDLY motif), cannot form the HCC-FLASH complex. The LDLY-AAAA mutant of FLASH, containing residues 51-137 (cyan trace), cannot form the complex either. Binding of FLASH to HCC did not cause a change in the elution volume, because it is much smaller than HCC. (**B**) Cleavage assays showing that deleting the C-terminal extension of Lsm10 (ΔC, residues 100-C terminus) and both extensions of Lsm10 (ΔN/ΔC, residues 1-7, 100-C terminus) greatly reduces the cleavage activity. The Lsm10 ΔN/ΔC mutant shows mis-processing of the pre-mRNA, primarily generating a product of 27 nucleotides in comparison with 26 nucleotides

for the authentic product. Therefore, these extensions may also have a crucial role in correctly positioning CPSF73 and the pre-mRNA for the cleavage reaction, consistent with the structural observations that they contact CPSF73 (fig. S10A) and contribute to the binding of the pre-mRNA (Fig. 3D). Deleting the N-terminal extension of Lsm11 (ΔN, residues 1-152) abolished cleavage. The symplekin NTD is present in *cis* in these assays. (**C**) Deletion of the N- and C-terminal extensions of Lsm10 does not abrogate the formation of the U7 snRNP. The mutant with both extensions removed does have reduced ability to form the snRNP, and higher concentrations of the mutant was included in the experiment. (**D**) Deletion of the N- and C-terminal extensions of Lsm10 does not abrogate the formation of the processing machinery, based on a pulldown experiment with H2a* pre-mRNA. In contrast, deletion of the N-terminal extension of Lsm11 impairs HCC recruitment, consistent with the structural observations and previous biochemical data (*43*). (**E**) Cleavage assays showing that mutating as few as two basic symplekin NTD residues at the interface with the HDE-U7 duplex (fig. S6C) greatly reduces the cleavage activity. WT: wild-type. Mut1: K127E/K128E. Mut2: H182E/K185E, Mut3: K90E/H182E/K185E, Mut4: K127E/K128E/H182E/K185E. The symplekin NTD is present in *trans* in these assays. (**F**) The Lsm11-FLASH complex partially co-migrates with the SLBP-H2a* complex on a gel filtration column, and some quaternary complex is observed (second lane, red box, compared to the blue box). Lsm11: residues 23-79, which bind the FLASH dimer (*14*). This indicates that the quaternary complex can be formed but is likely more stable in the context of the entire machinery, similar to our observations with the symplekin NTD.

Table S1.

Alphabetical listing of abbreviations in this paper

Abbreviation	Definition/Explanation
β-CASP	Metallo-β-lactamase-associated CPSF Artemis SNM1/PSO2. A domain in CPSF73 and CPSF100. Interacts with the metallo-β-lactamase domain and regulates substrate access in CPSF73.
CPSF	Cleavage and polyadenylation specificity factor. Contains CPSF30, CPSF73, CPSF100, CPSF160, WDR33 and Fip1 subunits, which form the mPSF and mCF modules.
CPSF73	73 kDa subunit of CPSF. Endoribonuclease for the cleavage reaction for canonical and histone pre-mRNAs. Contains metallo-β-lactamase and β-CASP domains for the catalytic module at the N-terminus, and CTDs (Fig. 1B).
CPSF100	100 kDa subunit of CPSF. Weak sequence homolog of CPSF73. Lacks several of the conserved residues in the active site compared to CPSF73. The β-CASP domain contains inserted segments of highly hydrophilic residues (gray in Fig. 1B).
CstF	Cleavage stimulation factor. Contains CstF50, CstF64 and CstF77 subunits.
CstF64	64 kDa subunit of CstF. Contains an RNA recognition module (RRM) at the N-terminus, followed by a hinge region that interacts with CstF77 or symplekin, and a CTD that interacts with Pcf11 (*73*).
CTD	C-terminal domain. Present in CPSF73, CPSF100, symplekin, CstF64 and RNA polymerase II largest subunit.
FLASH	FLICE-associated huge protein. Has 1982 residues. Residues 51-140 involved in histone pre-mRNA processing.
HCC	Histone pre-mRNA cleavage complex. Contains CPSF73, CPSF100, symplekin and CstF64 subunits (Fig. 1A).
HDE	Histone downstream element. A purine-rich sequence element recognized by base pairing with the U7 snRNA.
HDE-U7 duplex	A duplex formed between the histone downstream element (HDE) 3′ to the cleavage site and the 5′ end of the U7 snRNA.
Lsm10	Unique subunit in U7 snRNP. Replaces SmD1 of the spliceosomal snRNPs. Has small extensions at the N- and C-termini beyond the Sm core (Fig. 1B).
Lsm11	Unique subunit in U7 snRNP. Replaces SmD2 of the spliceosomal snRNPs. Has a long N-terminal extension beyond the Sm core, and an insert in the Sm core (gray in Fig. 1B).
MβL	Metallo-β-lactase. A domain in CPSF73 and CPSF100. The domain in CPSF73 contains residues that coordinate zinc and carry out endonuclease activity.
mCF	Mammalian cleavage factor. Contains CPSF73, CPSF100 and symplekin subunits.
mPSF	Mammalian polyadenylation specificity factor. Contains CPSF160, WDR33, CPSF30 and Fip1 subunits. Recognizes the AAUAAA polyadenylation signal and recruits the poly(A) polymerase. Forms the core of the canonical machinery.
NTD	N-terminal domain. Present in symplekin.
SL	Stem-loop. An RNA secondary structure present 5′ to the cleavage site in histone pre-mRNAs and at the 3′ end of U7 snRNA (Fig. 1A).

SLBP	Stem-loop binding protein. Also known as hairpin binding protein (HBP). Recognizes the stem-loop 5′ to the cleavage site in histone pre-mRNAs. Has other central functions in histone pre-mRNA 3′-end processing and mRNA translation.
U7 snRNA	U7 small nuclear RNA. Part of U7 snRNP.
U7 snRNP	U7 small nuclear ribonucleoprotein. Contains U7 snRNA and Sm ring with Lsm10, Lsm11, SmB, SmD3, SmE, SmF and SmG subunits. Crucial for replication-dependent histone pre-mRNA 3′-end processing.
WDR33	WD repeat-containing protein 33. A subunit of CPSF (and mPSF).

Table S2.

Cryo-EM data collection, refinement and validation statistics for the core of the machinery

	(PDB 6V4X, EMDB-21050)
Data collection and processing	
Camera	Gatan K2 Summit
Nominal magnification	22,500
Magnification (calibrated at detector level)	46,729 (Data-Krios1), 47,847 (Data-Krios3)
Pixel size (Å)	1.07 (rescaled from 1.045 for Data-Krios1)
Voltage (kV)	300
Exposure rate (e^-/pixel/s)	8
Exposure time (s)	10
Total exposure ($e^-/Å^2$)	70 (Data-Krios1), 73 (Data-Krios3)
Frame rate (s)	0.2
Number of frames	50
Automation software	Leginon
Defocus range (μm)	1.2~2.5
Symmetry imposed	C1
Image stacks (No.)	7,103
Initial particle images (No.)	3,158,925
Final particle images (No.)	325,282
Relion translations accuracy	0.539
Relion rotations accuracy	0.917
Map resolution at FSC=0.5 (Å)	3.6 (masked), 4.2 (unmasked)
Map resolution at FSC=0.143 (Å)	3.2 (masked), 3.7 (unmasked)
Local resolution range (Å)	3.0-4.0
Map sharpening B factor (A^2)	–96
Refinement	
Refinement protocol	PHENIX real space refinement
Number of protein residues	1931
Number of RNA nucleotides	52
Number of metal ions	2
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.3
MolProbity score	1.86
B factors	
Protein	48.3
Nucleic acid	71.5
Model-map score	0.81
Main chain CC; side chain CC	0.79; 0.75
EMRinger score	3.71
C-beta deviations (%)	0
CaBLAM outliers (%)	2.74
PDB validation	
Clash score	9
Poor rotamers (%)	1
Ramachandran plot	
Favored (%)	94.25
Allowed (%)	5.69
Disallowed (%)	0.05

Table S3.

Cryo-EM data processing and reconstruction statistics for entire machinery and masked refinements

	Entire machinery (EMDB-21047)	Core of HCC (EMDB-21046)	Symplekin CTD (EMDB-21045)
Data collection and processing			
Camera	Gatan K2 Summit	–[1]	–[1]
Nominal magnification	22,500	–[1]	–[1]
Magnification (calibrated at detector level)	46,729 (Data-Krios1), 47,847 (Data-Krios3)	–[1]	–[1]
Pixel size (Å)	1.07 (rescaled from 1.045 for Data-Krios1)	–[1]	–[1]
Voltage (kV)	300	–[1]	–[1]
Exposure rate (e^-/pixel/s)	8	–[1]	–[1]
Exposure time (s)	10	–[1]	–[1]
Total exposure ($e^-/Å^2$)	70 (Data-Krios1), 73 (Data-Krios3)	–[1]	–[1]
Frame rate (s)	0.2	–[1]	–[1]
Number of frames	50	–[1]	–[1]
Automation software	Leginon	–[1]	–[1]
Defocus range (μm)	1.2~2.5	–[1]	–[1]
Symmetry imposed	C1	–[1]	–[1]
Image stacks (No.)	7103	–[1]	–[1]
Initial particle images (No.)	3,158,925	–[1]	–[1]
Final particle images (No.)	19,315	19,315	19,315
Relion translations accuracy	0.991	1.041	4.702
Relion rotations accuracy	1.729	2.314	2.074
Map resolution at FSC=0.5 (Å)	6.2 (masked), 9.0 (unmasked)	6.3 (masked), 9.0 (unmasked)	7.3 (masked), 8.8 (unmasked)
Map resolution at FSC=0.143 (Å)	4.1 (masked), 6.5 (unmasked)	4.1 (masked), 6.6 (unmasked)	4.7 (masked), 7.2 (unmasked)
Local resolution range (Å)	3.5-8.0	3.5-8.0	4.0-8.5
Map sharpening B factor (A^2)	0	-98	-160

1. The value is the same as that for the entire machinery.

Movie S1

A movie showing the composite cryo-EM map and atomic model of the overall histone pre-mRNA 3′-end processing machinery.

Movie S2

A movie showing the putative histone pre-mRNA 3′-end processing cycle. The subunits in the machinery are represented by their structures. Conformational changes in the HCC and CPSF73 for their activation are also indicated.