ARTICLE

https://doi.org/10.1038/s41586-019-1523-6

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

Xueni Li[1,8], Shiheng Liu[2,3,8], Lingdi Zhang[1], Aaron Issaian[1], Ryan C. Hill[1], Sara Espinosa[1], Shasha Shi[1], Yanxiang Cui[3], Kalli Kappel[4], Rhiju Das[4,5,6], Kirk C. Hansen[1], Z. Hong Zhou[2,3]* & Rui Zhao[1,7]*

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 *EFM5* 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 dominate[3]. On the other hand, exon definition[4] 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 composition[5,6], 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 pathway[7]. CircRNAs are involved in the regulation of their host genes or microRNAs, ageing, and other disease processes[8]. Although canonical splicing signals and the spliceosome are needed for production of circRNAs[9], 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 heterogenous 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 being 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.

[1]Department of Biochemistry and Molecular Genetics, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [2]Department of Microbiology, Immunology, and Molecular Genetics, UCLA, Los Angeles, CA, USA. [3]Electron Imaging Center for Nanomachines, UCLA, Los Angeles, CA, USA. [4]Biophysics Program, Stanford University, Stanford, CA, USA. [5]Department of Biochemistry, Stanford University, Stanford, CA, USA. [6]Department of Physics, Stanford University, Stanford, CA, USA. [7]RNA Bioscience Initiative, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [8]These authors contributed equally: Xueni Li, Shiheng Liu. *e-mail: hong.zhou@ucla.edu; rui.zhao@cuanschutz.edu

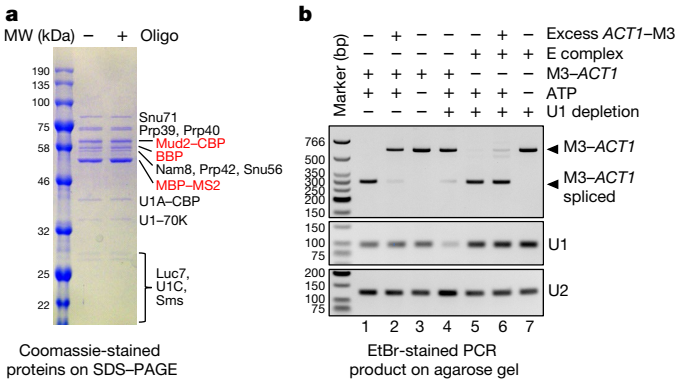


Fig. 1 | In vitro-assembled E complex is functional. a, The E complex assembled on M3–*ACT1* (with or without DNA oligo-directed RNase H treatment to cleave between the 5′ SS and BPS) is affinity purified and its protein components shown. **b**, Yeast splicing extract with or without U1 snRNA depletion is incubated with in vitro-transcribed M3–*ACT1* or E complex assembled on M3–*ACT1* in the presence or absence of ATP or excess *ACT1*–M3. The splicing outcome is monitored using RT–PCR with primers located in the MS2 binding site region and exon 2 of *ACT1* (top gel). The middle and bottom gels show levels of U1 and U2 snRNA in each sample. Experiments in Fig. 1 were repeated two additional times with similar results. For gel source data for all Figures, see Supplementary Fig. 1.

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 structures[14,15]. 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 introns[16,17]. 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 Nam8[18].

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 snRNA[15]. Our *ACT1* complex is assembled 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* reporter[19], 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 complexes[20] 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.

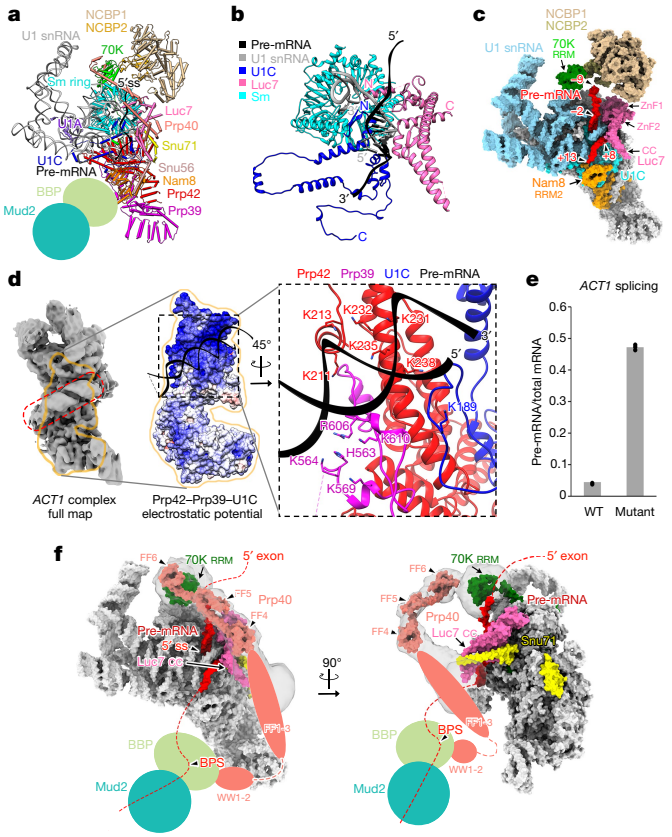


Fig. 2 | Cryo-EM structure of the E complex. a, The overall E complex structure containing features from both the *ACT1* and *UBC4* complexes. BBP and Mud2 are not modelled owing to their weak density, but their locations are indicated. **b**, Ribbon diagrams of protein and RNA models immediately around the 5′ SS. **c**, Surface representation of proteins that are in close proximity to the 5′ SS (coloured), other proteins (grey), and U1 snRNA (cyan). Pre-mRNA is shown in red and nucleotide positions relative to the 5′ SS are labelled (−1 and +1 denote the last nucleotide of the exon and the first nucleotide of the intron, respectively). **d**, Secondary structure in pre-mRNA. Left, cryo-EM density map (filtered to 6 Å) of the entire E complex showing density (in red dashed box) for the pre-mRNA double helix. Middle, electrostatic potentials of the binding surface for the pre-mRNA double helix. Right, the binding surface formed by Prp39, Prp42, and U1C in ribbon diagrams. Positively charged residues on Prp39 and Prp42 that interact with this double helix are shown in sticks. **e**, Splicing efficiency of wild-type (WT) *ACT1* intron and the mutant that disrupts the secondary structure in the 5′ SS-to-BPS region in an *ACT1–CUP1* reporter plasmid, as evaluated by quantitative RT–PCR. Dots represent three technical replicates. This experiment was repeated two additional times with similar results. **f**, Surface representation of proteins that interact or possibly interact with Prp40 in different colours. To simplify the figure, NCBPs are not shown. Locations of proteins or protein domains that are not modelled owing to weak densities are indicated by shapes. Transparent grey areas are 8 Å low-pass filtered densities showing likely contacts between Prp40 and U1–70K. Red dashed lines represent hypothetical paths of the pre-mRNA.

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 process[13]. 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 helices[21,22] (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

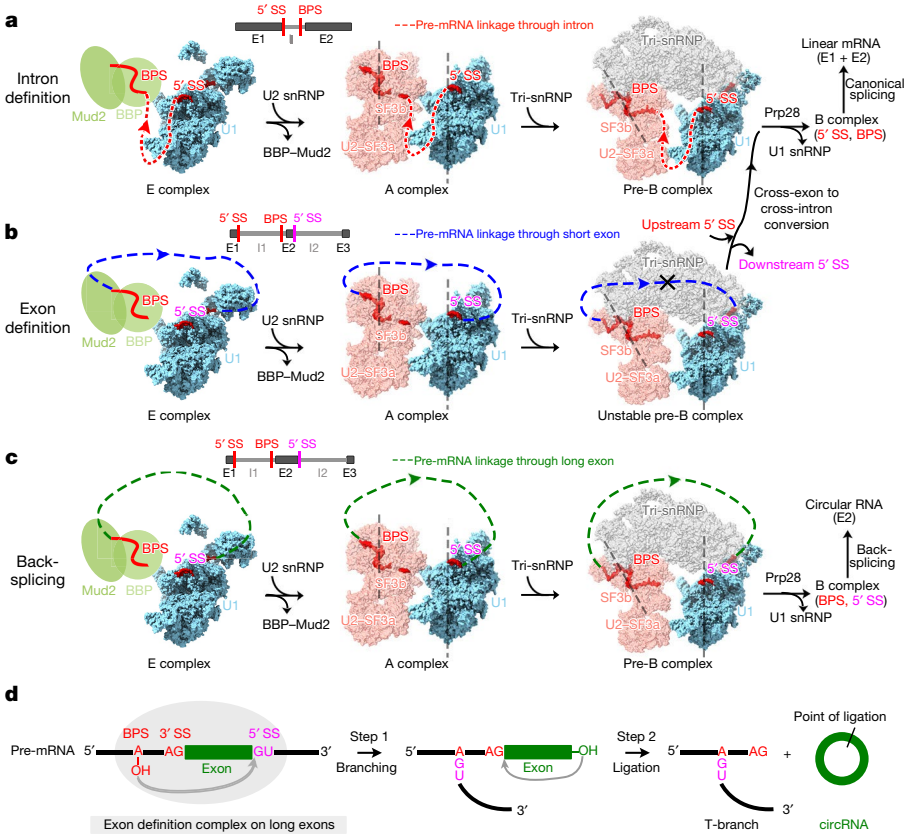


Fig. 3 | A unified model for intron definition, exon definition, and back-splicing. a, Structures of the E, A, and pre-B complexes in surface representations with U1, U2, and tri-snRNPs in different colours, illustrating the canonical assembly pathway across an intron. Pre-mRNA is shown in red with an arrow indicating the 5′-to-3′ direction. Red dashed line indicates the hypothetical path of intron connecting the 5′ SS and downstream BPS. Vertical dashed lines denote the orientation of U1 snRNP and U2 SF3b in the A complex. In the pre-B complex, the orientation of U1 snRNP remains the same but that of U2 SF3b is tilted by about 30°. **b**, The same spliceosomal E and A complexes as in **a** can assemble across an exon, but cannot form the pre-B complex on short exons owing to steric hindrance. Blue dashed line indicates the hypothetical path of an exon connecting the BPS and downstream 5′ SS. **c**, As in **b**, but with a long exon (green dashed line), illustrating that the EDC on long exons can catalyse back-splicing. **d**, A schematic showing how the EDC on a long exon carries out back-splicing and generates circRNA through the same transesterification reactions used by canonical splicing.

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–Luc7[23–25], and the Prp40 WW domains interact directly with the N-terminal domain of BBP[13,26]. BBP also forms non-exclusive interactions with both Prp40 and Mud2[13], and binds directly to the BPS of pre-mRNA[27]. 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 structure[14] (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 that are large enough to fill the volume of this density. This density is not obvious in the *UBC4* complex structure, potentially because *UBC4* lacks the pre-mRNA helix that brings the BPS close to the 5′ SS. Prp40 therefore bridges both ends of the intron by interacting with U1–70K and Snu71–Luc7 of the U1 snRNP through its FF domains, and interacting with BBP through its WW domains. The entire approximately 60-residue linker region between the WW and FF domains is predicted to be disordered (Extended Data Fig. 7c), explaining why density corresponding to BBP–Mud2 is difficult to observe.

Exon definition occurs in yeast

The architecture of the E complex, in particular the relative positions between the 5′ SS and the BPS where BBP binds, suggests that the same E complex can form across an exon. Instead of connecting the upstream 5′ SS to a downstream BPS through an intron (Fig. 3a), the BPS can be connected to a downstream 5′ SS through an exon (Fig. 3b). Similarly, the A complex structure[14] suggests that the same A complex could also form across exons (Fig. 3b). Modelling using the Rosetta RNP-denovo method[28] 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 SF3b[14,15] (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). This may signal to the spliceosome that an EDC has formed and provide an opportunity for the upstream 5′ SS to interact with the tri-snRNP to form an intron-spanning B complex (Fig. 3b). We use EDC to refer to spliceosomal complexes assembled across an exon, which can be an exon-defined E, A, or unstable pre-B complex.

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* IEI, Extended Data Fig. 8b).

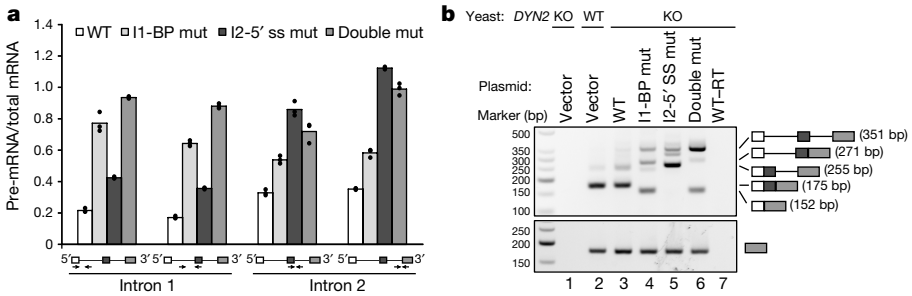


Fig. 4 | Exon definition occurs in yeast. a, A plasmid containing the wild-type *DYN2* gene or various mutants was transformed into a *DYN2* KO strain. The splicing efficiency of introns 1 and 2 was evaluated using quantitative RT–PCR with primers specific for intron 1 or intron 2 (indicated by arrows in the schematics under the bar diagram) normalized to total mRNA. Dots represent three technical replicates. **b**, RT–PCR of RNA extracted from yeast strain carrying indicated plasmids, using primers located in exons 1 and 3 of *DYN2*. Schematics of the splicing products and their expected sizes are shown on the right. RT–PCR products using primers in exon 3 (bottom gel) serve as an internal quality control of the samples. Experiments in **a** and **b** were repeated two additional times with similar results.

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 vertebrates[4]. 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 impaired 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 observations[29]. 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 skipping[29]. Together, our results demonstrate that exon definition occurs for a fraction of *DYN2* transcripts in vivo in yeast.

The EDC catalyses back-splicing 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 back-spliced 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 products[7].

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 release[30]. As expected, purified spliceosomes contained spliced mRNA and lariat for the yeast single-intronic gene *RPP1B*, as well as the unique T-branch and circRNA for the multi-intronic genes *EFM5* and *HMRA1* (Fig. 5a, Extended Data Fig. 9a). These results show that Cef1-purified spliceosome contains both canonical and back-splicing products.

Further supporting this model (Fig. 3c, d), we showed using PCR with reverse transcription (RT–PCR) that the *EFM5* 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 *EFM5* 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 IEI-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 *EFM5* or *HMRA1*, and that this catalyses back-splicing 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 opportune point 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 complex[6]. 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

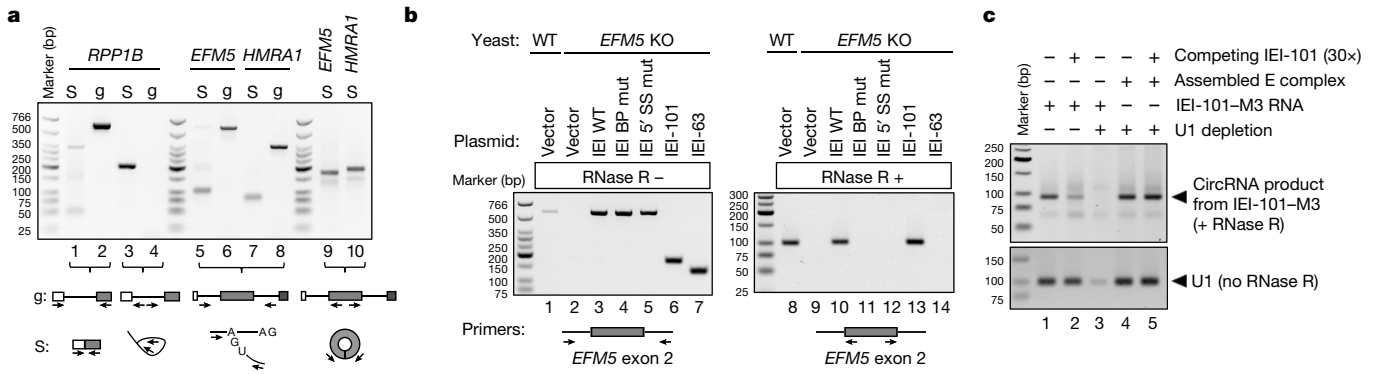


Fig. 5 | The EDC catalyses back-splicing and produces circRNA. a, RT–PCR of RNA isolated from spliceosome purified from the Prp22(H606A) yeast strain (S) and PCR using yeast genomic DNA (g, as negative control) for the single-intron gene *RPP1B* and multi-intron genes *EFM5* and *HMRA1* demonstrates 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 *EFM5* 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 *EFM5* 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.

in most or all species, even on the same pre-mRNA. Whether intron or exon definition is dominant in vivo is likely to be determined by gene architecture (such as the length of introns or exons) and other factors (such as exonic or intronic enhancers or suppressors and their associated proteins, RNA secondary structures, transcription processivity, and nucleosome positioning).

CircRNA generated by back-splicing of exons has attracted increasing attention, but its origin and biogenesis have largely remained a mystery[8]. Although exon definition has been speculated to play a role in back-splicing[31,32], it is unclear which of the canonical spliceosomal components are required and what exact signals are being recognized that make an exon form circRNA instead of participating in canonical splicing. Our results demonstrate that back-splicing is catalysed by exon-definition complexes on long exons (or multiple exons) that are not remodelled to intron-spanning complexes, suggesting that circRNA is a natural byproduct of spliceosome-mediated splicing in all eukaryotic species. This model is consistent with a previous proposal based on competition between back-splicing and canonical splicing[31] and with the observation that long but not short exons (when flanked by the same intronic sequences) can form circRNAs in human cells[33]. Indeed, the average exon length in circRNA is 690 nt[34], much longer than the median length of 120 nt for human exons[35]. The long exon inevitably lowers the efficiency of initial exon definition, contributing to the low frequency of back-splicing and circRNA production. Intronic complementary sequences flanking the exon and RNA-binding proteins potentially increase the efficiency of initial exon definition and facilitate circRNA production[8]. These RNA elements or proteins may also bring opposite ends of different exons close together for back-splicing, generating circRNAs that contain multiple exons. It is worth noting that, accurately speaking, the distance between the upstream branch point and downstream 5′ SS across an exon, instead of the exon length in yeast pre-mRNAs, determines the fate of the EDC, because the 3′ SS is not recognized in early yeast spliceosomes. However, given the generally short distance between yeast branch points and 3′ SS (14 nt for *EFM5* and 10 nt for *HMRA1* intron 1 not including the branch point and 3’ SS themselves)[36], exon lengths ultimately play a major role in determining the outcome of EDC remodelling.

In summary, our E complex structure enabled us to propose a simple model that unifies intron definition, exon definition, and back-splicing, without needing a different spliceosome for each process. This model is supported by our biochemical analyses, performed exclusively in yeast, which is now positioned to serve as a well-defined model system to understand exon definition or back-splicing. The core of this model is likely to hold true for all eukaryotes, although many *cis* or *trans* factors (including RNA, protein, transcription, nucleosome and so on) may act as modulators to promote or suppress a particular process. For example, most vertebrate exons are short, which is likely to be the main signal for EDC remodelling, but other factors may facilitate remodelling of EDCs assembled on long exons and lower the efficiency of back-splicing. This model should inspire experiments in many other systems to understand the mechanism and regulation of exon definition and back-splicing, some of the most fundamental unanswered questions in pre-mRNA splicing.

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: 04 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**, 339–351 (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. *J. Biol. Chem.* **270**, 2411–2414 (1995).
5. Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C. & Black, D. L. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.* **15**, 183–191 (2008).
6. Schneider, M. et al. Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic 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. Starke, 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. 5′ 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 spliceosome 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* spliceosome 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. Qiu, Z. R., Chico, L., Chang, J., Shuman, S. & Schwer, B. Genetic interactions of hypomorphic mutations in the m7G cap-binding pocket of yeast nuclear cap binding complex: an essential role for Cbc2 in meiosis via splicing of *MER3* pre-mRNA. *RNA* **18**, 1996–2011 (2012).
18. 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).
19. 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).
20. Liu, S. et al. Structure of the yeast spliceosomal postcatalytic P complex. *Science* **358**, 1278–1283 (2017).
21. 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).
22. Liu, J., Fan, S., Lee, C. J., Greenleaf, A. L. & Zhou, P. Specific interaction of the transcription elongation regulator TCERG1 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).
23. Li, X. et al. CryoEM structure of *Saccharomyces cerevisiae* U1 snRNP offers insight into alternative splicing. *Nat. Commun.* **8**, 1035 (2017).
24. Görnemann, J. et al. Cotranscriptional spliceosome assembly and splicing are independent of the Prp40p WW domain. *RNA* **17**, 2119–2129 (2011).
25. Ester, C. & Uetz, P. The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. *BMC Biochem.* **9**, 29 (2008).
26. Wiesner, S., Stier, 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).
27. Jacewicz, A., Chico, L., Smith, P., Schwer, B. & Shuman, S. Structural basis for recognition of intron branchpoint RNA by yeast Msl5 and selective effects of interfacial mutations on splicing of yeast pre-mRNAs. *RNA* **21**, 401–414 (2015).
28. Kappel, K. & Das, R. Sampling native-like structures of RNA-protein complexes through Rosetta folding and docking. *Structure* **27**, 140–151.e145 (2019).
29. Howe, K. J., Kane, C. M. & 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., Shirokikh, 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. Jeck, W. R. et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
35. Mokry, M. et al. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res.* **38**, e116 (2010).
36. Spingola, M., Grate, L., Haussler, D. & Ares, M., Jr. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5**, 221–234 (1999).

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 *BBP* and *MUD2* were amplified by PCR using genomic *S. cerevisiae* DNA as templates. *BBP* fused to an N-terminal protein A (protA) tag was inserted between a *GPD* promoter and a *CYC1* terminator, and the resulting expression cassette was cloned into pRS414 to generate the pRS414/GPD-protA-*BBP* 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 OD_{600} = 3–4. The cells were flash-frozen in liquid nitrogen to form yeast 'popcorns' and cryogenically ground using a SPEX 6870 Freezer/Mill. 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, 0.1% NP-40, 1 mM DTT) with protease inhibitor cocktails (Sigma-Aldrich) and 1 mM benzamidine. The cell lysate was first centrifuged at 27,485*g* for 1 h in a GSA rotor (Sorvall) and the supernatant was further centrifuged at 167,424*g* in a 45Ti rotor (Beckman) for 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 BBP–Mud2 dimer was released by TEV protease in TEV150 buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.02% 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[37]. The *UBC4* pre-mRNA consisted of a 20-nt 5′ exon, a 95-nt intron, and a 32-nt 3′ exon[38]. 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* IEI pre-mRNA consisted of intron 1 without the first 9 nt, the middle exon, and intron 2 truncated before the BPS. The *EFM5* IEI-101 pre-mRNA consisted of intron 1 without the first 10 nt, the middle exon shortened to 101 nt, and intron 2 truncated to 9 nt upstream of the BPS. The *HMRA1* IEI-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* IEI-246 pre-mRNA was generated after mutating the underlined portion of the last 41 nt of its middle exon from 5′-CAAAGAAATGTGGCATTACTCCACTTCAAGTAAGAGTTTGG-3′ to 5′-ACTAATGCCACTACTTTACTCCACTTCAAGTAAGAGTTTGG-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*, *EFM5*, and *HMRA1* genes. Pre-mRNA substrates 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 substrate was bound to the MBP–MS2 fusion protein and mixed with purified U1 snRNP[23] and BBP–Mud2 dimer (or BBP–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, pH7.9, 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, pH7.9, 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, pH8.0.

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 Vitrobot. A FEI TF20 cryo-EM instrument was used to screen grids. CryoEM grids with optimal particle distribution and ice thickness were obtained by varying the gas source (air using PELCO easiGlow, target vacuum of 0.37 mbar, target current of 15 mA; or H_2/O_2 using Gatan Model 950 advanced plasma system, target vacuum of 70 mTorr, target power of 50 W) and time for glow discharge, the volume of applied samples, chamber temperature and humidity, blotting time and force, as well as wait time before blotting. Our best grids for the *ACT1* complex were obtained with 50 s glow discharge using air and with the Vitrobot sample chamber set at 12 °C temperature, 100% humidity, 2.5 s blotting time, –3 blotting force and 20 s wait time. The best grids for the *UBC4* complex were obtained with 60 s glow discharge using air and with the Vitrobot sample chamber set at 12 °C temperature, 100% humidity, 3 s blotting time, 1 blotting force and 60 s wait time.

Optimized cryo-EM grids were loaded into an FEI Titan Krios electron microscope with a Gatan Imaging Filter (GIF) Quantum LS device and a post-GIF K2 Summit direct electron detector. The microscope was operated at 300 kV with the GIF energy-filtering slit width set at 20 eV. Movies were acquired with Leginon[39] by electron counting in either super-resolution mode at a pixel size of 0.68 Å/pixel (*ACT1* complex) or counting mode at a pixel size of 1.36 Å/pixel (*UBC4* complex). A total of 40 frames were acquired in 8 s for each movie, giving a total dose of ~30 $e^-/Å^2$/movie.

Drift correction for movie frames. Frames in each movie were aligned for drift correction with the GPU-accelerated program MotionCor2[40]. 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 used only 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 CTFFIND4[41] 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/kzhang). 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) before further processing by the GPU accelerated RELION2.1. The reported U1 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 this 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 CTFFIND4 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) before further processing by the GPU accelerated RELION2.1. The reported U1 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 ('Step1' 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 Prp40, NCBP1–NCBP2, and Nam8; 'Step 2' in Extended Data Fig. 3c). During each round of 3D classification, only one class showed features corresponding to the intact E complex monomer (probably owing to the cross-linking reagent used for the *UBC4* complex, one class from the second round of 3D classification exhibited features characteristic of the E complex dimer). These good particles from the final round of 3D classification were unbinned to 384 × 384 square pixels (pixel size of 1.36 Å). We re-centred these particles and removed duplications based on the unique index of each particle given by RELION.

The resulting 124,825 unbinned, unique particles (6.5% of all particles) were subjected to a final step of 3D auto-refinement ('Step 3' in Extended Data Fig. 3c). 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 3.6 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 substitution 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 procedures[44].

Local resolution was estimated using ResMap[45]. The overall quality of the maps for the *ACT1* and *UBC4* 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: 5UZ5, 3.7 Å), which was fitted into the *UBC4* complex density map by UCSF CHIMERA[46]. The central regions of the *UBC4* 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 *UBC4* 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 assignment 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 assignment.

For the RNA region near the 5′ SS (nt −2 to +8 of pre-mRNA with respect to the exon–intron junction; nt 1–10 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 RCrane[48] and ERRASER[49].

Models built for the protein and RNA subunits in these central regions include: U1–70K (amino acids (aa) 1–91), U1C (aa 3–197), U1A (aa 2–46, 55–125, 133–148), Prp42 (aa 1–544), Prp39 (aa 288–553, 561–627), Nam8 (aa 291–425, 432–449, 492–523), Snu56 (aa 43–170, 185–295), Snu71 (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 Snu71 since deletion of the coiled-coil domain of Luc7 reduces its interaction with Snu71 and Prp40 (Extended Data Fig. 7b), but only Snu71 has isolated long helices.

Resolutions for the periphery of the *UBC4* complex were more varied, ranging from 6 Å to 25 Å, insufficient for de novo atomic modelling. The following proteins were built with homology modelling using the I-TASSER server and rigidly docked into the low-pass filtered map of the 3.6 Å map using CHIMERA: RRM domain of U1–70K (aa 94–188), N-terminal region of Prp39 (aa 43–285), and RRM2 domain of Nam8 (aa 161–242). The homology model of the NCBP1–NCBP2 heterodimer (NCBP1:616 aa 36–861; NCBP2: aa 19–156) was rigidly fitted into its local refinement map. In addition, the periphery region has a boomerang-shaped density which matches well with that of the crystal structures of tandem FF domains connected by long helices[21,22]. We assigned this density as the Prp40 FF4–6 domains, considering that DSSO crosslinking in the *UBC4* complex and mass spectrometry analyses demonstrate that the C-terminal FF domains crosslink to U1–70K (Extended Data Fig. 7a) and that there is weak density connecting the boomerang-shaped density and U1–70K, although we cannot rule out the possibility of this density being the other tandem FF domains such as FF3–5. The FF4 (aa 355–413, from I-TASSER), FF5 (aa 427–488, from I-TASSER) and FF6 (aa 491–552, PDB: 2KFD) domains were rigidly docked into the low-pass filtered map using CHIMERA, and manually connected using COOT with the long helix between FF domains (Fig. 2f).

Except for nucleotides 27–33 at the tip of stem loop 1 and the last three nucleotides 566–568, the entire U1 snRNA is now modelled with DRRAFTER[50]. The estimated root mean square deviation (r.m.s.d.) accuracies for the DRRAFTER models are: 0.4 Å for residues 39–41, 4.3 Å for residues 97–103, 0.7 Å for residues 175–177, 3.5 Å for residues 202–236, 3.0 Å for residues 289–294, and 4.9 Å for residues 325–516. The median structures of the best ten scoring models are shown in Fig. 2a. Using the low-pass filtered map, we could also manually trace the main chain for nt −9 to −3 and nt +9 to +13 of pre-mRNA. Combined with the previous atomic model, 23 nucleotides of the pre-mRNA were manually built. The upstream region of this pre-mRNA could directly insert into NCBP1–NCBP2 heterodimer and its downstream region could interact with the RRM2 domain of Nam8.

The model of the pre-RNA helix and the putative localization of the BBP–Mud2 binding region were based on the 3.2 Å resolution structure of the *ACT1* complex. The modelling procedure is similar to that used for modelling the *UBC4* complex except for the following differences. First, we could observe the density for a ~25-bp double RNA helix with clear major and minor grooves on a binding surface formed by the C-terminal tail of Prp39, the N-terminal domain of Prp42, and the C-terminal domain of U1C. This double helix density was also observed in the pre-B complex structure and was tentatively modelled as part of U2 snRNA. As our *ACT1* complex was assembled from in vitro-transcribed *ACT1* pre-mRNA, purified U1 snRNP, BBP, and Mud2 proteins, there is no U2 snRNA present in our sample (Extended Data Fig. 5b). Although some bases can be separated in the density map of this double helix, we were unable to model specific nucleotides. Nonetheless, there is weak density connecting it to the 5′ SS, suggesting that this double helix belongs to the pre-mRNA region downstream of the 5′ SS. Second, 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 bound to pre-mRNA, given that its location corresponds roughly to where U2 SnRNP is in the A complex structure (Extended Data Fig. 4j).

The above models were refined using PHENIX in real space[51] with secondary structure and geometry restraints. Refinement statistics of the E complex are summarized in Extended Data Table 1. These models were also evaluated based on Molprobity scores[52] and Ramachandran plots (Extended Data Table 1). Model–map FSC validation is shown in Extended Data Figs. 2g, 3g. Representative densities for the proteins and RNA are shown in Extended Data Fig. 4. All structure-related images in this paper were generated using UCSF CHIMERA[46] and CHIMERAX[53].

To determine the minimum number of nucleotides needed to connect an upstream BPS to a downstream 5′ SS in the A complex[14], we modelled connection lengths ranging from 21 to 30 nucleotides between nucleotides 74 and −1 of pre-mRNA (chain I in PDB ID 6g90, nucleotide 70 is branch point and +1 is 5′ SS) using the Rosetta RNP-denovo method with full-atom refinement[28,50]. Nucleotides 75–78 of pre-mRNA were excised from the structure and all other nucleotides were kept fixed. We modelled 21–30 uridines de novo to connect nucleotide 74 to nucleotide −1. During the initial low-resolution stages of the modelling, score terms that rewarded favourable RNA–protein interactions, RNA base pairing, and compact RNA structures were turned off. Score terms that penalized clashes within the RNA and between the RNA and protein were included during this stage. During the final all-atom refinement, the complete all-atom RNA–protein score function was used. The weight on the score term that penalized chainbreaks was increased to 50.0 and models with a chainbreak score less than 0.5 were considered fully connected. The top scoring model (at least 250 models were built for each connection length) by full Rosetta score was used as a representative model for each connection length. We found that the representative model for 22-nucleotide connection length is fully connected and has similar total Rosetta score as models for longer connection lengths. These results indicate that 22 nucleotides are sufficient for connecting nucleotide 74 to −1 (which is equivalent to 28 nucleotides connecting branch point 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 spliceosome E complex was crosslinked with 10 mM DSSO (disuccinimidyl sulfoxide) 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[54]. In brief, ~100 μg of crosslinked sample was reduced, alkylated, and digested at 1:50 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 using 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 min), 100% B (23–25.5 min) with solvent A (10 mM KH_2PO_4, 25% acetonitrile, pH 3.00) and solvent B (10 mM KH_2PO_4, 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 (UHP)LC–MS/MS (Easy-nLC1200, Orbitrap Fusion LumosTribrid, 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 CORTECS C18 resin (2.7 μm, spherical solid core). Samples were run at 400 nl/min over a 90-min linear gradient from 4 to 32% acetonitrile with 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 frequency was determined by a 5-s 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 30% in the orbitrap. Data acquisition was performed using Xcalibur (version 4.1) software.

Instrument raw files were directly loaded in to 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/hydrolysed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for crosslink detection against DSSO. Precursor mass tolerance was set to 10 p.p.m., with MS/MS mass tolerance set to 20 p.p.m. Results were manually validated and visualized using xVis[55].

Oligo-directed RNase H digestion of pre-mRNA in purified E complex. Purified E complex with *ACT1* or *DYN2* IEI pre-mRNA as substrate was incubated with RNase H (New England Biolabs) in the presence or absence of 5 μM DNA oligo, at 25 °C for 30 min. The complex was then bound to amylose resin pre-washed with buffer G120. The resin was washed with buffer G120 and eluted in buffer G120 containing 10 mM maltose. The eluted samples were analysed on SDS–PAGE and stained with Coomassie to visualize the proteins. For RNA detection, the samples were digested with 1 μg/μl proteinase K, separated on 7 M urea denaturing polyacrylamide gel and stained with EtBr, or on native agarose gel and stained with SYBR gold (Life Technologies). The DNA oligo used for digestion of *ACT1* was 5′-AAAATAAACGATGACACAG-3′, and for *DYN2* was 5′-TCATGGAAGAAAACCTCAC-3′.

Chase experiment with assembled E complex in U1-depleted yeast extract. BSY82 (GAL-U1) yeast strain[56] obtained from M. Rosbash's laboratory was maintained in yeast extract peptone (YEP) medium containing 2% galactose. For U1 snRNA depletion cultures, log phase cells growing in 2% galactose were diluted into medium containing 2% glucose to an OD_{600} of 0.03 and grown for 17 h to an OD_{600} of 2.5. Yeast splicing extracts were prepared from 2 l of yeast cells cultured in medium containing either galactose or glucose. Splicing reactions were carried out at 23 °C for 15 min in a 25-μl reaction containing 2.5 nM M3–*ACT1* pre-mRNA alone or purified E complex, with or without 50-fold (in molar quantity) of *ACT1*–M3 (*ACT1* pre-mRNA containing 54 nt 3′ exon fused with three copies of MS2 stem loops at the 3′-end), 40% yeast extract, and splicing buffer (60 mM potassium phosphate, pH 7.4, 3% PEG 8000, 2.5 mM $MgCl_2$, 2 mM ATP). RNAs were then phenol/chloroform extracted and precipitated with 2.5 volumes of ethanol. After DNase I (Roche) treatment, first strand cDNAs were synthesized from 1 μg of RNA using ProtoScript II reverse transcriptase with reverse primers specific to M3–*ACT1*, U1 snRNA, or U2 snRNA. PCR was performed using cDNA transcribed from 25 ng of RNA as template and the following primers: MS2 forward 5′-TCCGATATCCGTACACCATC-3′; *ACT1* exon 2 reverse 5′-TGATACCTTGGTGTCTTGGTCT-3′; yeast U1 forward 5′-AAACATGCGCTTCCAATAGT-3′, reverse 5′-TATGTGTGTGT GACCAAGGAG-3′; and yeast U2 forward 5′-AACTGAAATGACCTCAATG AGGCTC-3′, reverse 5′- AGACCTGACATTAGCGGAAAACAAC-3′. The products were analysed on 3% low melting point (LMP) agarose gel stained with EtBr.

The same experiment was also performed using *EFM5* IEI-101–M3 or *HMRA1* IEI-246–M3 pre-mRNA and E complex assembled on both pre-mRNAs. A second *EFM5* IEI-101 pre-mRNA was designed to remove the primer binding sites so that it was invisible in the RT–PCR reaction and was transcribed in vitro to be used as competing RNA. For *HMRA1*, wild type IEI-246 pre-mRNA was used as competing RNA. Primers used to detect circRNA formed specifically from *EFM5* IEI-101 were: IEI-101 cir forward 5′- CCTTGAAGAATTCAAAAGAGAGGATAG -3′ and cir reverse 5′- CGAGGGCATTAGCAGAAAG -3′. Primers used to detect circRNA formed specifically from *HMRA1* IEI-246 were: IEI-62 cir forward 5′-TCCACTTCAAGTAAGAGTTTGG-3′ and cir reverse 5′-GAGTAAAGTAGTGGCATTAGTCA-3′.

Analysing the role of *ACT1* intronic secondary structure in splicing. We mutated multiple stretches of sequences in the 5′ SS-to-BPS region of *ACT1* so that it no longer formed extensive secondary structures. The final sequence of this region (mutated nucleotides underlined) was: 5′-GTATGTTCTAGCGCTTGCACCATCCCATTTAACTGTAAAAAAAAATGC ACGGTCCCAATTGCTCGAGAGATTTCTCTTTTACAAAAAAATACTATTAA AAAAAGAGAAAAAACCTCCTATATTGACTGATCTGTAATAACCACGATA TTATTGGAATAAATAGGGGCTAAAAATTTGGAAAAAAAGAAAAACTGA AATATTTTCGTGATAAGTGATAGTGATAAAAAAAAATTATTTGCTACTGT GTCTCATGTACTAA-3′. We synthesized this mutant region through Genscript and replaced the 5′ SS-to-BPS region of the *ACT1–CUP1* reporter plasmid[19] with this sequence to generate the mutant reporter plasmid pMA6.

The wild-type and mutant *ACT1–CUP1* reporter plasmids were transformed into yeast strain BY4741, and grown in synthetic complete (SC) –Leu medium until $OD_{600} = 1.0$. RNA was extracted from 5 ml culture, treated with DNase I (Roche), and reverse transcribed using the ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs) and random primers. qPCR was performed using the iTaq Universal SYBR Green Supermix (Biorad) with cDNA transcribed from 10 ng RNA as template. The primers for detecting pre-mRNAs are located in the *ACT1* intron and *CUP1*: Act intron forward 5′-TTATTTGCTACTGTGTCTCATG-3′; YAC7 reverse 5′- GCATTGGCACTCATGACCTT-3′. The primers for detecting total reporter mRNA are located in *ACT1* exon 2 and *CUP1*: ActEx2 forward 5′- GTTCTGGTATGTGTAAAGCC-3′; CUP1end reverse 5′-CCAGAGCAGCATGATTTCTT-3′.

Co-purification assay in yeast. The coding regions of full-length or truncated yeast Prp40, Snu71, and Luc7 were amplified by PCR using genomic *S. cerevisiae* DNA as templates, and ligated into pRS414, pRS416 and pRS317 vectors. The final plasmids constructed were: pRS414/GPD-protA-Prp40 (full-length, FF1–3 and FF4–6), pRS317/GPD-Prp40, pRS414/GPD-protA-Snu71, pRS416/GPD-CBP-Luc7 or Luc7ΔCC (Luc7 coiled-coil domain (residues 123–190) deleted). Yeast BCY123 cells were transformed with different combinations of the plasmids and selected on appropriate selective media. Clones from the transformation were cultured in 50 ml of liquid selective medium to $OD_{600} = 3–4$. Cells were harvested and lysed in lysis buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.05% NP40, 1 mM DTT, 2.5 mM $CaCl_2$, 1.5 mM $MgCl_2$) using the bead-beating method. The lysates were incubated with IgG resin for 3 h at 4 °C. The resins were washed with the lysis buffer. The proteins were cleaved off IgG resin using TEV protease in buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.01% NP40, 0.5 mM DTT). The proteins were separated by SDS–PAGE and transferred to a nitrocellulose membrane. Western blot was performed using an anti-CBP tag antibody (GenScript, A00635).

***DYN2* splicing analyses.** The *DYN2* gene was PCR amplified from *S. cerevisiae* genomic DNA along with the 5′-UTR (356 bp) and 3′-UTR (305 bp) and cloned into the pRS415 vector. The BPS of the first intron was mutated from TACTAAC to TAGTACC and the 5′ SS of the second intron from GT to CG separately or together, to generate the I1-BP, I2-5′ SS, and double mutants. The wild-type and mutant plasmids were transformed into a *DYN2* deletion yeast strain (Open Biosystem), and transformants were selected on SC –Leu plates. Cells were grown in SC –Leu medium to an OD_{600} of ~1.0. Total RNA was isolated from 10 ml of cells using a hot-phenol extraction method and dissolved in 100 μl diethylpyrocarbonate (DEPC)-treated water. A total of 1 μg of RNA was treated with DNase I and reverse transcribed into cDNA. qPCR was performed using the following primers: E1 forward 5′-CCAAAATGAGCGATGAAAATAAGAG-3′ and I1 reverse 5′-TCATGGAAGAAAACCTCACTC-3′ to detect exon 1–intron 1 product; I1 forward 5′-TATGTCAGTTAATCTCAGTCACAAT-3′ and E2 reverse 5′-TATGTCAGACGCCTTAACAATAG-3′ to detect intron 1–exon 2 product; E2 forward 5′-CTATTGTTAAGGCGTCTGACATA-3′ and I2 reverse 5′-GGTCTAAGTTTTCTCCTTGTTAG-3′ to detect exon 2–intron 2 product; I2 forward 5′-CATGTTTTGTGTGTGTACATTTG-3′ and E3 reverse 5′-CAGGTATTGCCGTATTTGAC-3′ to detect intron 2–exon 3 product; E3 forward 5′-CGACAAGCTGAAAGAGGATA-3′ and E3 reverse to detect exon 3.

CircRNA detection from purified spliceosome. Spliceosome was purified from 3 l of yeast cells carrying the Prp22(H060A) mutant to enrich the post-catalytic complex[20] and increase our chance of detecting branching and ligation products before their release. RNA from the purified spliceosomal complex was purified and reverse-transcribed into cDNA. PCR was performed to detect the presence of T-branches and circRNAs from the yeast multi-intronic genes *EFM5* (YGR001C) and *HMRA1* (YCR097W), using circle-specific primers[7] and the following primers for T-branches: *EFM5* I1 forward 5′-TTTTCAACACAGTAACGTAGAATTAC-3′, I2 reverse 5′-AACAGTTAGTAAGATGAAAAGATACTGG-3′; *HMRA1* I1 forward 5′-GTATGTTTTCATTTCAAGGATAG-3′, I2 reverse 5′-TGTTAGTATAGGATATATTTAAGTTTGA-3′. PCR products were analysed on 3.5% LMP agarose gel stained with EtBr, and cloned into pMiniT vectors (New England Biolabs) for sequencing.

CircRNA detection from *EFM5* and *HMRA1* IEI constructs in vivo. The region from +47 to +630 of *EFM5* (containing *EFM5* partial intron 1, exon 2, and partial intron 2) was PCR amplified from *S. cerevisiae* genomic DNA and was inserted between a *GPD* promoter and a *CYC1* terminator, and the resulting expression cassette was cloned into a pRS317 vector, generating the *EFM5* IEI plasmid. The BPS of the first intron was mutated from TACTAACTAAC to TAGTACCTACC (mutations underlined)[57] and the 5′ SS of the second intron from GT to CG separately, to generate the BP and 5′ SS mutants. The IEI-101 truncation was generated by fusing the first 49 nt to the last 52 nt of the middle exon. The IEI-63 truncation was engineered on IEI-101 to shorten the middle exon to 63 nt in length, with the following sequence 5′-GACACTTTCTGCCC TTGAAGAATTCAAAAGAGAGGATAGATTGTTAATTGACCCAAACAAAGT-3′.

The *EFM5* IEI plasmid and the empty pRS317 vector were transformed into an *EFM5* deletion yeast strain (Open Biosystem). Cells were grown in SC –Lys medium to OD_{600} of ~1.0. Total RNA was isolated and reverse transcribed into cDNA. For RNase R treatment, 1 μg total RNA was incubated at 37 °C for 30 min with 5 U/μg RNase R (Epicentre Technologies) and used directly for reverse transcription without further purification. PCR was performed using specific primers[7] to detect circRNA formed from exon 2 of *EFM5* using the following primers: *EFM5* cir forward 5′-GAGAGGATAGATTGTTAATTGACCC-3′ and *EFM5* cir reverse 5′-CTTTTGAATTCTTCAAGGGCA-3′. The primer pair used to detect un-spliced *EFM5* pre-mRNA was: *EFM5* I1 f5′-TTTTCAACACAGTAACGT AGAATTAC-3′ and I2 reverse 5′-GAGTAGGGATATGTTTATGATATACATAC-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 *EFM5* IEI plasmid. The IEI-62 truncation was engineered to shorten the middle exon to 62 nt in length, with sequence 5′-TTTATAATGGAAAGTAATTTGACTAATGCCAC TACTTTACTCCACTTCAAGTAAGAGTTTGG-3′. Primers used to detect circRNA were IEI-62 cir forward and reverse, which are the same as those used for IEI-246. The primer pair used to detect un-spliced *HMRA1* pre-mRNA was: *HMRA1* I1 forward 5′-CCAAGAACTTAGTTTCGACTCTAGATTTCAAGGAT AGCCTTTGAATC-3′, I2 reverse 5′-AACTAATTACATGATGGGCCCGGATAT ATTTAAGTTTGATTCTCATATTACATAC-3′.

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 (6N7P for the *UBC4* complex and 6N7R 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. Leginon: 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. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
42. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9**, 853–854 (2012).
43. 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).
44. 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).
45. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
46. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
47. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
48. Keating, K. S. & Pyle, A. M. RCrane: semi-automated RNA model building. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 985–995 (2012).
49. Chou, F. C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. & Das, R. Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat. Methods* **10**, 74–76 (2013).
50. Kappel, K. et al. De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. *Nat. Methods* **15**, 947–954 (2018).
51. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
52. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
53. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
54. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
55. Grimm, M., Zimniak, T., Kahraman, A. & Herzog, F. xVis: a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints. *Nucleic Acids Res.* **43**, W362–W369 (2015).
56. Seraphin, B. & Rosbash, M. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell* **59**, 349–358 (1989).
57. Qin, D., Huang, L., Wlodaver, A., Andrade, J. & Staley, J. P. Sequencing of lariat termini in *S. cerevisiae* reveals 5′ splice sites, branch points, and novel splicing events. *RNA* **22**, 237–253 (2016).
58. Li, Z. & Brow, D. A. A rapid assay for quantitative detection of specific RNAs. *Nucleic Acids Res.* **21**, 4645–4646 (1993).
59. 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 GM126157 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 (1S10OD018111, 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, with support from NIGMS P41-GM103311 (Chimera, ChimeraX) and NIH R01-GM129325 (ChimeraX). We also thank M. Ares, D. Black, and D. Brow for comments on early versions of the manuscript.

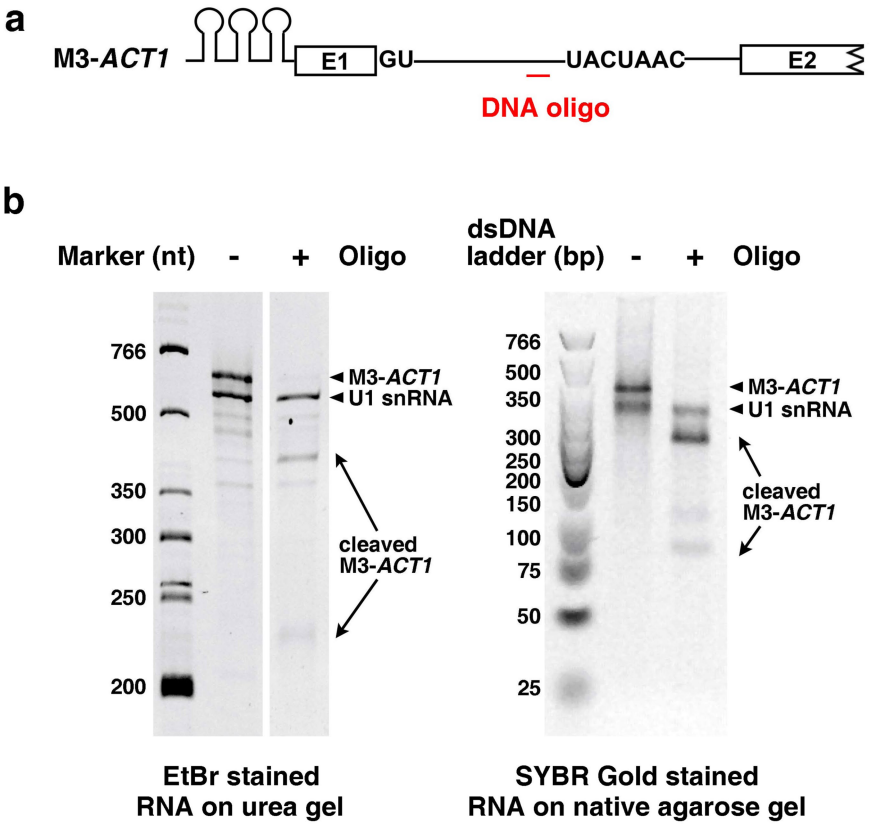
Author contributions X.L. and S.L. contributed equally to the work and are listed alphabetically in the author list. R.Z. and Z.H.Z. conceived the project; X.L. prepared and optimized the sample; X.L., L.Z., S.E. and S.S. performed biochemical analyses; S.L. and Y.C. recorded and processed the EM data; A.I., R.C.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.

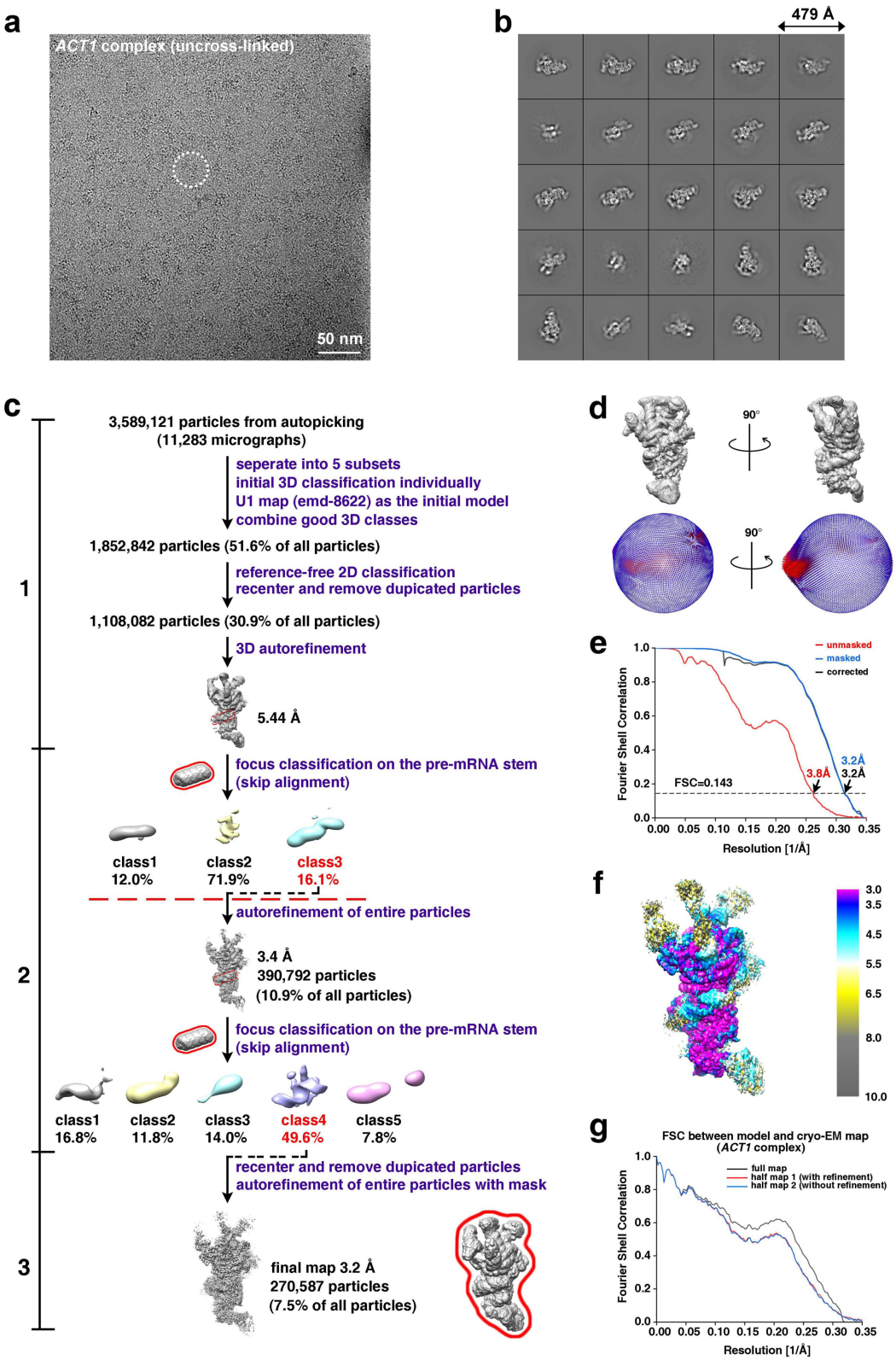
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1523-6.

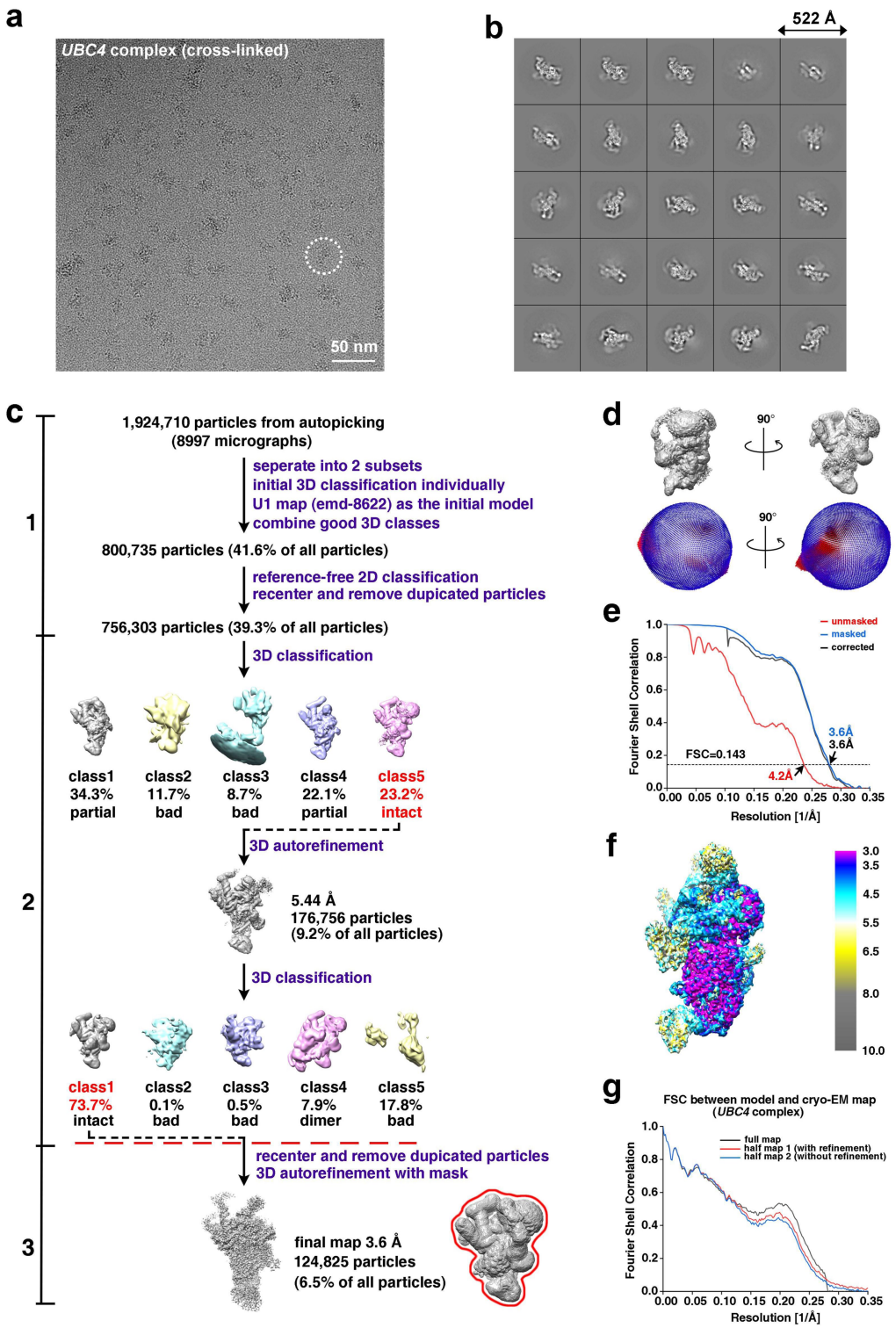
Correspondence and requests for materials should be addressed to Z.H.Z. or R.Z.



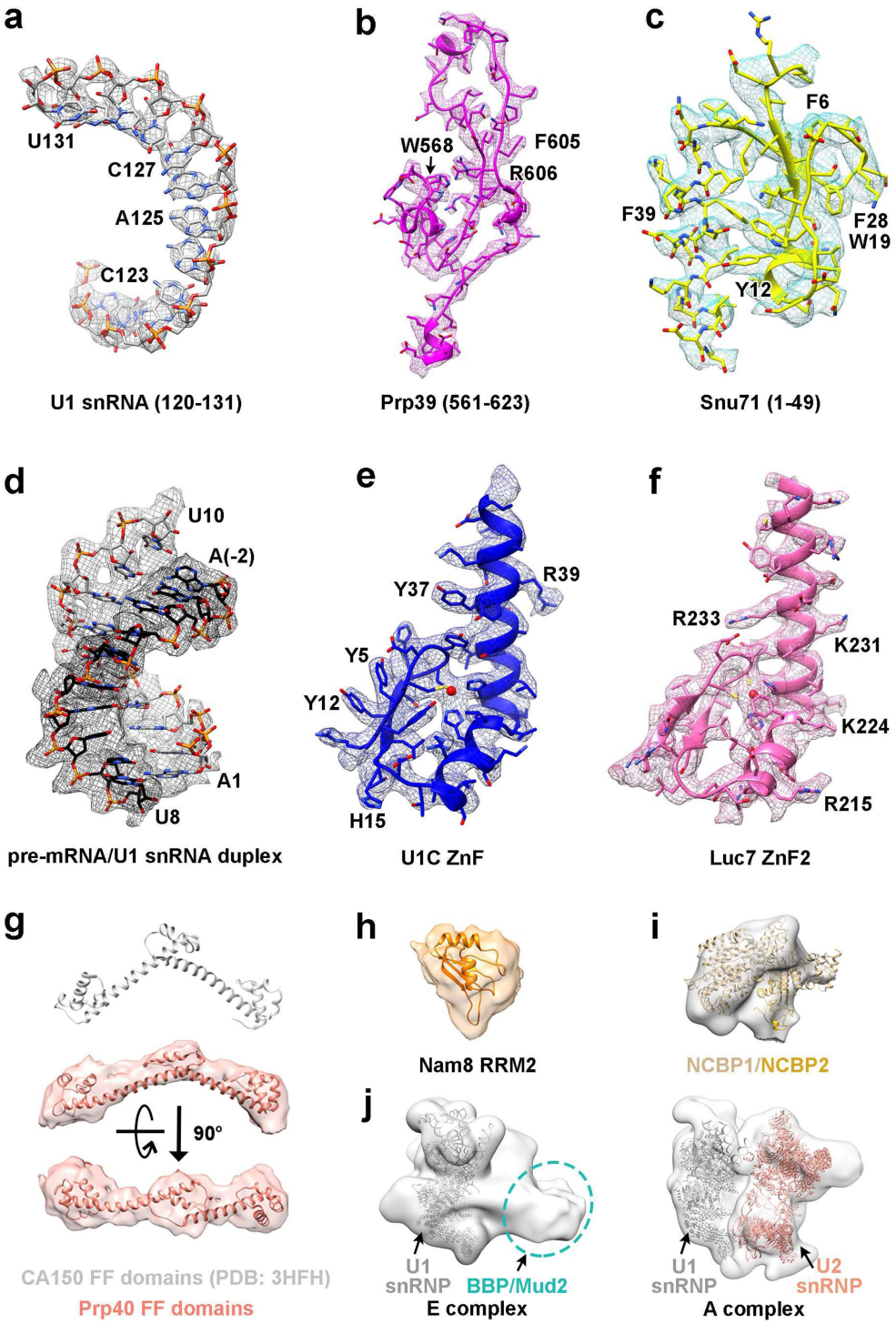
Extended Data Fig. 1 | In vitro assembly and purification of the *ACT1* complex. a, Schematic representation of the *ACT1* pre-mRNA tagged with three MS2-binding sites (M3–*ACT1*) used for E complex assembly and purification. Boxes represent exon 1 (E1) and truncated exon 2 (E2). The 5′ SS (GU) and BPS (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.



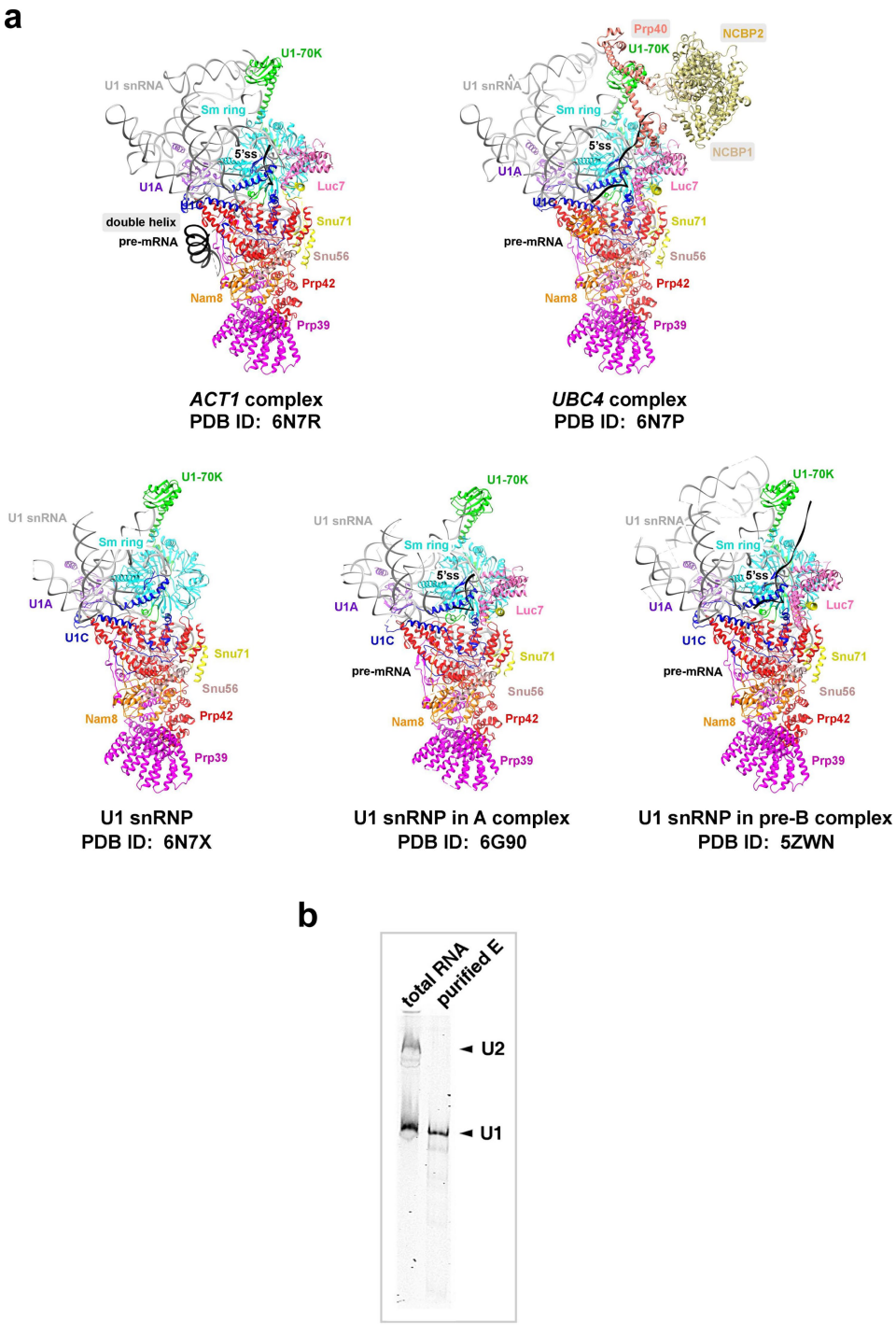
Extended Data Fig. 2 | The cryo-EM structural determination process for the *ACT1* complex. a, Representative drift-corrected cryo-EM micrograph (out of 11,283 micrographs) of the E complex assembled on the *ACT1* pre-mRNA. A representative particle is shown in a white dotted circle. **b**, Representative 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.



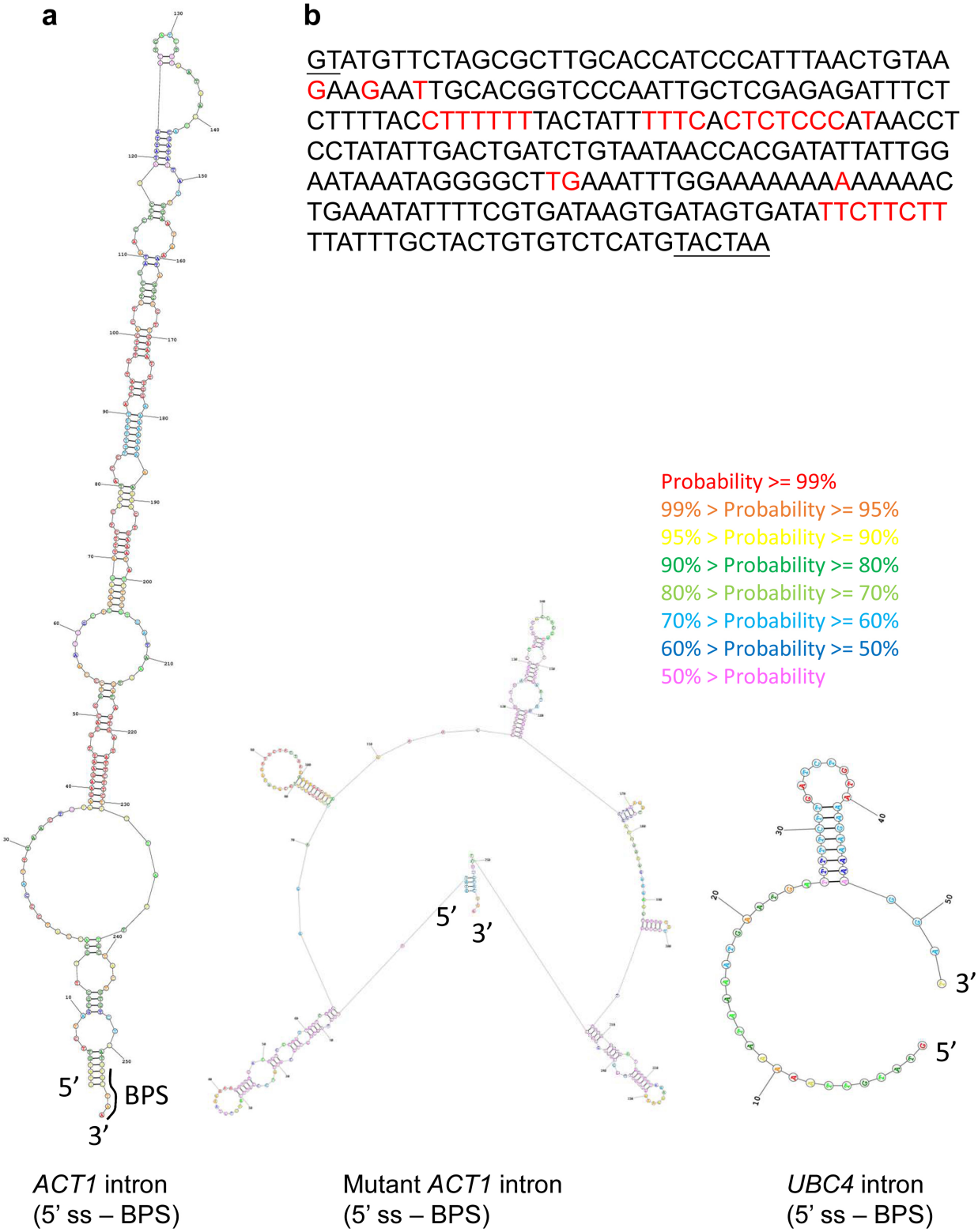
Extended Data Fig. 3 | The Cryo-EM structural determination process for the *UBC4* complex. a, Representative drift-corrected cryo-EM micrograph (out of 8,997 micrographs) of the E complex assembled on the *UBC4* pre-mRNA. A representative particle is shown in a white dotted circle. **b**, Representative 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.



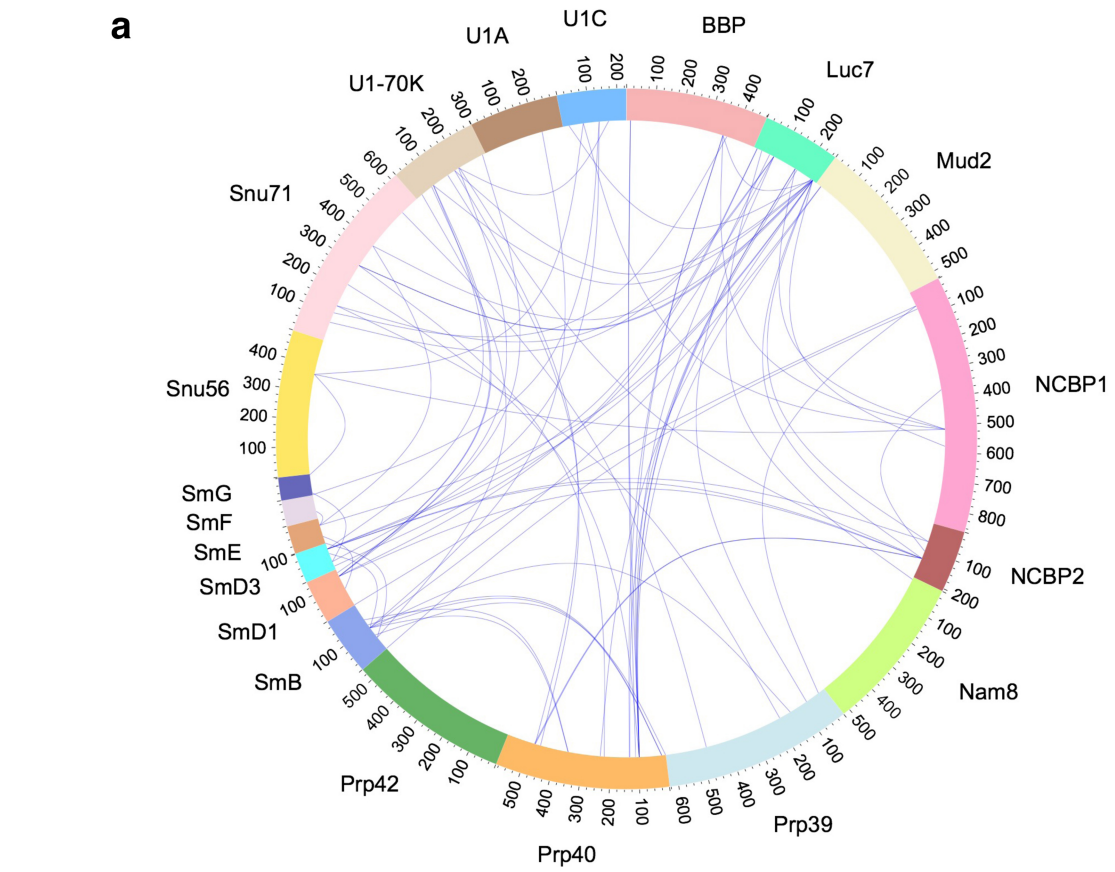
Extended Data Fig. 4 | Representative cryo-EM density maps of the E complex. **a**–**i**, Densities for the *UBC4* complex; **j**, density for the *ACT1* complex. Cryo-EM density maps are shown as follows. **a**, Selected regions of U1 snRNA. **b**, C-terminal region of Prp39. **c**, N-terminal domain 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.

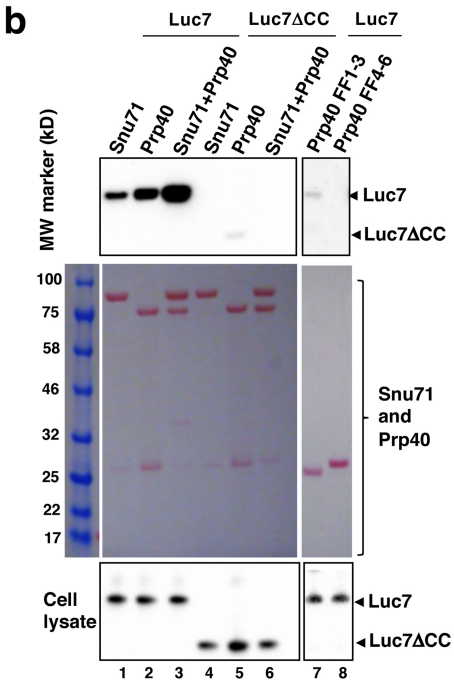


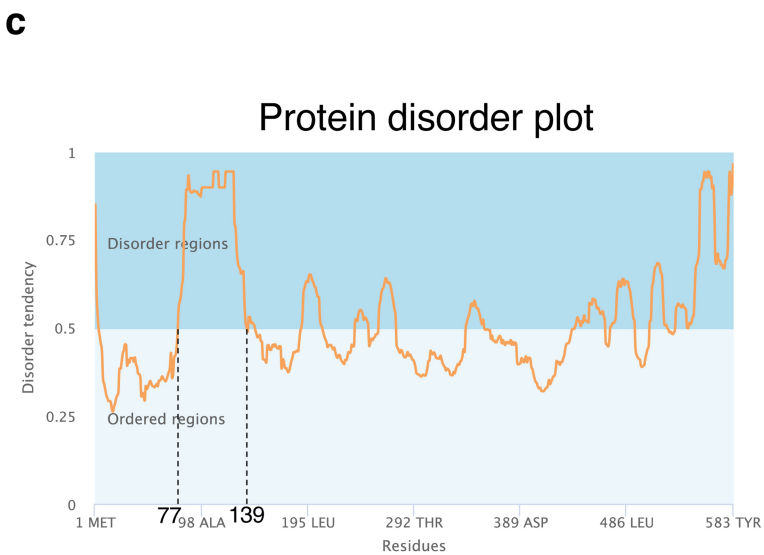
Extended Data Fig. 5 | Structural and biochemical characterization of the *ACT1* and *UBC4* complexes. a, Comparison of the ribbon models of the *ACT1* complex, the *UBC4* complex, and U1 snRNP from other previously determined structures (the U1 snRNP, A, and pre-B complexes). Labels with shading indicate 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[58] 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.



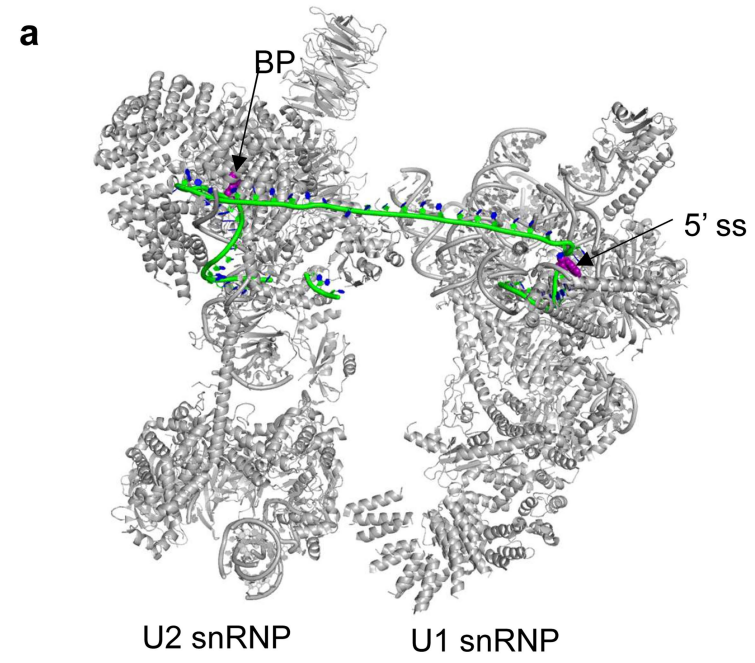
Extended Data Fig. 6 | Secondary structures in the region between the 5′ SS and BPS in the wild-type and mutant *ACT1* and *UBC4* pre-mRNAs. a, Secondary structures predicted by RNAstructure 6.0 (https://rna.urmc.rochester.edu/RNAstructureWeb/). **b**, Sequence between the 5′ SS and BPS (underlined) 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.





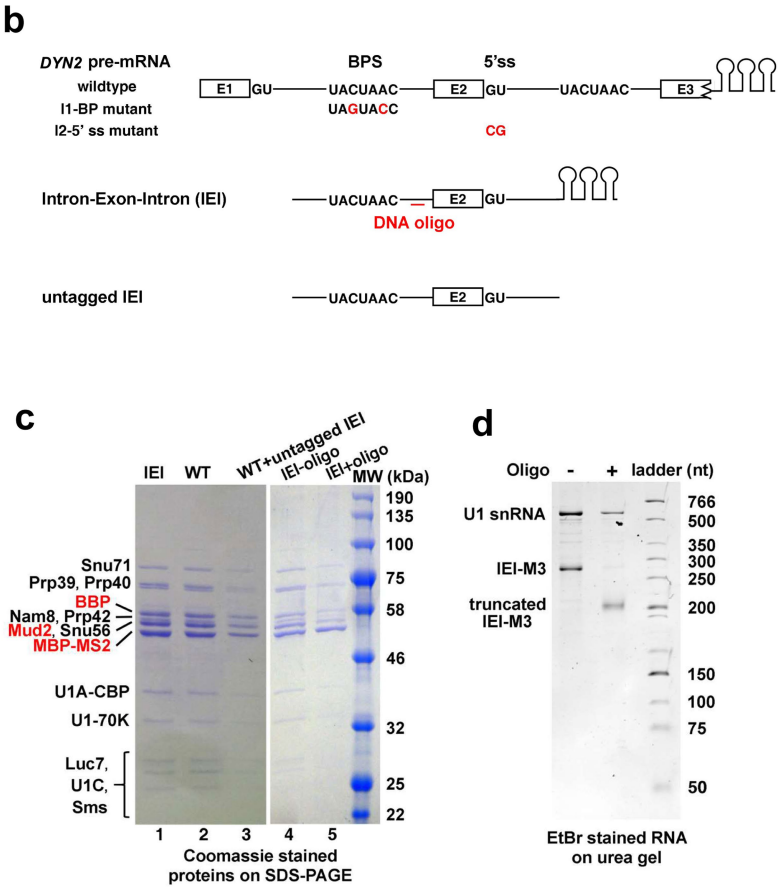


Extended Data Fig. 7 | Protein interactions in the *UBC4* complex. **a**, DSSO crosslinking and mass spectrometry analyses of the *UBC4* complex. Each blue line indicates a crosslink between a pair of Lys residues. Note that BBP–Mud2 are crosslinked to Luc7, Prp40, Snu56, and Snu71. **b**, Co-purification 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 MetaDisorderMD2[59].

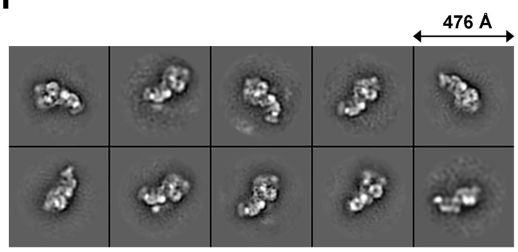


e

Proteins	IEI/WT
SmB	1.14
SmD1	0.78
SmD2	0.86
SmD3	0.98
SmE	0.64
SmF	1.01
SmG	0.96
U1A	0.99
U1C	0.89
U1-70K	1.18
Luc7	1
Nam8	1.11
Prp39	1.01
Prp40	1.08
Prp42	1.02
Snu56	1.02
Snu71	0.98
BBP	1.08
Mud2	1.08
NCBP1	0.24
NCBP2	0.3



f



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.

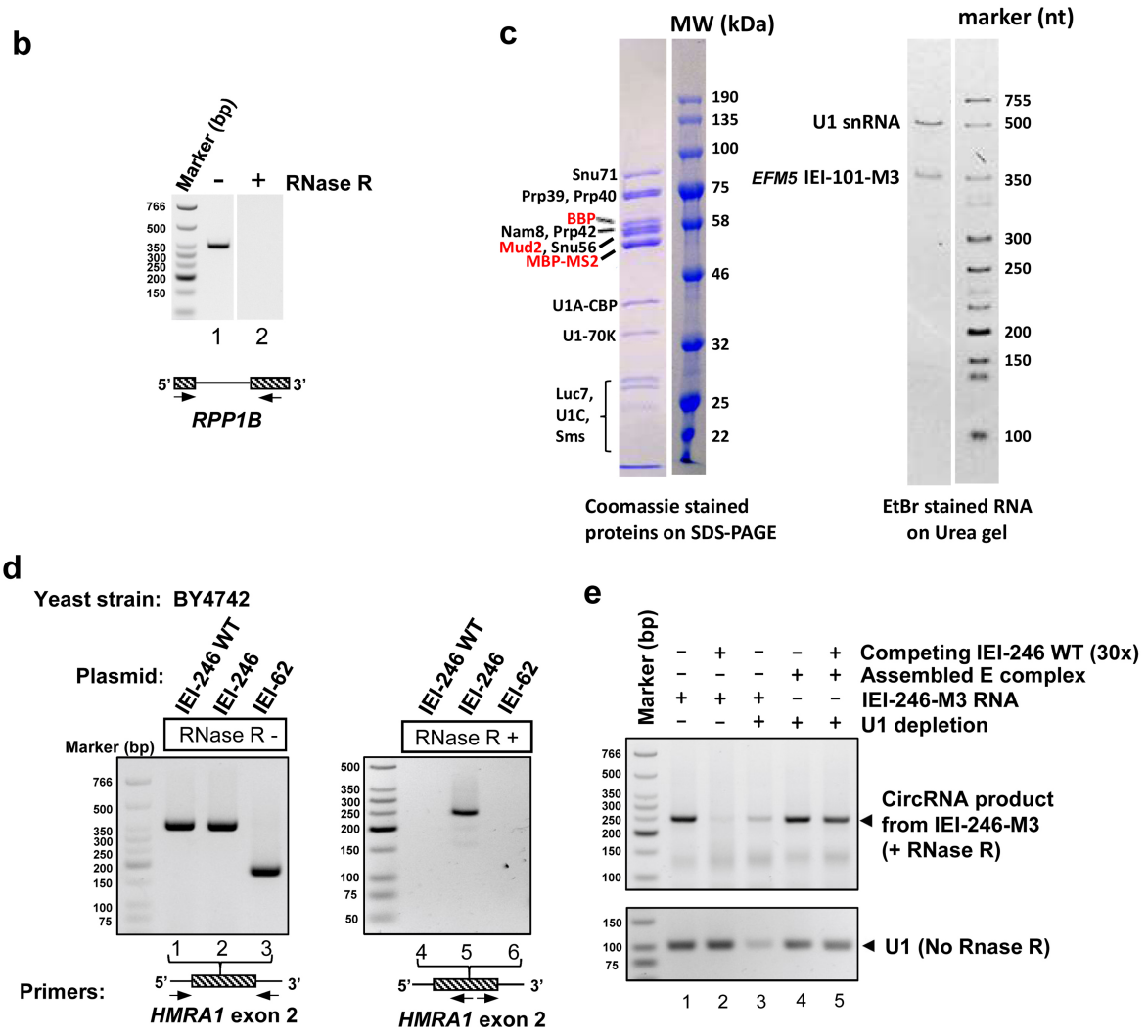
a

EFM5 PCR product generated from circRNA using outward facing primers on exon2 (Fig. 5a, lane 9):
…tatcagAACCACTGGACTTCAGTGATGAAATTAAAGGAAAAGTTGATAGATTGTTAATTGACCCACCTTTTT
TAAATGAAGATTGTCAAACAAAGT/GACACTTTCTGCTAATGCCCTCGCTGCCCTTGAAGAATTCAAAAGAG
AGGAACAACAACATCAAGAAGCCTTTCAAAAGCTTTACGACGatcatg…

EFM5 PCR product generated from T-branch (Fig. 5a, lane 5):
…catgatTTTTCAACACAGTAACGTAGAAT**TACTAACT--IGT**ATGTATATCATAAACAT
ATCCCTACTCATTTTTTAATCTTTTTTCCAGTATCTTTTCATCTTACTAACTGTTctgata…

HMRA1 PCR product generated from circRNA using outward facing primers on exon2 (Fig. 5a, lane 10):
…tatcagGAAAGCAAAGCCTTAATTCCAAGGAAAAAGAAGAAGTTGCAAAGAAATGTGGCATTACTCCACTT
CAAGTAAGAGTTTGG/TTTATAATGGAAAGTAATTTGACTAAAGTAGAGCAACATACATTACACAAAAATATT
TCTAACAATAGGTTAGAAATATACCACCACATTAAAAAAGAGAAGAGCCCAAAGGGAAAatcatg…

HMRA1 PCR product generated from T-branch (Fig. 5a, lane 7):
…catgatGTATGTTTTCATTTCAAGGATAGCCTTTGAATCAATT**TAC--AIGT**ATGTAAT
ATGAGAATCAAACTTAAATATATCCTATACTAACActgata…



Extended Data Fig. 9 | See next page for caption.

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) (PDB 6N7P)	**Act1 complex** (EMD-0361) (PDB 6N7R)
Data collection and processing		
Magnification	105,000	105,000
Voltage (kV)	300	300
Electron exposure (e^{-}/Å^2)	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	---
NCBPs	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 (Å^2)	-147.1	-94.0
Model composition		
Non-hydrogen atoms	41487	35784
Protein residues	4839	3568
Ligands	3	1
B factors (Å^2)		
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

nature research

Corresponding author(s): Rui Zhao and Hong Zhou

Last updated by author(s): Mar 31, 2019

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.

n/a	Confirmed
☐	☒ 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 *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ 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	Leginon2
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

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

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.