A unified mechanism for intron and exon definition and back-splicing

Xueni Li1,8, Shiheng Liu2,3,8, Lingdi Zhang1, Aaron Issaian1, Ryan C. Hill1, Sara Espinosa1, Shasha Shi1, Yanxiang Cui1, Kalli Kappel3, Rhiju Das4,5,6, Kirk C. Hansen1, Z. Hong Zhou2,3, and Rui Zhao1,7,8

The molecular mechanisms of exon definition and back-splicing are fundamental unanswered questions in pre-mRNA splicing. Here we report cryo-electron microscopy structures of the yeast spliceosomal E complex assembled on introns, providing a view of the earliest event in the splicing cycle that commits pre-mRNAs to splicing. The E complex architecture suggests that the same spliceosome can assemble across an exon, and that it either remodels to span an intron for canonical linear splicing (typically on short exons) or catalyses back-splicing to generate circular RNA (on long exons). The model is supported by our experiments, which show that an E complex assembled on the middle exon of yeast EF5 or HMRA1 can be chased into circular RNA when the exon is sufficiently long. This simple model unifies intron definition, exon definition, and back-splicing through the same spliceosome in all eukaryotes and should inspire experiments in many other systems to understand the mechanism and regulation of these processes.

The spliceosome sequentially forms the E, A, pre-B, B, Bact, B”, C, C”, P, and ILS complexes during the splicing cycle. Cryo-electron microscopy (cryo-EM) structures of all but one spliceosomal complex from Saccharomyces cerevisiae (yeast)1,2 provided valuable information on later stages of the splicing cycle. There is, however, a lack of structural and mechanistic understanding of the formation of the E complex, the earliest event that initiates the splicing cycle. Thus, it remains unclear how the splicing machinery accurately defines introns and exons. In yeast (which typically contain small introns and large exons), intron definition, where the spliceosome initially recognizes and assembles across an intron, seems to dominate4. On the other hand, exon definition4 prevails in vertebrates, where small exons and large introns are prevalent. In the exon definition model, the spliceosome first recognizes and assembles across an exon. However, it has been assumed that, in order to splice out introns, the exon definition complex (EDC) needs to be remodelled to a cross-intron complex. Support for the exon definition model is largely circumstantial, and biochemical and structural analyses of the exon definition process are limited. Although the EDC seems to be similar to the intron definition complex (IDC) in composition5,6, we do not know whether the two complexes differ in their structural organization, or how an EDC remodels to span an intron.

In addition to canonical splicing, a peculiar back-splicing reaction generates a class of circular RNAs (circRNAs) in diverse eukaryotic species, prompting speculation that back-splicing is also an ancient and conserved feature of the eukaryotic gene expression pathway7. CircRNAs are involved in the regulation of their host genes or microRNAs, ageing, and other disease processes8. Although canonical splicing signals and the spliceosome are needed for production of circRNAs9, the exact players and mechanism of back-splicing remain unknown.

To fill these gaps, we set out to obtain molecular details of the earliest step in the yeast splicing cycle, which commits a pre-mRNA to splicing. In yeast, intron recognition is initiated by the recognition of the 5’ splice site (SS) by the U1 small nuclear ribonucleoprotein (snRNP)10–12, and of the branch point sequence (BPS) by the BBP–Mud2 heterodimer (the 3’ SS is not recognized until much later), forming the E complex (also referred to as the CC2 complex)13. Here we report the cryo-EM structure of the yeast E complex assembled on either the ACT1 pre-mRNA or the UBC4 pre-mRNA. These structures and subsequent biochemical analyses reveal a unified mechanism for intron definition, exon definition and remodelling, and back-splicing-mediated production of circRNA.

In vitro-assembled E complex is functional

After discovering that the E complex purified from yeast is too heterogeneous for structural determination, we assembled the E complex in vitro using uncapped ACT1 pre-mRNA fused with three copies of MS2 stem loops at the 5’ end (M3–ACT1) and purified U1 snRNP, BBP and Mud2 proteins (referred to as the ACT1 complex). The complex was purified sequentially using the MS2 tag and the calmodulin binding peptide (CBP) tag on U1A and Mud2. After cleavage of M3–ACT1 into two fragments using RNase H (Extended Data Fig. 1), the MS2 tag still pulled down all U1 snRNP proteins, BBP, and Mud2 (Fig. 1a), confirming that U1 snRNP and BBP–Mud2 interact instead of simply tethered through M3–ACT1. In addition, the assembled ACT1 complex can be chased into spliced M3–ACT1 in yeast extract lacking U1 snRNA (Fig. 1b, lane 5). Although excess ACT1 pre-mRNA fused with three copies of MS2 stem loops at the 3’ end (ACT1–M3) can effectively compete with free M3–ACT1 for splicing (Fig. 1b, lane 2), it cannot compete with the assembled ACT1 complex (Fig. 1b, lane 6). These data indicate that our assembled E complex has not fallen apart substantially in the splicing extract and is functional.

Protein–RNA components facilitate 5’ SS recognition

We determined the cryo-EM structure of the ACT1 complex to 3.2 Å resolution (Extended Data Fig. 2, Extended Data Table 1). After observing low resolution in several key areas, we also assembled the E complex on a capped UBC4 pre-mRNA, crosslinked the complex with BS3, and determined its structure to 3.6 Å resolution (Extended Data Figs. 3, 4). The overall structures of the two complexes are similar (Extended Data Fig. 5a) and subsequent discussions refer to their common features unless otherwise stated.
In these structures, the 5′ SS base-pairs with the 5′ end of the U1 snRNA (Fig. 2a, b), which is stabilized by the U1C and Luc7 proteins (Fig. 2b), as observed in the yeast A and pre-B structures4,15. In addition, a homology model of the yeast nuclear cap binding protein (NCBP) complex can be fitted as a rigid body into the density upstream of nucleotide −9 of UBC4 (Fig. 2c), probably binding to the pre-mRNA cap. The RNA recognition motif (RRM) domain of U1–70K is shifted towards NCBP in the UBC4 complex compared to the uncapped ACT1 complex (Extended Data Fig. 5a), suggesting that NCBP interacts directly with U1–70K RRM and providing a possible mechanism by which NCBP recruits the U1 snRNP and facilitates splicing of cap-proximal introns6,17. In both complexes, the RRM2 domain of Nam8 is positioned to bind to the intronic region immediately downstream of nucleotide +13 (Fig. 2c), illustrating the structural basis of the facilitation of 5′ SS recognition by Nam818.

A striking feature in the ACT1 complex is an approximately 25-bp double helix on a binding surface formed by many positively charged residues on the C-terminal tail of Prp39 and the N-terminal domain of Prp42, as well as the C-terminal domain of U1C (Fig. 2d). A similar double helix density is also observed in the Pre-B complex structure and has been tentatively modelled as part of the U2 snRNP13. Our ACT1 complex is assembles in vitro and contains no U2 snRNA (Extended Data Fig. 5b). Furthermore, no such double helix exists in the UBC4 complex, suggesting that this helix is part of the ACT1 pre-mRNA. Although we were unable to model specific nucleotides, a weak density connects this helix to the 5′ SS, suggesting that it belongs to the region downstream of the 5′ SS. The 5′ SS-to-BPS region (265 nt) of the ACT1 intron is predicted to form long stem-like structures, whereas the same region in UBC4 (58 nt) contains a much shorter possible secondary structure (Extended Data Fig. 6a), potentially explaining why a stem-like structure is observed in the ACT1 complex but not the UBC4 complex. Mutation of this region in the ACT1–CUP1 reporter19, which abolishes extensive secondary structures (Extended Data Fig. 6), leads to substantial pre-mRNA accumulation compared to the wild type (Fig. 2e), suggesting that this secondary structure facilitates splicing. Our structures of the E and P complexes20 therefore provide direct evidence that the intronic regions of pre-mRNA can form highly ordered secondary structures, which may help to bring key intronic elements close together and may also facilitate spliceosomal assembly by interacting directly with proteins.

**The 5′ SS and BPS are bridged by flexible Prp40**

A key event in the first step of the splicing cycle is to define the intron by bringing together the 5′ SS and BPS, and the U1 snRNP protein Prp40 is important for this process13. Prp40 contains two N-terminal WW domains, an approximately 60-residue linker, and six C-terminal FF domains. In the region between U1–70K and Luc7 in the UBC4 complex structure, there is a boomerang-shaped density that matches well with the crystal structures of tandem FF domains connected by long helices21,22 (Fig. 2f, Extended Data Fig. 4g). (This density is not obvious in the ACT1 complex, possibly because the ACT1 complex is not cross-linked with BS3.) A weak density connects the boomerang-shaped density and U1–70K (Fig. 2f), and the C-terminal FF domain crosslinks to U1–70K in the UBC4 complex (while the N-terminal and middle FF domains crosslink to Luc7...
and Snu71; Extended Data Fig. 7a). These observations led us to assign the boomerang-shaped density as the FF4–6 domains of Prp40 (although we cannot rule out the possibility that this density represents other tandem FF domains, such as FF3–5), which is also consistent with our observation that the FF1–3 domains of Prp40 interact with Luc7 (Extended Data Fig. 7b).

Prior biochemical analyses have shown that Prp40 forms a stable dimer with Snu71 and a trimer with Snu71–Luc723–25 and that the Prp40 WW domains interact directly with the N-terminal domain of BBP13,26 BBP also forms non-exclusive interactions with both Prp40 and Mud221, and binds directly to the BPS of pre-mRNA27. In the ACT1 complex structure, there is a large volume of weak density close to the pre-mRNA double helix (Extended Data Fig. 4j). The density can be best interpreted as the BBP–Mud2 dimer for three reasons: its location corresponds roughly to where the U2 snRNP is in the A complex structure14 (Extended Data Fig. 4j); crosslinking and mass spectrometry analyses indicate that BBP–Mud2 is located in this region (Extended Data Fig. 7a); and BBP–Mud2 are the only proteins left in the E complex on long exons (Fig. 3b). Modelling using the Rosetta RNP-denovo method28 suggests that only 28 nt between the upstream branch point and downstream 5′ SS is needed to span the U2 snRNP and U1 snRNP in the A complex (Extended Data Fig. 8a). The minimal distance that connects the same branch point and 5′ SS is likely to be similar or smaller in the E complex, given the similar spatial position and smaller size of BBP–Mud2 compared to the U2 snRNP (Extended Data Fig. 4j). On the other hand, adding the tri-snRNP to form the pre-B complex forces an increase of about 30° in the angle between the U1 snRNP and U2 SF3b14,15 (Fig. 3b). A relatively short exon may hinder this conformational change and also create steric hindrance for the addition of the bulky tri-snRNP (Fig. 3b).

To test whether the E complex can form across a yeast exon, we truncated the multi-intronic DYN2 gene to contain only its middle exon and partial flanking introns (DYN2 IE1, Extended Data Fig. 8b).
Spliceosomal complexes that assemble on either DYN2 wild-type or IEI pre-mRNAs (using the same protocol as the ACT1 complex) contain the same protein components in similar quantities, even after cleavage by RNase H between the BPS and 5’ SS (Extended Data Fig. 8c–e). Furthermore, 2D classifications of negative-stain images of the DYN2 IEI complex resemble those of the ACT1 and UBC4 complexes (Extended Data Fig. 8f). These observations support the formation of an E complex across the DYN2 middle exon in vitro.

We next investigated whether exon definition occurs in vivo in yeast, by evaluating whether mutation of splice sites bordering the DYN2 middle exon negatively affected splicing of both flanking introns, a hallmark used to support the exon definition model in vertebrates6. We generated a BPS mutation in intron 1 (I1-BP mutant), a 5’ SS mutation in intron 2 (I2-5’ SS mutant), and a double mutation on DYN2 (Extended Data Fig. 8b). We demonstrated that the I2-5’ SS and I1-BP mutations impaire the splicing of intron 1 and intron 2, respectively (Fig. 4a). We further evaluated the splicing products of wild-type DYN2 and each mutant using PCR and primers located in exons 1 and 3 (Fig. 4b). If splicing of DYN2 is governed solely by intron definition, we would observe retention of the intron in which these mutations reside (with minimal effect on the distal intron), generating products containing a single intron (255- and 271-bp bands). On the other hand, if splicing of DYN2 is governed solely by exon definition, the mutations would lead to the retention of both introns (that is, accumulation of the 351-bp pre-mRNA band) or exon skipping (the 152-bp band), but not any product containing a single intron (indicating that the distal intron was successfully spliced). The fact that we observed both pre-mRNA accumulation and single-intron-containing products (Fig. 4b, lanes 4 and 5) suggest that both intron definition and exon definition contribute to splicing of DYN2 in vivo. We observed exon skipping for the I1-BP mutant but not the I2-5’ SS mutant, consistent with previous observations29. This observation differs from findings in the mammalian system, where exon definition mutations lead to predominantly exon skipping, probably because intron definition also contributes to splicing of DYN2, which would lead to co-transcriptional splicing of intron 1 in the I2-5’ SS mutant and prevent exon skipping29. Together, our results demonstrate that exon definition occurs for a fraction of DYN2 transcripts in vivo in yeast.

The EDC catalyses back-splncing on long exons

An intriguing prediction of our exon definition model is that if the exon that connects the branch point and downstream 5’ SS is long enough, it will not create much steric hindrance and will allow the tri-snRNP to join the pre-B complex and complete the rest of the splicing cycle (Fig. 3c). As a result, the 5’ SS downstream of the exon will be backspliced to the upstream 3’ SS, generating a circRNA through the same transesterification reaction that is used by canonical splicing (Fig. 3d). Consistent with this hypothesis, seven of the ten multi-intron genes in S. cerevisiae form circRNA products7.

To test this model, we purified the yeast spliceosome using TAP-tagged Cef1 (a strategy used to purify and determine the cryo-EM structures of multiple spliceosomal complexes) from the Prp22(H606A) mutant strain, which is defective in exon release30. As expected, purified spliceosomes contained spliced mRNA and lariat for the yeast single-intronic gene RPP18, as well as the unique T-branch and circRNA for the multi-intronic genes EF5 and HMRA1 (Fig. 5a, Extended Data Fig. 9a). These results show that Cef1-purified spliceosome contains both canonical and back-splncing products.

Further supporting this model (Fig. 3c, d), we showed using PCR with reverse transcription (RT–PCR) that the EF5S IEI construct on an expression plasmid generated an RNase R-resistant circRNA corresponding to exon 2 in vivo (Fig. 5b, lane 10, Extended Data Fig. 9b). Mutating the BPS or 5’ SS, or shortening exon 2 to 63 nt, abolished circRNA formation (Fig. 5b, lanes 11, 12, 14). An E complex assembled on in vitro-transcribed EF5S IEI-101-M3 (exon 2 shortened to 101 nt with 3 × MS2 at the 3’ end; Extended Data Fig. 9c) can be chased into circRNA in U1-depleted yeast extract in the presence of excess competing IE1-101 RNA (Fig. 5c). To ensure the generality of our observation, we carried out the same experiments using another yeast multi-intronic gene, HMRA1, and obtained the same results (Extended Data Fig. 9d, e). Together, these results support the idea that exon definition occurs in yeast across the middle exon of EF5S or HMRA1, and that this catalyses back-splncing and generates circRNA when the exon is sufficiently long.

Discussion

It was previously unclear whether the EDC is the same as or different from the IDC. The architecture of the E complex indicates that the same complex can form across either introns or exons without the need for additional components or structural rearrangement, and the same can be deduced for the A complex. The structures of the E and A complexes predict a minimal BP-to-5’ SS distance (28 nt for the A complex and probably a similar or smaller number for the E complex) in order for exon definition to occur. An exon that is above this minimum but still relatively short potentially makes it difficult for the tri-snRNP to join the spliceosome. This causes the spliceosome to stall at the pre-B stage and fail to handoff the 5’ SS from U1 to U6, providing an opportunity for the spliceosome to remodel into an intron-spanning B complex involving the upstream 5’ SS. This model is consistent with the observation in mammalian systems that tri-snRNP is loosely associated with the EDC, and becomes stably associated only when a 5’ SS-containing RNA oligonucleotide is added and the EDC is converted into a B-like intron-spanning complex6. In support of our exon definition model, we showed that both intron definition and exon definition contribute to yeast DYN2 splicing in vivo (Fig. 4). Although yeast has few multi-intronic genes and exon definition is clearly not the driving force of splicing, our results provide proof of principle that both intron and exon definition can occur through the same spliceosomal structure.
The EDC catalyses back-splicing and produces circRNA. a, RT–PCR of RNA isolated from spliceosome purified from the Prp22(H606A) yeast strain (S) and PCR using yeast genomic DNA (g, as negative control) for the single-intron gene RPP1B and multi-intron genes EFMS and HMRA1 demonstrates the presence of ligated exons (lane 1), lariat (lane 3), T-branches (lanes 5, 7) and circRNA (lanes 9, 10). Primer positions are indicated as arrows in the schematic diagrams below. All images in Fig. 5 are RT–PCR or PCR products on agarose gel with ethidium bromide (EtBr) staining. b, RT–PCR of RNA extracted from wild-type or EDC KO strain carrying indicated plasmid, with or without RNaseR treatment, using primers shown in the schematic diagrams below. Numbers 101 and 63 designate exon lengths. Lanes 1–7 indicate that all EFMS constructs are transcribed. c, IEI-101–M3 RNA or E complex assembled on IEI-101–M3 was incubated with splicing extract with or without U1 snRNA depletion in the absence or presence of 30-fold excess competing IEI-101 RNA. CircRNA products were monitored as in b. Competing IEI-101 was modified to remove the primer binding sites so it is invisible in the RT–PCR reaction. Experiments were repeated one (a) or two (b, c) additional times with similar results.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1523-6.

Received: 31 March 2019; Accepted: 1 August 2019; Published online 4 September 2019.

1. Zhang, L., Vielle, A., Espinosa, S. & Zhao, R. RNAs in the spliceosome: insight from cryoEM structures. *Wiley Interdiscip. Rev. RNA* 10, e1523 (2019).
2. Wan, R., Bai, R., Yan, C., Lei, J. & Shi, Y. Structures of the catalytically activated yeast spliceosome reveal the mechanism of branching. *Cell* 177, 359–391 (2019).
3. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA* 4, 49–60 (2013).
4. Berget, S. M. Exon recognition in vertebrate splicing. *Mol. Cell* 270, 2411–2414 (1995).
5. Sharma, S., Kohlstaedt, L. A., Damiano, A., Rio, D. C. & Black, D. L. Poly(A) polymerase and poly(A) 5′-end formation are trans-splicing defects in mutants of the vertebrate spliceosome. *Mol. Biol. Cell* 15, 183–191 (2004).
6. Schneider, M. et al. Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatagenic splicing complexes. *Mol. Cell* 38, 223–235 (2010).
7. Wang, P. L. et al. Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 9, e90859 (2014).
8. Wilusz, J. E. A. A 360° view of circular RNAs: from biogenesis to functions. *Wiley Interdiscip. Rev. RNA* 9, e1478 (2018).
9. Stark, S. et al. Exon circularization requires canonical splice signals. *Cell Rep.* 10, 103–111 (2015).
10. Séraphin, B., Kretzner, L. & Rosbash, M. A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5′ cleavage site. *EMBO J.* 7, 2533–2538 (1988).
11. Siliciano, P. G. & Guthrie, C. S′ splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev.* 2, 1258–1267 (1988).
12. Ruby, S. W. & Abelson, J. An early hierarchic role of U1 small nuclear ribonucleoprotein in spliceosome assembly. *Science* 242, 1028–1035 (1988).
13. Abovich, N. & Rosbash, M. Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. Cell 89, 403–412 (1997).

14. Plaschka, C., Lin, P. C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into splicesome assembly and regulation. Nature 559, 419–422 (2018).

15. Bai, R., Wan, R., Yan, C., Lei, J. & Shi, Y. Structures of the fully assembled Saccharomyces cerevisiae splicesome before activation. Science 360, 1423–1429 (2018).

16. Lewis, J. D., Izaurralde, E., Jarmolowski, A., McGuigan, C. & Mattaj, I. W. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5′ splice site. Genes Dev. 10, 1683–1698 (1996).

17. Li, X. et al. CryoEM structure of the three tandem FF domains of the yeast spliceosomal postcatalytic P complex. Structure 27, 140–151.e143 (2019).

18. Liu, J., Fan, S., Lee, C. J., Greenleaf, A. L. & Zhou, P. Specific interaction of the yeast nuclear cap-binding complex: an essential role for Cbc2 in meiosis via splicing of MER3 pre-mRNA. RNA 18, 1996–2011 (2012).

19. Puig, O., Gottschalk, A., Fabrizio, P. & Séraphin, B. Interaction of the U1 snRNP with nonconserved intronic sequences affects 5′ splice site selection. Genes Dev. 13, 569–580 (1999).

20. Lesser, C. F. & Guthrie, C. Mutational analysis of pre-mRNA splicing in Saccharomyces cerevisiae using a sensitive new reporter gene, CUP1. Genetics 133, 851–863 (1993).

21. Liu, S. et al. Structure of the yeast spliceosomal postcatalytic P complex. Science 358, 1278–1283 (2017).

22. Lu, M. et al. Crystal structure of the three tandem FF domains of the transcription elongation regulator CA150. J. Mol. Biol. 393, 397–408 (2009).

23. Liu, J., Fan, S., Lee, C. J., Greenleaf, A. L. & Zhou, P. Specific interaction of the transcription elongation regulator TCEG1 with RNA polymerase II requires simultaneous phosphorylation at Ser2, Ser5, and Ser7 within the carboxyl-terminal domain repeat. J. Biol. Chem. 288, 10890–10901 (2013).

24. Li, X. et al. CryoEM structure of Saccharomyces cerevisiae U1 snRNP offers insight into alternative splicing. Nat. Commun. 8, 1035 (2017).

25. Görnemann, J. et al. Cotranscriptional spliceosome assembly and splicing are independent of the Prp40p WW domain. RNA 17, 2119–2129 (2011).

26. Estor, C. & Uetz, P. The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. BMC Biochem. 9, 29 (2008).

27. Wiesner, S., Stehr, G., Sattler, M. & Macias, M. J. Solution structure and ligand recognition of the WW domain pair of the yeast splicing factor Prp40. J. Mol. Biol. 324, 807–822 (2002).

28. Jacewicz, A., Chico, L., Smith, P., Schwer, B. & Shuman, S. Structural basis for recognition of intron branchpoint RNA by yeast MsIS and selective effects of interfacial mutations on splicing of yeast pre-mRNAs. RNA 21, 401–414 (2015).

29. Cheng, Y., Kane, C. & Ares, M., Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. RNA 9, 993–1006 (2003).

30. Campodonico, E. & Schwer, B. ATP-dependent remodeling of the spliceosome: intragenic suppressors of release-defective mutants of Saccharomyces cerevisiae Prp22. Genetics 160, 407–415 (2002).

31. Liang, D. et al. The output of protein-coding genes shifts to circular RNAs when the pre-mRNA processing machinery is limiting. Mol. Cell. 68, 940–954.e943 (2017).

32. Ragan, C., Goodall, G. J., Shiromik, N. E. & Preiss, T. Insights into the biogenesis and potential functions of exonic circular RNA. Sci. Rep. 9, 2048 (2019).

33. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. Genes Dev. 28, 2233–2247 (2014).

34. Spingola, M., Grate, L., Haussler, D. & Ares, M., Jr. Genome-wide bioinformatic analysis of introns in Saccharomyces cerevisiae. RNA 5, 221–234 (1999).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019
METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Yeast E complex assembly and purification. The coding regions of yeast BPB and MUD2 were amplified by PCR using genomic S. cerevisiae DNA as templates. BPB fused to a N-terminal protein A (protA) tag was inserted between a GPD promoter as a C-terminal terminator, and the resulting expression cassette was cloned into pRS414 to generate the pRS414/GPD-protA BPB plasmid. Similarly, MUD2 with or without a C-terminal CBP tag was cloned into pRS416 vectors to generate the pRS416/GPD-MUD2-CBP or pRS416/GPD-MUD2 plasmid. Six litres of BCY123 cells harbouring both plasmids were grown in URA–TRP–selective medium to OD600 = 3–4. The cells were flash-frozen in liquid nitrogen to form yeast ‘pocorns’ and cryogenically ground using a SPEX 6870 FreezeMill. The frozen cell powder was thawed at room temperature and re-suspended in lysis buffer (50 mM Tris- HCl, pH 8.0, 400 mM NaCl, 0.01% NP-40, 1 mM DTT) with protease inhibitor cocktails (Sigma–Aldrich) and 1 mM benzonamide. The cell lysate was first centrifuged at 27,485g for 1 h in a GSA rotor (Sorvall) and the supernatant was further centrifuged at 167,42g in a 45T rotor (Beckman) at 1.5 h at 4 °C. The supernatant was incubated with 2 ml IgG Sepharose–6 Fast Flow resin (GE Healthcare) overnight at 4 °C. The resin was first washed with IgG washing buffer (20 mM Tris- HCl, pH 8.0, 350 mM NaCl, 0.05% NP-40, 0.5 mM DTT, 1 mM benzamidine and protease inhibitor cocktails), then with buffer containing 250 mM and 150 mM NaCl. The BPB–Mud2 dimer was released by TEV protease in TEV 150 buffer (20 mM Tris- HCl, pH 8.0, 150 mM NaCl, 0.62% NP-40, 0.5 mM DTT). The ACT1 pre-mRNA used consisted of a 73–nt 5’ exon, the 302-nt intron that lacks a cryptic BPS, and a 167–nt 3’ exon. The UBCL pre-mRNA consisted of a 20–nt 5’ exon, a 95–nt intron, and a 32–nt 3’ exon38. The DYN2 wild-type pre-mRNA consisted of three exons (22 nt, 23 nt, and 35 nt in length) separated by two introns (96 nt and 80 nt). The DYN2 IE1 pre-mRNA consisted of intron 1 without the first 9 nt, the middle exon, and intron 2 truncated before the BPS. The EF5M IE1-101 pre-mRNA consisted of intron 1 without the first 11 nt, the middle exon, and intron 2 truncated to 9 nt upstream of the BPS. The HMRA1 IE1-246 wild-type pre-mRNA consisted of intron 1 without the first 10 nt, the entire middle exon, and intron 2 truncated to 2 nt upstream of the BPS. The HMRA1 IE1-246 pre-mRNA was generated after mutating the underlined portion of the last 41 nt of its middle exon from 5’-CAAAGAAATGCTGATCAATTCTCCACTGAATAGGTGTTG-3’ to 5’-ACTAATTGCACATTCTTTCAACCTGAATAGGTGTTG-3’. This modification enables us to use specific primers to detect only exogenous but not endogenous HMRA1 in a wild-type yeast strain. DNA templates for in vitro transcription were generated after the addition of three copies of MS2 stem loops to the 5’-end of the ACT1 gene or to the 3’-end of the UBC4, DYN2, EF5M, and HMRA1 genes, and the templates were generated by in vitro transcription from linearized plasmid DNA templates, and capped using Vaccinia Capping System (New England Biolabs) if indicated.

To obtain the yeast complex E for structural studies, the ACT1 or UBC4 pre-mRNA substrates were bound to the MBP–MS2 fusion protein and mixed with purified U1 snRNP39 and BPB–Mud2 dimer (or BPB–Mud2–CBP in the case of ACT1), then applied to amylose resin (New England Biolabs) pre-washed with buffer G120 (20 mM HEPES, pH 7.9, 120 mM KCl, 0.01% NP-40). After 3 h incubation at 4 °C, the resin was washed and eluted with buffer G120 containing 10 mM maltose. Elutions were pooled and applied to 100 μl calmodulin resin (Agilent) pre-washed with washing buffer (20 mM HEPES, pH 7.9, 120 mM KCl, 2 mM CaCl2, 1 mM imidazole, 0.01% NP-40), and incubated for 3 h at 4 °C. The resin was washed with washing buffer, and eluted six times with 100 μl eluting buffer (20 mM HEPES, pH 7.9, 120 mM KCl, 2 mM EGTA) each time. The elutions containing the most concentrated E complex were used for cryo-EM imaging. Crosslinked samples were prepared by treating the complex with 1 mM BS3 (Thermo Fisher) on ice for 30 min, and subsequent quenching with 50 mM Tris, pH 8.0.

Cryo-EM sample preparation and imaging. For cryo-EM sample optimization, an aliquot of 3 μl of sample (~0.2–0.5 μM) was applied onto a glow-discharged lacey carbon film-coated copper grid (400 mesh, Ted Pella). The grid was blotted with grade 595 filter paper (Ted Pella) and flash-frozen in liquid ethane with a FEI Mark IV IVT. A FEI TF20 cryo-EM instrument was used to record drift correction images for each sample. Drift correction is performed using a RELION-based cryo-EM drift correction method. For the final drift correction step, we used the Relion 3.1 drift correction module. The drift correction parameters were: drift correction for电影 frames. Frames in each movie were aligned for drift correction with the GPU-accelerated program MotionCor249. The first frame was skipped during drift correction because of concern about more severe drift/charging of this frame. Two averaged micrographs, one with dose weighting and the other without dose weighting, were generated for each movie after drift correction. The averaged micrographs have a calibrated pixel size of 1.36 Å on the specimen scale. The averaged micrographs without dose weighting were only used for defocus determination and the averaged micrographs with dose weighting were used for all other steps of image processing.

Structure determination for the ACT1 complex. For the ACT1 complex, the defocus value of each averaged micrograph was determined by CTFFinder41 to range from –1.5 to ~3 μm. Initially, 3,589,121 particles were automatically picked from 11,283 averaged micrographs without reference using Gautomatch (http://www.mrc-lmb.cam.ac.uk/zhhang). The particles were boxed out in dimensions of 352 × 352 square pixels and binned to 176 × 176 square pixels (pixel size of 2.72 Å) for initial 3D classification by the GPU-accelerated RELION2.1. The reported U1 snRNP23 and BBP–Mud2 dimer (or BBP–Mud2–CBP in the case of ACT1) pre-mRNA model (EMD-8622) was low-pass filtered to 60 Å to serve as an initial model for 3D classification. After one round of 3D classification, only the classes exhibiting features characteristic of the E complex (for example, 5’ SS and pre-mRNA helix bound to U1 snRNP) were kept, which contained 1,852,842 particles. Several iterations of reference-free 2D classification were subsequently performed to remove ‘bad’ particles (that is, classes with fuzzy or uninterpretable features), yielding 1,108,069 ‘good’ particles. Auto-refinement of these particles by RELION yielded a map with an average resolution of 5.44 Å (Step 1 in Extended Data Fig. 2c).

Next, we performed two rounds of focused classification on the pre-mRNA helix region of the E complex to further eliminate particles without the pre-mRNA helix (Step 2 in Extended Data Fig. 2c). The first round of this focused classification generated one good class containing 390,792 particles. These particles were unbinned to 352 × 352 square pixels (pixel size of 1.36 Å) and subjected to another round of focused classification. We re-centred the particles from the best class and removed duplications based on the unique index of each particle given by RELION. The 270,587 unbinned, unique particles (7.5% of all particles) resulting from the focused classification were subjected to a final step of 3D auto-refinement (Step 3 in Extended Data Fig. 2c). The two half maps from the final auto-refinement step were subjected to RELION’s standard post-processing procedure. The final map of the ACT1 complex has an average resolution of 3.2 Å based on RELION’s gold-standard Fourier shell correlation (FSC; see Resolution assessment below).

Structure determination for the UBC4 complex. For the UBC4 complex, the defocus value of each averaged micrograph was determined by CTFFinder44 to range from ~1.5 to ~3 μm. Initially, 1,924,710 particles were automatically picked from 8,997 averaged micrographs without reference using Gautomatch. The particles were boxed out in dimensions of 384 × 384 square pixels and binned to 192 × 192 square pixels (pixel size of 2.72 Å) for further processing by the GPU-accelerated RELION2.1. The reported U1 snRNP23 and BBP–Mud2 dimer (or BBP–Mud2–CBP in the case of ACT1) pre-mRNA model (EMD-8622) was low-pass filtered to 60 Å to serve as an initial model for 3D classification. After one round of 3D classification, only classes showing features corresponding to the E complex (for example, 5’ SS binding to U1 snRNP) were kept, which contained 800,735 particles. Several iterations of reference-free 2D classification were subsequently performed to remove bad particles (that is, classes with fuzzy or uninterpretable features), yielding 756,303 good particles (Step 1 in Extended Data Fig. 3c).

Next, we performed another two rounds of 3D classification to further improve the ratio of the intact E complex (that is, particles containing TRP40, NCBP1–NCPB2, and Nam8; ‘Step 2’ in Extended Data Fig. 3c). During each round of 3D classification, only classes showing features corresponding to the intact E complex (for example, SS binding to U1 snRNP) were kept, which contained 689,147 particles. The two half maps from this auto-refinement step were subjected to RELION’s standard post-processing procedure. The final map of the UBC4 complex has an average resolution of 3.2 Å based on RELION’s gold-standard Fourier shell correlation (FSC; see Resolution assessment below).
average resolution of 3.6 Å based on RELION’s gold-standard FSC (see Resolution assessment below).

**Resolution assessment.** All resolutions reported above are based on the gold-standard FSC 0.143 criterion.42 FSC curves were calculated using soft masks and high-resolution noise subtraction was used to correct for convolution effects of the masks on the FSC curves.43 Prior to visualization, all maps were sharpened by applying a negative B-factor, which was estimated using automated procedures44. Local resolution was estimated using ResMap. The C250-pi all-atom models for the ACT1 and UB4 complexes is presented in Extended Data Figs. 2d–f, 3d–f. Data collection and reconstruction statistics are presented in Extended Data Table 1.

**Model building and refinement.** To aid subunit assignment and model building, we took advantage of the reported U1 structure (PDB code: SUZ5, 3.7 Å), which was fitted into the UB4 complex density map by UCSF CHIMERA.46 The central regions of the UB4 complex have resolutions ranging from 3.0 to 4.5 Å (Extended Data Fig. 3f); thus, protein and RNA components in these regions were rebuilt manually using COOT.47 In brief, for protein subunits that matched well with the densities in the UB4 complex structure, we manually adjusted their side chain conformation and, when necessary, moved their main chains to match the density map. For protein subunits that exhibit substantial main chain mismatches or have not been identified, we built atomic models de novo. To do so, sequence assign- ment was mainly guided by visible densities of amino acid residues with bulky side chains, such as Trp, Tyr, Phe, and Arg. Other residues including Gly and Pro also helped the assignment process. Unique patterns of sequence segments containing such residues were used for validation of residue assign- ment.

For the RNA region near the 5′ SS (nt +2 to +8 of pre-mRNA with respect to the exon–intron junction of the U1 snRNA), well-defined nucleotide densities, along with the base pairs between U1 snRNA and pre-mRNA, facilitated the RNA model building process. RNA model building in these regions was performed de novo in COOT. For the central regions of U1 snRNA, the previous U1 snRNA model was adjusted for their base conformation and, when necessary, for their main chains to match the density map. The RNA components were subsequently adjusted using RCrane48 and ERRASER49. Models built for the protein and RNA subunits in these central regions include: U1–70K (mRNA (aa) 1–91), U1C (aa 2–3), U1A (2a – 2b, 125, 133–138), Prp69 (aa 1–544), Prp39 (aa 288–553, 561–627), Nam8 (aa 291–425, 432–449, 492–523), Sm56 (aa 43–170, 185–295), Sm71 (aa 1–52, Luc7 (aa 4–19, 38–140, 172–244), Sm ring; the core regions of U1 snRNA, pre-mRNA (nt –2 to 133–148), Prp42 (aa 1–544), Prp39 (aa 288–553, 561–627), Nam8 (aa 291–425, 432–449, 492–523), Sm56 (aa 43–170, 185–295), Sm71 (aa 1–52, Luc7 (aa 4–19, 38–140, 172–244), Sm ring; the core regions of U1 snRNA, pre-mRNA (nt –2 to +8 with respect to the exon–intron junction) (Extended Data Fig. 4). The long helix interacting with ZnF2 and the coiled-coil domain of Luc7 was traced with poly-alanine, which probably belongs to Sm71 since deletion of the coiled-coil domain of Luc7 reduces its interaction with Sm71 and Prp40 (Extended Data Fig. 7b), but only Sm71 has isolated long helices.

Resolutions for the periphery of the UB4 complex were more varied, ranging from 20 Å (which is equivalent to 28 nucleotides connecting branch point and 5′ SS at nucleotide 70 and 5′ SS at nucleotide +1, not including the branch point and 5′ SS themselves) without highly unfavourable interactions such as clashes. Crosslinking and mass spectrometry. The purified yeast splicedosome E complex was crosslinked with 10 mM DSSO (disuccinimidyl sulfosuccinate) for 45 min at 4°C, and the reaction was quenched by adding ammonium bicarbonate to a final concentration of 50 mM. The crosslinked complex was proteolytically digested according to the FASP (filter-aided sample preparation) protocol as previously described.51 In brief, 100 μg of crosslinked sample was reduced, alkylated, and digested at 150 with sequencing grade trypsin (Promega) by incubating at 37°C for 18 h. Peptides were eluted and acidified to 0.1% formic acid. Enrichment of crosslinked peptides was performed by using strong cation exchange chromatography (SCX) with a Dionex UltiMate 3000 system (Thermo Fisher Scientific). A Proteomix SCX-NP1.7 column (4.6 mm inner diameter, 150 mm length, Sepax Technologies) was used. In brief, peptides were separated in the following gradient: 0% B (0–3.5 min), 0–22.5% B (3.5–18.5 min), 22.5–50% B (18.5–21.5 min), 50–100% B (21.5–23.5 min), 100% B (23–25.5 min) with solvent A (10 mM KH2PO4, 25% acetonitrile, pH 3.00) and solvent B (10 mM KH2PO4, 25% acetonitrile, 500 mM KCl, pH 3.00) at a flow rate of 0.7 mL/min. Fractions were collected every minute. Fractions 6–26 were pooled into groups of three and desalted using StageTips for subsequent liquid chromatography with tandem mass spectrometry (LC–MS/MS) analysis.

Crosslinked peptides were then analysed by nano-ultrahigh performance (UHPLC)-LC–MS/MS (Easy-nLC1200, Orbitrap Fusion Lumos Tribrid, Thermo Fisher Scientific). Sample (14 μL) was loaded directly onto an in-house packed 100 μm i.d. × 250 mm fused silica column packed with CORECTES C18 resin (2.7 μm, spherical solid core). Samples were run at 400 nL/min over a 90–min linear gradient from 0% to 100% solvent B containing 0.1% formic acid. The mass spectrometer was operated in positive ion mode with two sequential experiments per duty cycle. For crosslink peptide identification, MS1 scans were run in the orbitrap from 375 to 1,500 m/z at 60,000 resolution. MS2 was performed in a stoichiometric fashion.
on top ions from each precursor scan and fragmented at a CID collision energy of 22%. MS2 scan mass frequency was determined by a s+5 total duty cycle. MS3 was triggered by the targeted mass difference of 31.9721 Da represented by the cleavage of the DSSO sulfoxide bond, and was performed as a stepped HCD collision energy of 33 ± 3%. For linear peptide identification, a second precursor scan was performed at 120,000 resolution in a scan range of 350–1,000 m/z. Stoichiometric sampling of ions for MS2 fragmentation was capped at 2 s and performed at an HCD collision energy of 29% in the orbitrap. Data acquisition was performed using Xcalibur (version 4.1 software).

Instrument raw files were directly loaded into Proteome Discoverer 2.2 and were searched against 22 proteins making up the E complex of U1 snRNP from S. cerevisiae of the Swiss-prot database (update 2018_08_08) using the XlinkX plugin. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydroxylsed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for plug-in.
The EFMS IEI plasmid and the empty pRS317 vector were transformed into an EFMS deletion yeast strain (Open Biosystem). Yeast cells were grown in SC-Lys medium to OD600 of ~1.0. Total RNA was isolated and reverse transcribed into cDNA. For RNase T1 treatment, 1 μg total RNA was incubated at 37°C for 30 min with 5 U/μg RNase T1 (Epitope Technologies) and used directly for reverse transcription without further purification. PCR was performed using specific primers to detect circRNA formed from exon 2 of EFMS using the following primers: EFMS cDNA forward 5′-GAGAGGATAGATTGTAAATTGACCC-3′ and EFMS reverse 5′-CTTTTGAACTTTCTCAGAGG-3′. The primers used to un-splice EFMS pre-mRNA was: EFMS h forward 5′-TTTCAAACAGTTAGCTAGAATTACATG-3′ and EFMS h reverse 5′-GAGAGGATAGATTGTAAATTGACCC-3′. The products were analysed on 3% LMP agarose gel stained with EtBr.

The region from +116 to +439 of HMRA1 (containing exon 2 flanked by partial intron 1 and intron 2) was PCR amplified from S. cerevisiae genomic DNA and cloned into a pRS317 vector in the same way as the EFMS IEI plasmid. The IEI-62 truncation was engineered to shorten the middle exon to 62 nt in length, with sequence 5′-TTTATAXTGGAAGTAAATTGACCTTGCCAC TACTTTACTCCATCCTAAGTTTGATTCTCATATTACATG-3′. Primers used to detect circRNA were IEI-62 cDNA forward and reverse, which were as the same used as those used for IEI-246. The primer pair used to detect un-spliced HMRA1 pre-mRNA was: HMRA1 h forward 5′-CCCAAGAATGTTGAGCTCAGTTGAGCCAATGAGGGTGGG-3′, HMRA1 h reverse 5′-AAGCCTCAGAGGCGACCTCTCTTCTCAGAGGAGAT AATTTAAGTTTGATTCTCATATTACATG-3′. The products were analysed on 3% LMP agarose gel stained with EtBr.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The coordinate files have been deposited in the Protein Data Bank (EMD-0360 for the EUB4 complex and EMD-0361 for the ACT1 complex). The cryo-EM maps have been deposited in the Electron Microscopy Data Bank (EMD-0360 for the UBC4 complex and EMD-0361 for the ACT1 complex).

37. Li, X. et al. Comprehensive in vivo RNA-binding site analyses reveal a role of Prp8 in spliceosomal assembly. Nucleic Acids Res. 41, 3805–3818 (2013).
38. Abelson, J. et al. Conformational dynamics of single pre-mRNA molecules during in vitro splicing. Nat. Struct. Mol. Biol. 17, 504–512 (2010).
39. Carragher, B. et al. Legionon: an automated system for acquisition of images from vitreous ice specimens. J. Struct. Biol. 132, 33–45 (2000).
40. Zheng, S. Q., Palovcak, E., Armache, J.-P., Cheng, Y. & Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods 14, 331–332 (2017).
41. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. Nat. Methods 9, 853–854 (2012).
42. Chen, S. et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy 135, 24–35 (2013).
43. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J. Mol. Biol. 333, 721–745 (2003).
44. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. Nat. Methods 11, 63–65 (2014).
45. Kappel, K. et al. De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. Nat. Methods 15, 947–954 (2018).
46. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 212–221 (2010).
47. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular cryocrystallography. Acta Crystallogr. D Biol. Crystallogr. 66, 12–21 (2010).
48. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. Protein Sci. 27, 14–29 (2018).
49. Grimm, M., Zimmniak, T., Kahraman, A. & Herzog, F. Xivis: a web server for the visualization and interpretation of crosslinker-dual spatial restraints. Nucleic Acids Res. 43, W632–W639 (2015).
50. Seraphin, B. & Rosbash, M. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell 59, 339–348 (1989).
51. Qin, D., Huang, L., Wlodawer, A., Andrade, J. & Staley, J. P. Sequencing of lariat termini in S. cerevisiae reveals 5′ splice site branches, branch points, and novel splicing events. RNA 22, 237–253 (2016).
52. Li, Z. & Brow, D. A. A rapid assay for quantitative detection of specific RNAs. Nucleic Acids Res. 21, 4645–4646 (1993).
53. Kozlowski, L. P. & Bujnicki, J. M. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13, 111 (2012).

Acknowledgements This work was supported by NIH grants GM121657 and GM130673 (R.Z.); GM071940 and AI094386 (Z.H.Z.); and GM122579, GM121487, and CA219847 (R.D.). S.E. is a Howard Hughes Medical Institute Gilliam Fellow. K.K. was supported by an NSF GRFP award and a Stanford Graduate Fellowship. We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines (supported by UCLA and by grants from the NIH (1S1000D018111.1U24GM116792) and NSF (DBI-1338135 and DMR-1548924)) as well as the CU Anschutz School of Medicine Cryo-EM and proteomics core facilities (partially supported by the School of Medicine and the University of Colorado Cancer Center Support Grant P30CA046934). Molecular graphics and analyses were performed with the UCSF Chimera and ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco from coordinates link from NIGMS P41-GM103311 (Chimera, ChimeraX) and NIH RO1-GM129325 (ChimeraX). We also thank M. Ares, D. Black, and D. Brow for comments on early versions of the manuscript.

Author contributions X.L. and S.L. contributed equally to the work and are listed alphabetically in the author list. R.Z. and Z.H.Z. conceived the project; X.L. prepared and optimized the sample; X.L., L.Z., S.E. and S.S. performed biochemical analyses; S.L. and Y.C. recorded and processed the EM data; A.L., R.G.H. and K.C.H. performed mass spectrometry analyses; S.L. built the atomic models; K.K. and R.D. built the partial U1 snRNA model and the minimal exon model in the A complex; R.Z., S.L., X.L., and Z.H.Z. analysed and interpreted the models; S.L., X.L. and R.Z. prepared the illustrations; R.Z., S.L. and Z.H.Z. wrote the paper; and all authors contributed to the editing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1523-6. Correspondence and requests for materials should be addressed to Z.H.Z. or R.Z. Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | In vitro assembly and purification of the ACT1 complex. a, Schematic representation of the ACT1 pre-mRNA tagged with three MS2-binding sites (M3–ACT1) used for E complex assembly and purification. Boxes represent exon 1 (E1) and truncated exon 2 (E2). The 5′ SS (GU) and BPS (UACUAAC) are also shown. The red line represents the DNA oligo complementary to a region 5 nt upstream of the BPS for the RNase H cleavage experiment. b, RNA components of the assembled E complex (with or without DNA oligo and RNase H treatment) after proteinase K digestion are shown on a denaturing urea gel or native agarose gel. These results demonstrate that RNase treatment cleaved M3–ACT1 into two fragments. Note that the sizes of RNA on the native gel do not match their linear length, possibly owing to the existence of secondary structures. This experiment was repeated two additional times with similar results.
Extended Data Fig. 2 | The cryo-EM structural determination process for the ACT1 complex. a, Representative drift-corrected cryo-EM micrograph (out of 11,283 micrographs) of the E complex assembled on the ACT1 pre-mRNA. A representative particle is shown in a white dotted circle. b, Representative 2D class averages of the ACT1 complex obtained in RELION. This experiment was repeated one additional time with similar results. c, Data processing workflow. For processing above the red dashed line, the particle images were binned to a pixel size of 2.72 Å. The rest of the processing was performed with a pixel size of 1.36 Å. The masks used in data processing are outlined with red solid lines (see Methods). d, Angular distribution of all particles used for the final 3.2 Å map of the ACT1 complex. e, FSC as a function of spatial frequency demonstrating the resolution of the final reconstruction of the ACT1 complex. f, Resmap local resolution estimation. g, FSC coefficients as a functional of spatial frequency between model and cryo-EM density maps. The generally similar appearances between the FSC curves obtained with half maps with (red) and without (blue) model refinement indicate that the refinement of the atomic coordinates did not suffer from severe over-fitting.
Extended Data Fig. 3 | The Cryo-EM structural determination process for the UBC4 complex. a, Representative drift-corrected cryo-EM micrograph (out of 8,997 micrographs) of the E complex assembled on the UBC4 pre-mRNA. A representative particle is shown in a white dotted circle. b, Representative 2D class averages of the UBC4 complex obtained in RELION. c, Data processing workflow. For processing above the red dashed line, the particle images were binned to a pixel size of 2.72 Å. The rest of the processing was performed with a pixel size of 1.36 Å. The masks used in data processing are outlined with red solid lines (see Methods). d, Angular distribution of all particles used for the final 3.6 Å map of the UBC4 complex. e, FSC as a function of spatial frequency demonstrating the resolution of the final reconstruction of the UBC4 complex. f, Resmap local resolution estimation. g, FSC coefficients as a functional of spatial frequency between model and cryo-EM density maps. The generally similar appearances between the FSC curves obtained with half maps with (red) and without (blue) model refinement indicate that the refinement of the atomic coordinates did not suffer from severe over-fitting.
Extended Data Fig. 4 | Representative cryo-EM density maps of the E complex. a–i, Densities for the UBC4 complex; j, density for the ACT1 complex. Cryo-EM density maps are shown as follows. a, Selected regions of U1 snRNA. b, C-terminal region of Prp39. c, N-terminal domain of Snu71. d, Pre-mRNA and U1 snRNA duplex. e, U1C ZnF domain. f, Luc7 ZnF2 domain. g, Tandem FF domains of Prp40 (the known structure of tandem FF domains from CA150 is also shown with the characteristic boomerang shape). h, RRM2 domain of Nam8. i, NCBP1 and NCBP2. j, Weak density in the ACT1 complex that is assigned as the putative BBP–Mud2 heterodimer. The A complex is also shown, with U1 snRNP in the same orientation as the ACT1 complex and U2 snRNP located in similar positions as the BBP–Mud2 heterodimer with respect to U1 snRNP. The map of the ACT1 complex was low-pass filtered to 40 Å.
Extended Data Fig. 5 | Structural and biochemical characterization of the ACT1 and UBC4 complexes. a, Comparison of the ribbon models of the ACT1 complex, the UBC4 complex, and U1 snRNP from other previously determined structures (the U1 snRNP, A, and pre-B complexes). Labels with shading indicate protein or RNA components that differ between the ACT1 and UBC4 complexes. These components and the RRM2 domain of Nam8 are also absent from previously determined structures. Note that U1–70K is shifted towards NCBP2 in the UBC4 complex. b, Purified E complex does not contain U2 snRNA. A native polyacrylamide gel shows the solution hybridization result of total cellular RNA or RNA from purified E complex hybridized with fluorescent probes specific for U1 and U2 snRNAs. This experiment was repeated one additional time with similar results.
Extended Data Fig. 6 | Secondary structures in the region between the 5′ SS and BPS in the wild-type and mutant ACT1 and UBC4 pre-mRNAs. a, Secondary structures predicted by RNAstructure 6.0 (https://rna.urmc.rochester.edu/RNAstructureWeb/). b, Sequence between the 5′ SS and BPS (underlined) of ACT1. Red nucleotides were mutated to A (other than the one A, which was mutated to G) in mutant ACT1 to disrupt predicted secondary structures.
Extended Data Fig. 7 | Protein interactions in the UBC4 complex. a, DSSO crosslinking and mass spectrometry analyses of the UBC4 complex. Each blue line indicates a crosslink between a pair of Lys residues. Note that BBP–Mud2 are crosslinked to Luc7, Prp40, Snu56, and Snu71. b, Co-purification assays probing the interaction between Snu71 (or Prp40) and Luc7. Various combinations of protein A–TEV–Prp40, protein A–TEV–Snu71, and CBP-tagged Luc7 or Luc7ΔCC (with coiled-coil domain (residues 123–190) deleted) were co-overexpressed in yeast (only Snu71 is protein A tagged in the Snu71 + Prp40 lanes), purified using IgG resin, eluted through TEV cleavage, analysed on SDS–PAGE, and visualized using western blot with an anti-CBP antibody to detect Luc7 (top) and Ponceau S stain to show Snu71 or Prp40 (middle). Western blot using the same anti-CBP antibody was used to demonstrate Luc7 expression levels in cell lysates (bottom). The faint band around 26 kD in all lanes of the middle gel is TEV. This experiment was repeated one additional time with similar results. c, The linker (residues 73–131) between the WW and FF domains of Prp40 is predicted to be disordered using program MetaDisorderMD259.
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | Computational, biochemical, and structural characterization of the EDC. a, The minimal length of RNA needed to connect the upstream branch point (BP) and downstream 5′ SS in the A complex is modelled using the Rosetta RNP-denovo method. The A complex (PDB ID 6G90) is shown in grey. The pre-mRNA is shown in green. The upstream branch point and downstream 5′ SS are shown as purple space-filling models. Twenty-eight nucleotides are sufficient to connect the upstream branch point and downstream 5′ SS (not including the branch point and 5′ SS themselves) without any chain break or clashes. b, Schematics of wild-type and mutant DYN2 pre-mRNA (mutated nucleotides shown in red), IEI, and untagged IEI used for the EDC assembly and in vivo exon definition experiments. Stem-loops represent the MS2 binding sites, and the red line represents the DNA oligonucleotide used for RNase H cleavage. c, SDS–PAGE shows protein components of complexes assembled on wild-type and IEI substrates (lanes 1, 2), on wild-type in the presence of competing untagged IEI (lane 3), and on IEI after RNase H treatment in the absence and presence of the DNA oligo (lanes 4, 5). This experiment was repeated one additional time with similar results. d, RNA components of the same complexes as in lanes 4, 5 of c, confirming that RNase H treatment in the presence of the oligonucleotide cleaves the pre-mRNA. The smaller cleaved fragment (61 nucleotides) is difficult to see because EtBr stains short single-stranded RNA with low efficiency. This experiment was repeated two additional times with similar results. e, Mass spectrometry analyses of spliceosome assembled on the IEI and wild-type DYN2 pre-mRNA indicate that the two complexes have the same components in similar quantities with the exception of NCBP1 and 2, which are absent from the IEI complex. f, 2D classification of negative-stain TEM images of the E complex assembled on DYN2 IEI pre-mRNA. This experiment was repeated one additional time with similar results.
**Extended Data Fig. 9** | See next page for caption.

**a**

EFM5 PCR product generated from circRNA using outward facing primers on exon2 (Fig. 5a, lane 9):

...tatcagAACACCTGAACATCCAGTGTAAGTGAACAAAGAGTTGATAGATTGTATAATGGATCCACCTTTT
TTAATGAAGATTTGGCAAACAAAGGTGACACTTTCTGCTTAAGCGCTCCTGGGCTTTGAAGAGATTTGAAGAG
AGGAAACACAAATCAAGAGCCCTTCTAACAAAGCTTTACGACGatcatg...

EFM5 PCR product generated from T-branch (Fig. 5a, lane 5):

...catgatTTTTCAACACATCAAGCTGAAATTACTAATCTIGATGTATATCATATAAACACAT
ATCCCTACTCAATTTTTAATCTTTTCCAGTATCTTTTCTCTTTACTAATCTTCTgata...

HMRA1 PCR product generated from circRNA using outward facing primers on exon2 (Fig. 5a, lane 10):

...tatcagGAAAGCAGCCAAGCTTAAATCAAGAGGAAAGATGTGGGCAATTACATTTGACTAAAGTAGAGCAACATACATTACACAAATATT
TCAACAAATAGGTAGAATATACCACCATATTTAAAAGAGAGGAGGCCAAGGAAAtcatg...

HMRA1 PCR product generated from T-branch (Fig. 5a, lane 7):

...catgatGTATGTGTTTCAACAGGATAAGGCTCCCTTGAATCTTTACACTTATC--AAATGATGTAT
ATGAGAATCCAACTTTAATATATCTTCTACTACActgata...

**b**

![Image of gel electrophoresis](image)

| Marker (bp) | - | + | RNase R |
|-------------|---|---|--------|
| 1           |   |   |        |
| 2           |   |   |        |
| 3           |   |   |        |

**RPP1B**

**c**

| MW (kDa) | 190 | 135 | 100 |
|----------|-----|-----|-----|

**EFM5 IEI-101-M3**

| U1 snRNA |
|----------|
| 750 |

**marker (nt)**

| U1-70K | 32 |
| U1A-CBP | 25 |
| Sno71 | 22 |
| Prp40 |  |
| Prp39 |  |
| BBP |  |
| Nm5 |  |
| Mps |  |
| MBP-M52 |  |
| Cm |  |
| Luc |  |
| U1C |  |
| Sm |  |

**Dye-stained proteins on SDS-PAGE**

**EtBr stained RNA on Urea gel**

**d**

Yeast strain: **BY4742**

**Plasmid:**

| Marker (bp) | RNase R - | RNase R + |
|-------------|-----------|-----------|
| 1           |           |           |
| 2           |           |           |
| 3           |           |           |

**Primers:**

| Primers | 5'–3' | 5'–3' |
|----------|-------|-------|
| HMRA1 exon 2 |       |       |
| HMRA1 exon 2 |       |       |

**e**

| Marker (bp) | - | + | - | + | - | + | - | + | - | + |
|-------------|---|---|---|---|---|---|---|---|---|---|
| 1           |   |   |   |   |   |   |   |   |   |   |
| 2           |   |   |   |   |   |   |   |   |   |   |
| 3           |   |   |   |   |   |   |   |   |   |   |
| 4           |   |   |   |   |   |   |   |   |   |   |
| 5           |   |   |   |   |   |   |   |   |   |   |

**Competing IEI-246 WT (30x)**

**Assembled E complex**

**IEI-246-M3 RNA**

**U1 depletion**

**CircRNA product from IEI-246-M3 (+ RNase R)**

**U1 (No Rnase R)**
Extended Data Fig. 9 | Characterization of circRNAs. a, Sanger sequencing confirmed that the PCR products in Fig. 5a were derived from T-branches and circRNAs of EFM5 and HMRA1. Solidus, site where two ends of exon 2 are ligated; vertical line, site where the 5′ SS of intron 2 is ligated to the BP of intron 1. The 5′ SS and BPS are shown in bold. The BPS contains deletions (shown as -) due to errors caused by reverse transcriptase reading through the branch. b, RT–PCR was carried out on RNA extracted from wild-type yeast cells with or without RNaseR treatment using primers indicated in the schematic diagrams below the gel, indicating that RNase R treatment eliminates linear RNAs. This experiment was repeated four additional times with similar results. c, Protein and RNA components of E complex assembled on EFM5 IEI-101–M3 pre-mRNA. d, RT–PCR of RNA extracted from BY4742 yeast strain carrying indicated HRMA1 plasmids, with or without RNaseR treatment, using primers shown in the schematic diagrams below the gel. Numbers 246 and 62 designate exon lengths. Lanes 1–3 indicate that all constructs were transcribed (endogenous HMRA1 pre-mRNA level is too low to be detected as indicated in lane 3). The HMRA1 middle exon was slightly modified to create a circRNA primer binding site so that only the modified exogenous (for example, IEI-246 in lane 5) but not wild-type HMRA1 circRNA (IEI-246 WT in lane 4) could be detected. e, IEI-246–M3 RNA or E complex assembled on IEI-246–M3 was incubated with wild-type or U1-depleted yeast extract in the absence or presence of 30-fold excess competing IEI-246 wild-type RNA. CircRNA products were monitored using RT–PCR as in d. Experiments in c–e were repeated one additional time with similar results.
Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics.

|                         | Ubc4 complex (EMD-0360) | Act1 complex (EMD-0361) |
|-------------------------|--------------------------|--------------------------|
|                         | (PDB 6N7P)               | (PDB 6N7R)               |
| **Data collection and processing** |                          |                          |
| Magnification           | 105,000                  | 105,000                  |
| Voltage (kV)            | 300                      | 300                      |
| Electron exposure (e^-/Å²) | 34.6                     | 29.4                     |
| Defocus range (μm)      | -1.5 ~ -3.0              | -1.5 ~ -3.0              |
| Pixel size (Å)          | 1.36                     | 1.36                     |
| Symmetry imposed        | C1                       | C1                       |
| Initial particle images (no.) | 1,924,710              | 3,589,121                |
| Final particle images (no.) | 124,825                | 270,587                  |
| Map resolution (Å)      | 3.6                      | 3.2                      |
| FSC threshold           | 0.143                    | 0.143                    |
| Map resolution range (Å) |                          |                          |
| Core                    | 3.0-4.5                  | 3.0-4.5                  |
| Pre-mRNA helix          | ---                      | 3.0-6.5                  |
| Prp40                   | 15-20                    | ---                      |
| Nam8                    | 15-20                    | ---                      |
| NCBP5s                  | 15-25                    | ---                      |
| U1 snRNA                | 6-15                     | 6-15                     |
| **Refinement**          |                          |                          |
| Initial model used (PDB code) | n/a                     | n/a                      |
| Model resolution (Å)    | 4.6                      | 4.3                      |
| FSC threshold           | 0.5                      | 0.5                      |
| Model resolution range (Å) | 4.6                     | 4.3                      |
| Map sharpening B factor (Å²) | -147.1                 | -94.0                    |
| Model composition       |                          |                          |
| Non-hydrogen atoms      | 41487                    | 35784                    |
| Protein residues        | 4839                     | 3568                     |
| Ligands                 | 3                        | 1                        |
| B factors (Å²)          |                          |                          |
| Protein                 | 57.5                     | 59.9                     |
| Ligand                  | 85.3                     | 79.6                     |
| R.m.s. deviations       |                          |                          |
| Bond lengths (Å)        | 0.01                     | 0.02                     |
| Bond angles (°)         | 1.35                     | 1.50                     |
| **Validation**          |                          |                          |
| MolProbity score        | 1.89                     | 1.99                     |
| Clashscore              | 5.09                     | 5.07                     |
| Poor rotamers (%)       | 1.50                     | 2.72                     |
| Ramachandran plot       |                          |                          |
| Favored (%)             | 92.05                    | 94.21                    |
| Allowed (%)             | 7.35                     | 5.37                     |
| Disallowed (%)          | 0.60                     | 0.42                     |
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
- Lego

Data analysis
- Gautomatch_v0.53, CTFFIND4, RELION2.1, UCSF Chimera and ChimeraX, Resmap1.95, MotionCor2, Coot0.8.3, RCrane, ERRASER, DRRAFTER, Xcalibur, Rosetta (2018.33.60351), PHENIX, Proteome Discoverer 2.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The atomic models and the cryoEM density maps have been deposited to the Protein Data Bank and the Electron Microscopy Data Bank, under the accession numbers of 6N7P and EMD-0360 for the Ubc4 complex, and 6N7R and EMD-0361 for the Act1 complex.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
3D reconstructions were calculated from 11,283 images (3.6 million particles) for the Act1 complex, and 8,997 images (1.9 million particles) for the Ubc4 complex. These are typical image sizes used to obtain high resolution cryoEM structures.

Data exclusions
For cryoEM analysis, particles that do not belong to the class of interest or have poor qualities based on well established cryoEM principle were excluded after rounds of 2D and 3D classification. This is standard practice required to obtain high resolution cryo EM structure of the class of interest. For functional studies, no data were excluded from any analysis.

Replication
All biochemical experiments were repeated two or more times and are all reproducible.

Randomization
No grouping required for our studies.

Blinding
Since there is no grouping, there is also no blinding with respect to the grouping.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a | Involved in the study |
| ☒ ☒ Antibodies | ☒ ChIP-seq |
| ☒ ☒ Eukaryotic cell lines | ☒ Flow cytometry |
| ☒ Palaeontology | ☒ MRI-based neuroimaging |
| ☒ Animals and other organisms | |
| ☒ Human research participants | |
| ☒ Clinical data | |

Antibodies

Antibodies used
Anti-CBP antibody, GenScript, catalog # A00635.

Validation
The antibody was also validated via knockdown by the manufacturer. We also verified it using yeast cells didn’t expressing a CBP tagged Luc7. The western blot result for it was negative. It was further verified by the result showing that the CBP-tagged truncated Luc7 gave bands smaller than the wild type Luc7.