

Abstract

Structural Insights into Group II Intron Splicing and Retrotransposition

Kevin Chung

2024

Pre-mRNA processing is a critical metabolic step in eukaryotes. RNA transcripts must be spliced, polyadenylated, and capped, among other modifications, to ensure their proper function. Group II introns are an archetypal splicing system that can catalyze their self-excision and retroelements that invade DNA. They provide a window into understanding the evolution of splicing and have played an important role in shaping the eukaryotic genome as the ancestors of spliceosomal introns, telomerases, and non-LTR retrotransposons. Group II introns consist of a well-folded, large-structured RNA that forms a ribonucleoprotein complex with their encoded maturases to carry out splicing and intron integration functions. Despite their central role in RNA metabolism, there are key questions that remained unanswered regarding their structure, function, and mechanisms of action.

In this work, I primarily use biochemical, biophysical, and structural techniques to understand the splicing pathway of group II introns. Using cryo-electron microscopy, we visualized group II intron RNPs as they proceed through the branching pathway, revealing the network of interactions that dictate branch helix positioning, branch site selection, and splice site exchange. There are striking similarities with the spliceosome that highlight the pattern of conservation from group II introns to the spliceosome. Continuing with the intron lifecycle, I then investigated their function as retroelements that target new DNA insertion sites, which occurs after maturase-mediated intron excision. I characterized the interactions

of the intron holoenzyme with its structured DNA targets, resulting in important findings regarding the type of molecular strategies that enable shape and sequence specific recognition. Overall, this work provides a comprehensive view of the intron life cycle, encompassing the forward splicing (intron excision) and reverse splicing (intron retrotransposition) reactions. This work sheds light on pre-mRNA splicing machines and RNP retroelement complexes, providing a foundation for understanding how ancient group II intron elements have and continue to impact eukaryotic genomes.

Structural Insights into Group II Intron Splicing and Retrotransposition

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

by
Kevin Chung

Dissertation Director: Anna Marie Pyle

May 2024

Copyright © 2024 by Kevin Chung

All rights reserved.

Acknowledgements

I would like to thank my advisor, Professor Anna Marie Pyle for providing the support and tools for me to pursue scientific problems that intrigued me. Thank you for taking me into your lab, even though I knew very little about RNA, and allowing me to learn structural biology and cryoEM. You encouraged me to be independent as a scientist and to think critically when approaching a problem, which have been crucial skills throughout my PhD. I would also like to thank my thesis committee members, Yong Xiong and Mark Hochstrasser for their guidance and advice on my thesis project throughout the years.

I was first introduced to research as an undergraduate in the lab of Dr. David Eisenberg under the mentorship of Dr. Lorena Saelices. David and Lorena gave me the opportunity to learn about research and science and introduced me to the world of structural biology. Ever since, I have been fascinated by the intricacies of structural biology and the amount of information it can reveal and evidently, this was one of the driving forces behind my thesis project. Thank you for taking a chance on me and for inspiring me to pursue a PhD.

My PhD would not have been possible without mentors who support, motivate, and offer invaluable insights throughout this journey. Dr. Chris Lim and Dr. Joshua Lees were my mentors during my rotation project as I navigated the decision of which lab to join, and I am grateful to them for their time, effort, and advice. I am especially thankful for my rotation mentor in the Pyle Lab, Dr. Ling Xu who introduced me to the lab and taught me so much about protein and RNA biochemistry, and structural biology. It was not an easy feat to learn cryoEM from sample preparation to grid optimization, data collection and data processing and to set up the infrastructure within the lab. We encountered so many problems along the way, but your enthusiasm and perseverance always ensured that we found a way pass those obstacles. Thank you for not only being a mentor, but also for being a friend that I can talk to about life.

Next, I would like to thank my colleagues and fellow scientists within the lab. Thank you to Dr. Tianshuo Liu, who is one of the most scientifically curious people that I have met, for your insightful advice and for never being afraid to ask tough questions. Thank you also to Drs. Shivali Patel, Wenshuai Wang, Olga Fedorova, Li-Tao Guo, and Ananth Kumar for their words of wisdom, patience, and fruitful discussions that undoubtedly helped progress my thesis project. Additionally, I would like to extend my gratitude towards my collaborators and co-authors Drs. Swapnil Devarkar and Junhui Peng who were instrumental to our cryoEM efforts. Pengxin Chai and the folks in what I call the ‘structural biology hallway’ (Xiong, Bleichert, and Zhang Labs) along with Marc Llaguno, Jianfeng Lin, and others from the YCRC core facility who also played key roles in helping with any cryoEM related problems.

I am incredibly thankful to the friends outside of the lab that have supported me throughout my time as a PhD student. Thank you to Jon Li, Brandon Yau, Wesley Hong, Jason Mak, Jon Cheung, and their partners. for always making the time to hang out with me and for keeping me entertained whenever I am back home. I would also like to thank Tony Jun, who I never expected to cross paths with again here in New Haven of all places, for being my workout buddy and for his friendship during my time here at Yale. It has been great sharing so many laughs and good times with this group of friends since high school. I am also extremely thankful for the folks that I have met and had the pleasure of working with at YGCC. I learned so much from the various teams that I have been a part of, and I am grateful for the opportunities that this student organization has provided, which have truly broadened my horizons. Thank you to Tayah Turocy and Aishwarya Iyer for being amazing case partners and colleagues, and more importantly wonderful, supportive friends. The last couple years of the PhD were most definitely happier with your friendship and silly puns. Finally, thank you for all the other friends that I have made here at Yale, who I may not have mentioned explicitly, but were important to my journey as a graduate student.

Last but not least, I would like to thank my parents and family. They have provided endless support and encouragement and have been constant cheerleaders for everything that I pursue. They make it difficult to leave every time I visit home and I hope that I can spend more time with them in the future. Thank you Mom, Dad, Tommy, Henry, and Jenny for everything and for always being there for me.

Abbreviations

bpA – branchpoint adenosine
cryoEM – cryogenic electron microscopy
DBD – DNA binding domain
EBS – exon binding sequence
IBS – intron binding sequence
IEP – intron encoded protein
IFD – insertion helix within the finger domain
LINE – long interspersed nuclear element
LTR – long terminal repeat
mRNA – messenger RNA
RNP – ribonucleoprotein
RT – reverse transcriptase
SS – splice site
TPRT – target primed reverse transcription
WC – Watson Crick

Contents

CHAPTER 1 – INTRONS AND SPLICING	1
1.1 Group II introns as splicing ribozymes	1
1.2 Maturase as splicing cofactors	2
1.3 Group II intron subclasses.....	3
1.4 Linear RNA intron structures	3
1.5 Lariat RNA intron structure	5
1.6 Group II intron maturase structures.....	5
1.7 Group II intron RNPs	6
1.8 Thesis Aims	8
1.9 References:.....	12
CHAPTER 2: INTRON CATALYSIS AND DYNAMICS DURING BRANCHING AND EXON LIGATION	15
2.1 Preface.....	15
2.2 Summary.....	15
2.3 Introduction	15
2.4 Results	18
2.4.1 Capturing the group II intron holoenzyme in action	18
2.4.2 Positioning of the branch helix.....	19
2.4.3 5' splice site prior to branching.....	20
2.4.4 Branchpoint A recognition and dynamics.....	21
2.4.5 Branch helix conformational dynamics.....	22
2.4.6 Catalytic mechanism of intron splicing.....	24
2.5 Discussion.....	25
2.5.1 Splicing at the RNP interface	25
2.5.2 Conservation of molecular recognition	27
2.6 Materials and Methods	45
2.7 References	55
CHAPTER 3: REVERSE SPLICING AND INTRONS AS RETROELEMENTS ..	58
3.1 Preface.....	58

3.2 Summary	58
3.3 Introduction	59
3.4 Results	61
3.4.1 Overall architecture of an ancient group II intron retroelement	61
3.4.2 Features of the catalytic RNP core.....	63
3.4.3 Unexpected functional coordination between RNA and protein.....	65
3.4.4 New tertiary interactions with a structured DNA	66
3.4.5 Retroelement primed for attack.....	70
3.5 Discussion.....	71
3.5.1 Insights into Protein Facilitated Ribozyme Catalysis	71
3.5.2 RNP Recognition of DNA Structure: an expanded recognition repertoire	72
3.5.3 Implications for Reverse Splicing and Reverse Transcription	73
3.5.4 Retroelement Poised to Attack	74
3.5.5 Implications for Modern Retroelements	75
3.5.6 Group II RNP lifecycle.....	76
3.6 Materials and Methods	102
3.7 References	113
CHAPTER 4: INSIGHTS INTO THE MECHANISM OF GROUP II INTRON SPLICING	117
Preface.....	117
4.1 Group II intron forward splicing.....	117
4.2 Group II intron reverse splicing.....	119
4.3 Target primed reverse transcription.....	120
4.4 Future Outlook	122
APPENDIX I: STRUCTURES OF A MOBILE RETROELEMENT POISED TO ATTACKS ITS STRUCTURED DNA TARGET.	125
APPENDIX II: STRUCTURAL INSIGHTS INTO INTRON CATALYSIS AND DYNAMICS DURING SPLICING.....	134

List of Figures

Chapter 1

Figure 1.1 Group II intron domain organization

Chapter 2

Figure 2.1 Group II intron splicing.

Figure 2.2 The maturase acts as a branching switch.

Figure 2.3 CryoEM sample preparation of the intron holoenzyme.

Figure 2.4 Local resolution and particle distribution of the RNP reconstructions.

Figure 2.5 CryoEM reconstructions of the splicing pathway intron RNPs.

Figure 2.6 Branch helix RNA and protein interactions.

Figure 2.7 Biochemical validation of branch helix interactions.

Figure 2.8 5' splice site recognition.

Figure 2.9 Branchpoint A recognition prior to branching.

Figure 2.10 Branchpoint A local motions and interactions.

Figure 2.11 Branch helix large scale movements.

Figure 2.12 Splice site interactions throughout the branching pathway.

Figure 2.13 Molecular mechanism of group II RNP branching and exon ligation.

Figure 2.14 Comparison of the heteronuclear metal ion core.

Figure 2.15 Mechanistic comparison of group II introns and the spliceosome.

Figure 2.16 Comparison of protein-branch helix interactions.

Chapter 3

Figure 3.1 Cartoon of the retrotransposition reaction.

Figure 3.2 Purification of a group II intron RNP.

Figure 3.3 Biophysical characterization of the intron RNP.

Figure 3.4 Reverse splicing activity assay.

Figure 3.5 Purification of a group II intron retroelement.

Figure 3.6 CryoEM reconstruction of the intron retroelement.

Figure 3.7 CryoEM structures of a group II intron holoenzyme.

Figure 3.8 Sequence and structure of the holoenzyme DNA target.

Figure 3.9 Catalytic core of the group II intron holoenzyme.

Figure 3.10 Cleavage of the DNA target.

Figure 3.11 Tertiary interactions within the holoRNP.

Figure 3.12 Mechanism of maturase facilitated ribozyme catalysis.

Figure 3.13 Maturase positioning in group II RNPs.

Figure 3.14 Interactions of the structured DNA target with the intron RNP.

Figure 3.15 Shape specific interactions with the structured DNA target.

Figure 3.16 Tertiary interactions with the DNA substrate.

Figure 3.17 Interactions of the maturase DBD with the DNA.

Figure 3.18 Stacking interactions with the intron holoenzyme.

Figure 3.19 CryoEM structures of the apoRNP.

Figure 3.20 An RNP poised to attack.

Figure 3.21 Mimicry of the DNA structural motif.

Figure 3.22 Predicted structure of L1 ORF2p.

Figure 3.23 Group II intron splicing cycle.

Chapter 1 – Introns and Splicing

1.1 Group II introns as splicing ribozymes

Group II introns are large autocatalytic ribozymes that are important for gene expression and metabolism. They are found in plants, fungi, yeasts, and many bacteria^{1,2}, where they catalyze the essential splicing reaction, joining adjacent exons and removing the intervening intron in two consecutive transesterification steps³, which results in mature messenger RNA (mRNA) transcripts^{4,5}. Group II introns share a conserved secondary structure, and they adopt a tertiary fold that centers around a ribozyme active site⁶. Specifically, six structurally defined domains are positioned around a circular hub, with each serving a critical function in stabilization, catalysis, or conformational dynamics⁷⁻¹⁰. To splice, group II introns use a bulged adenosine located in the branch helix as the first nucleophile, and the subsequently exposed 3' OH at the splice site (SS) as the second attacking group to stitch together flanking exons¹¹. The hallmark of this branching reaction is a distinctive 2',5' phosphodiester linkage known as a lariat^{6,12}. Of note, in a competing splicing pathway, the intron can be released as a linear product when a water molecule acts as the first step nucleophile instead, although this is a less physiologically relevant pathway¹⁰. An identical branching mechanism has prevailed in the modern spliceosome and pre-mRNA splicing, which is hypothesized to have originated from the same common ancestor as the group II intron⁵.

Another important reaction that group II introns are implicated in is retrotransposition or reverse splicing, the process of intron integration into a new genomic location¹³. Introns liberated from the branching reaction roam the genome for insertion sites, invading novel targets in a two-step reaction that is exactly the reversal of intron excision¹³. These two

reactions combined enable introns to proliferate within a host genome by copying and pasting themselves^{14,15}. Through evolutionary time, group II introns likely degenerated into the common eukaryotic elements including retrotransposons, nuclear introns and the spliceosome, and telomerases that continue to shape our genome^{6,16}.

1.2 Maturase as splicing cofactors

An important component of the splicing reaction is a protein known as the maturase^{17,18}. While the intron RNA contains the essential catalytic elements of the splicing reaction, the maturase cofactor, or intron encoded protein (IEP), is equally important in helping the intron “mature”¹⁹⁻²¹. The maturase has coevolved with its parent intron and is encoded for within domain 4 (D4) of the intron, which contains a high affinity hairpin structure to which the maturase binds, and an ORF from which the protein is translated¹⁰. The maturase forms a ribonucleoprotein (RNP) holoenzyme with the intron RNA and is a key element in both splicing and retrotransposition reactions²².

Maturase proteins are not only splicing factors, but also potent reverse transcriptase (RT) enzymes^{23,24}. The RT activity is critical after reverse splicing, as the maturase is well positioned to generate a cDNA copy of the RNA-DNA chimeric insertion product, in a process known as target primed reverse transcription (TPRT)^{13,25,26}. In addition to the reverse transcriptase domain, intron maturases are composed of a thumb and DNA binding domain (DBD), which are important for coordinating and stabilizing intron components, and in certain intron classes, an endonuclease that can nick the antisense strand to generate a primer for TPRT¹⁸. By forming highly specific RNP complexes with their maturases, group II introns are primed for splicing and poised for retrotransposition.

1.3 Group II intron subclasses

Group II introns can be primarily classified into three subclasses, IIA, IIB, and IIC based on phylogenetic and secondary structure analysis²⁷. While the same conserved mechanism is used for chemical catalysis and splice site exchange, the intron classes differ at peripheral regions which contain additional components such as stems and hairpins that can form tertiary interactions that bolster the intron core⁶. Of the three families, group IIC introns are one of the most well studied and have served as a model system for investigating splicing, RNA tertiary folding, and RNA catalysis²⁸⁻³². Group IIC introns are the most ancient, and they have the most compact and minimalistic architecture, lacking the more ornate features that decorate higher order introns⁶. They are largely dependent on their maturases to facilitate splicing through lariat formation⁶ and they are often located downstream of transcription terminators, which have structures important for target recognition³³⁻³⁵. Importantly, group IIC introns differ in their 5' exon recognition strategy, with an abbreviated sequence that recognizes a shorter stretch of nucleotides compared to higher order IIA and IIB introns^{10,13}. Nonetheless, group IIC introns have proved to be a crucial resource for investigation of splicing and RNA structure, providing a wealth of information to the intron field.

1.4 Linear RNA intron structures

In truncated intron constructs or in the absence of the maturase cofactor, group IIC introns typically undergo splicing via the linear pathway⁶. As full-length intron constructs were difficult to work with, most of the initial work used a minimal construct that was amenable to crystallization^{6,28,29,32,36}. From these studies, we began to understand the

structural organization and tertiary folds that enable the ribozyme activity of group II introns⁷. At the heart of the ribozyme machine is the catalytic domain, D5, which uses the 2-nt bulge and catalytic triplex, in conjunction with the J2/3 linker, to coordinate the two Mg^{2+} ions necessary for catalysis^{28,32} (Figure 1.1). D1 acts as a scaffold, much like an ‘inverted cup’, surrounding D5, and engaging in several tertiary interactions with the core domain to fasten it within the overall structure^{6,32}. D1 is the largest domain within the intron and is an important folding intermediate that organizes all subsequent domains, using intricate elements such as kissing loops and a Z-anchor to stabilize the structure⁷. Additionally, it contains the D1d coordination loop which specifies the substrate recognition sequences known as the exon binding sites 1 and 3 (EBS1 and EBS3). D2 is coaxially stacked with the last stem of D1 while D3 projects outward, although it can twist back to brace the backside of D1, as in IIA and IIB introns^{18,37}. The shortened D4 stem is directed outwards to allow the encoded ORF to be translated^{6,10}. Importantly, D6 was not present in these structures, owing to instability either from hydrolysis of the branch helix or the absence of stabilization from the encoded maturase³².

Apart from revealing the architectural details of the intron structure, this earlier work also crucially identified the heteronuclear metal ion core responsible for ribozyme catalysis. From anomalous scattering and heavy atom soaking data, it was determined that the metal ion cluster consists of a pair of monovalent K^+ ions (K1 and K2), that are interchangeable with NH_4^+ , and a second pair of divalent Mg^{2+} ions (M1 and M2)^{28,32}. K1 interacts with the D5 phosphate backbone to facilitate M1 docking and K2 stabilizes the 5' splice site (5' SS) for nucleophilic attack²⁸. The divalent metal ions activate the nucleophile (M1) and stabilize the leaving group (M2) respectively²⁸. These observations explained the

specific metal requirements for splicing activity during both transesterification steps, and indeed, the same catalytic elements were found to be present in the spliceosome^{38,39}.

1.5 Lariat RNA intron structure

Another structural advance occurred with the visualization of the *Pylaiella littoralis* (*P. li*) intron from brown algae, which was the first lariat structure that captured the positioning of the branch helix³⁷. The structure offered insights into the mechanism of branch site attack during the first step of splicing and revealed an additional exon binding site (EBS2) that confers base pairing specificity, which is adjacent to a newly observed kissing loop present only in higher order introns^{6,37}. The larger IIB intron also resolved novel tertiary interactions that stem from the more elaborate D2 and D3, which act to brace the periphery of the intron³⁷. Most significantly, the D6 branch helix was found to be docked onto D2, engaged by a pair of tetraloop receptor interactions, thereby pulling the lariat away from the ribozyme active site^{6,10,37}. This structure hinted at branch helix dynamics that could potentially facilitate splice site exchange, as mutation of the tetraloop receptors, inhibited the second step of splicing³⁷. While certainly informative, important questions remained regarding the states immediately prior to both steps of splicing and the role of the maturase protein, which was not present in the structure.

1.6 Group II intron maturase structures

For a long time, the maturase proteins of group II introns resisted purification due to solubility and aggregation problems¹⁷. Eventually structures of group IIC maturases emerged, allowing a glimpse into the domain organization and biochemical characteristics

of these RT enzymes^{17,22,40-42}. Specifically, it was found that IIC maturases had remarkably similar RT domain scaffolds to the spliceosomal Prp8 protein and RNA-dependent RNA polymerases from flaviviruses, reinforcing the evolutionary connection between introns, spliceosomal proteins, and RNA viruses¹⁷. In fact, this implied that group II maturases are likely ancestrally linked to non-LTR retroelement RTs, which have similar conserved finger, helical thumb, and insertion in the finger domains^{10,17,22,41-43}.

These maturases also have unique surface charges and RT elements, thereby enabling them to simultaneously function as splicing cofactors and highly processive RT enzymes^{23,24}. The outer surfaces of the maturase, especially around the N-terminus, have a highly basic charge, which mediates high affinity interactions with D4, while the inner face is reserved for the RT catalytic residues, suggesting that the two functions can occur simultaneously^{17,41}. This strategy maximizes usage of the surface area of the multifunctional maturase protein. From the structures, IIC maturases were found to have alpha-helical insertions that mediate distinctly high processivity and fidelity, which is imperative for ensuring accurate copying of the intron sequence after reverse splicing, such that the intron tertiary structure remains intact^{41,42}. These structures provided the first glimpse into the evolution of specialized RT enzymes into splicing factors, but in isolation, they lack information on how a RNP splicing complex may assemble and function.

1.7 Group II intron RNPs

Advances in cryogenic electron microscopy (cryoEM) allowed capture of intron-maturase holoenzymes, and with it, a new set of structures elucidating the role of the maturase within the RNP complex and its function in retrotransposition^{18,44,45}. The first

RNP structure was of a group IIA LtrA maturase bound to its intron in the post-catalytic state, after both steps of splicing had occurred¹⁸. Here, the maturase was anchored within the RNP by interactions of the outer surface of the RT domain with D4a and by binding subdomains of the D1 scaffold, securing the protein against the ligated exons and the EBSs^{10,18}. The intron portion of the complex adopted folds much like those observed in earlier crystal structures^{28,29,32,37,46}. The singular structure also did not allow for interpretation of branch helix dynamics that might occur during splice site exchange¹⁸. Despite being a major step forward in understanding group II holoenzymes, the exact mechanistic role of the maturase protein and its ability to promote branching remained unclear.

More recently, a second series of structures of a group IIB intron bound to its DNA target and undergoing the first two steps of reverse splicing revealed further mechanistic details⁴⁴. To capture the intron holoenzyme during retrotransposition, the splicing reaction was driven to completion before the spliced products were captured by affinity chromatography, using a tagged DNA substrate⁴⁴. When vitrified on cryoEM grids and imaged, two structures were resolved, corresponding to the step immediately before intron insertion and the state after the holoenzyme has ligated its 3' end to and attacked the DNA at the insertion site. This work captured several important interactions that answered longstanding questions within the field. It had been established that group II introns maintain the same catalytic core throughout the steps of splicing, but this meant that the splice sites, which differ from the first step to the second step and are ~20Å apart, had to be exchanged⁶. These structures demonstrated that a swinging motion of the branch helix accomplished splice site exchange during retrotransposition, with similar motions observed

in the spliceosome when transitioning from the B* to C* complex^{44,47}. This motion was seemingly mediated by an expansion in the D5 helix that triggers D6 helix disengagement from D2, pulling the branch helix upwards against the maturase DBD to transition the RNP to the second step of reverse splicing. The maturase protein was also observed to mechanistically contribute to stabilization of the branch helix during reverse splicing, suggesting an indirect role in promoting branching⁴⁸.

Taken together, the intron holoenzyme structures informed our understanding of the role of the maturase protein within the RNA assembly and established the role of branch helix dynamics in splice site exchange. CryoEM enabled high-resolution structures of intron RNP structures, but more thorough biochemical and structural characterization had to be conducted to examine the steps along the branching pathway, the role of the branchpoint adenosine, and the DNA targets of intron holoenzymes, to gain a comprehensive understanding of the full intron splicing cycle which encompasses the forward and reverse splicing reactions.

1.8 Thesis Aims

It is clear that an abundance of research in the fields of group II introns, RNA folding, and ribozymes in the past decade has led to significant developments and greater understanding of the structures and mechanisms of splicing. Various classes of introns, in both linear, lariat and intron holoenzyme forms have been visualized using both x-ray crystallography, and more recently, cryoEM. The structural motifs and catalytic elements that make up group II introns are well defined and the maturase proteins that associate with introns have been characterized. The role of group II RNPs as not only holoenzymes that

act as splicing machines, but also retroelements capable of inserting into new DNA sites is better appreciated. Despite all this work, there are still a number of key questions that remained unanswered. For example, how is the intron holoenzyme organized prior to branching or prior to exon ligation? What conformational changes are necessary for splice site exchange after branching? Are these dynamics the same from branching to retrotransposition? Why is the branchpoint adenosine essential for lariat formation? How do group IIC introns engage in reverse splicing? How do they recognize their structured DNA targets? To address these lingering questions, my thesis sought to investigate the structural and mechanistic basis of group IIC intron splicing and retrotransposition.

Chapter 2 focuses on the conformational states along the group II intron RNP splicing pathway. By capturing the RNP in the state immediately before branching, we visualized how the branchpoint adenosine (bpA) and the splice site are held in place through molecular interactions between the branch helix and conserved regions of the RNP. The RNP structures preceding and following the exon ligation step allow us to visualize large movements of the branch helix, local movements of the branchpoint, and the splice site exchange that occurs between the two steps of splicing.

Chapter 3 discusses the role of group II intron RNPs as retroelements. We resolved the RNP bound to its structured DNA target primed to initiate the reverse splicing reaction. By comparing with the apoRNP structure, we can understand the molecular basis of RNP recognition of both the shape and sequence of their substrates, rationalizing the unique hairpin structures observed at their DNA insertion sites.

Chapter 4 summarizes the work done in this thesis and puts this work into a broader context within the field. This includes speculating as to what important questions remain unanswered and future directions that this field might progress towards.

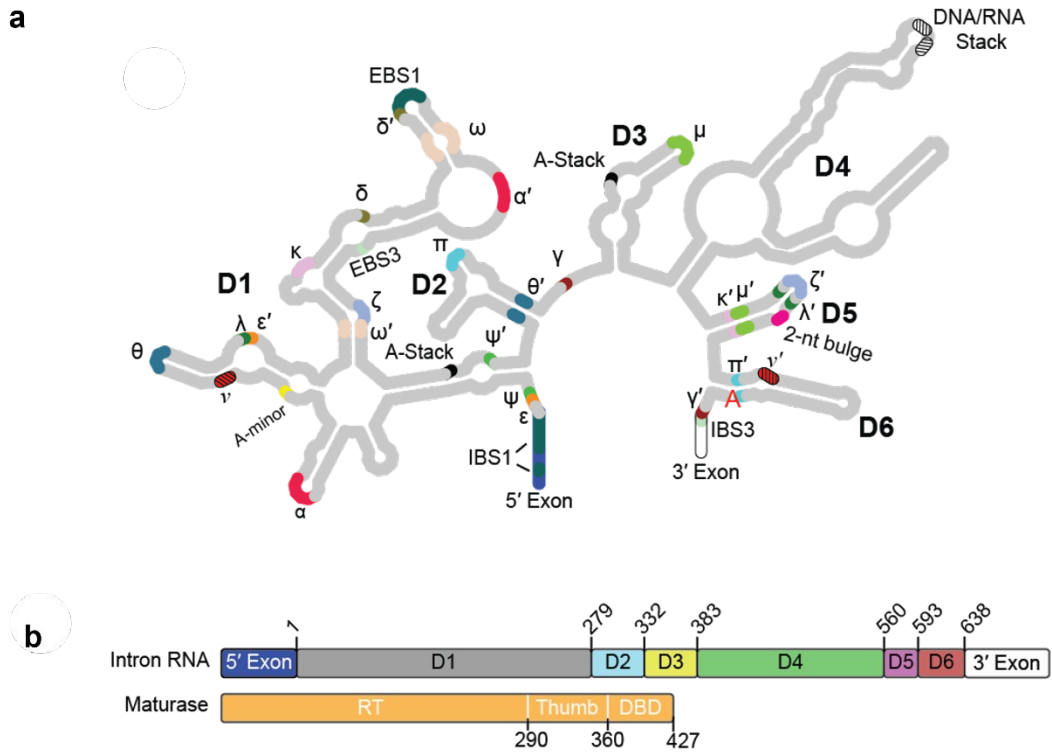


Figure 1.1 Group II intron domain organization

a. Secondary structure diagram of the group IIC intron with key elements and long-range tertiary interactions annotated. b. domain organization of the group II intron and its maturase cofactor.

1.9 References:

- 1 Ferat, J. L. & Michel, F. Group II self-splicing introns in bacteria. *Nature* **364**, 358-361, doi:10.1038/364358a0 (1993).
- 2 Bonen, L. & Vogel, J. The ins and outs of group II introns. *Trends Genet* **17**, 322-331, doi:10.1016/s0168-9525(01)02324-1 (2001).
- 3 Peebles, C. L. *et al.* A self-splicing RNA excises an intron lariat. *Cell* **44**, 213-223, doi:10.1016/0092-8674(86)90755-5 (1986).
- 4 Shi, Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat Rev Mol Cell Biol* **18**, 655-670, doi:10.1038/nrm.2017.86 (2017).
- 5 Galej, W. P., Toor, N., Newman, A. J. & Nagai, K. Molecular Mechanism and Evolution of Nuclear Pre-mRNA and Group II Intron Splicing: Insights from Cryo-Electron Microscopy Structures. *Chem Rev* **118**, 4156-4176, doi:10.1021/acs.chemrev.7b00499 (2018).
- 6 Pyle, A. M. Group II Intron Self-Splicing. *Annu Rev Biophys* **45**, 183-205, doi:10.1146/annurev-biophys-062215-011149 (2016).
- 7 Zhao, C., Rajashankar, K. R., Marcia, M. & Pyle, A. M. Crystal structure of group II intron domain 1 reveals a template for RNA assembly. *Nat Chem Biol* **11**, 967-972, doi:10.1038/nchembio.1949 (2015).
- 8 Fedorova, O. & Pyle, A. M. A conserved element that stabilizes the group II intron active site. *RNA* **14**, 1048-1056, doi:10.1261/rna.942308 (2008).
- 9 Fedorova, O., Mitros, T. & Pyle, A. M. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol* **330**, 197-209, doi:10.1016/s0022-2836(03)00594-1 (2003).
- 10 Zhao, C. & Pyle, A. M. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci* **42**, 470-482, doi:10.1016/j.tibs.2017.03.007 (2017).
- 11 Pyle, A. M. The tertiary structure of group II introns: implications for biological function and evolution. *Crit Rev Biochem Mol Biol* **45**, 215-232, doi:10.3109/10409231003796523 (2010).
- 12 Xu, L., Liu, T., Chung, K. & Pyle, A. M. Structural insights into intron catalysis and dynamics during splicing. *Nature*, doi:10.1038/s41586-023-06746-6 (2023).
- 13 Chung, K. *et al.* Structures of a mobile intron retroelement poised to attack its structured DNA target. *Science* **378**, 627-634, doi:10.1126/science.abq2844 (2022).
- 14 Belfort, M. & Lambowitz, A. M. Group II Intron RNPs and Reverse Transcriptases: From Retroelements to Research Tools. *Cold Spring Harb Perspect Biol* **11**, doi:10.1101/cshperspect.a032375 (2019).
- 15 Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3**, MDNA3-0050-2014, doi:10.1128/microbiolspec.MDNA3-0050-2014 (2015).
- 16 Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* **12**, 187-215, doi:10.1146/annurev-genom-082509-141802 (2011).
- 17 Zhao, C. & Pyle, A. M. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol* **23**, 558-565, doi:10.1038/nsmb.3224 (2016).
- 18 Qu, G. *et al.* Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* **23**, 549-557, doi:10.1038/nsmb.3220 (2016).
- 19 Kennell, J. C., Moran, J. V., Perlman, P. S., Butow, R. A. & Lambowitz, A. M. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell* **73**, 133-146, doi:10.1016/0092-8674(93)90166-n (1993).
- 20 Matsuura, M., Noah, J. W. & Lambowitz, A. M. Mechanism of maturase-promoted group II intron splicing. *EMBO J* **20**, 7259-7270, doi:10.1093/emboj/20.24.7259 (2001).
- 21 Matsuura, M. *et al.* A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev* **11**, 2910-2924, doi:10.1101/gad.11.21.2910 (1997).
- 22 Zhao, C. & Pyle, A. M. The group II intron maturase: a reverse transcriptase and splicing factor go hand in hand. *Curr Opin Struct Biol* **47**, 30-39, doi:10.1016/j.sbi.2017.05.002 (2017).

- 23 Guo, L. T., Olson, S., Patel, S., Graveley, B. R. & Pyle, A. M. Direct tracking of reverse-
transcriptase speed and template sensitivity: implications for sequencing and analysis of long RNA
molecules. *Nucleic Acids Res* **50**, 6980-6989, doi:10.1093/nar/gkac518 (2022).
- 24 Guo, L. T. *et al.* Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-
Generation Reverse Transcriptase. *J Mol Biol* **432**, 3338-3352, doi:10.1016/j.jmb.2020.03.022
(2020).
- 25 Zimmerly, S., Guo, H., Perlman, P. S. & Lambowitz, A. M. Group II intron mobility occurs by
target DNA-primed reverse transcription. *Cell* **82**, 545-554, doi:10.1016/0092-8674(95)90027-6
(1995).
- 26 Moran, J. V. *et al.* Mobile group II introns of yeast mitochondrial DNA are novel site-specific
retroelements. *Mol Cell Biol* **15**, 2828-2838, doi:10.1128/MCB.15.5.2828 (1995).
- 27 Zimmerly, S. & Semper, C. Evolution of group II introns. *Mob DNA* **6**, 7, doi:10.1186/s13100-
015-0037-5 (2015).
- 28 Marcia, M. & Pyle, A. M. Visualizing group II intron catalysis through the stages of splicing. *Cell*
151, 497-507, doi:10.1016/j.cell.2012.09.033 (2012).
- 29 Chan, R. T., Robart, A. R., Rajashankar, K. R., Pyle, A. M. & Toor, N. Crystal structure of a group
II intron in the pre-catalytic state. *Nat Struct Mol Biol* **19**, 555-557, doi:10.1038/nsmb.2270
(2012).
- 30 Keating, K. S., Toor, N., Perlman, P. S. & Pyle, A. M. A structural analysis of the group II intron
active site and implications for the spliceosome. *RNA* **16**, 1-9, doi:10.1261/rna.1791310 (2010).
- 31 Toor, N. *et al.* Tertiary architecture of the *Oceanobacillus ihayensis* group II intron. *RNA* **16**, 57-
69, doi:10.1261/rna.1844010 (2010).
- 32 Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II
intron. *Science* **320**, 77-82, doi:10.1126/science.1153803 (2008).
- 33 Toor, N., Robart, A. R., Christianson, J. & Zimmerly, S. Self-splicing of a group IIC intron: 5'
exon recognition and alternative 5' splicing events implicate the stem-loop motif of a
transcriptional terminator. *Nucleic Acids Res* **34**, 6461-6471, doi:10.1093/nar/gkl820 (2006).
- 34 Robart, A. R., Seo, W. & Zimmerly, S. Insertion of group II intron retroelements after intrinsic
transcriptional terminators. *Proc Natl Acad Sci U S A* **104**, 6620-6625,
doi:10.1073/pnas.0700561104 (2007).
- 35 Haack, D. B. & Toor, N. Recognition of transcription terminators during retrotransposition: How
to keep a group II intron quiet. *Mol Cell* **83**, 332-334, doi:10.1016/j.molcel.2022.12.027 (2023).
- 36 Costa, M., Walbott, H., Monachello, D., Westhof, E. & Michel, F. Crystal structures of a group II
intron lariat primed for reverse splicing. *Science* **354**, doi:10.1126/science.aaf9258 (2016).
- 37 Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a
eukaryotic group II intron lariat. *Nature* **514**, 193-197, doi:10.1038/nature13790 (2014).
- 38 Wilkinson, M. E., Fica, S. M., Galej, W. P. & Nagai, K. Structural basis for conformational
equilibrium of the catalytic spliceosome. *Mol Cell* **81**, 1439-1452 e1439,
doi:10.1016/j.molcel.2021.02.021 (2021).
- 39 Fica, S. M., Mefford, M. A., Piccirilli, J. A. & Staley, J. P. Evidence for a group II intron-like
catalytic triplex in the spliceosome. *Nat Struct Mol Biol* **21**, 464-471, doi:10.1038/nsmb.2815
(2014).
- 40 Mohr, G. *et al.* A Highly Proliferative Group IIC Intron from *Geobacillus stearothermophilus*
Reveals New Features of Group II Intron Mobility and Splicing. *J Mol Biol* **430**, 2760-2783,
doi:10.1016/j.jmb.2018.06.019 (2018).
- 41 Stamos, J. L., Lentzsch, A. M. & Lambowitz, A. M. Structure of a Thermostable Group II Intron
Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary Implications.
Mol Cell **68**, 926-939 e924, doi:10.1016/j.molcel.2017.10.024 (2017).
- 42 Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a
metazoan group II intron. *RNA* **24**, 183-195, doi:10.1261/rna.063479.117 (2018).
- 43 Wilkinson, M. E., Frangieh, C. J., Macrae, R. K. & Zhang, F. Structure of the R2 non-LTR
retrotransposon initiating target-primed reverse transcription. *Science* **380**, 301-308,
doi:10.1126/science.adg7883 (2023).
- 44 Haack, D. B. *et al.* Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA. *Cell* **178**,
612-623 e612, doi:10.1016/j.cell.2019.06.035 (2019).

- 45 Liu, N. *et al.* Exon and protein positioning in a pre-catalytic group II intron RNP primed for
splicing. *Nucleic Acids Res* **48**, 11185-11198, doi:10.1093/nar/gkaa773 (2020).
- 46 Toor, N., Rajashankar, K., Keating, K. S. & Pyle, A. M. Structural basis for exon recognition by a
group II intron. *Nat Struct Mol Biol* **15**, 1221-1222, doi:10.1038/nsmb.1509 (2008).
- 47 Fica, S. M., Oubridge, C., Wilkinson, M. E., Newman, A. J. & Nagai, K. A human postcatalytic
spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science* **363**,
710-714, doi:10.1126/science.aaw5569 (2019).
- 48 Galej, W. P. *et al.* Cryo-EM structure of the spliceosome immediately after branching. *Nature* **537**,
197-201, doi:10.1038/nature19316 (2016).

Chapter 2: Intron catalysis and dynamics during branching and exon ligation

2.1 Preface

This chapter is adapted from my work entitled “*Structural insights into intron catalysis and dynamics during splicing*”, co-authored with Ling Xu, Tianshuo Liu, and Anna Pyle, published in *Nature*¹. The biochemical assays and cryoEM work were done jointly by me, Ling Xu and Tianshuo Liu.

2.2 Summary

The group II intron ribonucleoprotein is an archetypal splicing system with numerous mechanistic parallels to the spliceosome, including excision of lariat introns^{2,3}. Despite the importance of branching in RNA metabolism, structural understanding of this process has remained elusive. Here, we present a comprehensive analysis of three cryoEM structures captured along the splicing pathway. They reveal the network of molecular interactions that specifies the branchpoint adenosine and positions key functional groups to catalyze lariat formation and coordinate exon ligation. The structures also reveal conformational rearrangements of the branch helix and the mechanism of splice site exchange that facilitates the transition from branching to ligation. These findings shed light on the evolution of splicing and highlight the conservation of structural components, catalytic mechanism, and dynamical strategies retained through time in pre-mRNA splicing machines.

2.3 Introduction

Splicing lies at the heart of RNA metabolism in eukaryotes. During this indispensable stage of gene expression, introns are removed from pre-mRNA transcripts to generate

mature mRNAs^{2,4,5} (Figure 2.1). The modern spliceosome, the molecular machine that executes the splicing reaction, is thought to originate from the same ancestral molecule as the self-splicing group II introns that are still commonly found in bacteria and organelles of plants and fungi³. Group II introns are large ribozymes that catalyze their own excision from precursor RNA transcripts⁶. Both splicing machineries form a conserved active site that hosts the catalytically essential heteronuclear metal ion core^{7,8}. Moreover, they both branch using a bulged adenosine nucleophile, forming the distinctive lariat intron featuring a 2',5'-linked phosphodiester linkage. Intron D4 contains an open reading frame (ORF) that encodes a specialized multi-domain protein, the maturase, that shares strong structural similarity to Prp8, a central protein component of the U5 snRNP^{9,10}. Through formation of a ribonucleoprotein (RNP) holoenzyme with the parent intron RNA, the maturase facilitates intron splicing out of the transcript as well as retrohoming into novel genomic loci¹¹.

In light of these structural and mechanistic similarities, group II intron RNPs have become a prototypical system for studying the general biochemical principles of RNA splicing and the molecular evolution of splicing machines¹². Despite advances in the visualization of group II intron RNAs^{8,13-15} and RNPs¹⁶⁻¹⁸, the structural organization of group II intron systems as they undergo branching and then coordinate the two steps of splicing has remained elusive. Therefore, it is unclear how group II introns properly recognize the branchpoint and the 5' splice site (5'SS) and how maturases facilitate the branching reaction. These questions are of significance as they provide clues on the origin of intron branching, which is among the most fundamental reactions in RNA biology.

To visualize the conformational states along the group II intron RNP splicing pathway, we chose the group IIC intron from *Eubacterium rectale* (*E.r.*) and its encoded maturase, MarathonRT¹⁹, as the model system. The maturase acts as a branching switch that shifts the intron splicing pathway from hydrolysis to branching (Figure 2.2). Here we employed single-particle cryogenic electron microscopy (cryoEM) to obtain the structures of the RNP at each sequential stage during splicing. By capturing the RNP in the state immediately before branching (3.0Å overall), we visualized how the branchpoint adenosine (bpA) and the splice site (SS) are held in place through molecular interactions between the branch helix and conserved regions of the RNP. Our structural observations reveal a close resemblance between group II intron RNPs and the spliceosome in terms of branchpoint recognition and branch helix positioning. We also gained unique insights into the strategy by which the attacking 2' OH is precisely positioned within activation distance to the catalytic metal M1, ready for nucleophilic attack. This high-resolution view enables us to construct a complete and catalytically relevant molecular picture of the splicing active site before branching, which has largely eluded structural characterization despite earlier hints²⁰.

In addition to the pre-branching RNP structure, we present the RNP structures preceding and following the exon ligation step. These structures allow us to visualize large movements of the branch helix, local movements of the branchpoint, and the splice site exchange that occurs between the two steps of splicing. The conformational dynamics of the spliceosome branch helix recapitulates that of the group II intron, thereby demonstrating that branch helix dynamics is itself a conserved physical attribute that is codified within splicing machines.

2.4 Results

2.4.1 Capturing the group II intron holoenzyme in action

To gain insights into the mechanism of forward splicing (Figure 2.1), through the branching, lariat formation pathway, we sought to visualize structures of the group II RNP holoenzyme at each stage along the pathway. We incubated the intron RNA in the presence of the maturase (MarathonRT), which acts as a branching switch that shifts the intron splicing pathway from hydrolysis to branching, to obtain lariat holoenzyme complexes (Figure 2.2). The two steps of group II intron splicing are spontaneous and do not require energy input or step-specific factors, apart from the maturase⁹. Therefore, it is difficult to stall the reaction steps without disrupting active site arrangement or introducing artefacts. Previous attempts resorted to mutations to key elements, including the catalytic triplex or the branchpoint adenosine, that made it challenging to interpret the molecular mechanism of branching^{21,22}. A biochemical technique that circumvents this issue is replacing Mg^{2+} with Ca^{2+} when preparing the splicing reaction mixture⁸. This allowed us to obtain complexes stalled in the precursor and branching intermediate states. To obtain the post-ligation complex, we first assembled the lariat apoRNP¹⁸, before supplementing the mixture with an oligonucleotide equivalent to the ligated exons (Figure 2.3).

These samples were vitrified on cryoEM grids and appeared as monodispersed particles when imaged. From earlier work, it was known that group II RNP samples suffer from orientation bias, so we used a combined approach, preparing grids with the Chameleon and Vitrobot, and collecting non-tilted and tilted datasets, respectively²³⁻²⁵. The improved particle distribution (Figure 2.4) allowed us to generate two distinct maps that we could assign to the pre-branching (pre-1F) and pre-ligation (pre-2F) states. These two maps had

resolutions of 2.8Å and 2.9Å respectively for the catalytic core. We also obtained a 3D reconstruction for the post-ligation (post-2F) holoenzyme, and the resolution of the catalytic core was determined to be 2.9Å. These data represented the full molecular picture of the group II intron RNP as it proceeds along the branching pathway (Figure 2.5).

2.4.2 Positioning of the branch helix

To splice via the branching pathway, the group IIC holoenzyme uses a number of intramolecular RNA and intermolecular RNA-protein interactions that stabilize and precisely arrange the branch helix, D6, in the pre-1F, branching competent form (Figure 2.6a). The D1 scaffold contributes to this network by forming a newly discovered, extended, interlocked interaction between D1c and D6 (denoted $v-v'$) (Figure 2.6b). This features the long-range base pairing of G86 and C601, which are bulged nucleotides with strong conservation signature, suggesting their importance in arranging the branching reaction. Consistent with their role in D6 organization, deletion of G86, C601, or both leads to defects in branching. Meanwhile, substitution of the GC pair for an AU pair, which preserves base pairing, partially rescues branching. An adjacent wobble pair in the D1c stem, between G84 and U104, anchors a supporting molecular network to further lock the conformational sampling of the D6 branch helix. Here, a G84A/G86A dual mutation, which maintains the pairing within D1c, but now interferes with the interdomain D1c to D6 pairing, has a pronounced effect on branching. The structural observations and biochemical validation (Figure 2.7a), highlight the significance of this extensive interaction network in properly positioning D6 for lariat formation.

On the opposite face of the D6 helix, several clusters of interactions form an RNA-protein intermolecular interface with the thumb and DBD domains of the maturase protein (Figure 2.6c-e). One cluster, involving Trp310, Ser313, and Gln359 grasps onto the basal stem of D6. Mutations of the residues in this cluster to alanine disables intron branching. At the junction between the thumb and DBD domains of the maturase, lies a second group of amino acids (Thr362 and Asn365) that secures the central portion of the branch helix right next to the branchpoint adenosine and the ribozyme active site. These two clusters involve coordination to the phosphate backbone, but remarkably, there is a highly conserved lysine residue that protrudes into the major groove of D6, contacting the 5' SS, specifically N7 of G1. This interaction juxtaposes the first step nucleophile with the scissile phosphate and accordingly, a single Lys361Ala mutant abolishes branching as does the Lys361/Thr362/Asn365 triple alanine mutant. The final players in this network are Lys372 and Arg377 of the DBD, which hold the distal stem of D6. Again, mutations (Figure 2.7b) introduced at these sites interrupt branching, consistent with structural observations.

This expansive molecular network freezes D6 in the branching competent conformation and secures the branchpoint adenosine against the 5' SS, thereby explain why the maturase is indispensable for branching.

2.4.3 5' splice site prior to branching

The 5' SS must be meticulously arranged so that it has the proper orientation relative to the branchpoint adenosine. In the pre-1F structure, the 5' SS opens its Watson-Crick (WC) edges to engage in tertiary interactions with the branch helix. The first two nucleotides of the intron, G1 and U2, play key roles in priming the holoenzyme for branching.

Specifically, O6 of G1 engages the 2' OH of C633, the nucleotide downstream of the branchpoint, which effectively brings D6 close to the 5' SS. U2 participates in this network through a base triple interaction with the G599-C629 base pair. With the pre-1F structure, we now have a stronger understanding of the intron 5' SS and we can explain its role in branch helix positioning. These observations support the conservation signatures at the 5' SS seen in group II and spliceosomal introns²⁶ (Figure 2.8).

2.4.4 Branchpoint A recognition and dynamics

The branch site is almost invariantly an adenosine in spliceosomal and group II introns^{27,28}. Yet the interaction network that recognizes and activates the branchpoint nucleotide for catalysis has eluded structural characterization due to lack of resolution. In the pre-1F structure, we can structurally rationalize the branchpoint recognition strategy. With a local resolution of 2.8Å proximal to the branchpoint, we were able to confidently build in a model that reveals the details of branch site recognition (Figure 2.9a). The branchpoint adenosine (A632, bpA) forms a base triple interaction with the G598-C630 base pair. N6 of the bpA hydrogen bonds with O2 of C630, consistent with earlier chemogenetic studies²⁸. The interaction partner of the bpA is located two nucleotides upstream, and is almost exclusively a pyrimidine, as confirmed by covariation analysis²⁹. As part of this base triple, N1 of the bpA hydrogen bonds with the 2' OH of C630. Similar trends are observed in the spliceosome^{7,30}. Recent models of the yeast C complex⁷ revealed the same molecular interaction between the bpA and the highly conserved uridine located two nucleotides upstream (Figure 2.9b). A similar bpA recognition strategy has also been proposed in the C complex of the human spliceosome³⁰. Our structure therefore unveils a

novel mechanistic parallel for defining the branchpoint in splicing machines, which appears to be hard coded by molecular evolution.

Our structures also gave us insights into the local conformational dynamics of the bpA along the branching pathway. In the pre-1F state, the bpA adopts an unusual conformation that causes it to point toward the major groove of the branch helix, via the base triple interaction. This conformation leads to distortion of the bpA sugar-phosphate backbone, which places the 2' OH next to the scissile phosphate. After branching, as the complex transitions to the pre-2F, pre-ligation state, the bpA flips to the opposite side of the branch helix, and points toward the 3' end of D6 (Figure 2.10a). This dramatic conformational rearrangement of the bpA relaxes the backbone distortion^{6,31}, potentially releasing free energy that compensates for the energetic cost of disengaging the interactions that originally anchored the bpA⁶. Upon exon ligation, the 3' end of the intron (C635 and G636) moves further inwards and contacts the Hoogsteen face of the bpA in the post-2F structure (Figure 2.10b). These additional molecular interactions, formed after exon ligation, limit the conformational flexibility of the bpA and mark the termination of the splicing pathway.

2.4.5 Branch helix conformational dynamics

In addition to the local dynamics of the bpA, comparison of the intron RNP structures reveals a set of tertiary contact rearrangements that are needed to coordinate the sequential steps of splicing (Figure 2.11). In the pre-1F state, the intron recognizes the 5' exon through the EBS1-IBS1 interaction, and the branch helix adopts the D1c, and maturase docked, up conformation. In this arrangement, an array of long-range interactions form between J4/5 (A559 and A560), and J5/6 (U591, G592, U593) which participate in a coordinated series

of interactions. This network begins with a canonical base pair (A560-U591) and continues with a non-canonical pairing, in which the Hoogsteen edge of A559 interacts with the sugar edge of G592 and is capped by the final nucleotide of J5/6, U593, which stacks beneath A559. The phosphate backbone connecting D5 and D6 adopts a bent conformation, flipping adjacent nucleotides to opposing sides due to the constraints imposed by the J5/6 interaction network and the interactions that pull D6 into the up position.

After branching, D6 undergoes a structural rearrangement that involves a $\sim 90^\circ$ swing, to the down position. This motion pulls the 5' SS and the newly formed lariat bond $\sim 21 \text{ \AA}$ out of the active site and exchanges it for the 3' SS, thereby preparing the active site for exon ligation. During this transition, the ν - ν' tertiary interaction and D6-maturase contacts are disrupted. In the resulting pre-2F structure, D6 docks onto D2, engaging π - π' , which latches onto the branch helix, pulling D6 and the covalently linked 3' exon, into position. This allows formation of the EBS3-IBS3 base pairing that defines the 3' SS (U(+1)-A231). Comparison of the catalytic D5 helix in the pre-1F and pre-2F structures reveals that it remains stationary within the D1 scaffold (RMSD = 0.5 \AA). Instead, movement of the D6 helix hinges upon the J5/6 linker and appears as motion of the branch helix relative to a fixed RNP body. Swinging of D6 into the plane of the RNP relaxes the bent conformation of J5/6, enabling an exchange of substrates within the active site, and may be driven by the release of structural constraints from 5' SS cleavage, driving the branching reaction forward⁶. The structural importance of J5/6 in branching is consistent with mutational studies that investigated its biochemical function in positioning of the branch helix^{32,33}. Further movement of D6 is observed upon completion of splicing, where there is minor motion of the D6 3' end, which tucks inwards, allowing engagement of γ - γ' (A327-U638).

Our structures provide the first molecular insights (Figure 2.12) into the conformational rearrangements and sequential transitions that are required for branching and splice site specification during group II intron splicing.

2.4.6 Catalytic mechanism of intron splicing

Having revealed the dynamical strategies employed by the group II intron RNP throughout the branching pathway (Figure 2.13), we next sought to visualize the catalytic mechanisms for each step. Catalysis of the branching reaction is potentiated by a heteronuclear metal ion core³⁴ organized around the catalytic triplex and the 2-nt bulge of D5. Through precise positioning of D6 and formation of intra-D6 interactions, the first-step nucleophile (2' OH of the bpA) is positioned by catalytic metal M1 and placed within the activation distance (2.3Å), where it is poised for nucleophilic attack. Remarkably, the attacking 2' OH nucleophile in the pre-1F structure occupies an identical position as the water nucleophile in an earlier pre-hydrolytic structure, which provides an unambiguous explanation for the competitive nature of the two splicing pathways³⁵. In the absence of the maturase, without interactions with the protein to latch it into place in the branching competent form, conformational sampling of the branch helix allows water to outcompete the bpA, resulting in the linear hydrolytic pathway^{6,27}.

The scissile phosphate of the 5' SS, between U(-1) and G1, adopts the same sharply kinked conformation previously observed for the hydrolytic, pre-catalytic state⁸. The pro-Rp oxygen of the scissile phosphate coordinates both M1 and M2 while the 3' bridging oxygen is in direct contact with M2, facilitating departure of the 3' oxyanion leaving group. This high-resolution view of the active site in the pre-branching state hence provides the

first direct visualization of the two-metal ion mechanism for group II intron branching proposed three decades ago³⁴. Additionally, we identified two strong, globular densities around the divalent metal core, whose positions correspond to the previously identified monovalent ions, K1 and K2⁸. Our findings therefore highlight the formation of a heteronuclear metal ion core as a general catalytic strategy fundamental to RNA splicing^{7,8,36}.

Upon cleavage of the 5' SS, D6 movement brings the first-step nucleophile and the now covalently linked G1 out of the active site. The first-step leaving group, the U(-1) 3' OH, remains tightly coordinated with catalytic metal M2 and becomes the activated second-step nucleophile. The second-step scissile phosphate between U638 and U(+1) then becomes visible in the pre-2F state, adopting the same pre-cleavage kinked configuration. These data establish that the same active site is used for both splicing steps without modifying the catalytic ion configuration nor the metal-binding platform, as was seen in previous intron RNA structures⁸. Moreover, our structure of the post-2F state with the ligated exon bound shows that the metal core remains well organized. The 3' bridging oxygen of U(-1) remains associated with M2, and the 3' OH of U638 is coordinated with M1. With this series of snapshots, we now know that a common heteronuclear metal ion core persists during both steps of splicing.

2.5 Discussion

2.5.1 Splicing at the RNP interface

The spliceosome and group II introns not only share structural and chemical components (Figure 2.14), but also a conserved dynamical strategy for sequential rearrangement

between the steps of splicing. Direct parallels can be drawn between the motions of the D6 helix in the group II intron and the branch helix in the spliceosome. Comparison of their branching states reveals the same 90° swing of the U2-intron branch helix during the transition from the branching B* complex to the exon ligation C* complex. Analogous conformational dynamics are observed within the group II intron holoenzyme, as it swaps splice sites without disrupting the catalytic core, when transitioning between the steps of splicing. The branch helix swinging motion has equivalent centers of rotation to that of the spliceosome. The J5/6 linker in the group II intron acts as a hinge, much like the corresponding U2/U6 linker in the spliceosome³³. Intriguingly, as in the spliceosome (U2/U6), there are no conformational rearrangements of the catalytic triplex (Figure 2.15), which remains static through the stages of branching²⁰, unlike the minor motions seen in previous work^{17,22}. We now have evidence of a conserved dynamical mechanism of splice site exchange by group II introns that has direct parallels with the spliceosome, strengthening the argument that group II introns and the spliceosome share the same ancestry.

Despite the many features in common with group II introns, we identified a functional difference that provides an additional layer of regulation for the spliceosome. The first short α -helix located within the maturase DBD domain has a positively charged surface (Figure 2.16) that is indispensable for spontaneous group II intron RNP branching. In contrast, while the equivalent helix in the linker domain of Prp8 adopts a highly similar pose (Figure 2.15), the contact surface is negative to neutral in charge (Figure 2.16). This marked difference between the intron maturase and Prp8 has evolutionary implications. On one hand, the maturase is the lone protein cofactor necessary for proper positioning of the D6

branch helix. However, the spliceosome requires recruitment of step 1 specific factors, such as Yju2, to activate branching²⁰. We can now structurally rationalize the need for Yju2 by comparing the intron maturase and Prp8 surfaces. The N-terminus of Yju2 may compensate for the positive charges that are lost during molecular evolution from the maturase to Prp8 by forming a highly positively charged contact surface that interacts with the branch helix (Figure 2.16). The maturase side chain, Lys361, shown to be essential for intron branching in our study has no equivalent in Prp8; while a highly conserved residue (Arg3 in *S.cerevisiae* and *H.sapiens*)³⁷ at the N-terminus of Yju2 plays a similar role in contacting G1 of the 5'SS. Additionally, it is noteworthy that Yju2 occupies the same space as D1c, and is a first step specific factor. It may be reasonable to speculate that Yju2 evolved to function similarly to D1c, which is active within the holoenzyme as a stabilizing factor for the branch helix, only prior to branching. We therefore observe hints of a molecular evolutionary strategy that fragmented the single-protein RNP into a multi-protein splicing machine, which allows for fine tuning of RNA splicing as a regulated biological process.

2.5.2 Conservation of molecular recognition

The pre-branching RNP structure presented in this study reveals the 5' SS and branchpoint recognition strategy employed by group II introns, thereby providing critical new insights into how splicing machinery maintains precise splice sites and branchpoint definition during molecular evolution.

Most significantly, we present for the first time the pre-attack conformation of the branchpoint adenosine in group II introns (Figure 2.13). This high-resolution view unambiguously explains the branch site recognition strategy employed by group II introns.

Instead of canonical base pairing, the intron resorts to a base triple (*cis* WC/sugar edge interaction) formed between the bpA and a CG base pair located two nucleotides upstream. The interaction also serves to hold the bpA inwards, toward the major groove of the D6 branch helix, thereby limiting its conformational flexibility and correctly positioning its 2' OH relative to the catalytic core for activation. The same molecular recognition strategy is used by the spliceosome to anchor its bpA (Figure 2.9). This striking similarity provides molecular evidence that there is minimal change to the strategy for branchpoint definition during evolution from group II introns to the spliceosome.

Secondly, we revealed the molecular basis of group II intron 5' SS recognition. Through base-sugar interactions originating from G1 and a base triple interaction from U2 (Figure 2.8), the intron 5' SS interlocks with the branch helix and closely contacts the branchpoint, preparing the system for branching. The abundance of molecular interactions surrounding the 5' SS also enforces stringent nucleotide identity requirements. Given the mechanistic parallel with the spliceosome (Figure 2.8), we can now justify why the same 5' GU motif²⁶ has persisted through time, highlighting that the strategy to define the 5' SS is so robust that it has withstood the forces of molecular evolution.

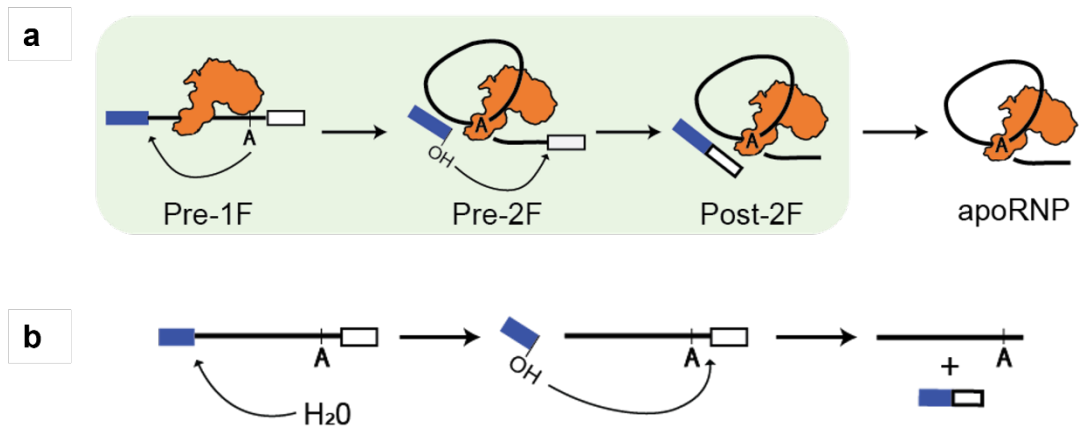


Figure 2.1 Group II intron splicing.

a. Cartoon of the forward splicing reaction that releases the intron holoenzyme as a lariat RNP. b. Cartoon of the competing hydrolytic pathway in which the water is the nucleophile, releasing the intron as a linear molecule.

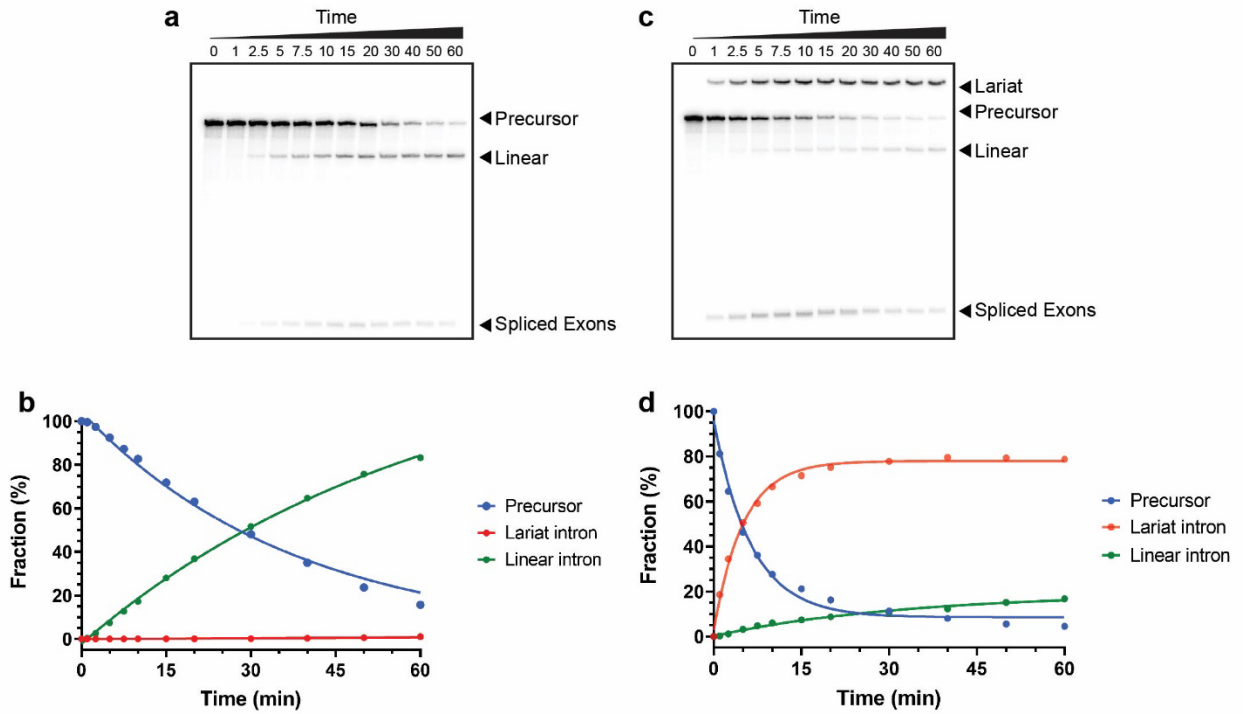


Figure 2.2 The maturase acts as a branching switch.

a. Splicing time course of the *E.r.* intron alone without its maturase cofactor. b. Quantification of the bands observed in the denaturing gel in (a). c. Splicing time course of the *E.r.* intron with its maturase cofactor (MarathonRT). d. Quantification of the bands detected in the denaturing gel in (c).

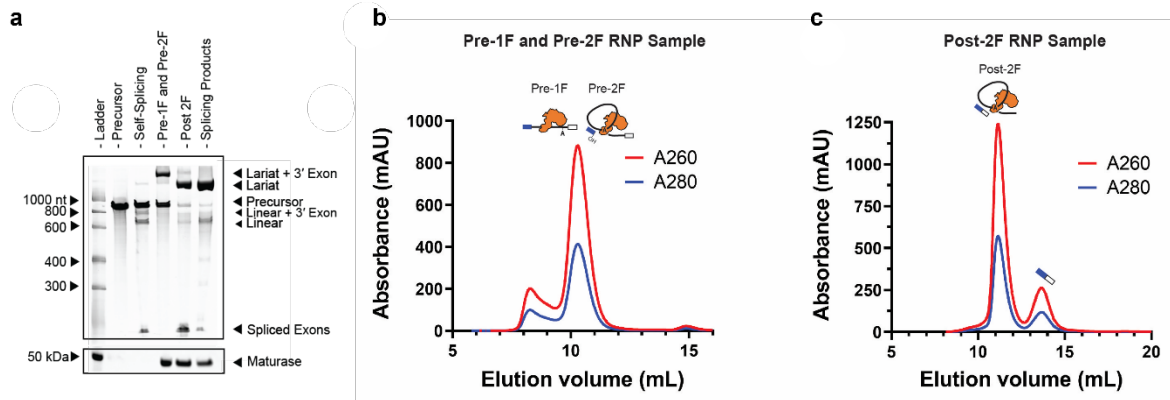


Figure 2.3 CryoEM sample preparation of the intron holoenzyme.

a. Denaturing gel showing the samples used for cryoEM grid preparation. Lanes 4 and 5 contain the samples used for preparation of the pre-1F and pre-2F (same grid) and post-2F samples respectively. All other lanes are control lanes used for band identification. b. Size exclusion chromatogram of the biochemical preparation for the pre-1F and pre-2F sample. c. Size exclusion chromatogram of the biochemical preparation for the post-2F sample. An RNA oligonucleotide that mimics the spliced exon was added back to the peak fraction as the native ligated exons dissociate during gel filtration.

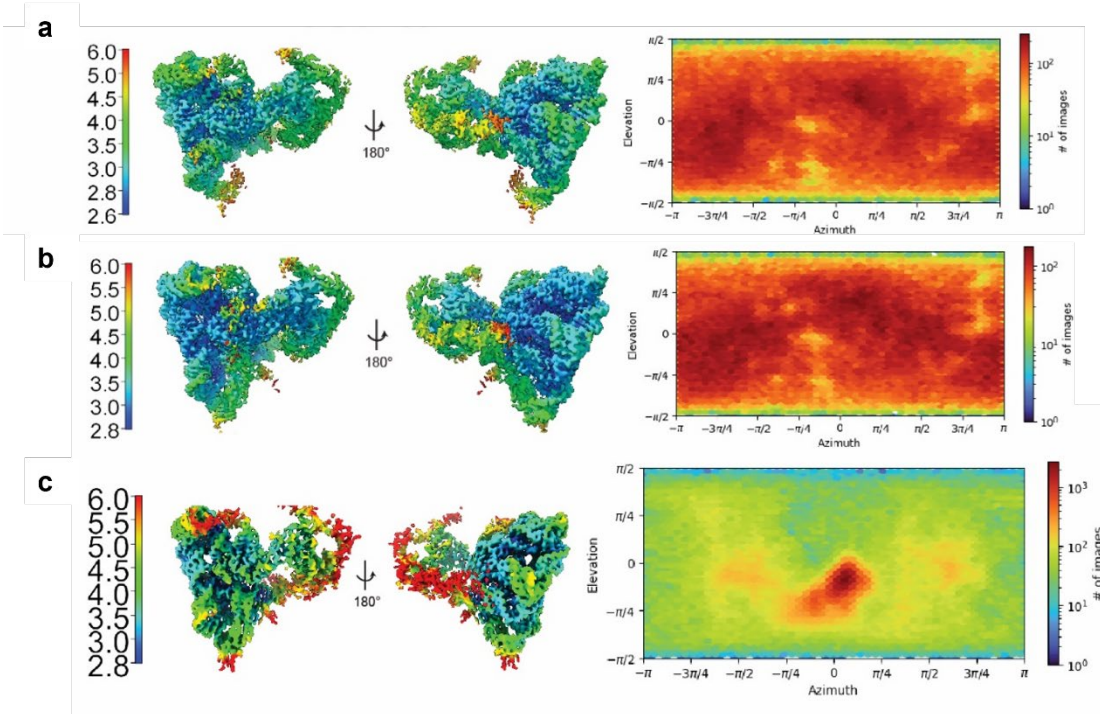


Figure 2.4 Local resolution and particle distribution of the RNP reconstructions.

a-c. Local resolution estimation (left) and particle distribution (right) of the three maps: (a) pre-1F, (b) pre-2F, and (c) post-2F intron RNPs. Focused refinement with separate masks was done on the left and right portions of each RNP reconstruction to improve resolution. A much-improved distribution is observed when the sample is prepared with the Chameleon or when the stage is tilted during data collection.

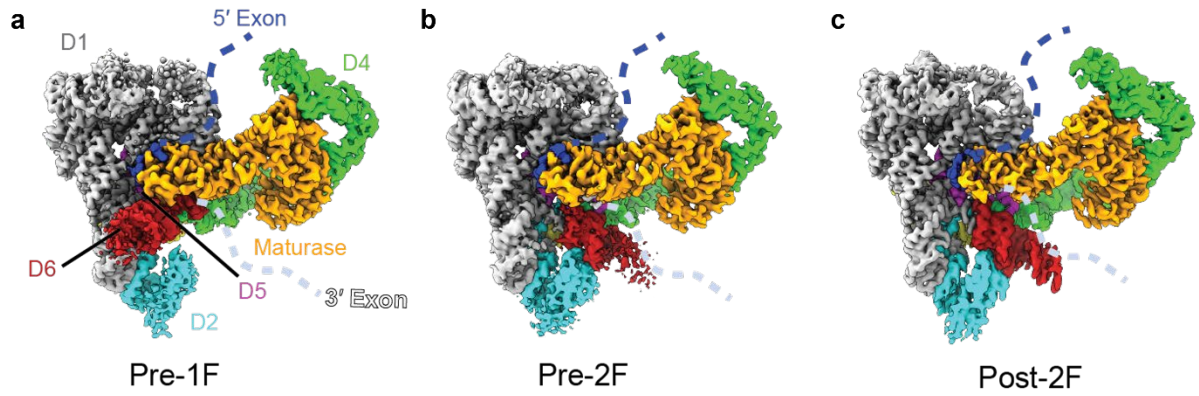


Figure 2.5 CryoEM reconstructions of the splicing pathway intron RNPs.

CryoEM composite maps of the (a). pre-1F, (b) pre-2F, and (c) post-2F RNP complexes.

The exons are depicted as dashed lines to indicate that while present in the construct, the distal sequences are not resolved by cryoEM.

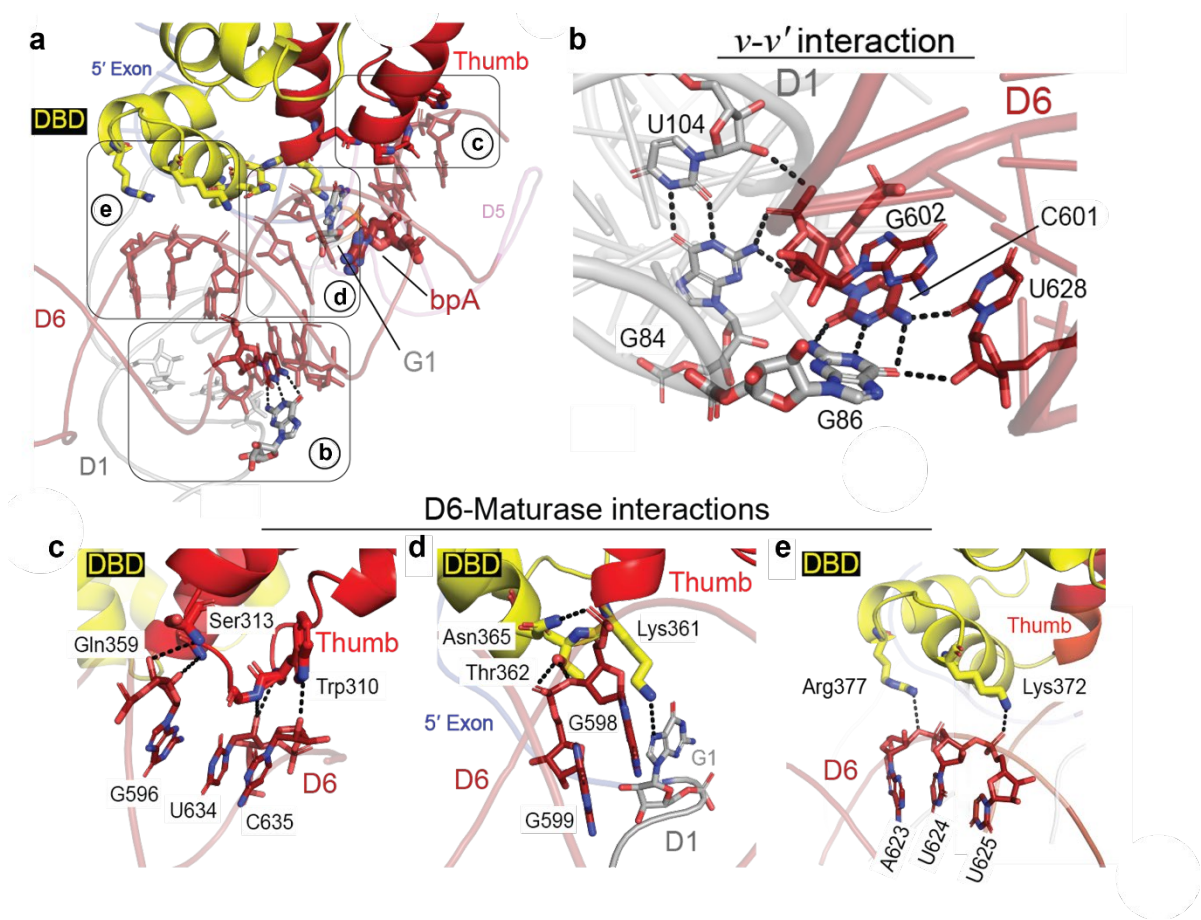


Figure 2.6 Branch helix RNA and protein interactions.

a. Interaction network surrounding the D6 helix prior to branching (boxed elements are labeled with the figure panel designations described below). b. Newly discovered long-range RNA interaction ($v-v'$) that positions D6. c-e. Intron-maturase interaction clusters that position D6.

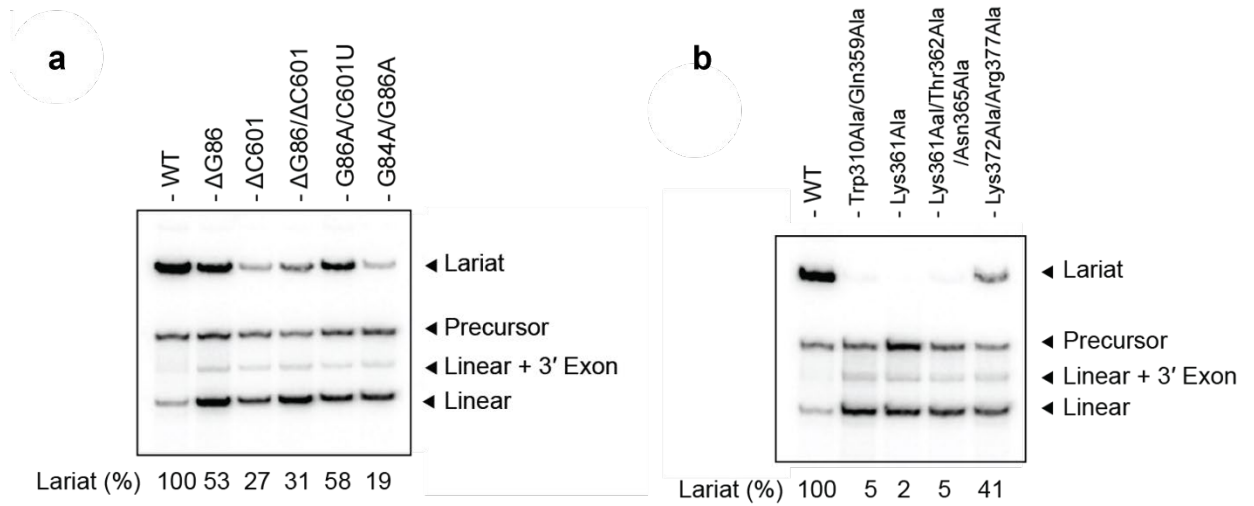


Figure 2.7 Biochemical validation of branch helix interactions.

a. Denaturing radioanalytical splicing gel showing effects of intron mutants in the presence of WT maturase protein. b. A denaturing radioanalytical splicing gel demonstrating effects of maturase mutants (see Figure 2.6c) on promoting branching of WT intron construct. The lariat percentage, obtained from $n = 4$ replicates, is indicated under the corresponding lane.

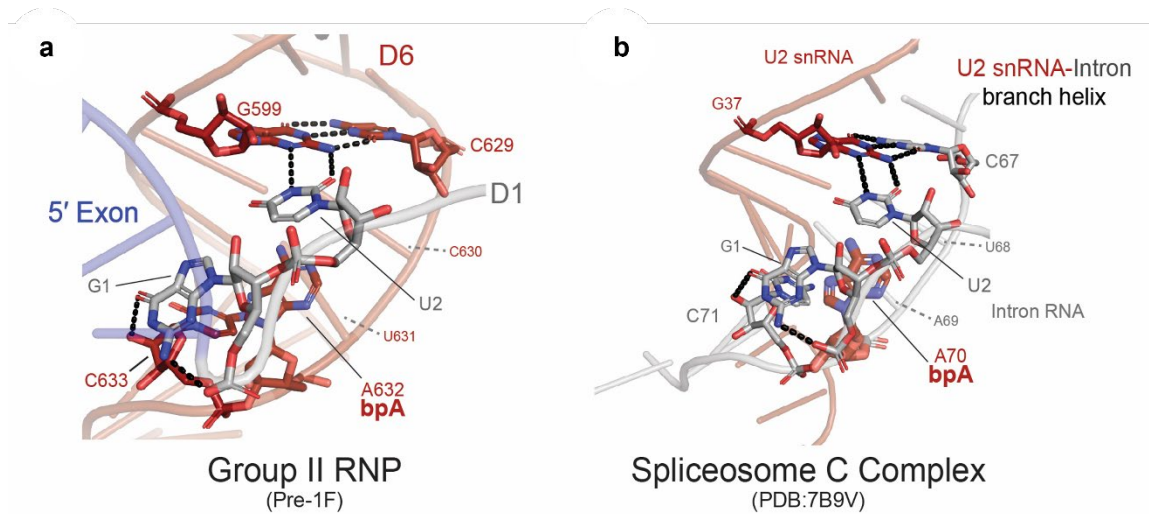


Figure 2.8 5' splice site recognition.

Comparison of the interactions surrounding the conserved 5' splice site GU nucleotides for a. the group II RNP in the pre-1F state and b. the yeast spliceosome C complex (PDB:7B9V). The position of bpA adjacent nucleotides (C630 and U631 in (a) and U68 and A69 in (b)) are indicated with dashed lines and drawn as cartoon batons.

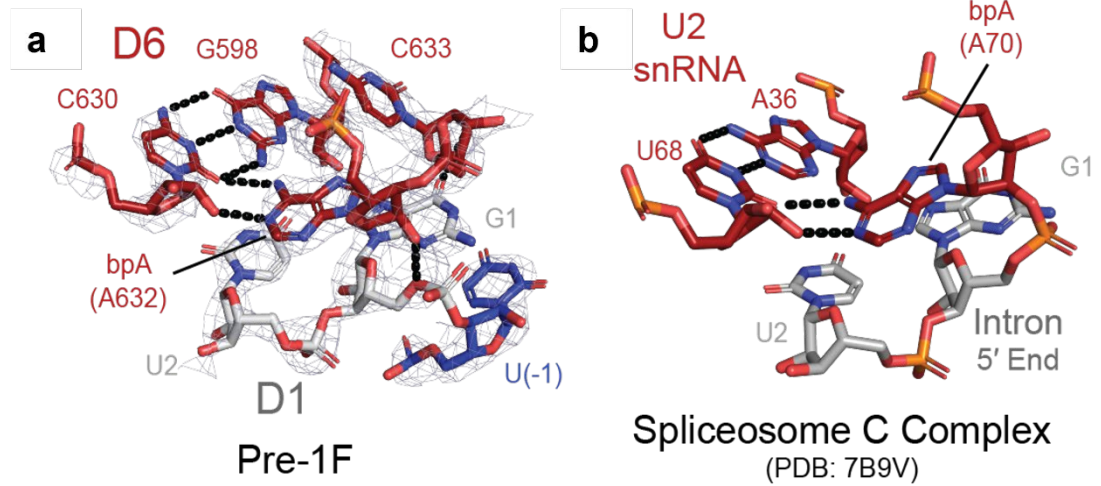


Figure 2.9 Branchpoint A recognition prior to branching.

a. Interactions that specify the branch site nucleotide identity prior to branching. b. Positioning of the branchpoint adenosine in the yeast spliceosome C complex (post-first step of branching).

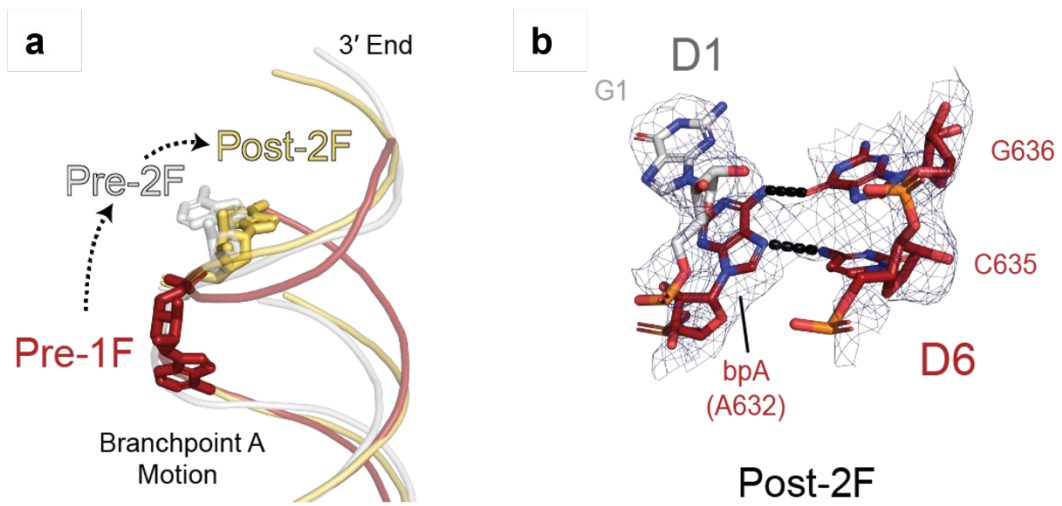


Figure 2.10 Branchpoint A local motions and interactions.

a. Aligned D6 helices showing conformational movement of the bpA during the stages of splicing. b. Interaction network surrounding the bpA post-ligation.

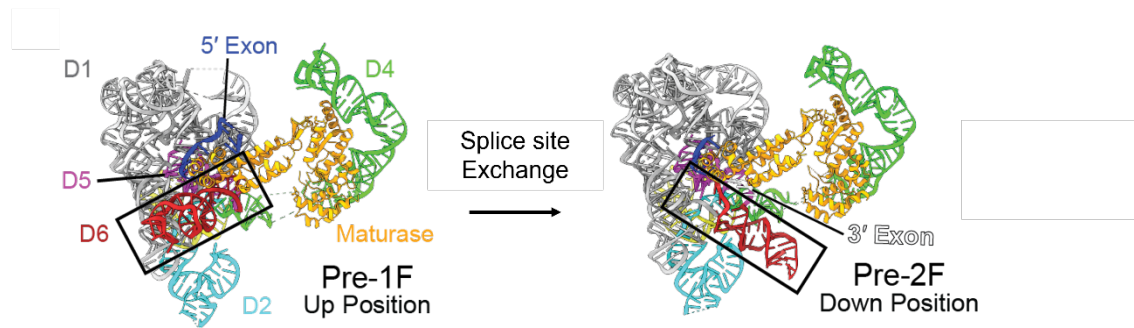


Figure 2.11 Branch helix large scale movements.

Conformational rearrangement of D6 from branching to exon ligation. The position of the helix in the pre-1F state is designated as the ‘up’ position, while the position of the helix in the pre-2F state is designated the ‘down’ position.

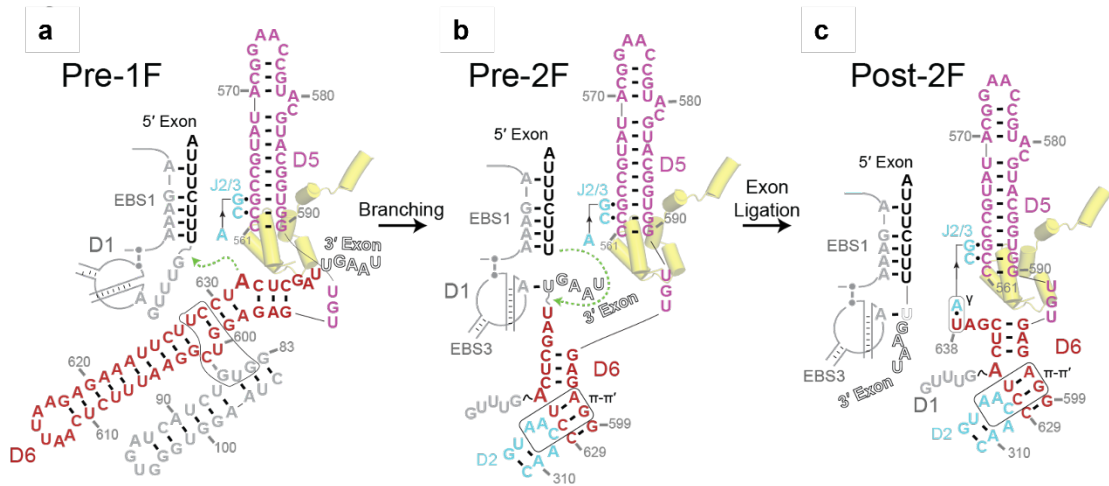


Figure 2.12 Splice site interactions throughout the branching pathway.

Secondary structure schematic with annotated tertiary contacts of the RNP in (a) the pre-1F, (b) the pre-2F, and (c) the post-2F state. Yellow cartoon batons indicate the position of the maturase DBD helices.

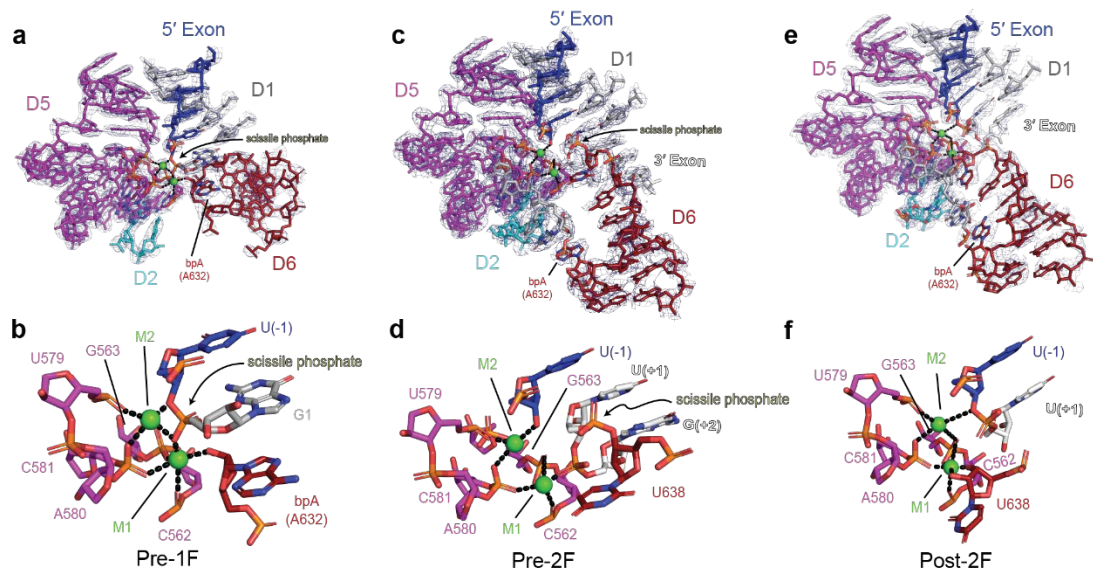


Figure 2.13 Molecular mechanism of group II RNP branching and exon ligation.

a-b. Organization of catalytic elements prior to branching. The bpA is juxtaposed to the 5' SS and poised for lariat formation. c-d. Active site configuration prior to exon ligation. The 5' SS is primed for attack to ligate the exons. e-f. Positioning of active site elements immediately after exon ligation. Divalent metal ions are shown as green spheres.

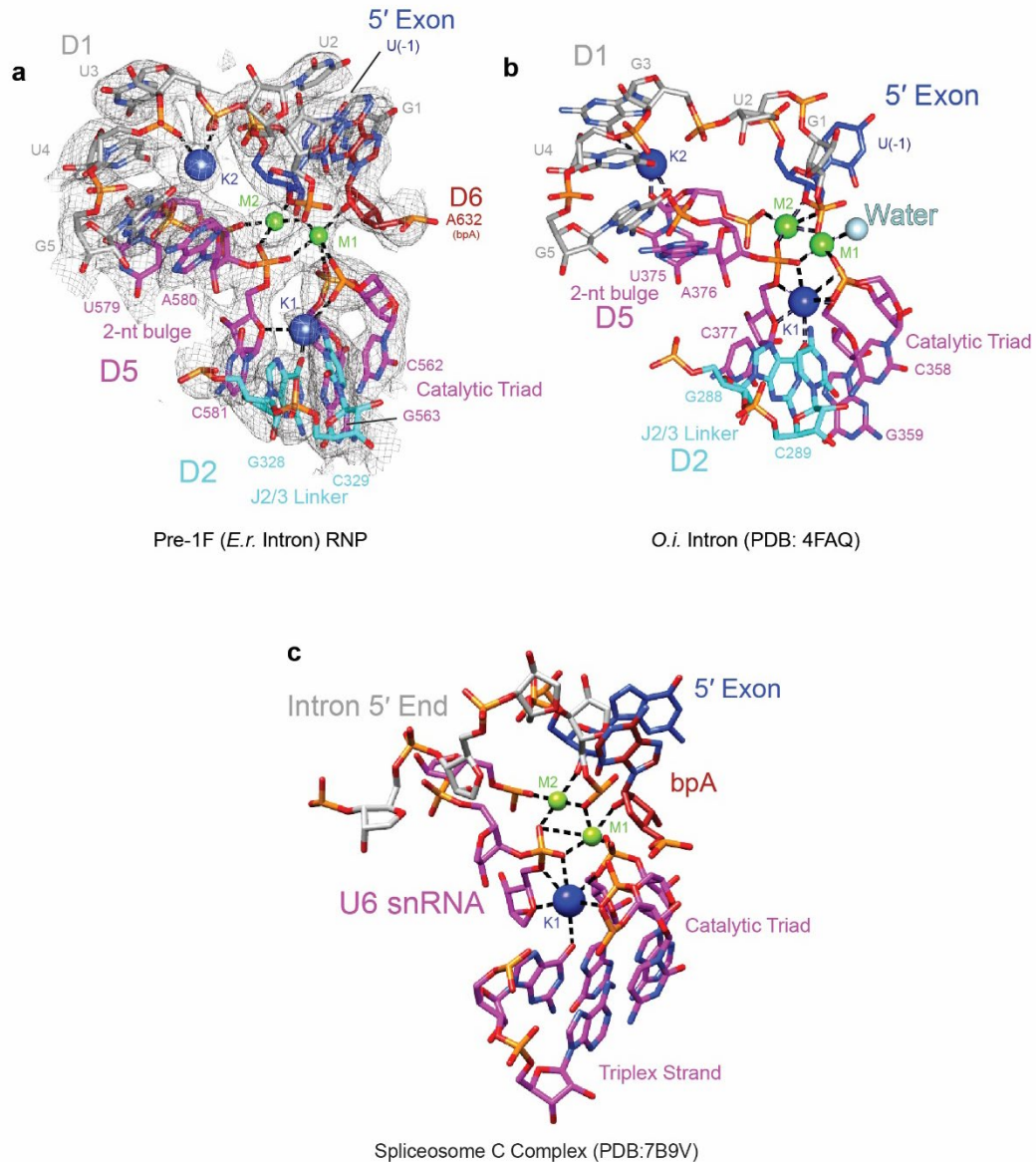


Figure 2.14 Comparison of the heteronuclear metal ion core.

a. The catalytic elements including the 5' SS, branchpoint A, J2/3 linker, catalytic triad, 2-nt bulge and the metal ions (M1, M2, K1 and K2 equivalents) for the pre-1F RNP are shown. Density from the cryoEM map of the corresponding regions is shown to indicate model fit. Catalytic core of b. the *O.i.* intron (PDB: 4FAQ), and c. the spliceosome C complex (PDB: 7B9V), and the key mechanistic elements as in (a) are displayed. The attacking water nucleophile is shown as a light blue sphere in (b).

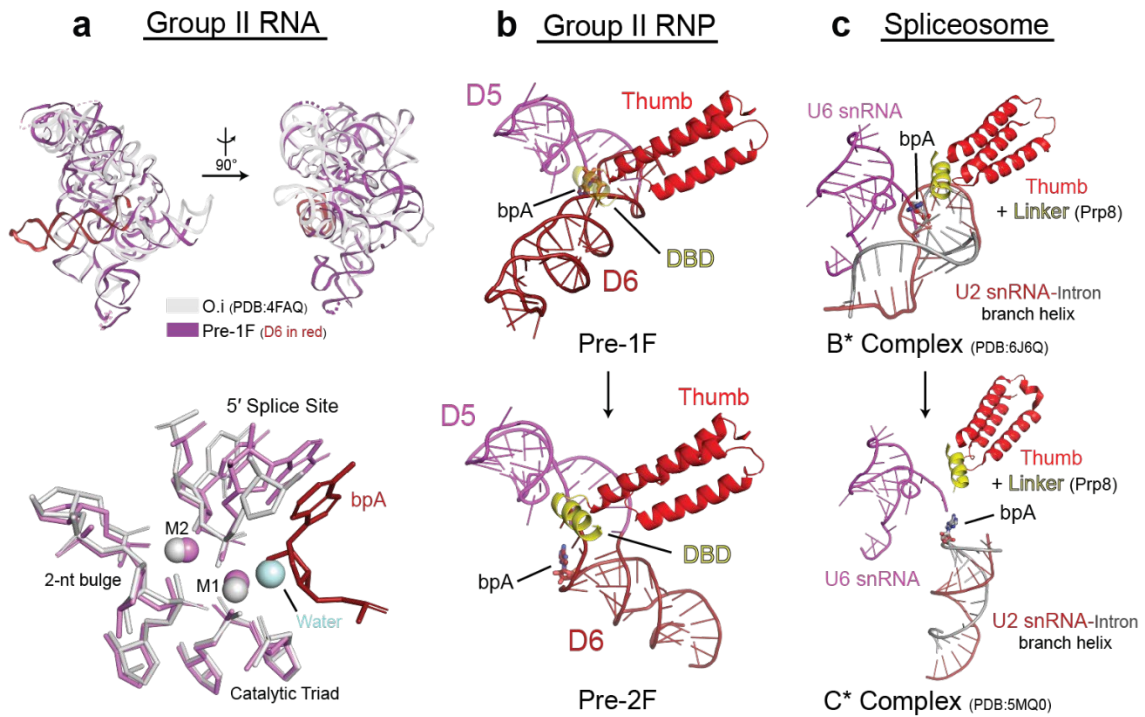


Figure 2.15 Mechanistic comparison of group II introns and the spliceosome.

a. Comparison of the overall fold of group IIC introns (top), and the aligned active sites of the *O.i.* intron prior to hydrolysis (nucleophilic water in light blue), and the *E.r.* intron prior to branching (bpA in dark red) (bottom). Conserved RNP interface and branch helix dynamics in b. group II RNPs and c. the spliceosome during the first to second step transition.

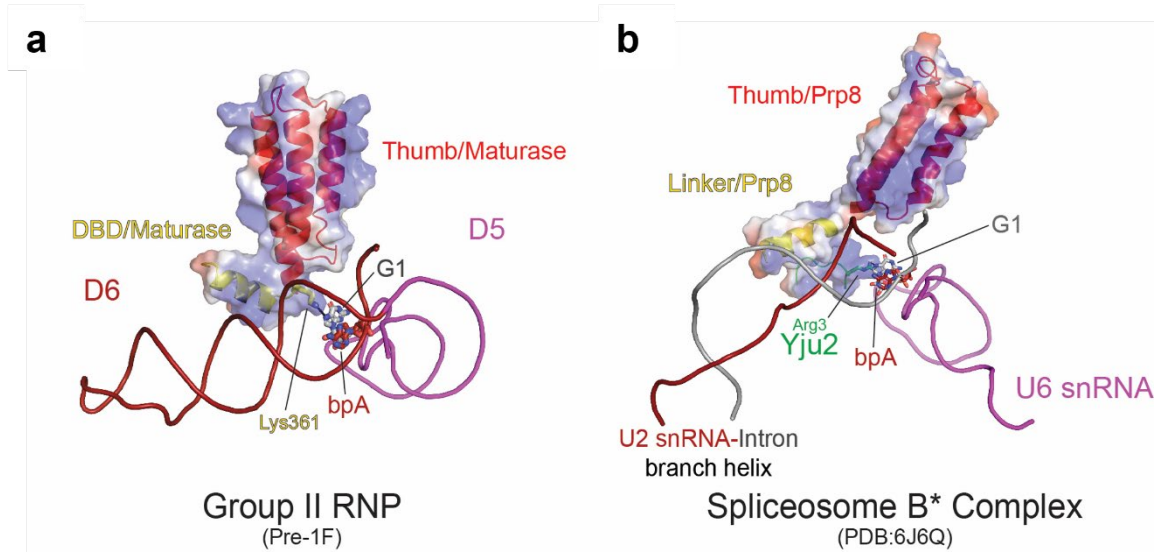


Figure 2.16 Comparison of protein-branch helix interactions.

a. Interactions between the DBD and thumb of the maturase protein with D6 in the group II intron. G1 and bpA (A632) are shown as sticks. Lys361 is shown interacting with G1. Surface charges around the protein are shown (blue-positive, white-neutral and red-negative). b. Interactions between the thumb (in red) and linker (in yellow) of Prp8 with the U2 snRNA-intron branch helix in the yeast spliceosome B* complex (PDB: 6J6Q). Surface charges of the proteins are shown with the same color code as in (a). G1 and bpA (A70) are shown as sticks. Arg3 from the Yju2 protein (in green) is shown interacting with G1.

2.6 Materials and Methods

Protein Purification

Wild-type maturase protein (MarathonRT) was purified as previously described¹⁸. Briefly, the recombinant protein was cloned into a pET-SUMO vector which has a N-terminal 6xHis-SUMO tag. The plasmid was transformed into Rosetta 2 (DE3) cells (MilliporeSigma), which were grown at 37°C in LB supplemented with kanamycin and chloramphenicol. Cells were grown to an OD of 1.5 in large 2L cultures before induction with IPTG and shaking at 16°C overnight. Cells were harvested by centrifugation and resuspended in a lysis buffer (25 mM Na-HEPES, pH 7.5, 1 M NaCl, 10% glycerol and 2 mM β -mercaptoethanol (β ME)) with dissolved protease inhibitor. Cells were lysed using a microfluidizer and the cell lysate was clarified to remove precipitants. The lysate was loaded onto a Ni-NTA column and was washed with lysis buffer, and wash buffer (25 mM Na-HEPES, pH 7.5, 150 mM NaCl, 10% glycerol, 2 mM β ME, and 25 mM imidazole) before elution (25 mM Na-HEPES, pH 7.5, 150 mM NaCl, 10% glycerol, 2 mM β ME, and 300 mM imidazole). The SUMO tag was removed using ULP1 SUMO protease by incubating at 4°C for 1 hour. After tag cleavage, the protein was loaded onto a HiTrap SP HP cation exchange column (Cytiva) equilibrated with buffer A (25 mM K-HEPES, pH 7.5, 150 mM KCl, 10% glycerol and 1 mM DTT). The protein was eluted by running a linear gradient to buffer B (25 mM K-HEPES, pH 7.5, 2 M KCl, 10% glycerol and 1 mM DTT). The peak fractions were pooled, concentrated to 5 mL and injected onto a HiLoad 16/600 Superdex 200 size exclusion column (Cytiva) and eluted using a SEC buffer (50 mM NH₄-HEPES, 150 mM NH₄Cl, 5 mM DTT and 10% glycerol). Peak fractions from the S200 column were pooled, concentrated to 5 mg/mL, flash frozen under liquid nitrogen

and stored at -80°C . Mutant proteins (Trp310A/Gln359Ala, Lys361Ala, Lys361Ala/Thr362Ala/Asn365Ala, and Lys372Ala/Arg377Ala) were also prepared in the same manner. The folding integrity of the WT maturase protein and all mutants were verified with an orthogonal activity assay involving the inherent RT capability³⁸ of these maturase proteins. All proteins were fully active for reverse transcription on long RNA templates.

RNA Transcription and Purification

The DNA sequence containing the T7 promoter, 100 nt of 5' exon, the intron (with ORF deletion) and 100 nt of 3' exon followed by the BamHI cleavage site were cloned into the pBlueScript vector (Invitrogen) to give the pLTS01 plasmid, which was used for structural studies. For splicing assays, a longer 5' exon (150 nt) construct (which was cloned into the pCZ26 plasmid) was used. The plasmids were linearized using BamHI (New England Biolabs) to generate the DNA template. Mutants of the construct were prepared using PfuUltra II Hotstart PCR Master Mix (Agilent) mixed with 20 ng of plasmid, 10 pmol of forward and reverse primers following the routine site-directed mutagenesis protocol. The sequences of the resulting plasmids were verified by Sanger sequencing (Quintara Biosciences).

RNA used for cryoEM sample preparation was prepared as previously described³⁹. Briefly, *in vitro* transcription of the intron precursor RNA was performed using in-house-prepared T7 RNA polymerase (P266L mutant) in a transcription buffer containing 40 mM Tris-HCl pH 8.0, 10 mM NaCl, 23 mM MgCl₂, 2 mM spermidine, 0.01% Triton X-100, 10 mM DTT and 5 mM each rNTP. 40 μg of linearized pLTS01 plasmid was added to each 1

mL of transcription and the reaction was incubated at 37°C for 6 hours before it was ethanol precipitated. The pellet was resuspended in water and mixed with an equal volume of 2x urea loading dye containing bromophenol blue and xylene cyanol, which was then loaded onto a 5% urea denaturing polyacrylamide gel to purify the intron precursor RNA. The band was visualized by UV shadow, cut with a sterile blade, crushed with a sterile syringe and eluted in a gel elution buffer overnight (10 mM Na-MOPS pH 6.0, 300 mM NaCl and 1 mM EDTA). The eluted RNA was then ethanol precipitated, resuspended in an RNA storage buffer (6 mM Na-MES pH 6.0) to a final concentration of 50 μ M and frozen at -80°C for preparation of RNP samples.

The radiolabeled intron precursor RNA for the splicing assay was prepared using the previously described body-labelling protocol⁴⁰. Briefly, the intron precursor RNA transcripts were body-labelled using in-house-prepared T7 RNA polymerase and 50 μ Ci of [α -³²P]-UTP (PerkinElmer) in a transcription buffer containing 40 mM Tris-HCl pH 8.0, 10 mM NaCl, 15 mM MgCl₂, 2 mM spermidine, 0.01% Triton X-100, 10 mM DTT and 3.6 mM each rNTP (except UTP, which was at 1 mM) and 5 μ g of linearized pCZ26 plasmid. Mutant intron precursor RNAs (Δ G86, Δ C601, Δ G86/ Δ C601, G86A/C601U and G84A/G86A) were prepared in the same manner. The reaction was incubated at 37°C for 1.5 hours and purified on a 5% urea denaturing polyacrylamide gel. The band corresponding to precursor RNA was visualized through brief exposure of the phosphor storage screen and the screen was subsequently imaged using an Amersham Typhoon RGB imager (Cytiva). The band was cut and eluted overnight in the gel elution buffer. The radiolabeled RNA was then ethanol precipitated, resuspended in the RNA storage buffer to a final concentration of 100 nM, and stored at -20°C for splicing assays.

***In vitro* Forward Splicing Assay**

Radiolabeled intron precursor RNA (and mutants) was mixed with purified maturase protein (and mutants) under near-physiological condition (50 mM K-HEPES pH 7.5, 150 mM KCl and 5 mM MgCl₂) to perform the *in vitro* forward splicing assay. To do so, radiolabeled intron RNA was first mixed with buffer and water, and heated to 95°C for 1 minute, after which it was returned to 37°C for 5 minutes. Potassium chloride and maturase protein stocks were added, and the sample was incubated at 37°C for another 5 minutes. Magnesium chloride stock was subsequently added to the mixture to initiate the splicing reaction. The final concentration of radiolabeled intron precursor RNA was 5 nM and maturase protein was 20 nM. After incubation at 37°C for 1 hour, 2 µL of the reaction mixture was taken out and quenched by mixing with an equal volume of 2x formamide loading dye (72% (v/v) formamide, 10% sucrose, 0.2% bromophenol blue dye, 0.2% xylene cyanol dye and 50 mM EDTA) pre-cooled on ice. Samples were analyzed on a 5% urea denaturing polyacrylamide gel. The gel was dried and used to expose the phosphor storage screen overnight. The screen was then imaged on an Amersham Typhoon RGB imager (Cytiva). The bands were quantified using ImageQuant TL 8.2 (Cytiva).

Group II Intron RNP Sample Preparation

To obtain the *E.r.* intron-maturase RNP complex stalled in the pre-1F and pre-2F states, a 0.9 mL reaction with 5 µM purified intron precursor RNA, 10 µM purified maturase protein was conducted in a buffer containing 50 mM NH₄-HEPES pH 7.5, 150 mM NH₄Cl, 10 mM CaCl₂ and 5 mM DTT. To do so, purified intron precursor RNA stock was mixed with buffer and water and heated to 95°C for 3 minutes. It was then incubated at 37°C for

5 minutes. The ammonium chloride and calcium chloride stocks were then added to refold the RNA at 37°C for 10 minutes. After that, the DTT and maturase stocks were added to the reaction mixture and the reaction was incubated at 37°C for 1 hour with shaking at 300 rpm on a thermomixer, after which it was centrifuged at 10,000 g for 2 minutes to pellet the precipitates. The supernatant was subsequently loaded onto a Superdex 200 Increase 10/300 GL column (Cytiva) pre-equilibrated with the buffer containing 50 mM NH₄-HEPES pH 7.5, 150 mM NH₄Cl, 10 mM CaCl₂ and 5 mM DTT. The elution peak was pooled together and concentrated to ~4 mg/mL using an Amicon concentrator (10 kDa MWCO) (MilliporeSigma). The concentrated sample was used to prepare cryoEM grids (see below).

To obtain the *E.r.* intron-maturase RNP complex in the post-2F state, the lariat intron apo-RNP was first purified as previously described¹⁸ the peak fraction was collected. 1.2x molar excess of a synthetic RNA oligonucleotide with the sequence 5'-AUUUCUUUUGAAU-3' (Integrated DNA Technologies) was then added to a final concentration of 600 nM, resulting in a final concentration of 500 nM for the RNP. The sample was incubated on ice for 10 minutes to allow formation of the ternary RNP complex, which was used to prepare cryo-grids (see below).

All samples for cryoEM, precursor RNA and reaction ladders generated from self-splicing and maturase-mediated splicing were loaded onto a 5% urea denaturing polyacrylamide gel run in a Mini-PROTEAN tetra-vertical electrophoresis cell (Bio-rad) at 180W for 50 minutes before staining with GelRed (Biotium) and imaging with the Cy3 channel on an Amersham Typhoon RGB imager (Cytiva).

Grid Preparation and Data Collection

For cryoEM analysis of the *E.r.* intron-maturase RNP complex stalled in the pre-1F and pre-2F states, 7 μL of purified sample was loaded onto the Chameleon system (SPT Labtech). 40 nL of the sample solution was dispensed to the glow-discharged Quantifoil Active Cu 300-mesh R1.2/2 grids with Cu nanowires (SPT Labtech) and the grids were plunged into liquid ethane and frozen in liquid nitrogen ~ 400 ms after. The grids were screened on a Talos Glacios microscope (ThermoFisher) operating at 200 keV. Grids with sufficient collectable squares and minimal crystalline ice contamination were selected for data collection on a Titan Krios microscope (ThermoFisher) operating at 300 keV with a K3 summit direct electron detector (Gatan) and the data were obtained in counting mode. SerialEM v3.9 was used for data collection. Two datasets of 4,612 micrographs (at 30° tilt angle) and 7,777 micrographs (at 0° tilt angle) were collected for the pre-1F/pre-2F sample. A nominal magnification of 81,000x and a defocus range of $-0.8 \mu\text{m}$ to $-2.5 \mu\text{m}$ was used, giving an effective pixel size of 0.844\AA at the specimen level. Each micrograph was dose-fractionated to 40 frames with a total exposure time of 3.482 s and a frame exposure time of 0.0865 s, resulting in a total dose of $60 \text{ e}^-/\text{\AA}^2$.

For cryoEM analysis of the *E.r.* intron-maturase RNP complex in the post-2F state, two separate approaches were taken. The first strategy used the Chameleon for grid preparation. Here 8 μL of purified sample was loaded onto the Chameleon system (SPT Labtech). 40 nL of the sample solution was dispensed to the glow-discharged Quantifoil Active Cu 300-mesh R1.2/2 grids with Cu nanowires (SPT Labtech) and the grids were plunged into liquid ethane and frozen in liquid nitrogen ~ 300 ms after. In parallel, grids were prepared using the Vitrobot. Here, 4 μL of the purified RNP sample was loaded onto plasma cleaned

QuantiFoil Cu R1.2/1.3 300-mesh grids prepped in-house with an extra layer of carbon. The grids were blotted and plunged into liquid ethane and frozen in liquid nitrogen using a condition of 100% humidity at 22°C. SerialEM v4.0 was used for data collections and micrographs were recorded on a Titan Krios microscope (ThermoFisher) operating at 300 keV, equipped with a K3 Summit direct electron detector (Gatan) operating in counting mode. Two datasets of 13,740 micrographs (Chameleon) and 3,852 micrographs (Vitrobot) were collected for the post-2F sample. For the Chameleon dataset, a nominal magnification of 105,000x and a defocus range of -1.0 μm to -2.5 μm was used, giving an effective pixel size of 0.832Å at the specimen level. Each micrograph was dose-fractionated to 50 frames with a total exposure time of 1.5s and a frame exposure time of 0.03 s, resulting in a total dose of 50.5 $\text{e}^-/\text{Å}^2$. For the Vitrobot dataset, a nominal magnification of 105,000x and a defocus range of -1.0 μm to -2.0 μm was used, giving an effective pixel size of 0.832Å at the specimen level. Each micrograph was dose-fractionated to 48 frames with a total exposure time of 1.92s and a frame exposure time of 0.04s, resulting in a total dose of 50 $\text{e}^-/\text{Å}^2$.

CryoEM Data Processing

Recorded movie frames were processed using cryoSPARC v3.4^{41,42}. Motion correction and CTF estimations were performed using default parameters in cryoSPARC. Exposures were curated and micrographs with obvious ice contamination, large motions, or damaged areas were removed.

For the pre-1F and pre-2F samples, particle picking was done with the automated blob picker and filtered using consecutive rounds of 2D classification. This subset of particles

was used for Topaz (Topaz 0.2.4) training and picking on the combined 11,783 micrographs. Topaz picking yielded 1,289,915 particles which were subject to three rounds of 2D classification, leaving 847,534 particles which were extracted with a box size of 384 x 384 pixels. 100,000 particles were selected for initial model creation, generating three reconstructions, of which one was selected as the reference for further classification. In the first round of 3D classification, eight classes were separated out, from which two groups of 422,728 particles and 234,625 particles corresponding to the pre-1F and pre-2F states were identified. Each group was subjected to another round of 3D classification, leaving a final subset of particles of 281,619 particles (pre-1F) and 234,625 particles (pre-2F, all particles were selected). Each group of particles was separately refined, yielding reconstructions of 3.0Å and 3.1Å respectively. For each reconstruction, separate masked local refinements, and local and global CTF refinement were conducted on the left and right halves to improve resolution of each. Two pairs of maps were obtained for the left and right halves of the pre-1F and pre-2F reconstructions which have resolutions of 2.9Å, 3.1Å, 3.0Å, and 3.3Å respectively, as evaluated by the GSFSC with a cutoff of 0.143.

A similar strategy was used to obtain the 3D reconstruction of the post-2F state. Briefly, particles were picked with an automated blob picker from the first dataset of 13,740 micrographs (Chameleon grid) and the results were filtered through several rounds of 2D classification and used as an input for Topaz training. This yielded 930,320 particles which were further cleaned through iterative rounds of 2D classification leaving 695,515 particles. Three initial models were generated with a subset of 100,000 particles out of which a single refinement with the best overall density was chosen. 3D classification was used to separate the particles into ten classes from which a single class with 217,454

particles was selected. These particles were used to train a Topaz model and after iterative rounds of 2D classification cleaning, 361,978 particles remained. These particles were combined with 342,031 particles that were Topaz picked and filtered from a separate dataset of 3,852 micrographs (Vitrobot grid). The combined 722,394 particles were 2D classified and manually separated into groups corresponding to the dominant and less represented views. 70,000 randomized particles were taken from the dominant view group and aggregated with all particles from the 'other' views. Particles were extracted with a box size of 384 x 384 pixels. After an initial 3D refinement, 3D classification into ten classes yielded two classes with good overall density, which were used to generate an overall reconstruction of 3.0Å. Similar to the other structures, masked local refinement was done on the separate halves along with local and global CTF refinement. This resulted in reconstructions of 2.8Å and 3.2Å for the left and right halves respectively as evaluated by the GSFSC with a cutoff of 0.143.

Model Building and Refinement

Model building was initiated by docking a previous group II RNP structure (PDB: 7UIN) into the generated reconstructions using UCSF Chimera⁴³⁻⁴⁵ (Chimera 1.15 and ChimeraX 1.2.5). NAMDINATOR⁴⁶ (<https://namdinator.au.dk/>) was used for flexible fitting of the docked models to obtain better starting models. The models were then manually rebuilt in COOT (COOT 0.9.6) to accommodate for the changes in branch helix position, the additional ligated exons, and the metal ion core. Density for the distal portion of D6 was weak in the reconstructions, but this region was modelled as a helix nonetheless based on data that demonstrated this domain forms a canonical helix. The three-way

junction of the D4 arm and portions of the α - α' interaction were not modeled as the density is difficult to interpret. The final pre-1F, pre-2F, and post-2F models were improved by iterative rounds of real-space refinement against the sharpened cryo-EM map in PHENIX (Phenix 1.20.1-4487) using secondary structure restraints for both RNA, protein and DNA, as well Ramachandran and rotamer restraints for protein chains, and subsequent rebuilding in COOT⁴⁷⁻⁴⁹. Directional resolution anisotropy analyses were performed using the 3DFSC²⁵ web server (<https://3dfsc.salk.edu/>).

Protein Conservation Analysis

Group IIC maturase protein sequences were obtained from the *Bacterial Group II Intron Database*⁵⁰ (<http://webapps2.ucalgary.ca/~groupii/>). Roughly 90 sequences were aligned with ClustalOmega. Alignments were analyzed and visualized using JalView.

RNA Conservation Analysis

Sequences corresponding to group IIC D1c and D6 sequences were obtained from the *Bacterial Group II Intron Database*⁵⁰ (<http://webapps2.ucalgary.ca/~groupii/>). Sequences were aligned with ClustalOmega and alignments were visualized and analyzed with JalView.

2.7 References

- 1 Xu, L., Liu, T., Chung, K. & Pyle, A. M. Structural insights into intron catalysis and dynamics during splicing. *Nature*, doi:10.1038/s41586-023-06746-6 (2023).
- 2 Shi, Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat Rev Mol Cell Biol* **18**, 655-670, doi:10.1038/nrm.2017.86 (2017).
- 3 Galej, W. P., Toor, N., Newman, A. J. & Nagai, K. Molecular Mechanism and Evolution of Nuclear Pre-mRNA and Group II Intron Splicing: Insights from Cryo-Electron Microscopy Structures. *Chem Rev* **118**, 4156-4176, doi:10.1021/acs.chemrev.7b00499 (2018).
- 4 Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nat Rev Genet* **23**, 697-710, doi:10.1038/s41576-022-00514-4 (2022).
- 5 Rogalska, M. E., Vivori, C. & Valcarcel, J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat Rev Genet* **24**, 251-269, doi:10.1038/s41576-022-00556-8 (2023).
- 6 Zhao, C. & Pyle, A. M. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci* **42**, 470-482, doi:10.1016/j.tibs.2017.03.007 (2017).
- 7 Wilkinson, M. E., Fica, S. M., Galej, W. P. & Nagai, K. Structural basis for conformational equilibrium of the catalytic spliceosome. *Mol Cell* **81**, 1439-1452 e1439, doi:10.1016/j.molcel.2021.02.021 (2021).
- 8 Marcia, M. & Pyle, A. M. Visualizing group II intron catalysis through the stages of splicing. *Cell* **151**, 497-507, doi:10.1016/j.cell.2012.09.033 (2012).
- 9 Zhao, C. & Pyle, A. M. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol* **23**, 558-565, doi:10.1038/nsmb.3224 (2016).
- 10 Galej, W. P., Oubridge, C., Newman, A. J. & Nagai, K. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* **493**, 638-643, doi:10.1038/nature11843 (2013).
- 11 Belfort, M. & Lambowitz, A. M. Group II Intron RNPs and Reverse Transcriptases: From Retroelements to Research Tools. *Cold Spring Harb Perspect Biol* **11**, doi:10.1101/cshperspect.a032375 (2019).
- 12 Eickbush, T. H. Mobile introns: retrohoming by complete reverse splicing. *Curr Biol* **9**, R11-14, doi:10.1016/s0960-9822(99)80034-7 (1999).
- 13 Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II intron. *Science* **320**, 77-82, doi:10.1126/science.1153803 (2008).
- 14 Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* **514**, 193-197, doi:10.1038/nature13790 (2014).
- 15 Chan, R. T. *et al.* Structural basis for the second step of group II intron splicing. *Nat Commun* **9**, 4676, doi:10.1038/s41467-018-06678-0 (2018).
- 16 Qu, G. *et al.* Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* **23**, 549-557, doi:10.1038/nsmb.3220 (2016).
- 17 Haack, D. B. *et al.* Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA. *Cell* **178**, 612-623 e612, doi:10.1016/j.cell.2019.06.035 (2019).
- 18 Chung, K. *et al.* Structures of a mobile intron retroelement poised to attack its structured DNA target. *Science* **378**, 627-634, doi:10.1126/science.abq2844 (2022).
- 19 Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183-195, doi:10.1261/rna.063479.117 (2018).
- 20 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 e313, doi:10.1016/j.cell.2019.02.006 (2019).
- 21 Liu, N. *et al.* Exon and protein positioning in a pre-catalytic group II intron RNP primed for splicing. *Nucleic Acids Res* **48**, 11185-11198, doi:10.1093/nar/gkaa773 (2020).
- 22 Costa, M., Walbott, H., Monachello, D., Westhof, E. & Michel, F. Crystal structures of a group II intron lariat primed for reverse splicing. *Science* **354**, doi:10.1126/science.aaf9258 (2016).
- 23 Dandey, V. P. *et al.* Spotiton: New features and applications. *J Struct Biol* **202**, 161-169, doi:10.1016/j.jsb.2018.01.002 (2018).
- 24 Noble, A. J. *et al.* Reducing effects of particle adsorption to the air-water interface in cryo-EM. *Nat Methods* **15**, 793-795, doi:10.1038/s41592-018-0139-3 (2018).
- 25 Tan, Y. Z. *et al.* Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat Methods* **14**, 793-796, doi:10.1038/nmeth.4347 (2017).

- 26 Zimmerly, S. & Semper, C. Evolution of group II introns. *Mob DNA* **6**, 7, doi:10.1186/s13100-015-0037-5 (2015).
- 27 Pyle, A. M. Group II Intron Self-Splicing. *Annu Rev Biophys* **45**, 183-205, doi:10.1146/annurev-biophys-062215-011149 (2016).
- 28 Liu, Q. *et al.* Branch-site selection in a group II intron mediated by active recognition of the adenine amino group and steric exclusion of non-adenine functionalities. *J Mol Biol* **267**, 163-171, doi:10.1006/jmbi.1996.0845 (1997).
- 29 Chu, V. T., Adamidi, C., Liu, Q., Perlman, P. S. & Pyle, A. M. Control of branch-site choice by a group II intron. *EMBO J* **20**, 6866-6876, doi:10.1093/emboj/20.23.6866 (2001).
- 30 Bertram, K. *et al.* Structural Insights into the Roles of Metazoan-Specific Splicing Factors in the Human Step 1 Spliceosome. *Mol Cell* **80**, 127-139 e126, doi:10.1016/j.molcel.2020.09.012 (2020).
- 31 Retraction: RNAi-Dependent and Independent Control of LINE1 Accumulation and Mobility in Mouse Embryonic Stem Cells. *PLoS Genet* **11**, e1005519, doi:10.1371/journal.pgen.1005519 (2015).
- 32 Boulanger, S. C. *et al.* Length changes in the joining segment between domains 5 and 6 of a group II intron inhibit self-splicing and alter 3' splice site selection. *Mol Cell Biol* **16**, 5896-5904, doi:10.1128/MCB.16.10.5896 (1996).
- 33 Fica, S. M. *et al.* Structure of a spliceosome remodelled for exon ligation. *Nature* **542**, 377-380, doi:10.1038/nature21078 (2017).
- 34 Steitz, T. A. & Steitz, J. A. A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* **90**, 6498-6502, doi:10.1073/pnas.90.14.6498 (1993).
- 35 Daniels, D. L., Michels, W. J., Jr. & Pyle, A. M. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. *J Mol Biol* **256**, 31-49, doi:10.1006/jmbi.1996.0066 (1996).
- 36 Genna, V., Colombo, M., De Vivo, M. & Marcia, M. Second-Shell Basic Residues Expand the Two-Metal-Ion Architecture of DNA and RNA Processing Enzymes. *Structure* **26**, 40-50 e42, doi:10.1016/j.str.2017.11.008 (2018).
- 37 Liu, Y. C., Chen, H. C., Wu, N. Y. & Cheng, S. C. A novel splicing factor, Yju2, is associated with NTC and acts after Prp2 in promoting the first catalytic reaction of pre-mRNA splicing. *Mol Cell Biol* **27**, 5403-5413, doi:10.1128/MCB.00346-07 (2007).
- 38 Guo, L. T., Olson, S., Patel, S., Graveley, B. R. & Pyle, A. M. Direct tracking of reverse-transcriptase speed and template sensitivity: implications for sequencing and analysis of long RNA molecules. *Nucleic Acids Res* **50**, 6980-6989, doi:10.1093/nar/gkac518 (2022).
- 39 Liu, T., Patel, S. & Pyle, A. M. Making RNA: Using T7 RNA polymerase to produce high yields of RNA from DNA templates. *Methods Enzymol*, doi:10.1016/bs.mie.2023.06.002 (2023).
- 40 Liu, T. & Pyle, A. M. Discovery of highly reactive self-splicing group II introns within the mitochondrial genomes of human pathogenic fungi. *Nucleic Acids Res* **49**, 12422-12432, doi:10.1093/nar/gkab1077 (2021).
- 41 Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296, doi:10.1038/nmeth.4169 (2017).
- 42 Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat Methods* **17**, 1214-1221, doi:10.1038/s41592-020-00990-8 (2020).
- 43 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).
- 44 Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J Struct Biol* **157**, 281-287, doi:10.1016/j.jsb.2006.06.010 (2007).
- 45 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).
- 46 Kidmose, R. T. *et al.* Namdinator - automatic molecular dynamics flexible fitting of structural models into cryo-EM and crystallography experimental maps. *IUCrJ* **6**, 526-531, doi:10.1107/S2052252519007619 (2019).

- 47 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221, doi:10.1107/S0907444909052925 (2010).
- 48 Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr D Struct Biol* **74**, 531-544, doi:10.1107/S2059798318006551 (2018).
- 49 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877, doi:10.1107/S2059798319011471 (2019).
- 50 Dai, L., Toor, N., Olson, R., Keeping, A. & Zimmerly, S. Database for mobile group II introns. *Nucleic Acids Res* **31**, 424-426, doi:10.1093/nar/gkg049 (2003).

Chapter 3: Reverse splicing and introns as retroelements

3.1 Preface

This chapter is adapted from my work entitled “*Structures of a mobile intron retroelement poised to attack its structured DNA target*”, co-authored with Ling Xu, Pengxin Chai, Junhui Peng, Swapnil Devarkar, and Anna Pyle, published in *Science*¹. The biochemical assays and cryoEM work were done jointly by me and Ling Xu. This paper was published prior to the manuscript highlighted in Chapter 2, although the work in this paper describes the latter half of the group II intron splicing lifecycle, namely how they can function as a retroelement. Therefore, this chapter will describe some insights into group II intron RNP and assembly that were revealed prior to our work on the branching pathway.

3.2 Summary

Group II introns are ribozymes that catalyze their self-excision, functioning as retroelements that invade DNA. As retrotransposons, group II introns form ribonucleoprotein complexes that roam the genome for insertion sites, integrating by reversal of forward splicing. Here we show that retrotransposition is achieved by a tertiary complex between a structurally elaborate ribozyme, its protein mobility factor, and a structured DNA substrate. We solved high resolution cryoEM structures of an intact group IIC intron-maturase retroelement poised for integration into a DNA stem-loop motif. By visualizing the RNP before and after DNA targeting, we show that it is primed for attack, and fits perfectly with its DNA target. This work reveals design principles of a prototypical

retroelement and reinforces the hypothesis that group II introns are ancient elements of genetic diversification.

3.3 Introduction

Group II introns are self-splicing retroelements that have played a key role in shaping eukaryotic genomes as the ancestors of spliceosomal introns and non-LTR retroelements². They remain important for gene expression in plants, fungi, yeasts, and many bacteria^{3,4}. Group II introns encode a specialized reverse transcriptase, called a maturase or intron-encoded protein (IEP), that binds its parent intron and facilitates self-splicing, releasing a well-folded lariat ribonucleoprotein (RNP) complex⁵. The liberated RNP functions as a retrotransposon, targeting DNA that contains spliced exon junction sequences and inserting via a two-step transesterification reaction known as reverse splicing (Figure 3.1)⁶. The resulting DNA-RNA chimera is copied into cDNA by the reverse transcriptase (RT) activity of the multifunctional IEP in a process known as target primed reverse transcription (TPRT)⁷. Host repair pathways complete the downstream DNA copy and paste steps needed to achieve total intron integration⁵.

There are three main classes of group II introns, IIA, IIB and IIC, which share a conserved secondary structure and a similar tertiary organization around a ribozyme active site⁸. Group IIC introns are an ancient class of bacterial introns that recognize both the sequence and three-dimensional structure of their DNA recognition sites, forming base-pairs with short DNA sequences adjacent to DNA stem-loop motifs that are found at target integration sites⁹. Unlike their larger IIA and IIB counterparts, group IIC introns are almost completely dependent on their IEPs to facilitate intron excision through lariat formation,

thereby forming the functional RNP that serves as the minimal element for retrotransposition⁹. Compared to their more evolved counterparts, group IIC IEPs lack an endonuclease domain for generating TPRT primers and instead exploit the lagging strands at DNA replication forks⁵.

Recent structural and biochemical studies of IIA and IIB introns have provided important insights into strategies for IEP recognition of intron RNA. However, available RNP structures are limited in resolution, and have not revealed a specific mechanistic role for the IEP during active RNP assembly, DNA recognition, or chemical catalysis. At present, the mechanism by which group IIC introns recognize DNA structures, and not just DNA sequences, remains unclear. Furthermore, there are no available structures of the free RNP retroelement before it has bound DNA. These open questions have precluded a clear understanding of group II intron retrotransposition and its evolutionary role in shaping modern genomes.

To address these problems, we solved cryoEM structures of a group IIC intron retroelement poised to undergo the first step of reverse splicing. This revealed an elaborate RNA-DNA-protein assembly in which functional coordination of retroelement components precisely positions the maturase to facilitate ribozyme catalysis. We show that DNA is not a passive player and that specific DNA structures can play an active role during molecular recognition and function of an RNP. Furthermore, having visualized the apo-retroelement, our work allows for comparison with the substrate-bound holoenzyme, enabling us to monitor structural rearrangements that occur upon DNA binding. Together, these findings reveal the intimate interdependence of retroelement components and the design principles of a prototypical retrotransposon. Our findings have important

implications for understanding mechanistic function of the evolutionarily related spliceosome and non-LTR retrotransposons, such as the long-interspersed nuclear element-1 (LINE-1 or L1), which continue to remodel and impact the function of human genomes^{10,11}.

3.4 Results

3.4.1 Overall architecture of an ancient group II intron retroelement

To investigate the mechanism of DNA insertion, we captured a group II intron retroelement prior to the first step of reverse-splicing into DNA (Figure 3.1). We first conducted *in vitro* splicing reactions of the IIC *Eubacterium rectale* (*E.r.*)¹² intron in complex with its encoded maturase (MarathonRT)^{13,14} and purified the reaction mixture to obtain a branched lariat-maturase complex (Figure 3.2). Purity and stability of this RNP complex were assessed using biophysical methods: sedimentation velocity analytical ultracentrifugation and size exclusion chromatography coupled to multi-angle light scattering indicated that the sedimentation coefficient and molecular mass of the RNP were larger compared to the those of the individual lariat or maturase components, suggesting complex formation (Figure 3.3).

To visualize the retroelement in action (Figure 3.4), we introduced a desthiobiotin-tagged DNA substrate to the intron-maturase RNP and isolated ternary complexes by affinity purification on an avidin column (Figure 3.5). The highly purified elution fraction was vitrified on grids and the holoenzyme molecules appeared as monodispersed particles on cryoEM micrographs, thereby allowing structure determination through standard methodologies (Figure 3.6).

Initial data analysis suggested preferred orientation of the sample, so a tilted data collection strategy was required to obtain additional projection views. After further classification and refinement, we obtained a 2.8Å resolution cryoEM structure of the *E.r.* group IIC intron in complex with its specific IEP and DNA target (Figure 3.7), thereby revealing the state immediately prior to the first step of reverse splicing.

The overall high-resolution 3D reconstruction was of sufficient quality to permit modelling of individual nucleotides and metal ions. The catalytic core formed by D5, the lariat branchpoint, EBS-IBS (exon binding site-intron binding site) sequences and the protein thumb and DNA binding domain (DBD) were resolved to < 3Å, whereas regions at the periphery of the structure, such as portions of D2 and D4 range from 4-8Å. A crystal structure of the free *Oceanobacillus iheyensis* (*O.i.*) intron was docked into the density map and used as a reference for building the central portion of the RNP¹⁵. DRRAFTER and molecular dynamics flexible fitting methods were used for modelling peripheral regions^{16,17}.

The overall structure reveals a compact assembly of intron RNA and maturase protein that is closely associated with the DNA substrate through a network of unusual interactions. The intron core adopts a fold similar to that of available crystal structures derived from truncated and modified group II intron constructs^{15,18}. Tertiary interactions previously identified in the *O.i.* IIC intron are present, along with additional novel interactions that are observed in this full-length intron construct that contains all six intact intron domains (Figure 3.8). The fold of the maturase resembles that of previously studied IIC proteins^{13,19}, although the thumb and DNA binding domains are now clearly resolved. The bound DNA

contains a structured linker region, a native DNA target sequence and a proximal 5' stem-loop motif that is unique to group IIC introns^{20,21}.

3.4.2 Features of the catalytic RNP core

Despite extensive efforts, a complete group II intron holoenzyme active site had not yet been visualized. In this work, we capture the complete ribozyme core architecture, which includes hallmark elements that were identified in earlier biochemical and structural studies^{9,22}. Given the improvement in resolution, the molecular interaction networks of these motifs are more well-defined and elaborate than previously described. For example, we see that the 2',5' lariat linkage, between the first intron nucleotide (G1) and the branchpoint A (A632), is a crucial structural motif for organizing the ribozyme core. The branch site actively engages the 3' end of the intron, positioning the terminal nucleotide (U638) for nucleophilic attack on DNA (Figure 3.9)^{23,24}. Facilitating this process, U638 stacks upon G1, and base-pairs with A327 to form the γ - γ' interaction (Figure 3.10)^{23,24}. The adjacent G328 and C329 nucleotides of the J2/3 linker form major-groove base triples with C562 and G563 (Figure 3.11), giving rise to the catalytic triplex that is common to all group II introns and the spliceosome^{25,26}. The 2-nt bulge (A580, C581) and catalytic triad (C562, G563, C564) in D5, along with U638, all serve to coordinate catalytic magnesium ion M1, placing it between the nucleophilic 3' OH and scissile phosphate, in an arrangement poised for the first step of reverse splicing (Figure 3.9)¹⁸. A second magnesium ion, M2, is located 4Å away from M1, consistent with the two-metal ion catalysis mechanism (Figure 3.9)^{18,27}. The high-resolution map allowed us to identify two additional, unambiguous densities at positions previously assigned to monovalent ions K1 and K2 in

crystallographic studies that employed anomalous scattering to establish sites of stable K^+ binding^{26,28}. In that case, as in this instance, NH_4^+ can functionally substitute for K^+ at these same positions, which is common in many enzymes (Figure 3.9). The specific coordination and placement of these monovalent ions is essential for positioning the catalytic divalent metal ions, forming a reactive, heteronuclear metal ion cluster. Presumably due to lower resolution, it was not possible to visualize the intact metal ion cluster in previous structural studies of the lariat intron^{29,30}. We see that numerous tertiary interactions stabilize the periphery of the catalytic core, with D3 bracing the backside of the D5 helix (μ - μ') and D2 contacting the D6 helix under the branchpoint A (π - π') to hold the lariat in place, as observed in IIA and IIB introns (Figure 3.11)^{23,31}. Although many of these active site elements have been observed independently in varying contexts, in linear introns or in introns of other classes, they have not been simultaneously captured in a single structure until now, thereby demonstrating that these active site elements function in concert. The *E.r.* holoenzyme structure provides a detailed view of a complete, reactive intron catalytic core, and reveals the coupling of RNA motifs with maturase protein residues.

Close inspection of the active site reveals structural interdependence between the intron RNA and its encoded maturase. Within the active site, the intron RNA forms short base pairings with its target DNA via the EBS-IBS interactions (EBS-IBS1 and EBS-IBS3) (Figure 3.9 and 3.10). These otherwise unstable short pairing interactions are buttressed and positioned by the maturase, which presses the middle α -helices of the DBD and the third α -helix of the thumb domain against the EBS1 and EBS3 recognition loops respectively, rigidifying them and helping to form a central cavity for engagement with DNA (Figure 3.9). These findings establish that the retroelement core does not consist

solely of RNA, rather it is a collaborative, ribonucleoprotein active site. This contrasts with results from previous studies in which the maturase thumb coordinated exclusively with the DNA substrate and was not observed to interact directly with the EBS sequences^{29,32}. The new roles we observe for the maturase thumb and DBD help to explain the strong maturase dependence for both RNA splicing and intron integration, particularly *in vivo*, and they highlight the symbiotic relationship between the intron RNA and its protein cofactor, which are known to have co-evolved^{33,34}.

3.4.3 Unexpected functional coordination between RNA and protein

A striking feature of the retroelement holoenzyme is the expansive D4 arm, which extends far from the core and then curves around to cradle the maturase along the periphery of the holoenzyme (Figure 3.12a). D4a, the high affinity maturase-binding subdomain^{13,35}, forms two anchor points with the basic surfaces of the protein (Figure 3.12a and b)³⁶. At the first anchor point, residues extending from the protein (R58, T156, R160, and R172) interact with RNA phosphate and ribose oxygens, securing the insertion helix within the finger domain (IFD) of the IEP against the minor groove interface in the middle of the long D4a hairpin (Figure 3.12b and c). A sharp turn places the distal portion of the D4a subdomain between α -helices 9 and 10 of the protein, where largely basic residues (R217, S234, S237, R240, N244 and R247) approach the RNA backbone from either side, fastening the palm to the D4a arm (Figure 3.12b and d). The surface of the finger domain (RT0) is not utilized for RNA recognition¹³. This contrasts with the behavior of group IIA and IIB RNPs, where the maturase associates with D4a primarily through RT0, leaving the backside of the palm exposed^{29,30} (Figure 3.13).

The distinct intron-maturase recognition strategy places the maturase thumb and DBD next to the intron core, allowing the protein to participate in catalysis by rigidifying the active-site (Figure 3.12a and e). The thumb and DBD grasp the EBS1 and EBS3 loops, directly coordinating substrate recognition elements within the retroelement active site (Figure 3.12e). One approach of this strategy involves specifically locking EBS nucleotides into a conformation conducive for substrate binding (ie. K388 with G187O6 of EBS1 and K358 with A231N7 of EBS3) (Figure 3.12e to g). A secondary tactic includes immobilizing the EBS3 phosphate backbone through interactions with a multitude of basic residues on the protein thumb (K300, K303, S309 and R351) (Figure 3.12g). A third strategy consists of DBD amino acids (R389 and main chain amines of I390 and A391), stabilizing the turn in EBS1 and enabling the formation of δ - δ' , thereby reinforcing this single base pair interaction (C183 with G158) that bridges the EBS loops. (Figure 3.12f). Interestingly, R308 of the protein thumb provides additional stabilization by simultaneously coordinating the phosphate backbone of EBS1 and 3 (through C183O5' and A230OP2), effectively joining and aligning the two substrate recognition loops (Figure 3.12h). These interactions demonstrate a specific mechanistic role for the maturase protein during catalysis, showing that it promotes proper formation of multiple active site components³⁷. These findings reveal the inextricable, functional coordination of intron and protein during the mechanism of splicing and retrotransposition.

3.4.4 New tertiary interactions with a structured DNA

Some of the most remarkable features of this structure involve the DNA target and its unusual strategies for molecular recognition by RNA and protein residues of the

holoenzyme. Here we show that the DNA is recognized through a combination of shape-selectivity and base-specific interactions, only a few of which involve canonical WC pairing. The DNA itself has distinct structural features that support this recognition strategy. Most prominent is an unusual, structurally conserved, DNA stem, which is comprised of a short B-form helix (7 bp) that is capped by an undertwisted duplex comprised of two G-C base pairs and two noncanonical, staggered G-A DNA base-pairs. Together, these extend the DNA stem to 11 bp, which approximates the consensus stem length for IIC insertion targets. The terminal DNA loop serves as a stacking platform for long-range interactions between the DNA and residues of the intron RNA. Adjacent to the DNA stem is a short linker region, which is followed by IBS1 nucleotides and the IBS3 nucleotide that flanks the DNA insertion site (Figure 3.14).

The DNA stem lies in a cleft that is formed by regions of both the protein (DBD and thumb) and the intron RNA (D1d and D4a). Two clusters of amino acids along the third α -helix of the protein thumb domain anchor the DNA stem by making contacts at both ends of the DNA helix, at positions separated by approximately one helical turn. The first cluster (S346, R349, N395, N405) secures the base of the stem through contacts with dG1 and non-bridging phosphate oxygens of dA2 and dG3 (Figure 3.15a). The second group (S336, K338, T339 and Y278) appears to locally deform the top base pairs of the stem at dC20 and dT21 (Figure 3.15a). This is the result of a striking DNA-protein interaction network that involves insertion of a prolyl-aromatic loop into the distorted, widened minor groove at the tip of the DNA stem. The complementary fit of this peptide loop is mediated by interactions between largely buried side chains (Y278, F279 and P281) and the methylene edges of DNA sugar moieties (Figure 3.15b). These protein-DNA interactions are

supported by contacts between the DNA and RNA backbone residues (dA403' and dG50P1 with G163 2' OH), reminiscent of ribose-zipper interactions observed within folded RNA molecules (Figure 3.16a)¹⁸. Collectively, these interactions enable the holoenzyme to coordinate and selectively identify the shape of a DNA helix.

This shape-selective recognition strategy of the DNA stem is complemented by sequence-specific interactions between the holoenzyme and single-stranded regions of the DNA target. Phylogenetically covarying base-pairs are formed between substrate-recognition regions of the intron (EBS1 and EBS3) and single-stranded DNA nucleobases downstream of the DNA stem (IBS1 and IBS3)³⁸. In the holoenzyme, we not only identify these critical WC pairings, but we also observe an unexpectedly complex network of interactions mediated by the spacer DNA that connects the stem with the IBS sequences. This sequential network of IBS-linker interactions begins with the nucleotide located immediately downstream of the insertion site (dA36), which forms a single-base-pair interaction (EBS3-IBS3) with a nucleotide extending from the D1d coordination loop within the intron (A231) (Figure 3.16b). Stacked atop this pair is a short helix formed via base-pairings between the subsequent stretch of DNA nucleotides (IBS1: dT35, dT34, dT33 and dC32) and a second substrate recognition loop that projects from the terminus of intron D1d (EBS1: A184, A185, A186 and G187) (Figure 3.15b). Like the short codon-anticodon helix in the ribosome³⁹, the EBS-IBS1 duplex is further stabilized via formation of an A-minor motif between A75 and the dC32-G187 base-pair (Figure 3.16c). Intriguingly, the structure reveals that EBS1-IBS1 is not limited to four contiguous base-pairs, rather, it is extended by an additional base-pair that is formed between the next consecutive nucleotide (A188) in the EBS1 loop and a discontinuous nucleotide from the

DNA spacer region (dT30). Indeed, the intervening DNA nucleotide (dT31) is extrahelical and stabilized by interactions with protein residues (*vide infra*) (Figure 3.17). Through these sequential stacking networks, supported by contacts with the protein (ie. dT36O4 with K361), the intron achieves stable, base-pairing specificity with the DNA target.

Nucleotides within the DNA spacer region actively participate in binding the RNP, as they adopt an ordered structure that engages in specific interactions with the holoenzyme. Rather than forming a helical stack, the spacer nucleotides (dA28, dT29, dT30 and dT31) form an unusual motif in which the nucleotides splay in alternating directions on either side of the central phosphate spine (Pauling-like DNA), thereby exposing a large interaction interface to the adjacent DBD (Figure 3.16 and 3.17). Amino acids from the DBD intercalate between the DNA spacer nucleotides while forming an abundance of interactions with both the bases and the phosphate backbone (Figure 3.17). For example, N3 of dT31 interacts with amide oxygens on both the main and sidechain of N378 while its adjacent phosphate oxygens interact with proximal arginine residues (R381 and R382). Together, these interactions stabilize an unusual backbone conformation that enables the dT30-A188 pair to form atop the EBS-IBS1 helix. In turn, these interactions with the DBD pull the protein domain into place, positioning a specialized barb-like structure formed by an α -helical bundle within the DBD at the base of DNA stem (Figure 3.17).

Capping the DNA stem-loop, a remarkable set of stacking interactions with RNA and protein clamp the loop terminus into position within the holoenzyme. One such interaction forms between the DNA and RNA loop nucleotides that project from D4, which effectively joins the DNA stem and RNA bases into one continuous stacking network. This extended stacking array consists of dG11, dC12, and two nucleotides from D4a (Figure 3.18). While

lower local resolution in this region prevents unambiguous identification of the D4a RNA nucleobases, there is obvious density for two such nucleotides, assigned here as A441 and A442. This DNA-RNA tertiary interaction is anchored in place by an adjacent stacking network that forms between the extrahelical dT13 residue and a series of conserved amino acid side chains, which form a sequential stack that merges with the hydrophobic core of the protein. Specifically, the aromatic plane of the dT13 nucleobase and W163 flank R268 on either side creating an arginine- π stacking sandwich configuration⁴⁰, in which each component is separated by a planar distance of 3.8Å (Figure 3.18).

3.4.5 Retroelement primed for attack

To better understand molecular rearrangements that might occur when the intron retroelement binds to DNA substrate, we solved the structure of the apo-RNP, visualizing the free intron-maturase complex at a resolution of 3.1Å (Figure 3.19). We were surprised to observe that the apo-RNP has an architecture that is almost identical to that of the complex bound to DNA, and that DNA binding induces only minor changes in the structure. Most remarkably, the RNP active site remains completely intact (Figure 3.19)^{29,30,41}. The maturase does not change its orientation in the absence of DNA, remaining coordinated at two anchor points along the D4a arm, with the thumb and DBD inserted into the active site to participate in catalysis (Figure 3.19). In this configuration, the binding interface for the target DNA is maintained, enabling the RNP to readily recognize an incoming DNA target and rapidly engage in retrotransposition. Upon recognition of the DNA stem-loop, the RNP (palm, fingers and D4a) appear to become stabilized, as we observe a concomitant increase in local resolution in the cryoEM map at these positions.

This is reminiscent of many protein enzymes, whereby docking of ligand into the active site freezes out local active-site motions and locks the substrate in place. In previous ligand-free intron structures, EBS nucleotides were found to be disordered or rearranged^{23,24,26}. Here, we observe that the precise positions of the EBS nucleotides are unchanged, likely due to the presence of the maturase. These findings reinforce the mechanistic role of the protein as a stabilizing catalytic component (Figure 3.20). Further evidence of a pre-formed catalytic core includes the persistence of the heteronuclear metal ion cluster, which remains organized around the lariat linkage, although M2 is not visible in this case (Figure 3.20).

3.5 Discussion

3.5.1 Insights into Protein Facilitated Ribozyme Catalysis

The structures presented here reveal how components of the holoenzyme help promote activity of the ribozyme core. The protein buttresses the catalytic residues responsible for specifically positioning the DNA substrate, providing a missing link in understanding how maturases unlock the full catalytic potential of group II intron ribozymes. Interactions with protein residues stabilize the intron substrate recognition loops and precisely arrange nucleotides for DNA binding. These interactions contribute to proper orientation of reaction components throughout ribozyme catalysis. Our findings provide a direct mechanistic role for the protein cofactor, and they help explain the lowered salt and magnesium requirements in its presence⁴².

Prior studies, due to resolution limitations and construct truncations, were unable to identify a specific function for the maturase protein, except in transient D6 stabilization²⁹. In those cases, a linear DNA substrate was sandwiched between the protein and intron

recognition loops, whereas, in our study, the DNA sits along the maturase thumb and DBD, allowing the protein to approach the substrate recognition loops within the ribozyme core (Figure 3.9). The same configuration is seen in the spliceosome where the Prp8 thumb is angled inward, toward the U5 recognition loop⁴³. For the first time, we can therefore make the distinction that not only are the protein thumb and DBD proximal to D1 catalytic residues, but also that they have critical roles in stabilizing substrate binding. Given its analogous spatial placement, it is possible that Prp8 may play a similar role during spliceosomal catalysis.

3.5.2 RNP Recognition of DNA Structure: an expanded recognition repertoire

The high-resolution cryoEM structures we provide here offer an unprecedented glimpse into RNP strategies for recognizing DNA (Figure 3.7). The holoenzyme structure reveals a stem-loop DNA nestled within the retroelement, bound to RNA and protein. Within this cleft, the protein assists in positioning the insertion site and aligning the DNA stem for steric fit against the complementary maturase surface (Figure 3.12). Additional aspects of the unusual recognition strategy include splayed Pauling-like DNA, an A-minor motif, and intermolecular stacking moieties that involve both RNA and protein (Figure 3.14). These interactions highlight the symbiotic nature of RNA and protein and underscore the multiplicity of strategies available to RNPs for achieving selective substrate recognition. Proteins have long been known to recognize DNA structures, but here we show that RNPs, with their expanded repertoire of DNA molecular recognition determinants, have this ability too⁴⁴. This is the first instance of a biomolecular machine in which a higher order

DNA structure guides a biochemical reaction and the first example of a ternary complex in which RNA, DNA, and protein structures each have a key mechanistic role.

The DNA stem loop motif is exclusive to IIC introns, which contain an abbreviated D1 scaffold and short exon recognition sequences⁸. Intriguingly, in the more highly evolved IIA and IIB introns, the RNP binding motif that we find occupied by the DNA stem in IIC introns is instead replaced by intron insertion motif D1d2, an RNA subdomain that includes EBS2, which is absent in the IIC class⁸. Comparison of this region across intron classes suggests that EBS2 evolved to imitate the target DNA stem (Figure 3.21). Indeed, the DNA stem motif structurally resembles the EBS2-IBS2 interactions typical of IIA and IIB introns and it functionally emulates EBS2 by anchoring the DNA substrate into the RNP. This mimicry suggests that initial non-specific recognition of a structured DNA motif by the more primitive IIC introns was replaced by novel RNA domains within the intron itself, resulting in longer target recognition sequences that subsequently provided greater base-pairing specificity for the retroelement.

3.5.3 Implications for Reverse Splicing and Reverse Transcription

A surprising feature of our structures is the way that the protein is positioned within the holoenzyme. Encircled by RNA, the exterior surfaces of the protein are occupied, but the concave interior of the protein, adjacent to the catalytic core, is conspicuously solvent accessible, which has functional implications. During reverse splicing, the D6 helix undergoes a conformational change that places the lariat linkage into the active site²⁹. To accomplish this, D6 disengages from D2 and swings 90° upward, contacting D1c and a basic patch on the protein thumb. Our structures do not preclude these D6 helix dynamics,

as there is ample space for a similar movement and the regions that D6 contacts remain accessible. The open architecture we observe provides a direct route for DNA to approach the RT active site, as it remains unobstructed by other intron domains and can readily accommodate an entire hybrid duplex for reverse transcription⁴⁵. This suggests that initiation of RT activity, within the current holoenzyme assembly, may be possible without significant conformational rearrangement.

3.5.4 Retroelement Poised to Attack

Group II intron retroelements are proliferative, invasive agents and our structures explain why. The apo-retroelement is completely poised to react and does not require any reorganization of structure upon target DNA binding. The arrangement of the active site, from substrate recognition nucleotides to the heteronuclear metal ion cluster to the DNA binding interface, is preserved despite the absence of DNA substrate (Figure 3.19). This prearranged organization is consistent with the biological role of group II introns as parasitic genetic elements⁴⁶. Use of the same catalytic core from splicing to integration eschews the need for major rearrangements or host cofactors and allows complete autonomy, which is highly advantageous for a genetic parasite.

Total integration of the RNP requires faithful and accurate reverse transcription of the intron sequence, including the long ORF that encodes the protein, after insertion. This is accomplished using the RT activity of the multifunctional maturase protein. MarathonRT, the protein within the holoenzyme visualized here, is a well characterized, robust, accurate and ultraprocessive RT enzyme capable of copying through long, structurally complex templates⁴⁷. Indeed, due to its ability to processively copy long, structured RNA templates,

this RT has become a common tool in research and biotechnology^{14,47}. The intimate association of the parent intron with this protein allows access to its exceptional reverse transcriptase properties and ensures that the intron sequence, which is pivotal to its tertiary architecture, is preserved, allowing the retroelement to continually propagate. The retroelement is prepared to attack and paired with MarathonRT, it is ready to retrotranspose.

3.5.5 Implications for Modern Retroelements

Study of group II intron complexes provides a window into our understanding of related non-LTR retrotransposons, such as the LINE-1 RNP, which is an active mobile element that continues to disperse in human genomes^{10,11}. Computationally predicted structures of ORF2p, the mobility factor of L1, show that its RT and thumb domain resemble that of the maturase protein, MarathonRT (Figure 3.22)⁴⁸. ORF2p contains an additional N-terminal endonuclease and a C-terminal extension, but these domains do not block the exterior basic surfaces of the RT and thumb. MarathonRT and ORF2p are evolutionarily related, and they are implicated in similar mobility mechanisms, so it is possible that the same surfaces are used for anchoring and substrate recognition⁴⁹. As the thumb plays a role in mediating DNA binding, it is tempting to speculate that mutations in this region may modulate substrate selection (Figure 3.9). Given the lack of structural information on L1 and the strong parallels between systems, our work provides a starting point for imagining how a similar retroelement like L1 might assemble.

3.5.6 Group II RNP lifecycle

By combining the cryoEM structures obtained in this study with previous mechanistic and structural work done on group II introns^{1,9,18,26,50}, we can now propose a mechanism for the group II intron splicing lifecycle (Figure 3.23), including excision from the flanking exons and retrohoming into DNA sites. After translation of the maturase from the ORF, the protein facilitates RNA folding by binding to the D4a arm and interacting with D1, which folds first and acts as a scaffold^{51,52} for assembly of downstream domains. The D6 branch helix docks onto D1 through the intramolecular ν - ν' interaction and engages the thumb/DBD domains of the maturase to stabilize the helix in the up position. Specific molecular interactions distinguish and lock the bpA and 5' SS into place, juxtaposing the 2' OH against the scissile phosphate for nucleophilic attack via the heteronuclear metal ion active site. During the first stage of branching, the 2',5' phosphodiester bond is formed, exposing the 3' OH of the 5' exon for ligation. To exchange substrates, the branch helix then disengages ν - ν' and the maturase-D6 interactions, permitting the bpA to pivot around its phosphate and the D6 helix to swing downwards, where it forms the π - π' tertiary interaction with D2. This pulls the lariat out of the active site and replaces it with the 3' SS, demarcated by the EBS3-IBS3 base pairing, thereby positioning the 3' exon for ligation. Using the same active site, the 3' OH is activated for nucleophilic attack, resulting in splicing of the exons. Following exon ligation, the D6 3' tail tucks inward, and the terminal nucleotide is secured by the γ - γ' interaction. The ligated exons are then released, and the liberated apoRNP retains its overall architecture, enabling it to remain primed for binding DNA substrates based on shape and sequence complementarity for engagement in reverse splicing¹.

To undergo retrotransposition, we postulate that equivalent, conserved D6 and branch site motions are employed²⁹ in order to achieve intron integration and substrate exchange using a persistent heteronuclear metal ion core. Given the proximity of the maturase RT active site to the 3' end of the integrated intron, a logical hypothesis is that the 3' end of the fully reverse spliced product is threaded into the maturase RT domain. Here the protein, using an exogenous primer, begins target primed reverse transcription, unraveling the base pairing, disassembling the elaborate intron tertiary structure⁵³, and generating a cDNA strand to effectively copy and paste the RNA sequence into a new genomic site, thereby completing the intron lifecycle. Further biochemical and structural work will be needed to evaluate this hypothesis and address the remaining mechanistic aspects of the group II RNP lifecycle post-branching.

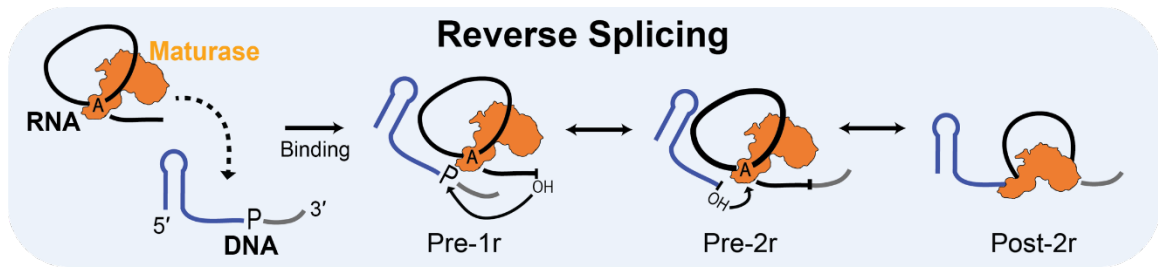


Figure 3.1 Cartoon of the retrotransposition reaction.

The liberated group II RNP can integrate into its structured DNA target, inserting via two transesterification steps that are exactly the reversal of the splicing reaction. The pre-1R species is the DNA bound form of the RNP, prior to reaction. The pre-2R results after the first transesterification step when the 3' end of the DNA has been covalently attached to the intron. After the intron holoenzyme has fully integrated, a chimeric RNA-DNA hybrid forms, which corresponds to the post-2R complex.

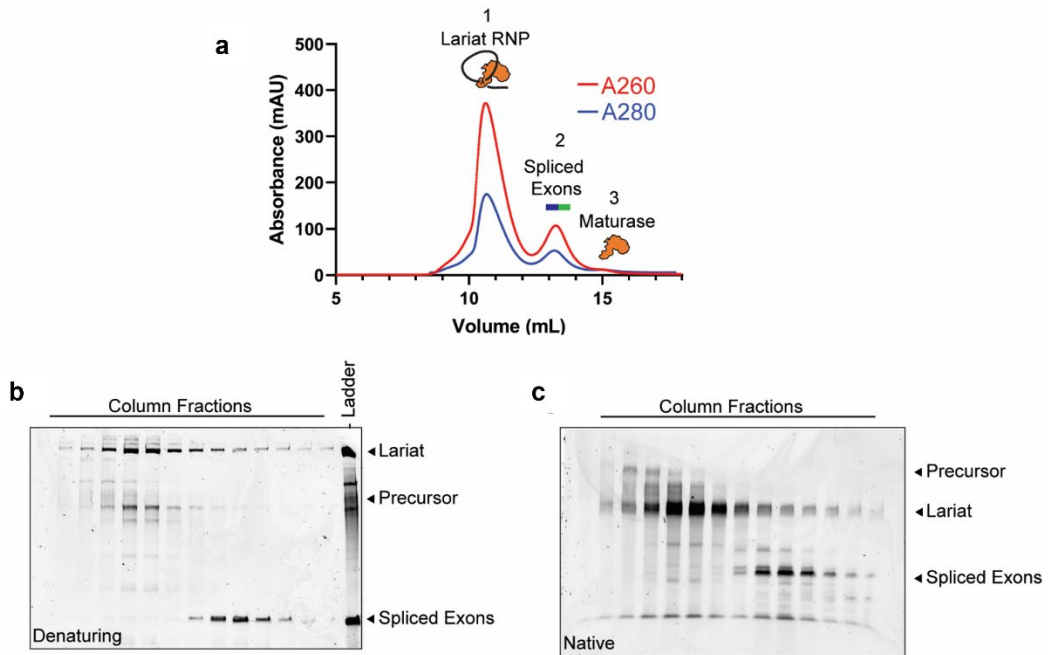


Figure 3.2 Purification of a group II intron RNP.

a. Size exclusion chromatogram of a group II intron forward splicing reaction. The three peaks are (1) lariat RNP (apo-complex), (2) spliced exons, and (3) excess maturase. A₂₆₀ and A₂₈₀ traces are shown in red and blue respectively. b. Denaturing gel of the column purification fractions from (a) with a splicing ladder reference. The three sets of bands correspond to the lariat, unreacted precursor, and spliced exons. c. Native gel of the column purification fractions from (a). The two dominant sets of bands correspond to lariat RNP and spliced exons.

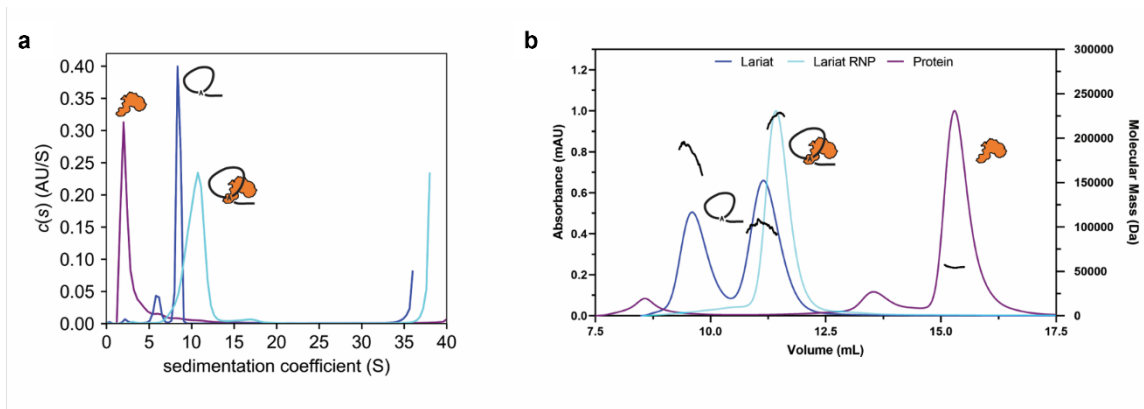


Figure 3.3 Biophysical characterization of the intron RNP.

a. Sedimentation profiles obtained from analytical ultracentrifugation. The three peaks from lowest to highest sedimentation coefficient correspond to the maturase protein, purified lariat alone, and purified lariat RNP. b. SEC-MALS chromatogram of purified lariat, purified lariat RNP and maturase. The purified lariat migrates as two separate species of 195 kDa and 110 kDa (degradation product). The purified lariat RNP appears as a monodispersed peak with a molecular weight of 225 kDa. The maturase protein elutes last and has a measured mass of 54 kDa.

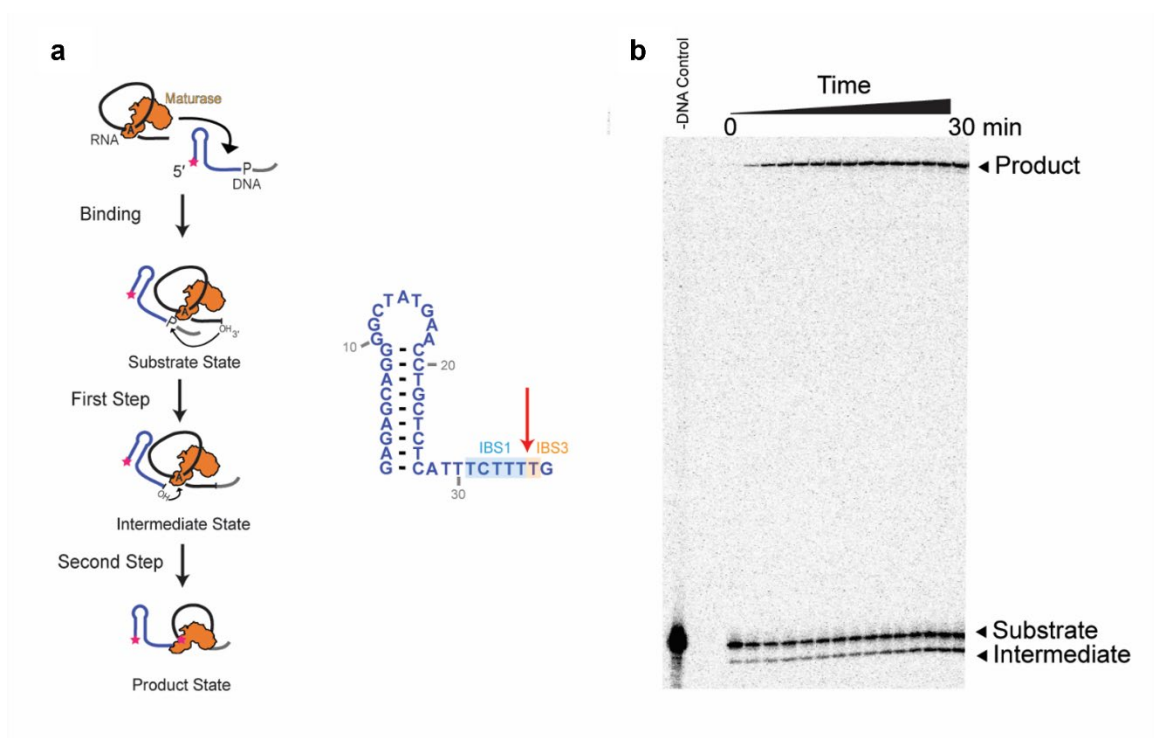


Figure 3.4 Reverse splicing activity assay.

a. Cartoon of the reverse splicing reaction (left) - the group II intron integrates into a DNA target (right). The lariat-maturase RNP binds and recognizes the DNA substrate before inserting into the exon junction followed by opening of the lariat. The red star indicates the 5' radiolabel used to visualize the substrate (37-nt), intermediate (35-nt) and product (~700-nt) states of the reaction. b. Denaturing gel of the reverse splicing reaction time course with WT DNA substrate.

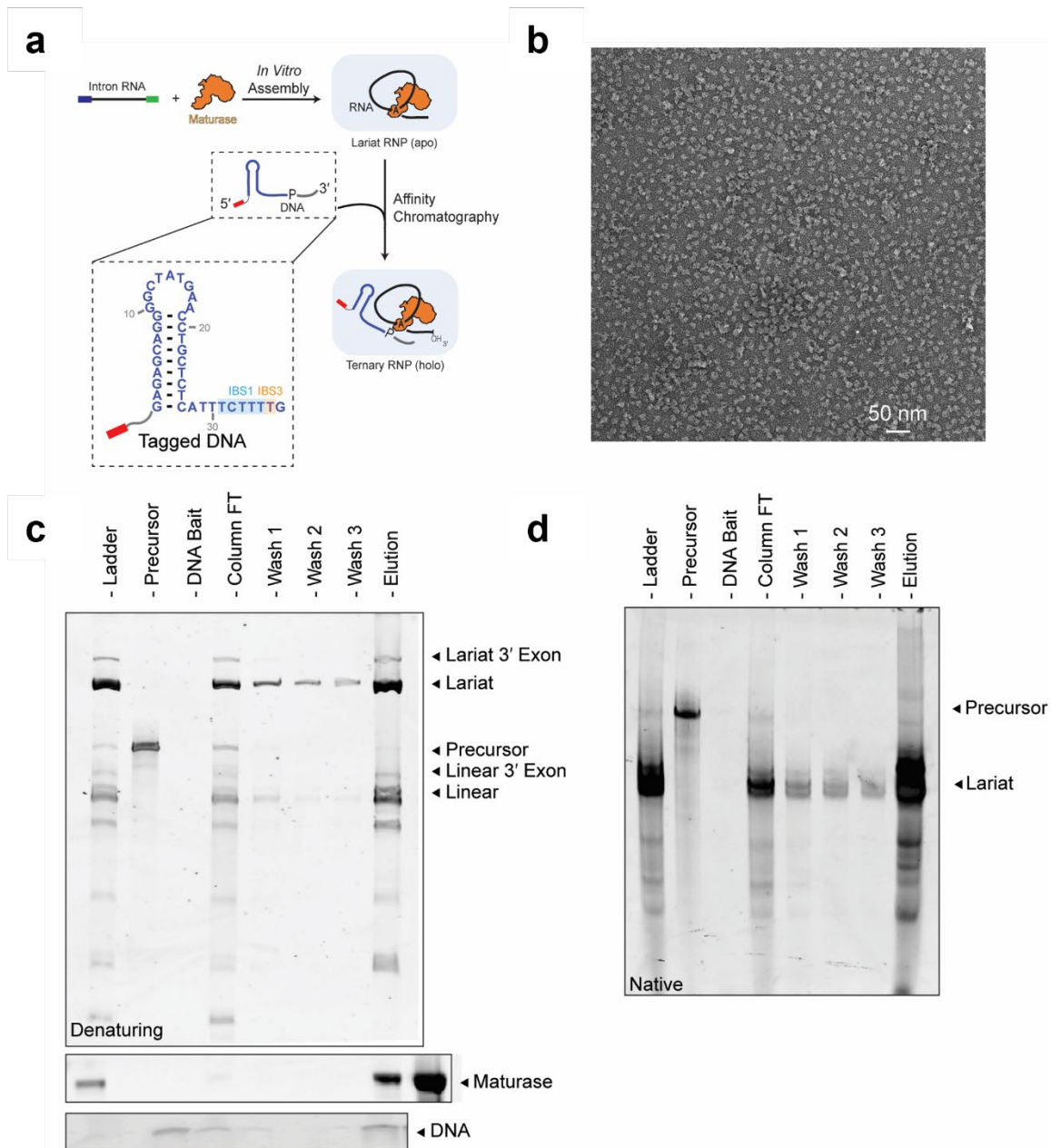


Figure 3.5 Purification of a group II intron retroelement.

a. Cartoon of the holoRNP purification process. Precursor intron RNA was incubated with maturase and reacted to form the lariat RNP complex. The *in vitro* assembled complex was purified and incubated with a tagged, structured DNA target. The secondary structure of the desthiobiotin tagged DNA is shown. Ternary complexes containing lariat RNA, maturase and the DNA hairpin were isolated by affinity chromatography by binding the

ternary complex to an avidin column and eluting with biotin. b. Negative stain electron micrograph of the elution fraction from affinity purification. c. Denaturing gel of the affinity chromatography fractions. The middle inset shows a SYPRO Ruby stained protein gel of the same fractions with an additional maturase control. The bottom inset shows a higher percentage denaturing gel of the same fractions. d. Native gel of the affinity purification fractions.

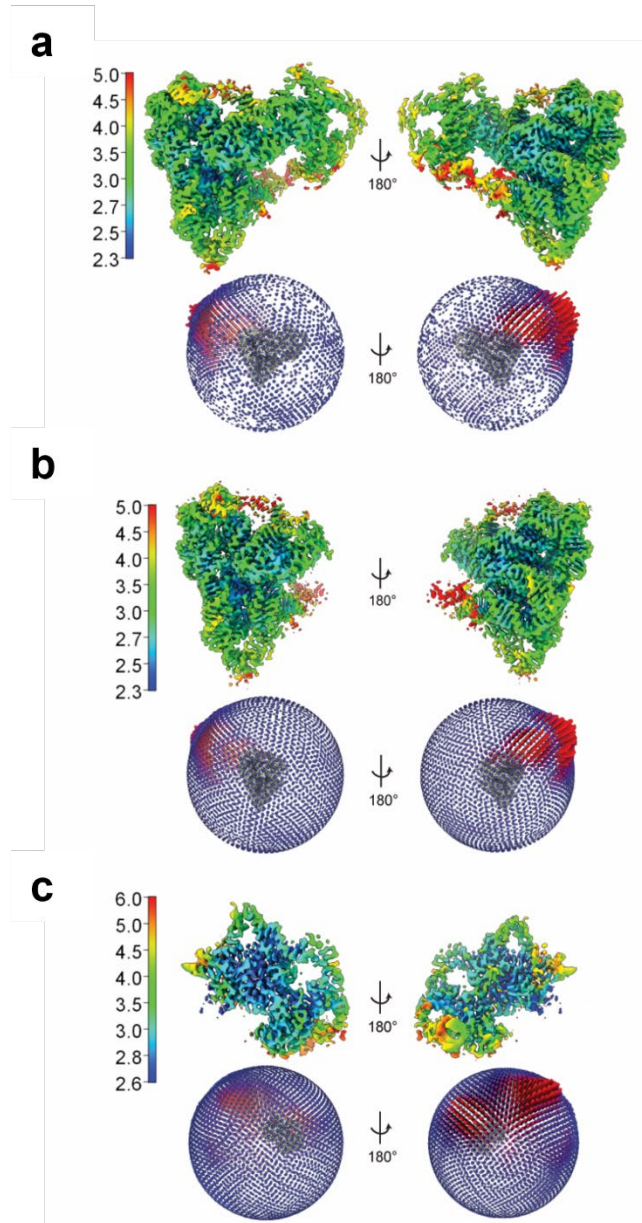


Figure 3.6 CryoEM reconstruction of the intron retroelement.

a-c. Local resolution map and particle distribution of the (a) holo-RNP full map, and (b) the left, (c) and right focused refined maps,

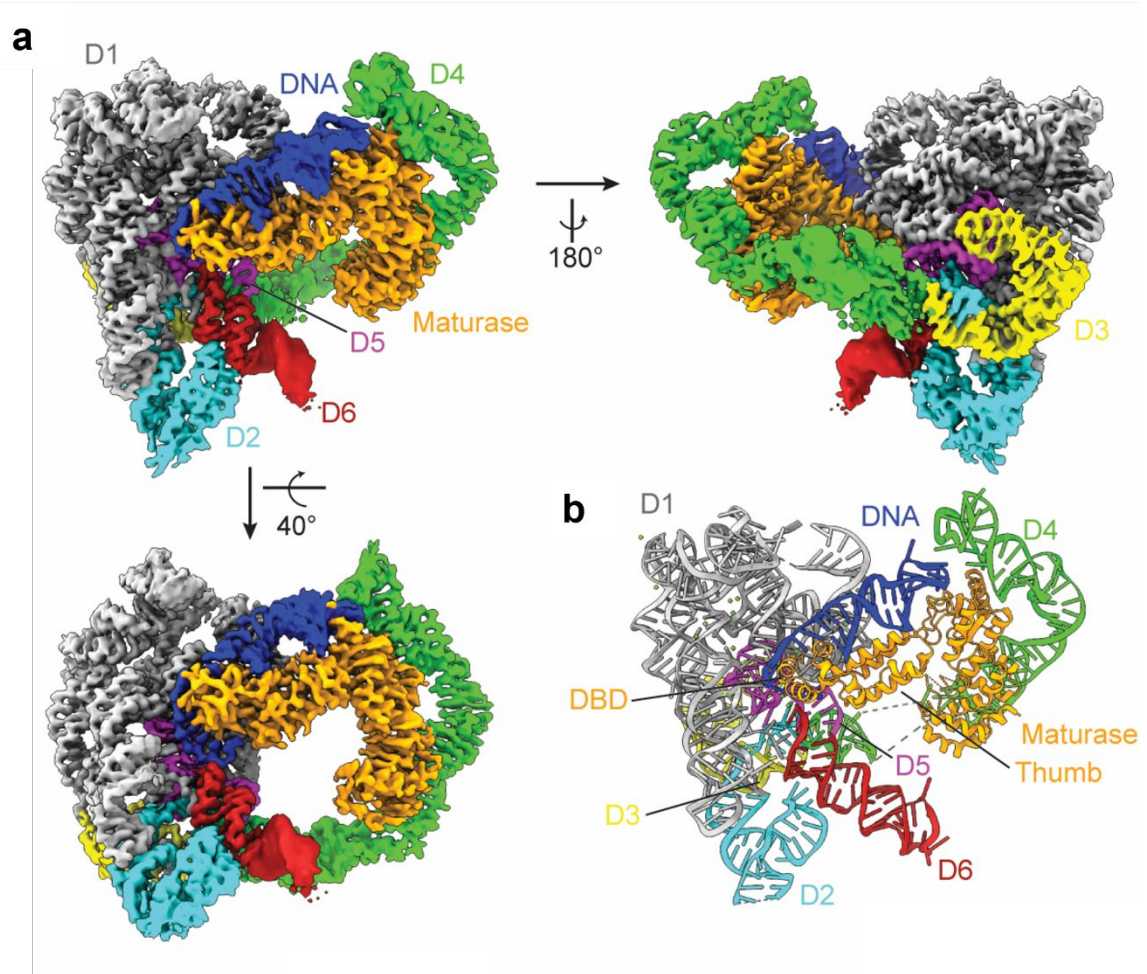


Figure 3.7 CryoEM structures of a group II intron holoenzyme.

a. Composite cryoEM map of the holo-RNP with bound DNA. b. Molecular model built into the group II intron holoenzyme reconstruction.

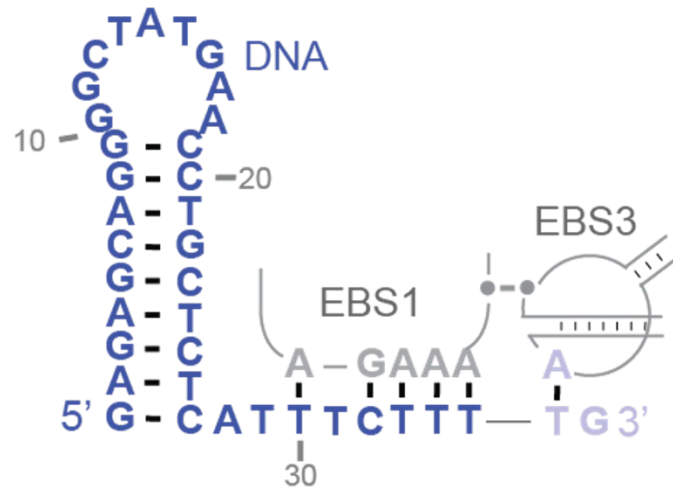


Figure 3.8 Sequence and structure of the holoenzyme DNA target.

The sequence of the WT DNA hairpin substrate is shown in blue. The sequences 5' of the cleavage site are shown in dark blue, while those 3' of it are shown in a light shade of blue.

The intron base pairing sequences are shown in grey and labeled as EBS1 and EBS3.

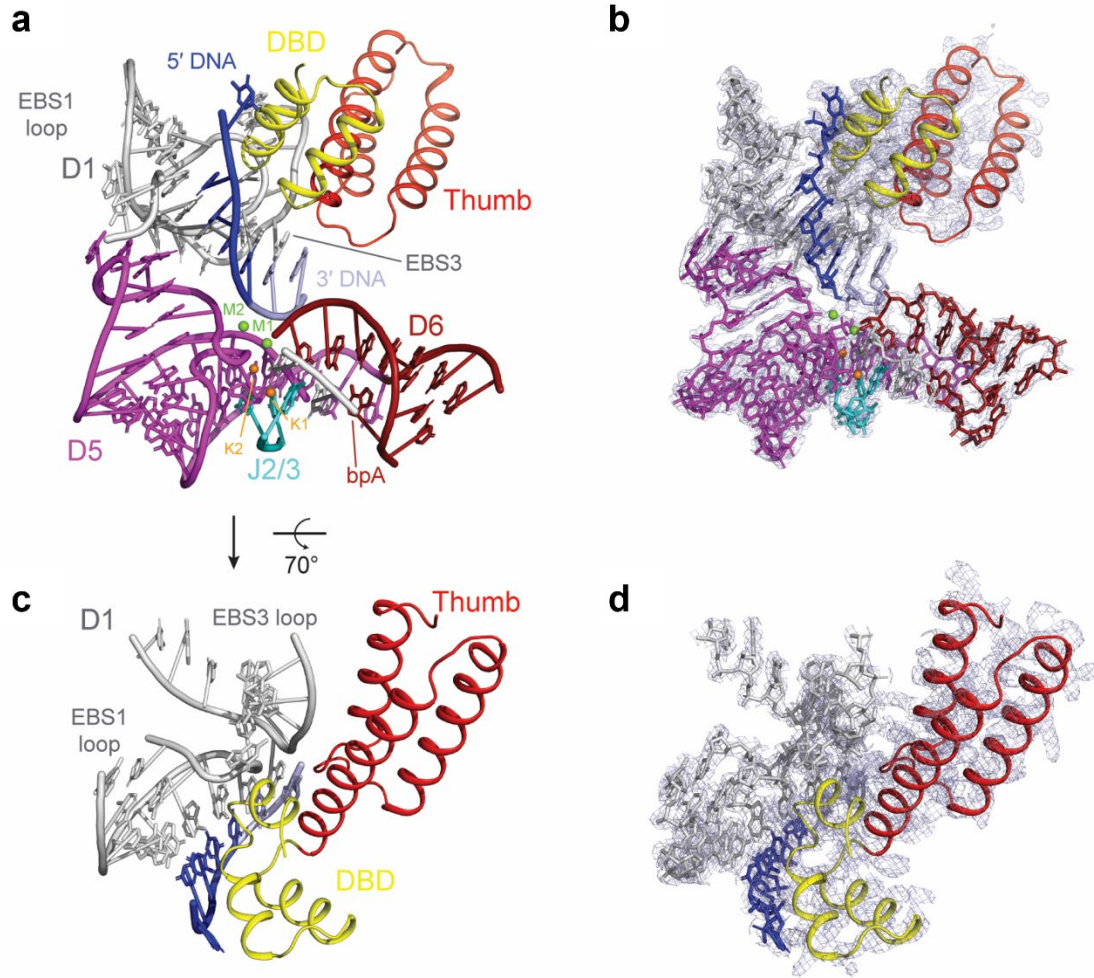


Figure 3.9 Catalytic core of the group II intron holoenzyme.

- a. Organization of the intron catalytic core around the central D5 helix and the metal ions.
- b. The key elements around the active site, with surrounding mesh, are shown in this RNP captured in the state just prior to DNA insertion.
- c. Zoomed in region around the active site demonstrating how the maturase DBD and thumb stabilize the EBS1 and 3 recognition loops.
- d. The same region shown in (c) with surrounding mesh.

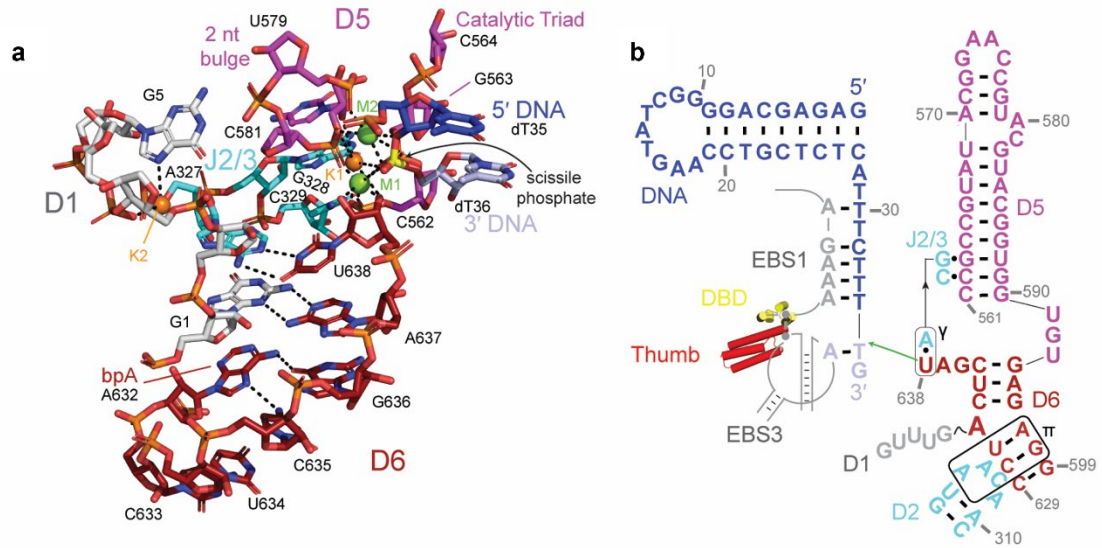


Figure 3.10 Cleavage of the DNA target.

a. Model of the terminal nucleotide poised to attack the scissile phosphate of the DNA target with the surrounding catalytic nucleotides. b. Secondary structure schematic of the intron retroelement prior to the first step of retrotransposition.

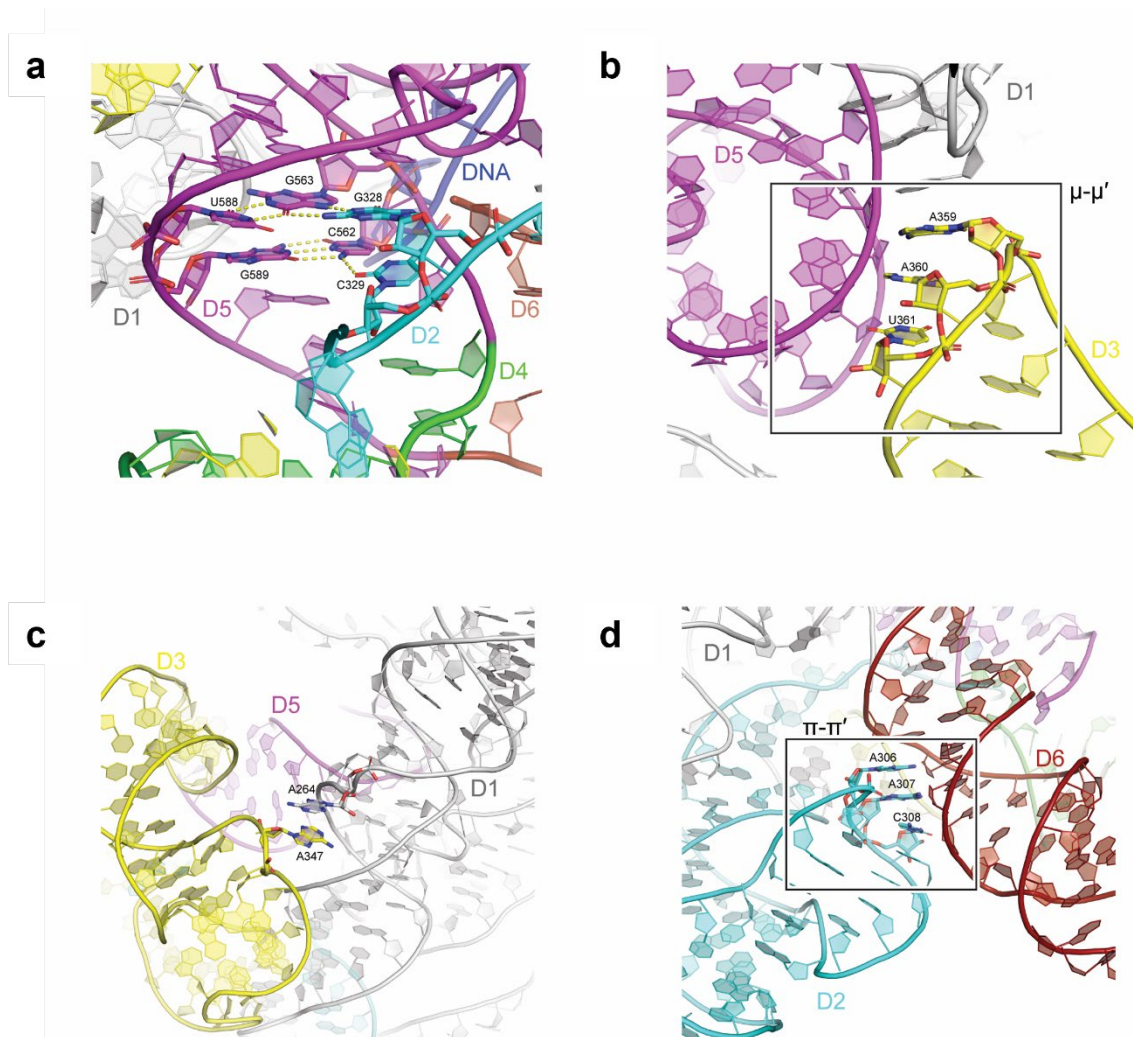


Figure 3.11 Tertiary interactions within the holoRNP.

a. Base triples formed by G328 and C329 of the J2/3 linker with G563-U588 and C562-G589 respectively. b. The μ - μ' tertiary interaction between the pentaloop of D3 (A359, A360, U361) and the minor groove of the D5 helix. c. An A-stacking interaction between A264 of D1 and A347 of D3. d. The π - π' tertiary interaction between D2b (A306, A307, C308) and the basal portion of the D6 helix.

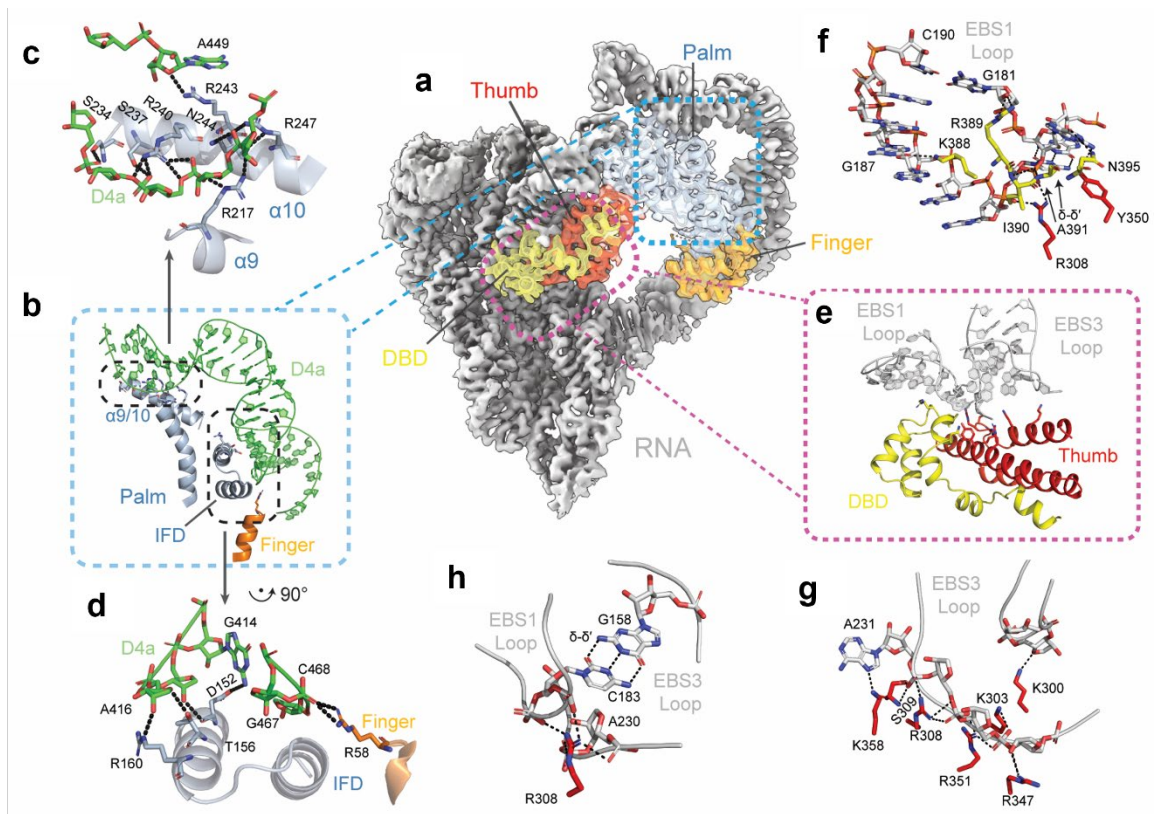


Figure 3.12 Mechanism of maturase facilitated ribozyme catalysis.

a. Protein positioning within the retroelement composite map. b. Protein-D4a contact points. c-d. Interactions that form the RNA-protein anchor points. e. Protein stabilization of EBS1 and EBS3 loops. f-g. Amino acids that rigidify the EBS1 and EBS3 loops. h. R308 joins EBS1 and EBS3 together.

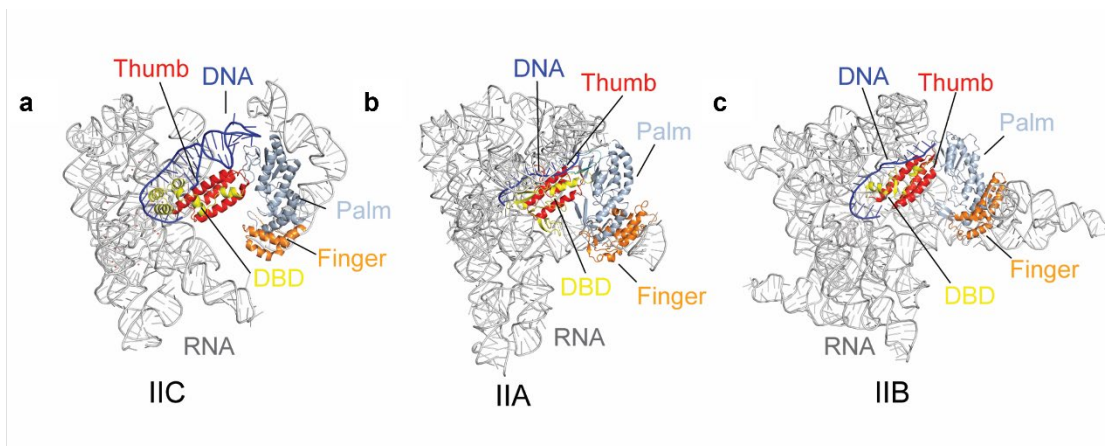


Figure 3.13 Maturase positioning in group II RNPs.

a-c. The maturase is coordinated by D4a in group II RNPs. The fingers, palm, thumb, DBD and DNA/mRNA substrate are indicated in orange, light blue, red, yellow, and dark blue respectively. The maturase thumb exclusively coordinates the DNA substrate and is not observed to interact directly with the EBS sequences in IIA and IIB RNPs. In IIA and IIB intron RNPs a linear DNA substrate is sandwiched between the protein and intron EBS loops, whereas in IIC intron RNPs, the DNA sits along the maturase thumb and DBD, allowing the protein to approach the substrate recognition loops within the ribozyme core. (a). Class IIC intron. (b). Class IIA intron (PDB: 5G2X). (c). Class IIB intron (PDB: 6ME0).

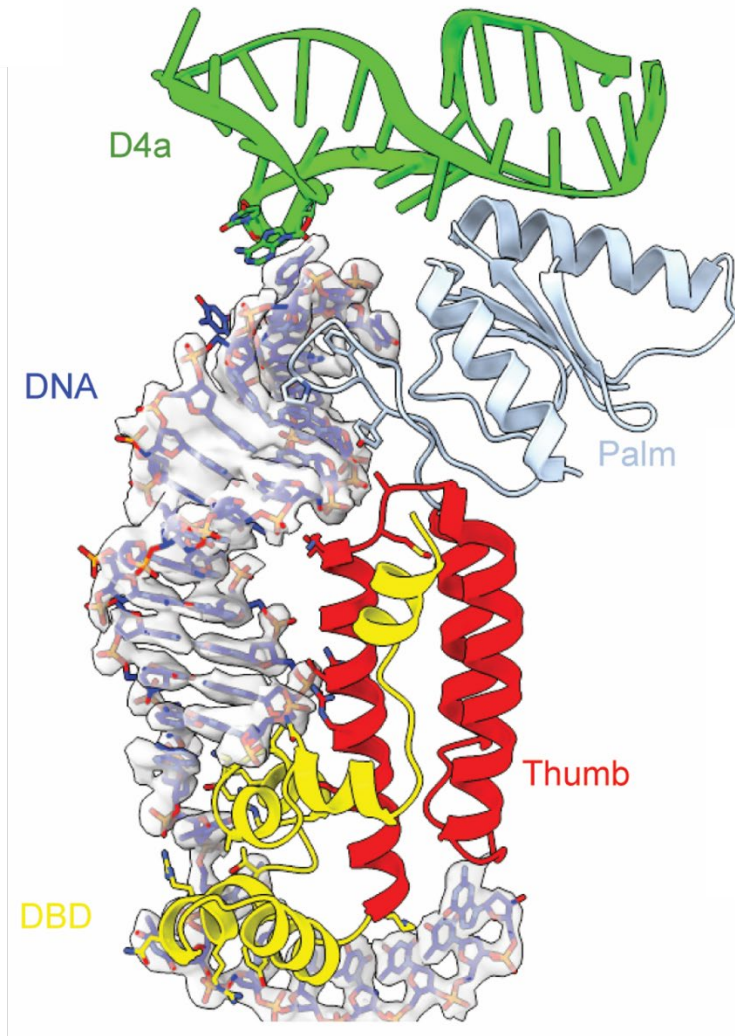


Figure 3.14 Interactions of the structured DNA target with the intron RNP.

The DNA hairpin target wraps around the protein DBD and thumb domains of the protein and fits against the D4a RNA loop. The cryoEM density surrounding the DNA is shown.

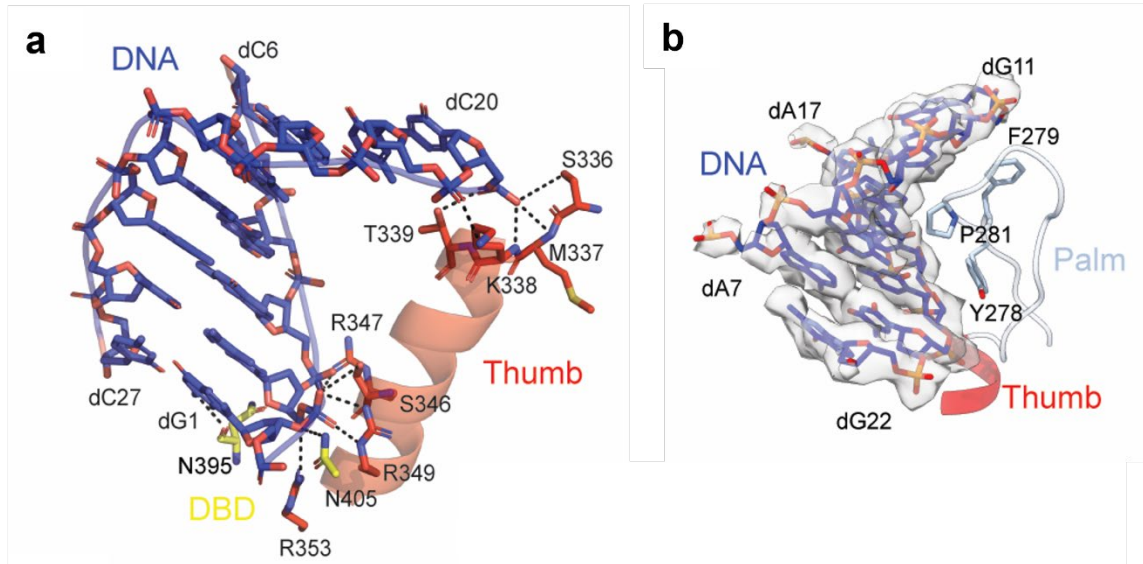


Figure 3.15 Shape specific interactions with the structured DNA target.

- a. Protein contacts with the DNA helical stem. The alpha-helix of the thumb domain contacts the DNA at sites separated by roughly one helical turn. b. Fit of the DNA groove against the protein palm linker.

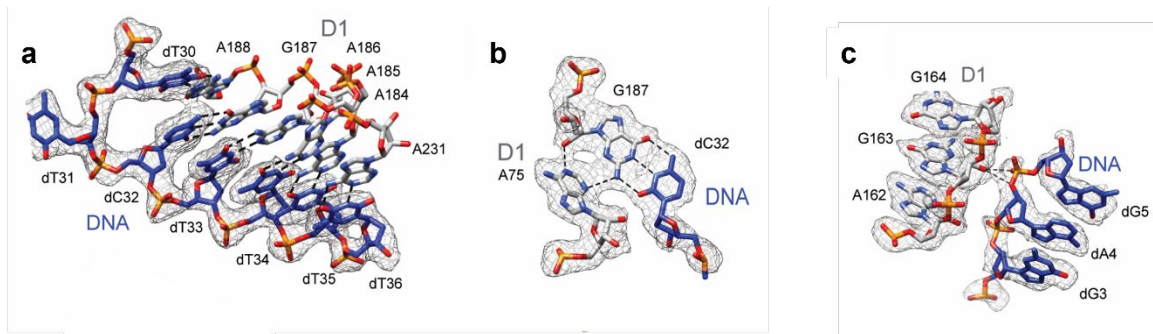


Figure 3.16 Tertiary interactions with the DNA substrate.

a. Sequence specific base pairing interactions between the intron EBS (gray) and the DNA (IBS) nucleotides. b. An A-minor motif interaction that stabilizes the G187-dC32 base pairing. c. A DNA-RNA backbone interaction that helps bind the DNA target to the intron RNP.

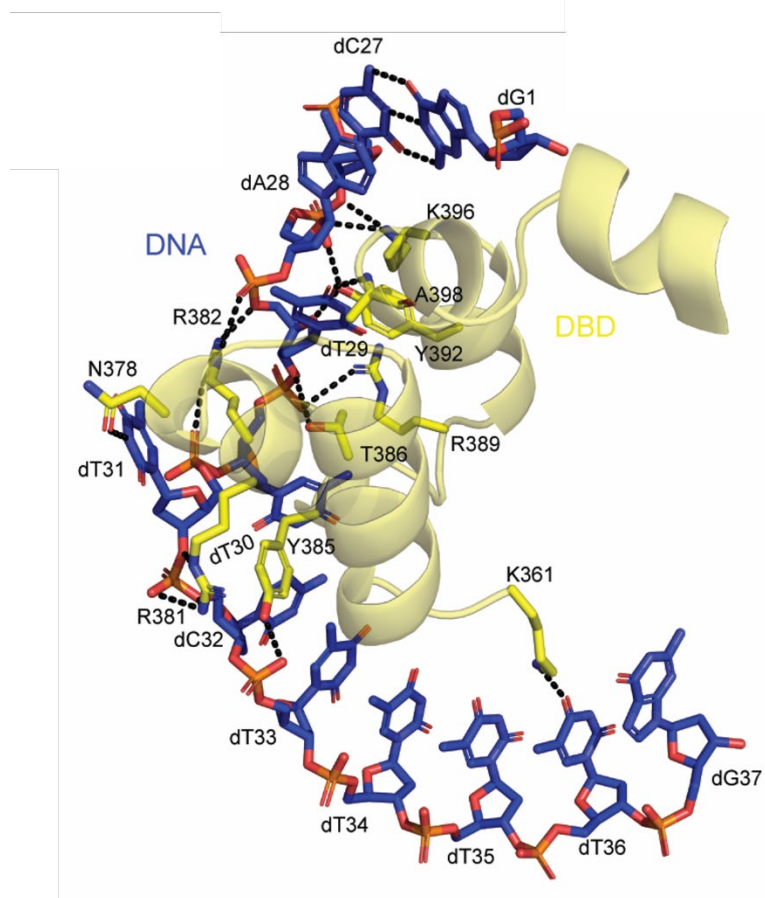


Figure 3.17 Interactions of the maturase DBD with the DNA.

a. The single stranded portion of the DNA curls around the DBD of the maturase and forms a number of interactions with both the backbone and specific bases to hold the DNA and the cleavage site in place for reverse splicing.

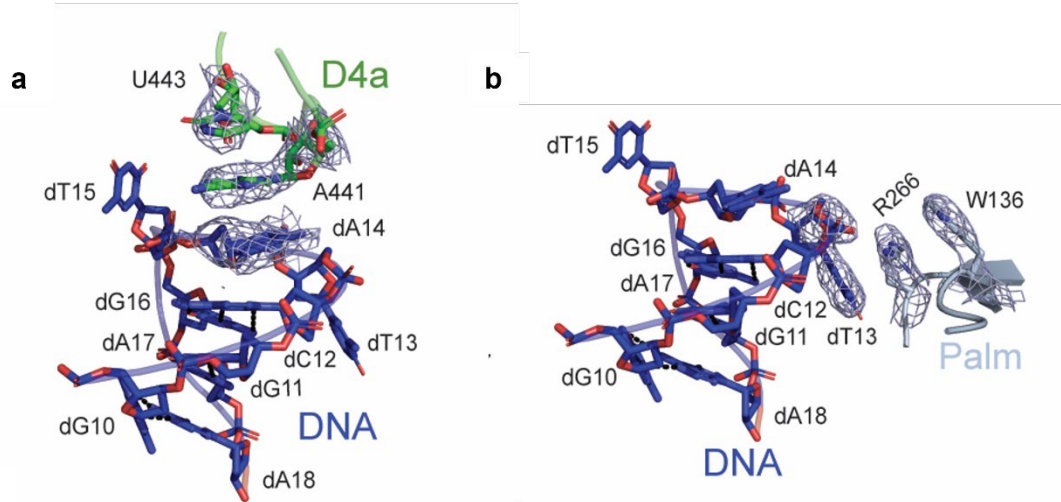


Figure 3.18 Stacking interactions with the intron holoenzyme.

a-b. Intermolecular stacking interactions between DNA and (a) RNA nucleotides, and (b) protein residues at the loop of the DNA hairpin.

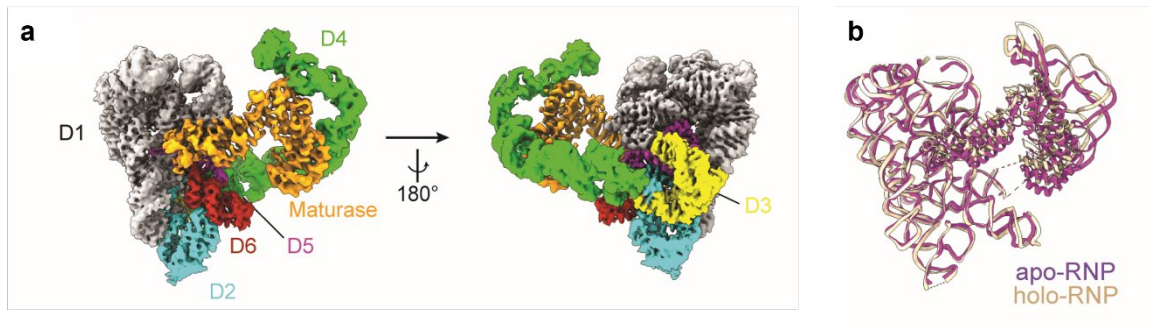


Figure 3.19 CryoEM structures of the apoRNP.

a. CryoEM reconstruction of the apo-retroelement without DNA substrate bound. b. comparison of the RNP backbone traces of the apo- and holo- forms of the group II intron retroelement.

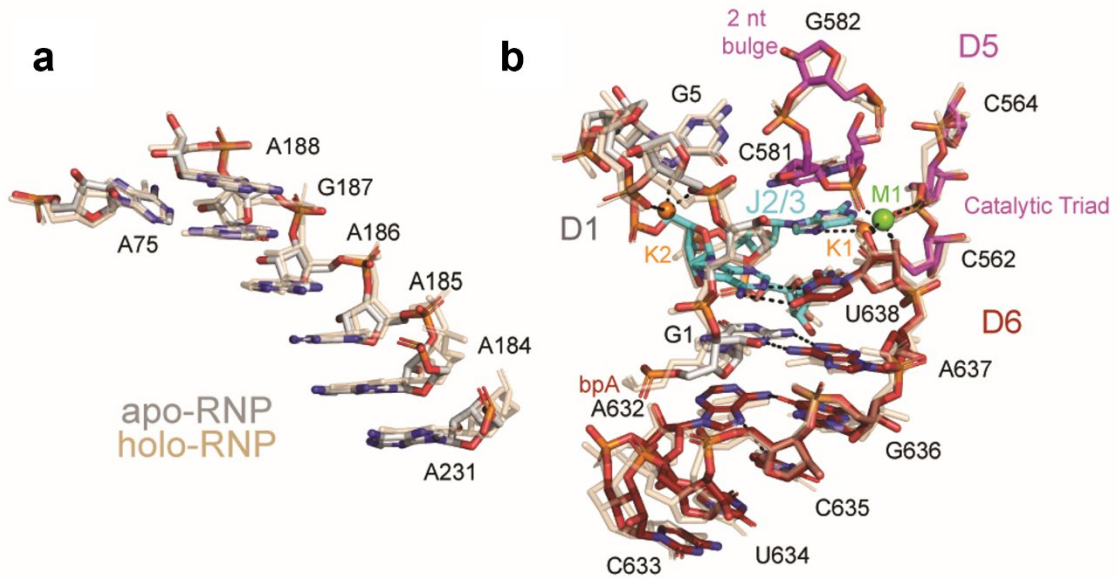


Figure 3.20 An RNP poised to attack.

a-b. Comparison of (a) the EBS nucleotides responsible for target recognition, and (b) the catalytic residues between the apo-RNP and holoRNP. The holoRNP residues are shown in wheat, while the corresponding elements in the apoRNP are coloured.

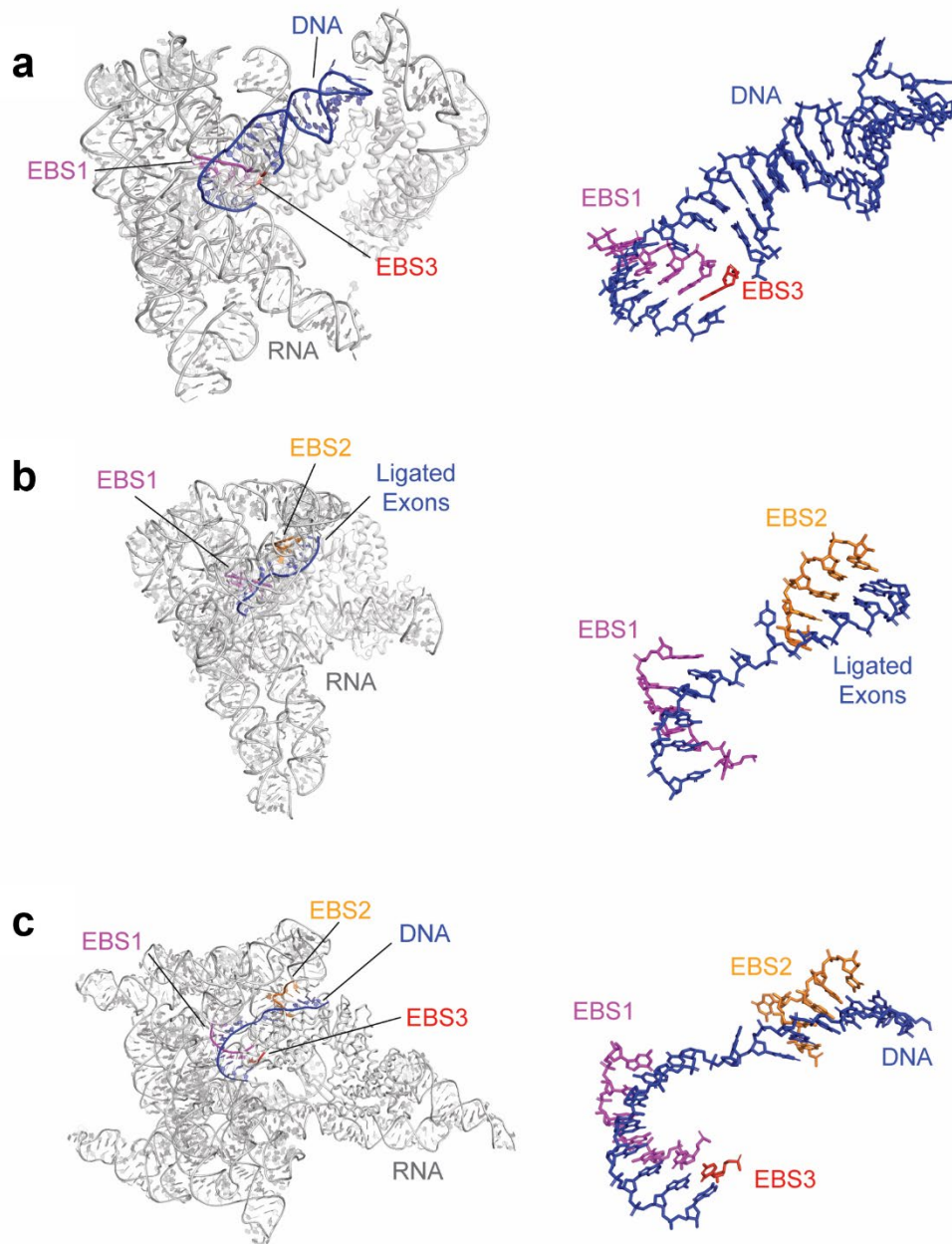


Figure 3.21 Mimicry of the DNA structural motif.

a - c. EBS-IBS interactions of three different classes of group II introns. DNA/mRNA substrate is shown in blue. EBS1 nucleotides are colored magenta. EBS3 nucleotides are colored red. EBS2 interactions are in orange. (a). Class IIC intron. (b). Class IIA intron (PDB: 5G2X). (c). Class IIB Intron (PDB: 6ME0).

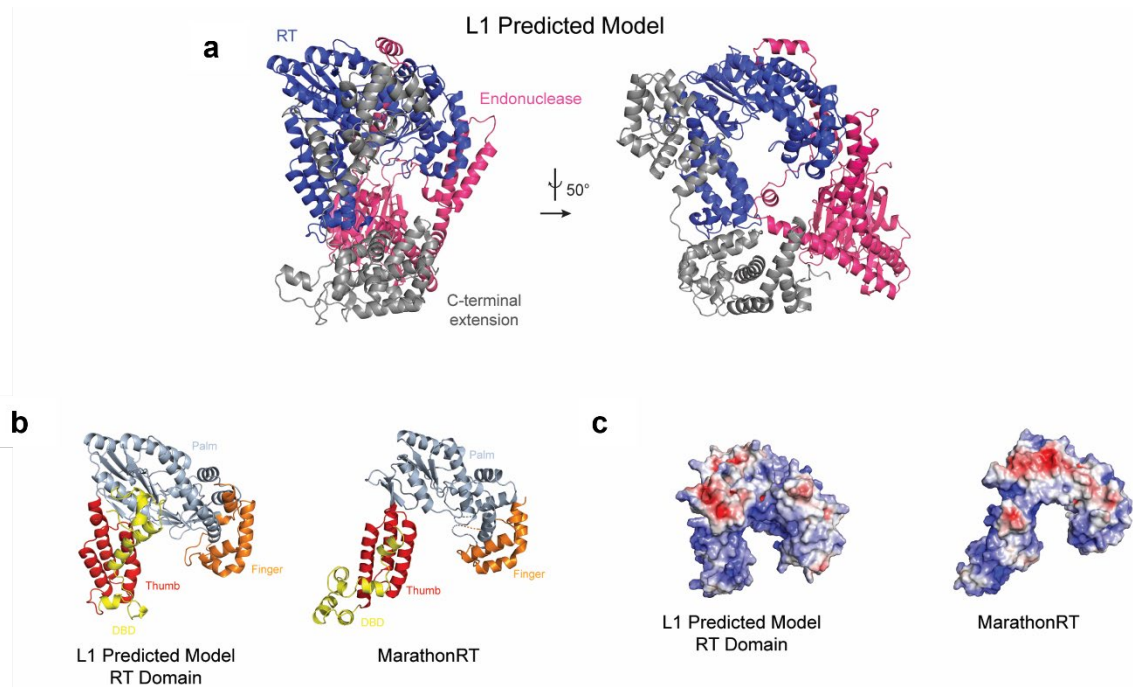


Figure 3.22 Predicted structure of L1 ORF2p.

a. Domain organization of L1 ORF2p. b. Comparison of the RT domain of L1 ORF2p with MarathonRT. c. Surface charges of the RT domain of L1 ORF2p (left) and MarathonRT (right). Blue indicates a positive charge while red indicates a negative charge.

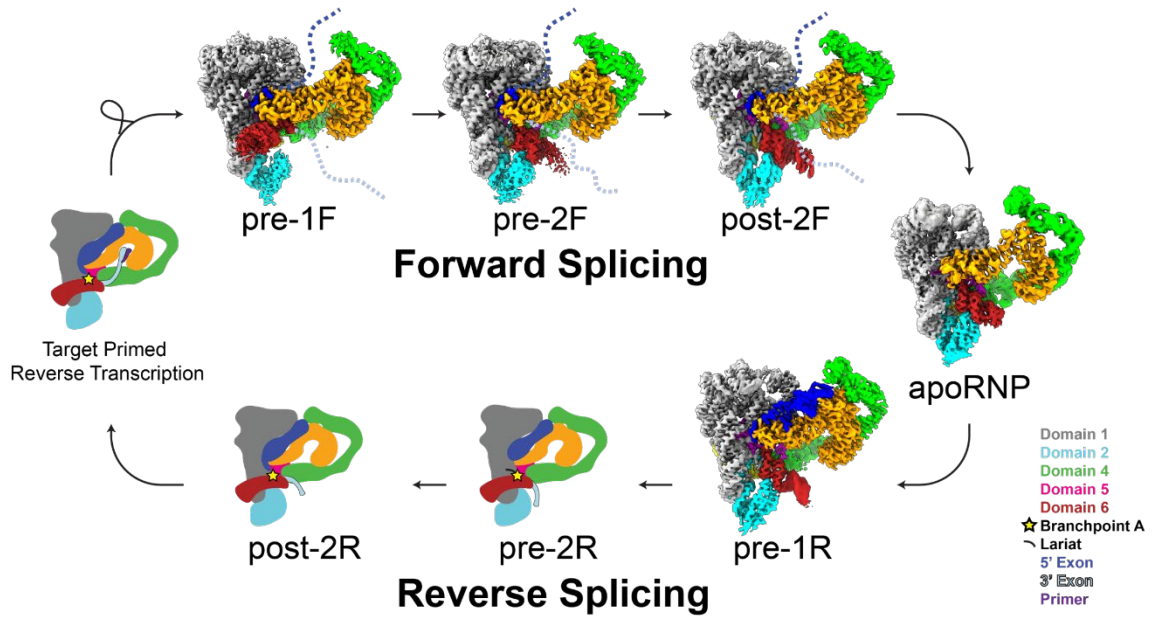


Figure 3.23 Group II intron splicing cycle.

Overview of the group II intron lifecycle including forward splicing, reverse splicing and target primed reverse transcription. The group II intron RNP first excises itself from the flanking exons before attacking its DNA target and integrating, likely using the same conserved motions observed during branching to retrotranspose. TPRT generates an RNA-DNA chimeric hybrid that is resolved by the host, effectively completing the copy-and-paste mechanism.

3.6 Materials and Methods

Protein Purification

The full-length protein (MarathonRT) was purified as previously described with minor changes¹². The recombinant protein was cloned into a pET-SUMO vector (ThermoFisher) and was directly fused to the C-terminus of a 6xHis-SUMO tag. The plasmid was transformed into Rosetta II (DE3) *Escherichia coli* cells (Millipore). The cells were grown at 37°C in LB medium supplemented with 50 µg/mL kanamycin and 17 µg/mL chloramphenicol, first in a small 200 mL culture overnight, before transferring to a large 2L culture, which was grown to an A₆₀₀ of 0.8 to 1.0. Protein expression was induced at 16°C overnight by adding 0.5 mM of isopropyl-β-d-thiogalactopyranoside (IPTG).

Cells were harvested and resuspended in lysis buffer (25 mM NaHEPES, pH 7.5, 1 M NaCl, 10% glycerol and 2 mM β-mercaptoethanol (βME)) containing dissolved protease inhibitor (Sigma). Cells were lysed by passing the resuspension through a microfluidizer and the lysate was centrifuged at 13K rpm to remove precipitates. The supernatant was loaded onto nickel-chelating columns (GE Healthcare) for purification. The column was washed with lysis buffer and then wash buffer (25 mM NaHEPES, pH 7.5, 150 mM NaCl, 10% glycerol, 2 mM βME, and 25 mM imidazole) and eluted with elution buffer (25 mM NaHEPES, pH 7.5, 150 mM NaCl, 10% glycerol, 2 mM βME, and 300 mM imidazole). The eluted protein was then incubated with ULP1 SUMO protease at 4°C for 1 h to cleave the N-terminal 6xHis-SUMO tag.

After tag cleavage, the protein was loaded onto a 5 mL HiTrap HP column (GE Healthcare), equilibrated with buffer A (25 mM K-HEPES, pH 7.5, 150 mM KCl, 10% glycerol and 1 mM DTT). The protein was eluted by running a gradient to buffer B (25

mM K-HEPES, pH 7.5, 2 M KCl, 10% glycerol and 1 mM DTT). The peak fractions were pooled, concentrated to 5 mL and injected onto a Superdex S200 gel filtration column (GE Healthcare) equilibrated with buffer A. After gel filtration, the peak fractions from the S200 column were pooled, concentrated to 50 mg/mL, flash frozen under liquid nitrogen and stored at -80°C .

RNA Transcription and Purification

The DNA sequence of the full-length group II intron was designed to include the T7 RNA polymerase promoter, the RNA coding sequence (with maturase ORF deleted), flanking exons, and a restriction site for BamHI in pBluescript (Invitrogen). The plasmid template was linearized using BamHI (Invitrogen) and then purified by ethanol precipitation. RNA samples were synthesized by *in vitro* transcription with purified T7 RNA polymerase in 1 mL reactions. The reaction mixture contained 10 mM Tris pH 7.5, 1 mM spermidine, 0.01% (v/v) Triton X-100, 10mM DTT, 5 mM NTPs, 15 mM MgCl_2 , 30 μg linearized DNA template, 1 μM T7 RNA polymerase. The reaction was performed at 37°C for 4h. After transcription, the reaction mixture was treated with 0.1 volumes of 0.5 M EDTA and separated on a 5% denaturing polyacrylamide gel after adding denaturing loading dye. The RNA band was eluted with a 10 mM MOPS pH 6.0 buffer at 4°C overnight and then filtered with a 0.2 μm vacuum filter (ThermoFisher). The RNA sample was further purified by adding 0.1V 5M NaCl, 3V absolute alcohol and incubated at -80°C overnight. Then, the transcription products were centrifuged, the supernatant was discarded, and the precipitate was air dried before being dissolved in 10 mM MOPS, pH 6.0.

Lariat-Maturase Complex Formation and Purification

To obtain lariat-maturase RNP complex, a forward splicing reaction was conducted in 0.5 mL reactions containing purified intron RNA, purified intron maturase, 50 mM NH₄-MES pH 6.0, 30 mM MgCl₂, 500 mM (NH₄)₂SO₄, 1 mM DTT, and RNase Inhibitor (ThermoFisher). The purified intron RNA was first mixed with buffer and water and allowed to fold by heating to 95°C for 2 minutes followed by cooling at 25°C for another 2 minutes. Next, MgCl₂, (NH₄)₂SO₄, DTT, and purified maturase were added. The reaction was incubated at 42°C for 1 hour after which it was centrifuged at 13K rpm for 2 minutes to remove the precipitant. The supernatant was loaded onto a HiLoad 10/300 S200 Increase column (GE Healthcare) equilibrated with buffer containing 50 mM NH₄-MES pH 6.0, 30 mM MgCl₂, 200 mM (NH₄)₂SO₄, and 1 mM DTT and separated. Fractions were run on a 5% denaturing gel and on a 5% native gel to check the contents. Those corresponding to the lariat-maturase complex were pooled and used in subsequent assays (see below).

Size Exclusion Chromatography-Multi Angle Light Scattering (SEC-MALS)

SEC-MALS was performed on samples (500 µL) using a Superdex 200 10/300 HR SEC column (GE Healthcare) connected to a High-Performance Liquid Chromatography System (HPLC) Agilent 1200 (Agilent). The elution from SEC was monitored using a DAWN Heleos-II spectrometer (Wyatt Technology) coupled to an Optilab T-rEX (Wyatt Technologies) interferometric refractometer. The SEC-UV/LS/RI system was equilibrated in a buffer containing 50 mM NH₄-MES pH 6.0, 30 mM MgCl₂, 200 mM (NH₄)₂SO₄, and 1 mM DTT. Chemstation software (Agilent) controlled the HPLC operation and data collection from the multi-wavelength UV/VIS detector, while the ASTRA software (Wyatt)

collected data from the refractive index detector, the light scattering detectors, and recorded the UV trace at 295 nm. The molecular weight (MW) was determined across the entire elution profile in intervals of 1 sec from static LS measurement using ASTRA software.

Analytical Ultracentrifugation

Sedimentation velocity analytical ultracentrifugation (SV-AUC) experiments were performed using a Beckman XL-A centrifuge with an An-60 Ti rotor (Beckman Coulter) at the Yale Chemical and Biophysical Instrumentation Center (CBIC) as in previous protocols¹². The lariat and lariat-maturase complex were in buffers containing 50 mM NH₄-MES pH 6.0, 30 mM MgCl₂, 200 mM (NH₄)₂SO₄, and 1 mM DTT. Purified protein was in 25 mM K-HEPES, pH 7.5, 150 mM KCl, 10% glycerol and 1 mM DTT. Sample concentrations were adjusted to an absorption of 0.5 at 260 nm for RNA containing samples and at 280 nm for protein. The samples were allowed to equilibrate at 20°C for 60 min in the instrument before collecting 150 radial scans in duplicate at 50,000 rpm. The entire data collection process took place over 16 hours. Data were analyzed using a continuous c(s) distribution model as implemented in Sedfit. Buffer density and viscosity values used were 1.0381 g/mL and 0.0111031 poise respectively.

Intron-Retroelement Complex Purification

To purify the ternary retroelement complex, we performed a pulldown experiment with tagged DNA. The DNA (IDT) target sequence contains a 5' desthiobiotin tag and has a sequence as follows: GAGAGCAGGGGCTATGAACCTGCTCTCATTCTTTTG. DNA was folded by heating to 95°C for 2 minutes and cooling at 25°C for 2 minutes. 2.5 µg of

annealed DNA was added to a 0.5 mL solution of purified lariat-maturase RNP and allowed to bind at 25°C for 30 minutes. Next, 100 µL of pre-washed Softlink avidin resin (Promega) was added and the mixture was incubated at 25°C for 30 minutes. The mixture was spun down, and the supernatant was discarded. The resin was washed three times with 0.5 mL of buffer (50 mM NH₄-MES pH 6.0, 30 mM MgCl₂, 200 mM (NH₄)₂SO₄, and 1 mM DTT). The retroelement complex was eluted with elution buffer (50 mM NH₄-MES pH 6.0, 30 mM MgCl₂, 200 mM (NH₄)₂SO₄, 1 mM DTT, and 5 mM biotin) by rotating at 25°C for 30 minutes. The eluant containing the assembled lariat-maturase-DNA complex was used for further structural studies.

***In vitro* Reverse Splicing Assays**

Purified lariat-maturase RNP complex was mixed with radiolabeled DNA to perform retrotransposition assays. Briefly, 5' end labelling was done by mixing [γ -³²P] ATP, PNK Buffer (NEB), T4 Kinase (NEB), and DNA. The mixture was incubated at 37°C for 30 minutes and the kinase enzyme was deactivated by heating at 65°C for 30 minutes. DNA was desalted by passing through a G25 column (ThermoFisher) by spinning at 700g for 2 minutes. 5' end labeled primers were then purified by ethanol precipitation, spun down, air dried and resuspended in water. For the intron insertion reaction, radiolabeled primer DNA was folded by heating at 95°C for 2 minutes and cooled at 25°C for 2 minutes. The DNA was mixed with purified lariat-maturase RNP complex in 50 mM K-HEPES pH 7.5, 4 mM MgCl₂, and 50 mM KCl and allowed to react at 42°C. Aliquots were taken from the mixture at various time points and quenched by combining with formamide dye containing xylene

cyanol and bromophenol blue. Samples were then analyzed on a 15% denaturing gel. Gels were imaged using a Typhoon phosphorimager.

Grid Preparation and Data Collection

For cryoEM analysis of the group II intron retroelement, 4 μl of the purified lariat-maturase complex (apo-RNP) or the assembled lariat-maturase-DNA complex (holo-RNP) was loaded onto plasma cleaned QuantiFoil Cu R1.2/1.3 300-mesh grids (Quantifoil) prepped with an extra layer of carbon. Using a condition of 100% humidity and 4°C, the grids were blotted before being plunged into liquid ethane and frozen in liquid nitrogen. SerialEM was used for data collection. Micrographs were recorded in a Titan Krios microscope (FEI) operating at 300 keV, equipped with a K3 Summit direct electron detector (Gatan) operating in counting mode. A tilt-specimen data collection was employed to address the preferred sample orientation issue. For holo-RNP 10,312, 1,321 and 2,851 micrographs were collected at 0°, 10°, and 30° using two different grids, one for the non-tilted data and one for the tilted dataset. For apo-RNP, 3,006, 609, 657, 3,184, and 549 micrographs were collected at 0°, 10°, 20°, 30° and 40° tilt angles respectively. A nominal magnification of 105,000x and a defocus range of -0.5 μm to -2.0 μm and -1.0 μm to -2.5 μm were used for imaging the holo-RNP and apo-RNP respectively, giving an effective pixel size of 0.832Å at the specimen level. For the holo-RNP, each micrograph was dose-fractionated to 40 frames under a dose rate of 17.6 $\text{e}^-/\text{pixel}/\text{s}$, with a total exposure time of 2 s and a frame exposure time of 0.05 s, resulting in a total dose of 50.85 $\text{e}^-/\text{Å}^2$. For the apo-RNP, each micrograph was dose-fractionated to 40 frames under a dose rate of 17.4 e^-

/pixel/s, with a total exposure time of 2 s and a frame exposure time of 0.05 s, resulting in a total dose of 50.27 e⁻/Å².

CryoEM Data Processing

CryoEM data processing workflows are outlined as indicated in Fig. S3A and Fig. S9A. Recorded movie frames were processed using cryoSPARC v3.0^{54,55}. Patch motion correction and CTF estimation were performed using default parameters in cryoSPARC. Exposures were curated and micrographs with ice contamination, excessive motion or damaged regions were removed. For the holo-RNP ternary complex, template picking was used to select particles from the untilted dataset (10,508 micrographs). For the tilted dataset (3,976 micrographs), 2D classification screened, template picked particles from 100 micrographs were used to train a Topaz⁵⁶ model for neural network-based picking from the entire tilted dataset. Particles from each dataset were extracted using a box size of 384 pixels, binned twice to 192 pixels, and were separately filtered through several rounds of 2D classification. Final 2D classes from the untilted dataset were manually separated into two subsets that represented (1) the dominant, front views and (2) the other, lesser represented views which had 3,136,194 particles and 702,685 particles, respectively. To improve particle distribution, particles from the dominant views were randomized into 10 subsets, from which one subset of particles was selected for furthering processing. This randomly selected subset of particles was combined with those from classes representing ‘other’ views and the particles from the tilted dataset, resulting in a total of 1,617,845 good particles. An initial reconstruction was performed using all good particles. Next, alignment free 3D classification, using unbinned, re-extracted full box size particles, was performed

using a loose mask covering the entire holo-RNP complex. Particles were separated into ten different classes from which one class, containing 300,344 particles, with the best, continuous density was selected. Non-uniform refinement⁵⁵ of the selected class was done to obtain the full holo-RNP map at 2.8Å. 3D variability analysis⁵⁷ was performed to identify continuous motions within the holo-RNP ternary complex. Subsequent focused refinement with a specified fulcrum position in the middle of the DNA hairpin was done on the left and right halves of the holo-RNP complex. After CTF refinement, a 2.7Å map for the holo-RNP (left) and 3.0Å map for the holo-RNP (right) were obtained as evaluated using a GSFSC criterion of 0.143.

A similar strategy was used for cryoEM reconstruction of the apo-RNP. 2D classification screened template picked particles from a subset of 100 micrographs were used to train a Topaz⁵⁶ model first for picking from the smaller subset of micrographs before picking from the full dataset of 7,500 micrographs. Picked particles were extracted with a box size of 384 pixels and binned twice to 192 pixels. Particles were filtered by several rounds of 2D classification, resulting in 454,725 good particles. After a consensus refinement with re-extracted, unbinned particles, focused classification without alignment was performed on this subset of particles with a mask on the protein and D4a arm. Two of the eight classes, totaling 136,222 particles, displayed good density for the maturase protein and RNA D4a arm. Particles from these two classes were selected and non-uniform refinement of these particles yielded a 3.6Å map of the full apo-RNP. 3D variability analysis indicated that there were multiple modes of motion in 1) the basal portion of the D4 arm connecting to the ribozyme core, and in 2) the D6 helix and protein/D4a arm.

Subsequently, focused refinement with a mask on the protein and D4a portions of the apo-complex was performed with a fulcrum placed at the C-terminus of the maturase protein. After CTF refinement, this allowed us to obtain a 3.9Å map of the protein and D4a arm of the apo-RNP complex. As the ribozyme core remains static relative to the protein and D4a arm as confirmed by focused classification, focused refinement and CTF refinement were conducted, generating a 3.0Å reconstruction of the ribozyme portion of the apo-RNP complex as evaluated by the GSFSC threshold of 0.143. Directional anisotropy analysis for the three maps of the holo-RNP complex and apo-RNP were performed using 3DFS⁵⁸. Composite maps were manually generated for the holo-RNP and apo-RNP using focused refined maps aligned to the consensus map.

Model Building and Refinement

Model building was initiated by docking an intron crystal structure (PDB: 3EOH) and an AlphaFold generated model of the maturase (MarathonRT) into the overall cryo-EM density map of holo-RNP using UCSF Chimera^{47,59,60}. The holo-RNP model was then manually rebuilt in COOT into the focused refined maps, aligned to the overall holoRNP complex map, to accommodate for sequence and structural changes, and novel segments⁶¹. The DNA in the model was built *de novo* in COOT. For the D4a arm, a stretch of high-resolution density along the backside of the maturase palm domain allowed assignment of RNA sequence based on the cryoEM densities and identification of purine and pyrimidine nucleobases. This allowed the D4a arm to be extended bidirectionally to the distal loop and towards the basal portion that connects to the ribozyme core. The three-way junction from the base of the D4 arm towards the D4a and D4b arm extensions exhibited lower density

and were not modeled. Density for the distal portion of D6 was weak, but this region was modelled in as a helix based on data from lower resolution classifications and 3D variability analysis that show that this domain forms a canonical but dynamic helix. The α - α' kissing loop, while visible in the cryoEM density, is dynamic as seen in 3D variability analysis and was not modeled. The maturase was modeled in save for an alpha-loop and beta-hairpin in the finger domain of the protein. The pentaloop of D2b was not visible in the cryoEM reconstructions. All other portions of the RNA and protein were modeled in. The generated holo-RNP model was docked into the apo-RNP maps, which was manually and computationally refined⁶². The final holo-RNP and apo-RNP model were improved by iterative rounds of real-space refinement against their respective composite cryo-EM map in PHENIX⁶³ using secondary structure restraints for both RNA, protein and DNA, as well as Ramachandran and rotamer restraints for protein chains, and subsequent rebuilding in COOT⁶³⁻⁶⁵. Molecular interfaces and interaction networks were analyzed for each of the models using PDBePISA⁶⁶.

Analysis of DNA Stem-Loop Motif

Group IIC sequences were collected from the Bacterial Group II Intron Database¹¹. Sequences 5' of the DNA insertion site were used for secondary structure prediction by RNAfold⁶⁷. As the DNA stem loop motif is located proximal to the exon junction, only the last 50 nt of the 5' flanking sequences were used. From this trimmed set of sequences, a further 5 nt were removed from the 3' end to account for IBS1 nucleotides, which are known to bind the intron EBS1 sequence and would not be involved in the DNA stem-loop secondary structure. The stem and loop length for each input sequence was recorded, and

the averages were calculated. Helix parameters for the DNA stem including major and minor groove width, and helical twist were calculated using 3DNA⁶⁸.

Protein Conservation

Group IIC maturase protein sequences were collected from the Bacterial Group II Intron Database¹¹. In total 91 sequences of functional and active maturase proteins were selected and used for alignment with ClustalOmega⁶⁹. Alignments were analyzed and visualized using JalView⁷⁰.

3.7 References

- 1 Eickbush, T. H. Mobile introns: retrohoming by complete reverse splicing. *Curr Biol* **9**, R11-14, doi:10.1016/s0960-9822(99)80034-7 (1999).
- 2 Ferat, J. L. & Michel, F. Group II self-splicing introns in bacteria. *Nature* **364**, 358-361, doi:10.1038/364358a0 (1993).
- 3 Bonen, L. & Vogel, J. The ins and outs of group II introns. *Trends Genet* **17**, 322-331, doi:10.1016/s0168-9525(01)02324-1 (2001).
- 4 Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* **3**, a003616, doi:10.1101/cshperspect.a003616 (2011).
- 5 Wank, H., SanFilippo, J., Singh, R. N., Matsuura, M. & Lambowitz, A. M. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell* **4**, 239-250, doi:10.1016/s1097-2765(00)80371-8 (1999).
- 6 Zimmerly, S., Guo, H., Perlman, P. S. & Lambowitz, A. M. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**, 545-554, doi:10.1016/0092-8674(95)90027-6 (1995).
- 7 Pyle, A. M. The tertiary structure of group II introns: implications for biological function and evolution. *Crit Rev Biochem Mol Biol* **45**, 215-232, doi:10.3109/10409231003796523 (2010).
- 8 Pyle, A. M. Group II Intron Self-Splicing. *Annu Rev Biophys* **45**, 183-205, doi:10.1146/annurev-biophys-062215-011149 (2016).
- 9 Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* **12**, 187-215, doi:10.1146/annurev-genom-082509-141802 (2011).
- 10 Kerachian, M. A. & Kerachian, M. Long interspersed nucleotide element-1 (LINE-1) methylation in colorectal cancer. *Clin Chim Acta* **488**, 209-214, doi:10.1016/j.cca.2018.11.018 (2019).
- 11 Dai, L., Toor, N., Olson, R., Keeping, A. & Zimmerly, S. Database for mobile group II introns. *Nucleic Acids Res* **31**, 424-426, doi:10.1093/nar/gkg049 (2003).
- 12 Zhao, C. & Pyle, A. M. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol* **23**, 558-565, doi:10.1038/nsmb.3224 (2016).
- 13 Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183-195, doi:10.1261/rna.063479.117 (2018).
- 14 Chan, R. T., Robart, A. R., Rajashankar, K. R., Pyle, A. M. & Toor, N. Crystal structure of a group II intron in the pre-catalytic state. *Nat Struct Mol Biol* **19**, 555-557, doi:10.1038/nsmb.2270 (2012).
- 15 Kappel, K. *et al.* Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. *Nat Methods* **17**, 699-707, doi:10.1038/s41592-020-0878-9 (2020).
- 16 Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**, 174-180, doi:10.1016/j.ymeth.2009.04.005 (2009).
- 17 Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II intron. *Science* **320**, 77-82, doi:10.1126/science.1153803 (2008).
- 18 Lentzsch, A. M., Stamos, J. L., Yao, J., Russell, R. & Lambowitz, A. M. Structural basis for template switching by a group II intron-encoded non-LTR-retroelement reverse transcriptase. *J Biol Chem* **297**, 100971, doi:10.1016/j.jbc.2021.100971 (2021).
- 19 Dai, L. & Zimmerly, S. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* **30**, 1091-1102, doi:10.1093/nar/30.5.1091 (2002).
- 20 Toor, N., Robart, A. R., Christianson, J. & Zimmerly, S. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem-loop motif of a transcriptional terminator. *Nucleic Acids Res* **34**, 6461-6471, doi:10.1093/nar/gkl820 (2006).
- 21 Zhao, C. & Pyle, A. M. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci* **42**, 470-482, doi:10.1016/j.tibs.2017.03.007 (2017).
- 22 Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* **514**, 193-197, doi:10.1038/nature13790 (2014).
- 23 Costa, M., Walbott, H., Monachello, D., Westhof, E. & Michel, F. Crystal structures of a group II intron lariat primed for reverse splicing. *Science* **354**, doi:10.1126/science.aaf9258 (2016).

- 24 Fica, S. M., Mefford, M. A., Piccirilli, J. A. & Staley, J. P. Evidence for a group II intron-like catalytic triplex in the spliceosome. *Nat Struct Mol Biol* **21**, 464-471, doi:10.1038/nsmb.2815 (2014).
- 25 Marcia, M. & Pyle, A. M. Visualizing group II intron catalysis through the stages of splicing. *Cell* **151**, 497-507, doi:10.1016/j.cell.2012.09.033 (2012).
- 26 Fica, S. M. *et al.* RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**, 229-234, doi:10.1038/nature12734 (2013).
- 27 Wilkinson, M. E., Fica, S. M., Galej, W. P. & Nagai, K. Structural basis for conformational equilibrium of the catalytic spliceosome. *Mol Cell* **81**, 1439-1452 e1439, doi:10.1016/j.molcel.2021.02.021 (2021).
- 28 Haack, D. B. *et al.* Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA. *Cell* **178**, 612-623 e612, doi:10.1016/j.cell.2019.06.035 (2019).
- 29 Qu, G. *et al.* Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* **23**, 549-557, doi:10.1038/nsmb.3220 (2016).
- 30 Fedorova, O. & Pyle, A. M. Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *EMBO J* **24**, 3906-3916, doi:10.1038/sj.emboj.7600852 (2005).
- 31 Piccirilli, J. A. & Staley, J. P. Reverse transcriptases lend a hand in splicing catalysis. *Nat Struct Mol Biol* **23**, 507-509, doi:10.1038/nsmb.3242 (2016).
- 32 Cousineau, B., Lawrence, S., Smith, D. & Belfort, M. Retrotransposition of a bacterial group II intron. *Nature* **404**, 1018-1021, doi:10.1038/35010029 (2000).
- 33 Toor, N., Hausner, G. & Zimmerly, S. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**, 1142-1152, doi:10.1017/s1355838201010251 (2001).
- 34 Singh, R. N., Saldanha, R. J., D'Souza, L. M. & Lambowitz, A. M. Binding of a group II intron-encoded reverse transcriptase/maturase to its high affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. *J Mol Biol* **318**, 287-303, doi:10.1016/S0022-2836(02)00054-2 (2002).
- 35 Zhao, C. & Pyle, A. M. The group II intron maturase: a reverse transcriptase and splicing factor go hand in hand. *Curr Opin Struct Biol* **47**, 30-39, doi:10.1016/j.sbi.2017.05.002 (2017).
- 36 Matsuura, M., Noah, J. W. & Lambowitz, A. M. Mechanism of maturase-promoted group II intron splicing. *EMBO J* **20**, 7259-7270, doi:10.1093/emboj/20.24.7259 (2001).
- 37 Michel, F., Umeson, K. & Ozeki, H. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**, 5-30, doi:10.1016/0378-1119(89)90026-7 (1989).
- 38 Lescoute, A. & Westhof, E. The A-minor motifs in the decoding recognition process. *Biochimie* **88**, 993-999, doi:10.1016/j.biochi.2006.05.018 (2006).
- 39 Flocco, M. M. & Mowbray, S. L. Planar stacking interactions of arginine and aromatic side-chains in proteins. *J Mol Biol* **235**, 709-717, doi:10.1006/jmbi.1994.1022 (1994).
- 40 Liu, N. *et al.* Exon and protein positioning in a pre-catalytic group II intron RNP primed for splicing. *Nucleic Acids Res* **48**, 11185-11198, doi:10.1093/nar/gkaa773 (2020).
- 41 Matsuura, M. *et al.* A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev* **11**, 2910-2924, doi:10.1101/gad.11.21.2910 (1997).
- 42 Fica, S. M., Oubridge, C., Wilkinson, M. E., Newman, A. J. & Nagai, K. A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science* **363**, 710-714, doi:10.1126/science.aaw5569 (2019).
- 43 Neidle, S. Beyond the double helix: DNA structural diversity and the PDB. *J Biol Chem* **296**, 100553, doi:10.1016/j.jbc.2021.100553 (2021).
- 44 Stamos, J. L., Lentzsch, A. M. & Lambowitz, A. M. Structure of a Thermostable Group II Intron Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary Implications. *Mol Cell* **68**, 926-939 e924, doi:10.1016/j.molcel.2017.10.024 (2017).
- 45 Lambowitz, A. M. & Zimmerly, S. Mobile group II introns. *Annu Rev Genet* **38**, 1-35, doi:10.1146/annurev.genet.38.072902.091600 (2004).
- 46 Guo, L. T. *et al.* Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J Mol Biol* **432**, 3338-3352, doi:10.1016/j.jmb.2020.03.022 (2020).

- 47 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).
- 48 Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**, 3353-3362 (1990).
- 49 Chung, K. *et al.* Structures of a mobile intron retroelement poised to attack its structured DNA target. *Science* **378**, 627-634, doi:10.1126/science.abq2844 (2022).
- 50 Xu, L., Liu, T., Chung, K. & Pyle, A. M. Structural insights into intron catalysis and dynamics during splicing. *Nature*, doi:10.1038/s41586-023-06746-6 (2023).
- 51 Zhao, C., Rajashankar, K. R., Marcia, M. & Pyle, A. M. Crystal structure of group II intron domain I reveals a template for RNA assembly. *Nat Chem Biol* **11**, 967-972, doi:10.1038/nchembio.1949 (2015).
- 52 Qin, P. Z. & Pyle, A. M. Stopped-flow fluorescence spectroscopy of a group II intron ribozyme reveals that domain I is an independent folding unit with a requirement for specific Mg²⁺ ions in the tertiary structure. *Biochemistry* **36**, 4718-4730, doi:10.1021/bi962665c (1997).
- 53 Guo, L. T., Olson, S., Patel, S., Graveley, B. R. & Pyle, A. M. Direct tracking of reverse-transcriptase speed and template sensitivity: implications for sequencing and analysis of long RNA molecules. *Nucleic Acids Res* **50**, 6980-6989, doi:10.1093/nar/gkac518 (2022).
- 54 Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296, doi:10.1038/nmeth.4169 (2017).
- 55 Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat Methods* **17**, 1214-1221, doi:10.1038/s41592-020-00990-8 (2020).
- 56 Bepler, T. *et al.* Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Res Comput Mol Biol* **10812**, 245-247 (2018).
- 57 Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J Struct Biol* **213**, 107702, doi:10.1016/j.jsb.2021.107702 (2021).
- 58 Tan, Y. Z. *et al.* Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat Methods* **14**, 793-796, doi:10.1038/nmeth.4347 (2017).
- 59 Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J Struct Biol* **157**, 281-287, doi:10.1016/j.jsb.2006.06.010 (2007).
- 60 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).
- 61 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501, doi:10.1107/S0907444910007493 (2010).
- 62 Kidmose, R. T. *et al.* Namdinator - automatic molecular dynamics flexible fitting of structural models into cryo-EM and crystallography experimental maps. *IUCrJ* **6**, 526-531, doi:10.1107/S2052252519007619 (2019).
- 63 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221, doi:10.1107/S0907444909052925 (2010).
- 64 Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr D Struct Biol* **74**, 531-544, doi:10.1107/S2059798318006551 (2018).
- 65 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877, doi:10.1107/S2059798319011471 (2019).
- 66 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-797, doi:10.1016/j.jmb.2007.05.022 (2007).
- 67 Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res* **36**, W70-74, doi:10.1093/nar/gkn188 (2008).
- 68 Lu, X. J. & Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* **3**, 1213-1227, doi:10.1038/nprot.2008.104 (2008).
- 69 Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* **1079**, 105-116, doi:10.1007/978-1-62703-646-7_6 (2014).

- 70 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

Chapter 4: Insights into the mechanism of group II intron splicing

Preface

This thesis has explored the group II intron splicing system and the coordination of the intron RNA components with their maturase cofactors. Chapter 1 provided an overview of the foundational work within the group II intron splicing field and provided context for the questions that I sought to answer throughout my thesis. Chapter 2 investigated the lariat forming splicing pathway and provided structural and mechanistic insights into branching. Chapter 3 focused on the role of group II introns as retroelements and highlighted how RNPs use unprecedented strategies to recognize the shape and sequence of their targets. Here, I will briefly summarize where the field stands considering these new studies and what lingering questions remain for group II introns.

4.1 Group II intron forward splicing

Advancements in cryoEM have paved the way for study of group II introns in their natural context, as an RNP bound to their maturase cofactors. We now have a thorough understanding of the intron holoenzyme as it proceeds through the stages of splicing to excise as a lariat¹. We not only know that the protein binds with high affinity to its parent intron, but we also know that the maturase plays a critical role in positioning the branch helix². The maturase uses its basic thumb and DBD to lock the intron RNP in a branching competent conformation³. Within D6, the branch site is selected for through base triple interactions, dictating the adenosine base identity. The transition from branching to exon ligation is coupled by relaxation of the RNA backbone at the 5' splice site after the first cleavage reaction⁴. The release of the backbone distortion allows the branch helix to

disengage and swivel downwards to remove the lariat from the active site and replace it with the 3' splice site. The intron holoenzyme completes the reaction by ligating the exons together before releasing them from its active site.

The biochemical and structural work done to investigate the mechanism of forward splicing have largely accounted for the remaining questions within the field⁴. Details regarding the structure, splice site exchange, active site arrangement, and maturase function have all been revealed. Admittedly, there is no post-branching structure of the intron RNP for the group IIC class, which would unequivocally complete the splicing gallery. However, the insight from the current available data suggests that the heteronuclear metal ion core is preserved throughout the splicing cycle, and the exon recognition sequences remain engaged, pointing to minimal change when compared with the pre-ligation state. The majority of this work has been conducted with the IIC intron class, the most primitive and ancient of the three predominant classes. The mechanistic details and structural features arising from this representative class likely applies to the higher order introns, although it should be noted that there may be some additional interactions present in their more complex architectures. One outstanding question is how particular introns can still branch efficiently despite the absence of their maturase cofactor, which may be compensated for by those additional tertiary interactions, although more work would have to be done to investigate this phenomenon.

A noteworthy aspect of the work done on forward splicing is the remarkable structural and compositional similarities to the modern spliceosome. Numerous elements have been preserved throughout evolutionary time from group II introns to the spliceosome. The catalytic triplex⁵, metal ion core⁶, branchpoint adenosine^{7,8}, splice site recognition⁷, and

now the mechanism of splice site exchange remain hard-coded strategies for pre-mRNA splicing. The observation that these splicing mechanisms have withstood evolutionary time speaks to how robust they are and their importance for RNA metabolism within the central dogma.

4.2 Group II intron reverse splicing

Group II introns do not only act as splicing ribozymes but are also important agents of genetic change. Introns spread to novel genomic loci by recognizing specific target sites and executing the reversal of splicing. Recent work has identified the DNA target recognition strategies^{9,10}, and revealed the conformational changes that parallel movements observed during intron excision. Together, these structures paint a picture of group II introns as potent retroelements primed to attack. Many of the same mechanistic features, including the heteronuclear metal ion core, nucleophilic attack, maturase coordination and target recognition strategies remain the same from forward to reverse splicing. To see if this is true for the final steps of reverse splicing, more work would have to be done to capture the pre-2R and post-2R steps for various intron classes to allow for comparison with current work.

The conserved, cyclical motion of the intron RNP provides a robust reaction mechanism that underlies the invasive nature of group II introns. While the rest of the RNP provides a sturdy scaffold that embraces the active site, the branch helix is the only moving component and its mechanical movements drive the transesterification reactions, which also explains the reversibility of splicing reactions. The intricate architecture dictates the positioning of the branch helix which aligns with the ability of the RNP to cycle through the steps of

forward and reverse splicing, analogous to the four-stroke cycle of an internal combustion engine. To continue the analogy, the modern spliceosome can be compared to modern engines, replete with auxiliary parts and additional technology, yet, built upon the minimalistic, but functional unit of the group II intron.

IIC introns have a peculiar hairpin motif located at their DNA target sites. The structures of the DNA bound intron holoenzyme demonstrated the significance of the short stem-loop structure, which plays an important structural role that complements the carefully crafted groove within the RNP³. Beyond its function in shape specific recognition, it is also intriguing to think of the biological role of these hairpin motifs¹¹⁻¹³. Stem loop structures are often located at the end of genes where they have a role in transcription termination. Targeting of these termination inducing structures could be one way to direct group II intron retrotransposons to specific sites that minimize damage to the host. Insertion at these sites would not disrupt coding regions and or compromise host fitness and allow for less stringent target site identification compared to IIA and IIB introns. Additionally, integration at these sites would ensure lower expression levels of group II intron retroelements, limiting their spread. The transcription terminator hairpin motif thus may have been one solution to silence group II introns¹¹.

4.3 Target primed reverse transcription

The next step along the splicing cycle is target primed reverse transcription, the process of copying the inserted intron into cDNA. Structural work has indicated that there is ample room to accommodate this reaction and suggests that it can occur simultaneously with reverse splicing^{3,14}. The 3' end of the intron points towards the cavity of the RT active site,

which can readily fit a primer-RNA duplex to begin cDNA synthesis. The coevolution of maturase and intron is significant as processive and accurate copying is essential to maintaining the sequence and tertiary fold of the intron RNA. This would explain why group II intron maturases make excellent reverse transcriptase enzymes. After intron insertion, the maturase would hypothetically extend a primer and unravel the entire group II intron RNP system. The extra insertion domains in the palm and fingers confer increased ability to disrupt structured regions such as hairpins and double stranded stretches¹⁵⁻¹⁸, which are found all throughout group II introns. All of this is possible while maintaining a largely basic exterior surface that can coordinate nucleic acids, which may play a role in substrate recognition in other retrotransposition systems^{19,20}. In light of TPRT and downstream processes, we can understand the interplay between intron and maturase that ensures seamless copying and pasting. To fully grasp the mechanistic role of TPRT, more effort would be required to examine whether additional conformational changes are necessary, whether the intron RNP regulates TPRT temporally relative to retrotransposition, and how the primer would be generated.

TPRT is not exclusive to group II introns and occurs in other non-LTR retroelements. Recent mechanistic studies into the mobility of the R2 retrotransposon has provided some perspective on the evolution of retroelements from RNA-dominant to protein centric machines^{19,20}. The proteins within those systems use the exterior surface along the backside of the thumb, to grasp onto their substrates, even recognizing structured regions. Those contact sites are important for specifying binding sequences and cleavage sequences, which is in stark contrast to the group II intron where such roles are exclusively for the EBS nucleotides. Comparison with these retroelements provides a window into understanding

another evolutionary route where many of the RNA functions in an intron RNP have gradually been taken on by increasingly complex proteins.

4.4 Future Outlook

From this work, I hope that one can gain an appreciation for group II introns as a model system for studying splicing, RNA folding, and RNA structure. They are not just simple splicing machines, but they have important implications for many of the eukaryotic genetic elements that persist today from the modern spliceosome to non-LTR retroelements. In some shape or form, introns, or their descendants, continue to impact eukaryotic genomes.

Increased structural understanding of the assembly and architecture of introns by cryoEM may perhaps pave the way for this approach as a proven technique. We are able to achieve high resolution for large, structured RNAs, and to gain mechanistic insights into ribozyme cores, challenging the prevailing narrative that RNAs are floppy, unstructured or difficult to work with. While group II intron RNPs still have their maturase cofactors present, increasingly we are shifting towards investigating RNAs in the absence of proteins²¹⁻²⁹. The range of RNA targets that scientists are pursuing is increasing, opening the possibilities to learn about challenging scientific questions involving RNA metabolism, processing, and regulation. Combined with new workflows, hybrid approaches and computational techniques²⁹, we are certainly in an exciting new era of RNA structural biology.

4.5 References

- 1 Liu, N. *et al.* Exon and protein positioning in a pre-catalytic group II intron RNP primed for splicing. *Nucleic Acids Res* **48**, 11185-11198, doi:10.1093/nar/gkaa773 (2020).
- 2 Xu, L., Liu, T., Chung, K. & Pyle, A. M. Structural insights into intron catalysis and dynamics during splicing. *Nature*, doi:10.1038/s41586-023-06746-6 (2023).
- 3 Chung, K. *et al.* Structures of a mobile intron retroelement poised to attack its structured DNA target. *Science* **378**, 627-634, doi:10.1126/science.abq2844 (2022).
- 4 Zhao, C. & Pyle, A. M. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci* **42**, 470-482, doi:10.1016/j.tibs.2017.03.007 (2017).
- 5 Fica, S. M., Mefford, M. A., Piccirilli, J. A. & Staley, J. P. Evidence for a group II intron-like catalytic triplex in the spliceosome. *Nat Struct Mol Biol* **21**, 464-471, doi:10.1038/nsmb.2815 (2014).
- 6 Fica, S. M. *et al.* RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**, 229-234, doi:10.1038/nature12734 (2013).
- 7 Liu, Q. *et al.* Branch-site selection in a group II intron mediated by active recognition of the adenine amino group and steric exclusion of non-adenine functionalities. *J Mol Biol* **267**, 163-171, doi:10.1006/jmbi.1996.0845 (1997).
- 8 Galej, W. P. *et al.* Cryo-EM structure of the spliceosome immediately after branching. *Nature* **537**, 197-201, doi:10.1038/nature19316 (2016).
- 9 Haack, D. B. *et al.* Cryo-EM Structures of a Group II Intron Reverse Splicing into DNA. *Cell* **178**, 612-623 e612, doi:10.1016/j.cell.2019.06.035 (2019).
- 10 Qu, G. *et al.* Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* **23**, 549-557, doi:10.1038/nsmb.3220 (2016).
- 11 Haack, D. B. & Toor, N. Recognition of transcription terminators during retrotransposition: How to keep a group II intron quiet. *Mol Cell* **83**, 332-334, doi:10.1016/j.molcel.2022.12.027 (2023).
- 12 Robart, A. R., Seo, W. & Zimmerly, S. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc Natl Acad Sci U S A* **104**, 6620-6625, doi:10.1073/pnas.0700561104 (2007).
- 13 Toor, N., Robart, A. R., Christianson, J. & Zimmerly, S. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem-loop motif of a transcriptional terminator. *Nucleic Acids Res* **34**, 6461-6471, doi:10.1093/nar/gkl820 (2006).
- 14 Stamos, J. L., Lentzsch, A. M. & Lambowitz, A. M. Structure of a Thermostable Group II Intron Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary Implications. *Mol Cell* **68**, 926-939 e924, doi:10.1016/j.molcel.2017.10.024 (2017).
- 15 Guo, L. T., Olson, S., Patel, S., Graveley, B. R. & Pyle, A. M. Direct tracking of reverse-transcriptase speed and template sensitivity: implications for sequencing and analysis of long RNA molecules. *Nucleic Acids Res* **50**, 6980-6989, doi:10.1093/nar/gkac518 (2022).
- 16 Guo, L. T. *et al.* Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J Mol Biol* **432**, 3338-3352, doi:10.1016/j.jmb.2020.03.022 (2020).
- 17 Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183-195, doi:10.1261/rna.063479.117 (2018).
- 18 Zhao, C. & Pyle, A. M. The group II intron maturase: a reverse transcriptase and splicing factor go hand in hand. *Curr Opin Struct Biol* **47**, 30-39, doi:10.1016/j.sbi.2017.05.002 (2017).
- 19 Deng, P. *et al.* Structural RNA components supervise the sequential DNA cleavage in R2 retrotransposon. *Cell* **186**, 2865-2879 e2820, doi:10.1016/j.cell.2023.05.032 (2023).
- 20 Wilkinson, M. E., Frangieh, C. J., Macrae, R. K. & Zhang, F. Structure of the R2 non-LTR retrotransposon initiating target-primed reverse transcription. *Science* **380**, 301-308, doi:10.1126/science.adg7883 (2023).
- 21 Su, Z. *et al.* Cryo-EM structures of full-length Tetrahymena ribozyme at 3.1 Å resolution. *Nature* **596**, 603-607, doi:10.1038/s41586-021-03803-w (2021).
- 22 Li, S., Palo, M. Z., Zhang, X., Pintilie, G. & Zhang, K. Snapshots of the second-step self-splicing of Tetrahymena ribozyme revealed by cryo-EM. *Nat Commun* **14**, 1294, doi:10.1038/s41467-023-36724-5 (2023).

- 23 Zhang, X., Li, S., Pintilie, G., Palo, M. Z. & Zhang, K. Snapshots of the first-step self-splicing of Tetrahymena ribozyme revealed by cryo-EM. *Nucleic Acids Res* **51**, 1317-1325, doi:10.1093/nar/gkac1268 (2023).
- 24 Li, S. *et al.* Topological crossing in the misfolded Tetrahymena ribozyme resolved by cryo-EM. *Proc Natl Acad Sci U S A* **119**, e2209146119, doi:10.1073/pnas.2209146119 (2022).
- 25 Kappel, K. *et al.* Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. *Nat Methods* **17**, 699-707, doi:10.1038/s41592-020-0878-9 (2020).
- 26 Bonilla, S. L. & Kieft, J. S. The promise of cryo-EM to explore RNA structural dynamics. *J Mol Biol* **434**, 167802, doi:10.1016/j.jmb.2022.167802 (2022).
- 27 Bonilla, S. L., Vicens, Q. & Kieft, J. S. Cryo-EM reveals an entangled kinetic trap in the folding of a catalytic RNA. *Sci Adv* **8**, eabq4144, doi:10.1126/sciadv.abq4144 (2022).
- 28 Bonilla, S. L., Sherlock, M. E., MacFadden, A. & Kieft, J. S. A viral RNA hijacks host machinery using dynamic conformational changes of a tRNA-like structure. *Science* **374**, 955-960, doi:10.1126/science.abe8526 (2021).
- 29 Ma, H., Jia, X., Zhang, K. & Su, Z. Cryo-EM advances in RNA structure determination. *Signal Transduct Target Ther* **7**, 58, doi:10.1038/s41392-022-00916-0 (2022).

Appendix I: Structures of a mobile retroelement poised to attack its structured DNA target.

Required copyright lines for reprint:

Science **378**, 627-634 (2022) | DOI: 10.1126/science.abq2844

Copyright © 2022 American Association for the Advancement of Science. All rights reserved.

STRUCTURAL BIOLOGY

Structures of a mobile intron retroelement poised to attack its structured DNA target

Kevin Chung¹†, Ling Xu^{2,3}†, Pengxin Cha¹, Junhui Peng⁴, Swapnil C. Devarkar¹, Anna Marie Pyle^{2,3*}

Group II introns are ribozymes that catalyze their self-excision and function as retroelements that invade DNA. As retrotransposons, group II introns form ribonucleoprotein (RNP) complexes that roam the genome, integrating by reversal of forward splicing. Here we show that retrotransposition is achieved by a tertiary complex between a structurally elaborate ribozyme, its protein mobility factor, and a structured DNA substrate. We solved cryo-electron microscopy structures of an intact group IIC intron-maturase retroelement that was poised for integration into a DNA stem-loop motif. By visualizing the RNP before and after DNA targeting, we show that it is primed for attack and fits perfectly with its DNA target. This study reveals design principles of a prototypical retroelement and reinforces the hypothesis that group II introns are ancient elements of genetic diversification.

Group II introns are self-splicing retroelements that have played a key role in shaping eukaryotic genomes as the ancestors of spliceosomal introns and non-long terminal repeat (LTR) retroelements (1). They remain important for gene expression in plants, fungi, yeasts, and many bacteria (2, 3). Group II introns encode a specialized reverse transcriptase (maturase) that binds its parent intron and facilitates self-splicing, which releases a well-folded lariat ribonucleoprotein (RNP) complex (4). The liberated RNP functions as a retrotransposon, targeting DNA that contains spliced exon junction sequences and inserting by means of a two-step transesterification reaction known as reverse splicing (Fig. 1A) (5). The resulting DNA-RNA chimera is copied into cDNA by the reverse transcriptase (RT) activity of the multifunctional maturase in a process known as target primed reverse transcription (TPRT) (6). Host repair pathways complete the downstream DNA copy-and-paste steps that are needed to achieve total intron integration (4).

There are three main classes of group II introns, IIA, IIB, and IIC, which share a conserved secondary structure and a similar tertiary organization around a ribozyme active site (7). Group IIC introns are an ancient class of bacterial introns that recognize both the sequence and three-dimensional (3D) structure of their DNA insertion sites (8). Unlike their larger IIA and IIB counterparts, group IIC introns are almost completely dependent on their maturases to facilitate intron excision through lariat formation, thereby forming the functional RNP that serves as the minimal

element for retrotransposition (8). Compared with their more evolved counterparts, group IIC RTs lack an endonuclease domain for generating TPRT primers and instead exploit the lagging strands at DNA replication forks (4).

Recent structural and biochemical studies of IIA and IIB introns have provided important insights into strategies for the maturase recognition of intron RNA (9–11). However, available RNP structures have not revealed a specific mechanistic role for the maturase during RNP assembly, DNA recognition, or chemical catalysis. At present, the mechanism by which group IIC introns recognize DNA structures and not just DNA sequences remains unclear. Furthermore, there are no available structures of the free RNP retroelement before it has bound DNA. These open questions preclude a clear understanding of group II intron retrotransposition and its evolutionary role in shaping modern genomes. To address these problems, we solved cryo-electron microscopy (cryo-EM) structures of a group IIC intron retroelement that was poised to undergo the first step of reverse splicing.

Results

Overall architecture of an ancient group II intron retroelement

To investigate the mechanism of DNA insertion, we captured a group II intron retroelement before the first step of reverse splicing into DNA (Fig. 1A). We first conducted *in vitro* splicing reactions of the IIC *Eubacterium rectale* (*E. r.*) intron (12) in complex with its encoded maturase (MarathonRT) (13, 14) and purified the reaction mixture to obtain a branched lariat-maturase complex (fig. S1, A to C). The purity and stability of this RNP complex were assessed by using biophysical methods: sedimentation velocity analytical ultracentrifugation and size exclusion chromatography coupled to multiangle light scattering (SEC-MALS) indicated that the sedimentation coefficient and molecular mass of the RNP

were larger compared with those of the individual lariat or maturase components, which suggests complex formation (fig. S1, D to F). To visualize the retroelement in action, we introduced a desthiobiotin-tagged DNA substrate to the intron-maturase RNP and isolated ternary complexes by affinity purification on an avidin column (fig. S2). The purified elution fraction was vitrified on grids, and the holoenzyme molecules appeared as monodisperse particles on cryo-EM micrographs, thereby allowing structure determination (fig. S2, B to D, and fig. S3).

The initial data analysis suggested a preferred orientation of the sample, so a tilted data collection strategy was required to obtain additional projection views (fig. S3). After further classification and focused refinement, we obtained a 2.8-Å resolution cryo-EM structure of the *E. r.* group IIC intron in complex with its specific maturase and DNA target (Fig. 1, B and C, and figs. S4 and S5), thereby revealing the state immediately before the first step of reverse splicing (Fig. 1A). The overall high-resolution 3D reconstruction was of sufficient quality to permit the modeling of individual nucleotides (movie S1) and metal ions. The catalytic core that was formed by D5, the lariat branchpoint, ERS-IBS (exon binding site–intron binding site) sequences, and the protein thumb and DNA binding domain (DBD) was resolved to <3 Å.

The overall structure reveals a compact assembly of intron RNA and maturase protein that is closely associated with the DNA substrate through an extensive network of interactions (movie S2). The intron core adopts a similar fold to that of intron structures that are derived from truncated and modified constructs (15, 16). The tertiary interactions that were identified in previous group II introns are present, along with several additional interactions that are observed in this full-length intron construct that contains all six intron domains (Fig. 1D). The fold of the maturase resembles that of previously studied IIC proteins (13, 17), although the thumb and DNA binding domains are now clearly resolved (Fig. 1C and fig. S5). The bound DNA contains a short spacer, the intron insertion site, and a 5' stem-loop motif that is exclusive to group IIC introns (Fig. 1D) (18, 19).

Features of the catalytic RNP core

Despite extensive efforts, a complete group II intron holoenzyme active site had not yet been visualized. In this work, we capture the complete ribozyme core architecture, which includes hallmark elements identified in earlier biochemical and structural studies (8, 20). For example, we see that the 2'-5' lariat linkage, between the first intron nucleotide (G1) and the branchpoint A (A632), is a crucial structural motif for organizing the ribozyme core.

¹Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. ²Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA. ³Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA. ⁴Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA.

*Corresponding author. Email: anna.pyle@yale.edu

†These authors contributed equally to this work.

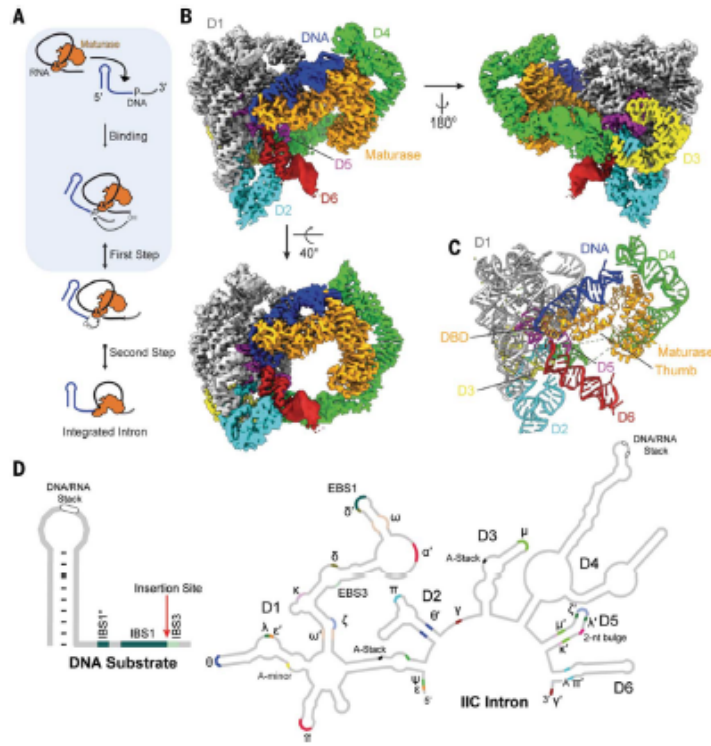


Fig. 1. Cryo-EM reconstruction of a group II intron retroelement. (A) Cartoon of the reverse splicing reaction. (B) Composite cryo-EM map of the holo-RNP with bound DNA. (C) Molecular model of the group II intron retroelement. (D) Secondary structure cartoon and tertiary interactions of the holo-RNP.

The branch site actively engages the 3' end of the intron (G1-A637 pairing), which helps to position the terminal nucleotide (U638) for nucleophilic attack on DNA (Fig. 2, A, B, and E) (21, 22). Facilitating this process, U638 base pairs with A327 to form the γ - γ' interaction (Fig. 2, E and F) (21, 22). The adjacent G328 and C329 nucleotides of the J2/3 linker form major-groove base triples with C562 and G563 (fig. S6A), which gives rise to the catalytic triplex that is common to all group II introns and the spliceosome (23, 24). The 2-nucleotide (nt) bulge (A580 and C581) and catalytic triad (C562, G563, and C564) in D6, along with U638, all serve to coordinate catalytic magnesium ion M1, placing it between the nucleophilic 3' OH and scissile phosphate in an arrangement poised for the first step of reverse splicing (Fig. 2, A, B, and E) (15). A second magnesium ion, M2, is located 3.9 Å away from M1, which is consistent with the two-metal ion catalysis mechanism (Fig. 2, A, B, and E) (15, 25). We identified two additional, unambiguous densities at positions that were previously assigned to the monovalent ions K1 and K2 in studies

that used anomalous scattering to establish sites of stable K^+ binding (24, 26). In that case, as in this instance, NH_4^+ can functionally substitute for K^+ at these same positions (Fig. 2, A, B, and E). The specific coordination and placement of these monovalent ions is essential for positioning the catalytic divalent metal ions, forming a reactive, heteronuclear metal ion cluster. Several tertiary interactions stabilize the periphery of the catalytic core, with D3, supported by an A-stacking interaction with D1, bracing the backside of the D5 helix (μ - μ') (fig. S6, B to C). D2 contacts D6 (π - π') to hold the lariat in place (fig. S6D) (21, 27). Although many of these active site elements have been observed independently, in linear introns or in introns of other classes, they have not been captured simultaneously in a single structure until now, thereby demonstrating that these active site elements function in concert and are conserved. The *Er.* holoenzyme structure provides a detailed view of a complete, reactive intron catalytic core (movie S3).

Close inspection of the active site reveals structural interdependence between the intron

and its encoded maturase. Within the active site, the intron RNA forms short base pairings with its target DNA through the EBS-IBS interactions (EBS1-IBS1 and EBS3-IBS3) (Fig. 2, A, B, E, and F). These otherwise unstable short pairing interactions are buttressed and positioned by the maturase, which presses the middle α helices of the DBD and the third α helix of the thumb domain against the EBS1 and EBS3 recognition loops, respectively (Figs. 1B and 2C), which rigidifies them and helps form a central cavity for engagement with DNA (Fig. 2, C and D). These findings establish that the retroelement core does not consist solely of RNA; rather, it is a collaborative, RNP-active site. The previously undescribed roles that we observe for the maturase thumb and DBD help explain the strong maturase dependence for both RNA splicing and intron integration, particularly *in vivo*, and they highlight the symbiotic relationship between the intron RNA and its protein cofactor, which are known to have coevolved (28, 29).

Functional coordination between RNA and protein

The retroelement holoenzyme has an expansive D4 arm, which extends far from the core and then curves around to cradle the maturase (Fig. 3A). D4a, the high-affinity maturase-binding subdomain (E3, 30), forms two anchor points with basic protein surfaces (Fig. 3, A and B) (31). At the first anchor point, residues extending from the protein [Arg²⁶⁶ (R58), Asp²⁵² (D152), Thr³⁶⁶ (T156), and R160] interact with RNA phosphate and ribose oxygens to secure the insertion helix within the finger domain (IFD) of the maturase against the minor groove interface in the middle of the long D4a hairpin (Fig. 3, B and D). A sharp turn places the distal portion of the D4a subdomain between α helices 9 and 10 of the protein, where largely basic residues [R217, Ser²³⁴ (S234), S237, R240, R243, Asn³⁴⁴ (N244), and R247] approach the RNA backbone from either side, fastening the palm to the D4a arm (Fig. 3, B and C). In contrast to other group II RNPs, the surface of the finger domain (RT0) is not used for RNA recognition (fig. S7) (9, 10, 13).

The distinctive intron-maturase recognition strategy places the maturase thumb and DBD next to the intron core, which allows the protein to participate in catalysis by rigidifying the active site (Fig. 3, A and E). The thumb and DBD grasp the EBS1 and EBS3 loops to directly coordinate substrate-recognition elements within the retroelement active site (Fig. 3E). One approach of this strategy involves locking EBS nucleotides into a conformation conducive for substrate binding [i.e., Lys³⁸⁸ (K388) with G187O6 of EBS1 and K358 with A231N7 of EBS3] (Fig. 3, E to G). A secondary tactic includes immobilizing the EBS3 phosphate backbone

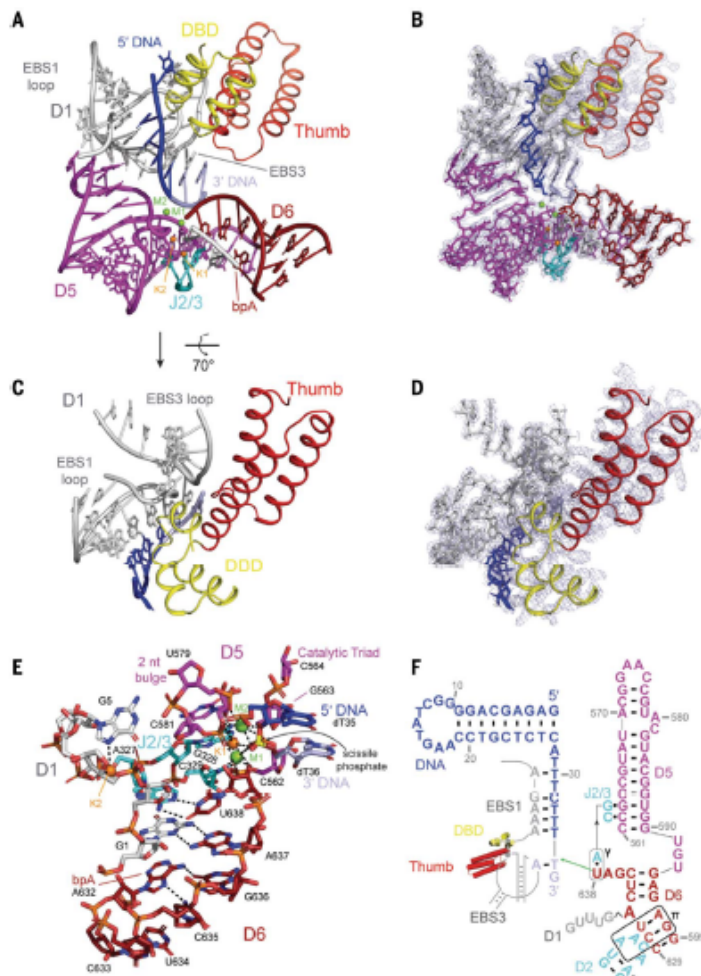


Fig. 2. Architecture of an intron retroelement active site. (A and B) Organization of the holo-RNP core domains. (C and D) Maturase DBD and thumb domains stabilize the DNA recognition loops. (E and F) Model and secondary structure schematic of the intron retroelement before the first step of retrotransposition.

through interactions with a multitude of basic residues on the protein thumb (K300, K303, S309, and R347) (Fig. 3G). A third strategy consists of amino acids [Tyr³⁵⁰ (Y350), R389, N395, and main-chain amines of Ile³⁹⁰ (I390) and Ala³⁹¹ (A391)] stabilizing the turn in EBS1 and enabling the formation of δ - δ' , thereby reinforcing this single-base pair interaction (C183 with G158) that bridges the EBS loops (Fig. 3F). R308 of the protein thumb provides additional stabilization by simultaneously coordinating the phosphate backbone of EBS1 and -3 (through C183 and A230) (Fig.

3H). These interactions demonstrate a specific mechanistic role for the maturase protein during catalysis, which shows that it promotes proper formation of multiple active site components (32). These findings reveal the inextricable, functional coordination of intron and protein during the mechanism of splicing and retrotransposition.

Tertiary interactions with a structured DNA

Our structure reveals unusual strategies for molecular recognition of the DNA target by the holoenzyme. The DNA is recognized through

a combination of shape selectivity and base-specific interactions (movie S4), only a few of which involve canonical Watson-Crick (WC) pairing. The DNA itself has distinct structural features that support this recognition strategy. Most prominent is an unusual, structurally conserved DNA stem, which is composed of a short helix [9 base pairs (bp)] that is capped by an undertwisted duplex composed of two non-canonical G-A DNA base pairs and a G-C base pair (Fig. 4, A and B). Together, these extend the DNA stem to 12 bp, which approximates the consensus stem length for IIC insertion targets. The terminal DNA loop serves as a stacking platform for long-range interactions. Adjacent to the DNA stem is a short spacer, which is followed by IBS1 nucleotides and the IBS3 nucleotide that flank the DNA insertion site.

The DNA stem lies in a cleft that is formed by regions of both the protein (DBD and thumb) and the intron RNA (D1d and D4a). Two clusters of amino acids along the third α helix of the protein thumb domain anchor the DNA stem by making contacts at both ends of the DNA helix, at positions separated by approximately one helical turn. The first cluster (S346, R347, R349, R353, N395, and N405) secures the base of the stem through contacts with dG1 and dA2 (Fig. 4C). The second group [S336, Met³³⁷ (M337), K338, and T339] appears to locally deform the top base pairs of the stem at dC20 and dT21 (Fig. 4C). This is the result of a DNA-protein interaction network that involves insertion of a prolyl-aromatic loop into the distorted, widened minor groove at the tip of the DNA stem. The complementary fit of this peptide loop is mediated by interactions between largely buried side chains [Y278, Phe²⁵⁹ (F279) and Pro²⁸¹ (P281)] and the methylene edges of DNA sugar moieties (Fig. 4D). These protein-DNA interactions are supported by contacts between the DNA and RNA backbone residues (dA4O3' and dC5OP1 with G163 2'OH), which is reminiscent of ribose-zipper interactions that are observed within folded RNA molecules (Fig. 4E) (15). Collectively, these interactions enable the holoenzyme to coordinate and selectively identify the shape of a DNA helix.

This shape-selective recognition strategy of the DNA stem is complemented by sequence-specific interactions between the holoenzyme and single-stranded regions of the DNA target. Phylogenetically covarying base pairs are formed between substrate-recognition regions of the intron and single-stranded DNA nucleobases downstream of the DNA stem (33). In the holoenzyme, we not only identify these critical WC pairings but also observe a complex network of interactions mediated by the spacer DNA that connects the stem with the IBS sequences. This sequential network of DNA IBS elements and the adjacent spacer

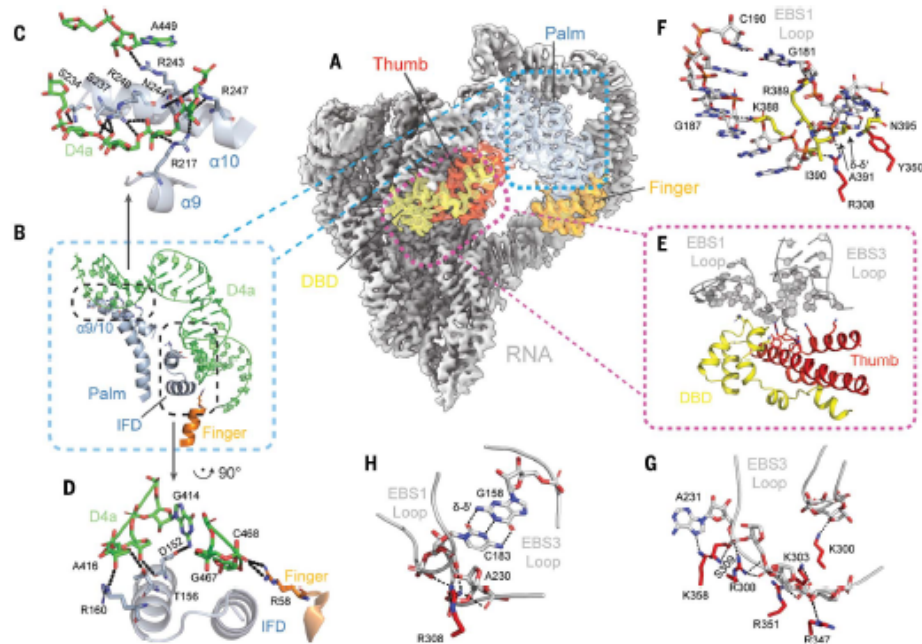


Fig. 3. Mechanism of maturase-facilitated ribozyme catalysis. (A) Protein positioning within the retroelement composite map. (B) Protein-D4a contact points. (C and D). Interactions that form the RNA-protein anchor points. (E) Protein stabilization of EBS1 and EBS3 loops. (F and G) Amino acids that rigidify the EBS1 and EBS3 loops. (H) R308 joins EBS1 and EBS3 together.

interactions begins with the nucleotide located immediately downstream of the insertion site (dT36), which forms a single-base pair interaction (EBS3-IBS3) with a nucleotide extending from the D1d coordination loop within the intron (A231) (Fig. 4F). Stacked atop this pair is a short helix formed through base pairings between the subsequent stretch of DNA nucleotides (IBS1: dT35, dT34, dT33, and dC32) and a second substrate-recognition loop that projects from the terminus of intron D1d (EBS1: A184, A185, A186, and G187) (Fig. 4F). Similar to the short codon-anticodon helix in the ribosome (34), the EBS-IBS1 duplex is further stabilized through the formation of an A-minor motif between A75 and the dC32-G187 base pair (Fig. 4G). The structure reveals that EBS1-IBS1 is not limited to four contiguous base pairs; rather, it is extended by an additional base pair that is formed between the next consecutive nucleotide (A188) in the EBS1 loop and a discontinuous nucleotide from the DNA spacer region (dT30). Indeed, the intervening DNA nucleotide (dT31) is extrahelical and stabilized by interactions with protein residues (Fig. 4H). Through these sequential stacking networks, which are supported by contacts with the protein (i.e., dT36O4 with K361), the

intron achieves stable, base-pairing specificity with the DNA target.

Nucleotides within the DNA spacer participate in binding the RNP, adopting an ordered structure that engages in specific interactions with the holoenzyme. Rather than forming a helical stack, the spacer nucleotides (dA28, dT29, dT30, and dT31) form an unusual motif in which the nucleotides splay in alternating directions on either side of the central phosphate spine (Pauling-like DNA), thereby exposing a large interaction interface to the adjacent DBD (Fig. 4, F and H). Amino acids from the DBD intercalate between the DNA spacer nucleotides while forming an abundance of interactions with both the bases and the phosphate backbone (Fig. 4H). For example, N3 of dT31 interacts with amide oxygens of N378, whereas its adjacent phosphate oxygens interact with proximal arginine residues (R381 and R382). Together, these interactions stabilize an unusual backbone conformation that enables the dT30-A188 pair to form atop the EBS-IBS1 helix. In turn, these interactions with the DBD pull the DNA into place, which positions the specialized barb-like structure formed by the α helical bundle within the DBD at the base of the DNA stem (Fig. 4, B and H).

By capping the DNA stem-loop, a set of stacking interactions clamp the loop terminus into position within the holoenzyme. One such interaction forms between the DNA and RNA loop nucleotides that project from D4, which effectively joins the DNA stem and RNA bases into one continuous stacking network. This extended stacking array consists of dA14 and A441 and U443 from D4a (Fig. 4I). This DNA-RNA tertiary interaction is anchored in place by an adjacent stacking network that forms between the extrahelical dT13 residue and a series of conserved amino acid side chains (Fig. S8), which form a sequential stack that merges with the hydrophobic core of the protein. The aromatic plane of the dT13 nucleobase and Trp¹⁵⁶ (W136) flank R266 on either side, which creates an arginine- π stacking sandwich configuration (35), in which each component is separated by a planar distance of 3.8 Å (Fig. 4J).

Retroelement primed for attack

To better understand molecular rearrangements that might occur when the intron retroelement binds to DNA substrate, we solved the structure of the apo-RNP, visualizing the free intron-maturase complex at a resolution of

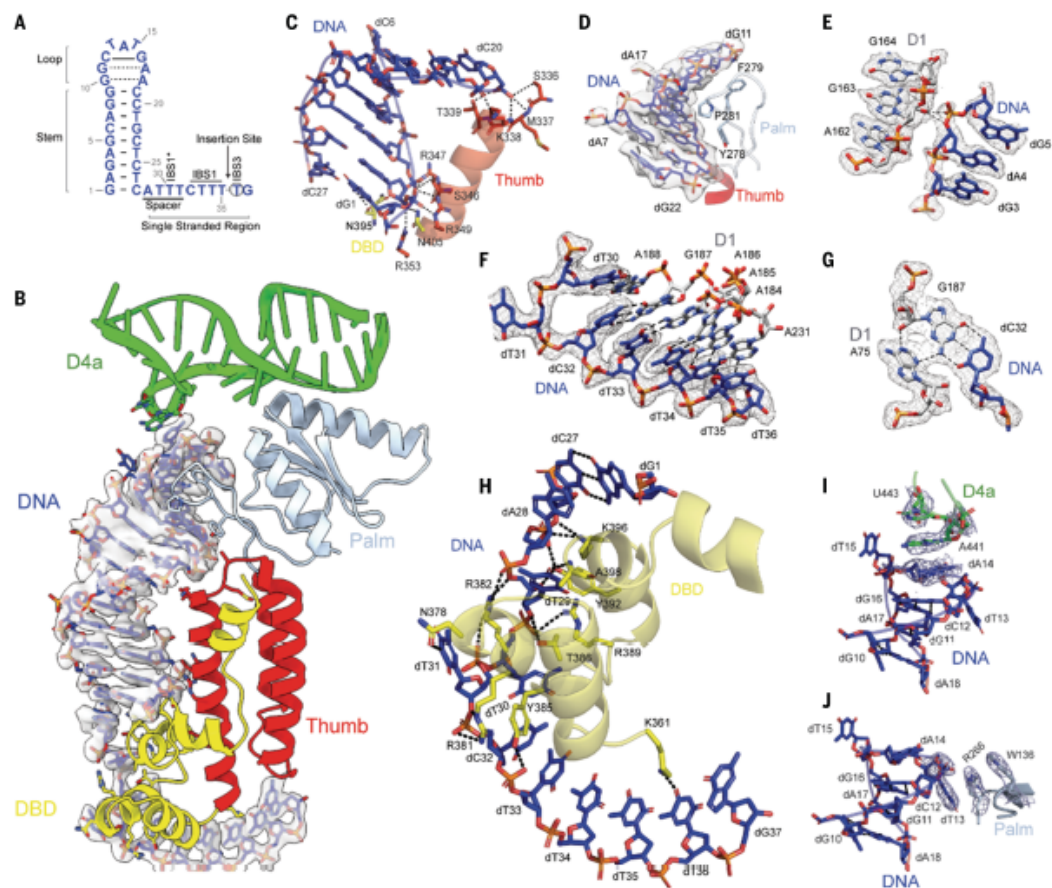


Fig. 4. Shape and sequence recognition of a DNA target. (A) Secondary structure of the DNA target. (B) Interactions of the structured DNA with holo-RNP. (C) Protein contacts with DNA helical stem. (D) Fit of the DNA groove against the protein palm linker. (E) DNA and D1 backbone interactions.

(F) EBS-IBS base-pairing interactions. (G) Stabilizing A-minor tertiary interaction. (H) Interactions between protein and single-stranded DNA. (I and J) Intermolecular stacking interactions between DNA and (I) RNA nucleotides and (J) protein residues.

3.6 Å (Fig. 5A and fig. S9). We observed that the apo-RNP has an architecture that is almost identical to that of the complex bound to DNA, and that substrate binding induces only minor changes in the structure. The RNP-active site remains completely intact (Fig. 5, A and B) (9–11). The maturase does not change its orientation in the absence of DNA and remains coordinated at two anchor points along the D4a arm, with the thumb and DBD inserted into the active site to participate in catalysis (Fig. 5A). In this configuration, the binding interface for the target DNA is maintained, which enables the RNP to readily recognize an incoming DNA target and rapidly engage

in retrotransposition (Fig. 5A). Upon recognition of the DNA stem-loop, the RNP (palm, fingers, and D4a) appears to become more rigid, as we observe a concomitant increase in local resolution at these positions (figs. S3B and S9B and movies S5 and S6). This is reminiscent of many protein enzymes, whereby docking of ligand into the active site freezes out local motions and locks the substrate in place. In previous ligand-free intron structures, EBS nucleotides were found to be disordered or rearranged (21, 22, 24). Here, we observe that the positions of the EBS nucleotides are unchanged, likely due to the presence of the maturase. These findings reinforce the mech-

anistic role of the protein as a stabilizing catalytic component (Fig. 5C). Further evidence of a preformed catalytic core includes the persistence of the heteronuclear metal ion cluster, which remains organized around the lariat, although M2 is not visible in this case (Fig. 5D).

Discussion Insights into protein-facilitated ribozyme catalysis

The structures presented here reveal how components of the holoenzyme help promote activity of the ribozyme core. The protein buttresses the catalytic residues responsible for specifically

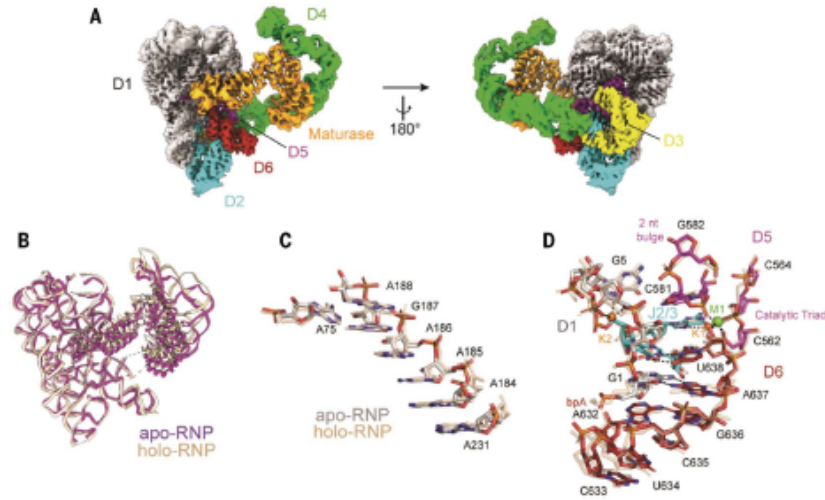


Fig. 5. Retroelement poised to attack. (A) Composite cryo-EM map of the apo-RNP. (B to D) Comparison of (B) the backbone traces, (C) the position of the EBS1 recognition sequences, and (D) the active site of the apo-RNP (colored) and holo-RNP (wheat).

positioning the DNA substrate, providing a missing link in understanding how maturases unlock the full catalytic potential of group II intron ribozymes (Fig. 2). Interactions with protein residues stabilize the intron substrate-recognition loops and precisely arrange nucleotides for DNA binding. These interactions contribute to the proper orientation of reaction components throughout ribozyme catalysis. Our findings provide a direct mechanistic role for the protein cofactor, and they help explain the lowered salt and magnesium requirements in its presence (36).

Prior studies, owing to resolution limitations or construct design, were unable to identify a specific function for the maturase protein, except in transient D6 stabilization (9). By contrast, we can now show that not only are the protein thumb and DBD proximal to D1 catalytic residues, they have critical roles in stabilizing substrate binding. Given its analogous spatial placement, it is possible that Prp8 may play a similar role during spliceosomal catalysis (fig. S10) (37).

RNP recognition of DNA structure: An expanded recognition repertoire

The high-resolution cryo-EM structures we provide here offer a glimpse into RNP strategies for recognizing DNA (Fig. 1B). The holoenzyme structure reveals a stem-loop DNA nestled within the retroelement, bound to RNA and protein. Within this cleft, the protein assists in positioning the insertion site and aligning the DNA stem for steric fit against the complementary maturase surface (Fig. 3). Additional aspects of the unusual recognition strategy include splayed Pauling-like DNA (38), a stabilizing A-minor motif, and intermolecular

stacking moieties that involve both RNA and protein (Fig. 4). These interactions highlight the symbiotic nature of RNA and protein and underscore the multiplicity of strategies available to RNPs for achieving selective substrate recognition.

The DNA stem-loop motif is exclusive to IIC introns, which contain an abbreviated D1 scaffold and short exon recognition sequences (7). In the more highly evolved IIA and IIB introns, the RNP binding motif that we find occupied by the DNA stem in IIC introns is instead replaced by intron insertion motif D1d2, an RNA subdomain that includes EBS2, which is absent in the IIC class (7). Comparison of this region across intron classes suggests that EBS2 evolved to imitate the target DNA stem (fig. S11). Indeed, the DNA stem motif structurally resembles the EBS2-IBS2 interactions typically of IIA and IIB introns, and it functionally emulates EBS2 by anchoring the DNA substrate to the RNP. This mimicry suggests that recognition of a structured DNA motif by the more primitive IIC introns was replaced by RNA domains within the intron itself which resulted in longer target-recognition sequences that provided greater base-pairing specificity for the retroelement.

Implications for reverse splicing and reverse transcription

Encircled by RNA, the exterior surfaces of the protein are enclosed, but the concave interior of the protein, adjacent to the catalytic core, is conspicuously solvent accessible, which has functional implications. During reverse splicing, the D6 helix undergoes a conformational change that places the lariat linkage into the active site (9). To accomplish this, D6 disen-

gages from D2 and swings 90° upward, contacting D1c and a basic patch on the protein thumb. Our structures do not preclude D6 helix dynamics, because there is ample space for a similar movement and the regions that D6 contacts are accessible. The open architecture we observe provides a direct route for DNA to approach the RT active site, because it remains unobstructed by other intron domains and can readily accommodate an entire hybrid duplex for reverse transcription (39). This suggests that initiation of RT activity, within the current holoenzyme assembly, may be possible without marked conformational rearrangement.

Retroelement poised to attack

Group II intron retroelements are proliferative, invasive agents, and our structures explain why. The apo-retroelement is poised to react and does not require any reorganization of structure upon target DNA binding. The arrangement of the active site, from substrate-recognition nucleotides to the heteronuclear metal ion cluster to the DNA binding interface, is preserved despite the absence of DNA substrate (Fig. 5). This prearranged organization is consistent with the biological role of group II introns as parasitic genetic elements (40). Use of the same catalytic core from splicing to integration eschews the need for major rearrangements or host cofactors and allows complete autonomy, which is highly advantageous for a genetic parasite.

Total integration of the RNP requires faithful and accurate reverse transcription of the intron sequence, including the long open reading frame (ORF) that encodes the protein, after insertion. This is accomplished by using

the RT activity of the multifunctional maturase. MarathonRT, the protein within the holoenzyme visualized here, is a well-characterized, robust, accurate, and ultraprocessive RT enzyme that is capable of copying through long, structurally complex templates (47). The intimate association of the parent intron with this protein allows access to its exceptional RT properties and ensures that the intron sequence, which is pivotal to its tertiary architecture, is preserved, allowing the retroelement to continually propagate.

Implications for modern retroelements

Study of group II intron complexes provides a window into our understanding of non-LTR retrotransposons, such as the L1 RNP, an active mobile element that continues to disperse in human genomes (42, 43). Computationally predicted structures of ORF2p, the mobility factor of L1, show that its RT and thumb domain resemble that of the maturase, MarathonRT (fig. S12) (44). ORF2p contains an additional N-terminal endonuclease and a C-terminal extension, but these domains do not block the exterior basic surfaces of the RT and thumb. MarathonRT and ORF2p are evolutionarily related, and they are implicated in similar mobility mechanisms, so it is possible that the same surfaces are used for anchoring and substrate recognition (45). Given the lack of structural information on L1 and the strong parallels between systems, our work provides a starting point for imagining how L1 might assemble and function.

This study reveals strategies for RNP interactions with DNA, having implications for mechanistic understanding of the spliceosome and non-LTR retrotransposons.

REFERENCES AND NOTES

1. T. H. Eickbush, *Curr. Biol.* **9**, R11–R14 (1999).
2. J. L. Feat, F. Michel, *Nature* **364**, 358–361 (1993).
3. L. Boren, J. Vogel, *Trends Genet. Evol.* **17**, 322–331 (2001).
4. A. M. Lambowitz, S. Zimmerly, *Cold Spring Harb. Perspect. Biol.* **3**, a003616 (2011).
5. H. Wank, J. Sanfilippo, R. N. Singh, M. Matsumura, A. M. Lambowitz, *Mol. Cell* **4**, 239–250 (1999).
6. S. Zimmerly, H. Guo, P. S. Perlmutter, A. M. Lambowitz, *Cell* **82**, 545–554 (1995).
7. A. M. Pyle, *Crit. Rev. Biochem. Mol. Biol.* **45**, 215–232 (2010).
8. A. M. Pyle, *Annu. Rev. Biophys.* **45**, 183–205 (2016).
9. D. B. Haack et al., *Cell* **178**, 612–623.e12 (2019).
10. G. Qu et al., *Nat. Struct. Mol. Biol.* **23**, 549–557 (2016).
11. N. Liu et al., *Nucleic Acids Res.* **48**, 1185–1198 (2020).
12. L. Dai, N. Toor, R. Olson, A. Keeping, S. Zimmerly, *Nucleic Acids Res.* **31**, 424–426 (2003).
13. C. Zhao, A. M. Pyle, *Nat. Struct. Mol. Biol.* **23**, 558–565 (2016).
14. C. Zhao, F. Liu, A. M. Pyle, *RNA* **24**, 183–195 (2018).
15. N. Toor, K. S. Keating, S. D. Taylor, A. M. Pyle, *Science* **320**, 77–82 (2008).
16. R. T. Chan, A. R. Robart, K. R. Rajashankar, A. M. Pyle, N. Toor, *Nat. Struct. Mol. Biol.* **19**, 595–597 (2012).
17. A. M. Lentzsch, J. L. Stamos, J. Yao, R. Russell, A. M. Lambowitz, *J. Biol. Chem.* **297**, 100971 (2021).
18. L. Dai, S. Zimmerly, *Nucleic Acids Res.* **30**, 1091–1102 (2002).
19. N. Toor, A. R. Robart, J. Christianson, S. Zimmerly, *Nucleic Acids Res.* **34**, 6461–6471 (2006).
20. C. Zhao, A. M. Pyle, *Trends Biochem. Sci.* **42**, 470–482 (2017).
21. A. R. Robart, R. T. Chan, J. K. Peters, K. R. Rajashankar, N. Toor, *Nature* **514**, 193–197 (2014).
22. M. Costa, H. Walbert, D. Monachello, E. Westhof, F. Michel, *Science* **354**, aa9258 (2016).
23. S. M. Fica, M. A. Metford, J. A. Piccirilli, J. P. Staley, *Nat. Struct. Mol. Biol.* **21**, 464–471 (2014).
24. M. Marcia, A. M. Pyle, *Cell* **151**, 497–507 (2012).
25. S. M. Fica et al., *Nature* **503**, 229–234 (2013).
26. M. E. Wilkinson, S. M. Fica, W. P. Galej, K. Nagai, *Mol. Cell* **81**, 1439–1452.e9 (2021).
27. O. Fedorova, A. M. Pyle, *EMBO J.* **24**, 3906–3916 (2005).
28. B. Cousineau, S. Lawrence, D. Smith, M. Belfort, *Nature* **404**, 1018–1021 (2000).
29. N. Toor, G. Hausner, S. Zimmerly, *RNA* **7**, 1142–1152 (2001).
30. R. N. Singh, R. J. Saldanha, L. M. D'Souza, A. M. Lambowitz, *J. Mol. Biol.* **318**, 287–303 (2002).
31. C. Zhao, A. M. Pyle, *Curr. Opin. Struct. Biol.* **47**, 30–39 (2017).
32. M. Matsumura, J. W. Noah, A. M. Lambowitz, *EMBO J.* **20**, 7259–7270 (2001).
33. F. Michel, K. Umeson, H. Ozeki, *Gene* **82**, 5–30 (1989).
34. A. Lescaute, E. Westhof, *Biochimie* **88**, 993–999 (2006).
35. M. M. Flocco, S. L. Mowbray, *J. Mol. Biol.* **235**, 709–717 (1994).
36. M. Matsumura et al., *Genes Dev.* **11**, 2910–2934 (1997).
37. S. M. Fica, C. Oubridge, M. E. Wilkinson, A. J. Newman, K. Nagai, *Science* **363**, 710–714 (2019).
38. L. Pauling, R. B. Corey, *Proc. Natl. Acad. Sci. USA* **39**, 84–97 (1953).
39. J. L. Stamos, A. M. Lentzsch, A. M. Lambowitz, *Mol. Cell* **68**, 926–939.e4 (2017).
40. A. M. Lambowitz, S. Zimmerly, *Annu. Rev. Genet.* **38**, 1–35 (2004).
41. L. T. Guo et al., *J. Mol. Biol.* **432**, 3338–3352 (2010).
42. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, *Annu. Rev. Genomes Hum. Genet.* **12**, 187–215 (2011).
43. M. A. Kerachian, M. Kerachian, *Clin. Chim. Acta* **488**, 209–214 (2019).
44. J. Jumper et al., *Nature* **596**, 583–589 (2021).
45. Y. Xiong, T. H. Eickbush, *EMBO J.* **9**, 3953–3962 (1990).

ACKNOWLEDGMENTS

We thank M. Uraguna, S. Wu, J. Lin, K. Zhou, and K. Gibson (YCRG) for help with grid preparation, sample screening, and data collection. We thank F. Bleichert for helping with cryoSPARC data processing and Y. Xiong, K. Zhang, and G. Wang for helpful suggestions. We also thank C. Zhao and O. Fedorova for insights and advice throughout this project. **Funding:** This work was supported by the Howard Hughes Medical Institute and the Gruber Foundation (Gruber Science Fellowship to K.C.). Cryo-EM data were collected with microscopes at the Yale CryoEM Resource Core that is funded in part by the NIH (S1000023603). Funding for open access charge was provided by the Howard Hughes Medical Institute. A.M.P. is an investigator and L.X. is a research associate with the Howard Hughes Medical Institute. **Author contributions:** K.C. and L.X. designed the protocol to purify the retroelement complexes, prepared the samples, made EM grids, and performed biochemical assays. S.C.D. conducted SEC-MALS experiments on purified intron complexes. K.C. and L.X. collected EM data. K.C. and L.X., with assistance from P.C. and S.C.D., processed the EM data. K.C. and L.X., with help from J.P. and S.C.D., built the atomic model. K.C., L.X., and A.M.P. analyzed the structure. K.C. and L.X. drafted and prepared the manuscript. A.M.P. supervised and coordinated the group II intron project. **Competing interests:** A patent application on MarathonRT has been filed by Yale University. **Data and materials availability:** All data are available in the main text and the supplementary materials. Cryo-EM maps are available in the Electron Microscopy Data Bank with codes EMD-26550 (holo-RNP) and EMD-26549 (apo-RNP). Structural models are available in the Protein Data Bank with PDB accession codes 7UIN (holo-RNP) and 7UIM (apo-RNP). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. <https://www.science.org/about/s-donor-licenses-journal-article-reuse>

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq2844

Materials and Methods

Figs. S1 to S2

Table S1

References (46–65)

MDAR Reproducibility Checklist

Movies S1 to S6

[View/request a protocol for this paper from Bio-protocol.](#)

Submitted 29 March 2022; resubmitted 10 June 2022

Accepted 17 October 2022

DOI: [10.1126/science.abq2844](https://doi.org/10.1126/science.abq2844)

Structures of a mobile intron retroelement poised to attack its structured DNA target

Kevin ChungLing XuPengxin ChaiJunhui PengSwapnil C. DevarkarAnna Marie Pyle

Science, 378 (6620), - DOI: 10.1126/science.abq2844

A group II intron ready to attack

By forming ribonucleoprotein (RNP) complexes with specialized reverse transcriptases, group II introns can splice out of RNA and insert themselves into new DNA sites. Chung *et al.* used cryo-electron microscopy to investigate how an ancient class of group II intron retroelements recognize the shape and sequence of a highly structured DNA target, thereby revealing new molecular recognition strategies between RNPs and DNA. Structural comparison of the isolated RNP with that of the DNA-bound holoenzyme reveals that the group II intron RNP is primed to attack its DNA target without major conformational rearrangements. The study sheds light on retroelement structure, function, and proliferation. —DJ

View the article online

<https://www.science.org/doi/10.1126/science.abq2844>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title Science is a registered trademark of AAAS.
Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works

Appendix II: Structural insights into intron catalysis and dynamics during splicing.

Required copyright lines for reprint:

Nature **624**, 682–688 (2023) | DOI: <https://www.nature.com/articles/s41586-023-06746-6>

Copyright © 2023 Nature America, Inc. All rights reserved.

Structural insights into intron catalysis and dynamics during splicing

<https://doi.org/10.1038/s41586-023-06746-6>


Ling Xu^{1,2,3,4}, Tianshuo Liu^{2,5}, Kevin Chung^{2,6} & Anna Marie Pyle^{1,2,4,6}

Received: 17 July 2023

Accepted: 13 October 2023

Published online: 22 November 2023

Open access

 Check for updates

The group II intron ribonucleoprotein is an archetypal splicing system with numerous mechanistic parallels to the spliceosome, including excision of lariat introns^{1,2}. Despite the importance of branching in RNA metabolism, structural understanding of this process has remained elusive. Here we present a comprehensive analysis of three single-particle cryogenic electron microscopy structures captured along the splicing pathway. They reveal the network of molecular interactions that specifies the branchpoint adenosine and positions key functional groups to catalyse lariat formation and coordinate exon ligation. The structures also reveal conformational rearrangements of the branch helix and the mechanism of splice site exchange that facilitate the transition from branching to ligation. These findings shed light on the evolution of splicing and highlight the conservation of structural components, catalytic mechanism and dynamical strategies retained through time in pre-messenger RNA splicing machines.

Splicing lies at the heart of RNA metabolism in eukaryotes. During this indispensable stage of gene expression, introns are removed from pre-messenger RNA transcripts to generate mature messenger RNAs (mRNAs)^{1,3,4} (Fig. 1a). The modern spliceosome, the molecular machine that executes the splicing reaction, is thought to originate from the same ancestral molecule as the self-splicing group II introns that are still commonly found in bacteria and organelles of plants and fungi⁵. Group II introns are large ribozymes that catalyse their own excision from precursor RNA transcripts⁶. Both splicing machineries form a conserved active site that hosts the catalytically essential heteronuclear metal ion core^{6,7}. Moreover, they both branch using a bulged adenosine nucleophile, forming the distinctive lariat intron featuring a 2' 5'-linked phosphodiester linkage. Intron D4 contains an open reading frame (ORF) that encodes a specialized multidomain protein (the 'maturase', Fig. 1b,c) which shares strong structural similarity to Prp8, a central protein component of the U5 snRNP^{8,9}. Through formation of a ribonucleoprotein (RNP) holoenzyme with the parent intron RNA, the maturase facilitates intron splicing out of the transcript as well as retrohomologing into new genomic loci¹⁰.

In light of these structural and mechanistic similarities, group II intron RNPs have become a prototypical system for studying the general biochemical principles of RNA splicing and the molecular evolution of splicing machines¹¹. Despite advances in the visualization of group II intron RNAs^{12–14} and RNPs^{15–17}, the structural organization of group II intron systems as they undergo branching and then coordinate the two steps of splicing has remained elusive. Therefore, it is unclear how group II introns properly recognize the branchpoint and the 5'-splice site (5'SS) and how maturases facilitate the branching reaction. These questions are of vital importance as they provide clues on the origin of intron branching, which is among the most fundamental reactions in RNA biology.

To visualize the conformational states along the group II intron RNP splicing pathway, we chose the group IIC intron from *Eubacterium*

rectale and its encoded maturase, MarathonRT¹⁸, as the model system. The maturase acts as a branching switch that shifts the intron splicing pathway from hydrolysis to branching (Extended Data Fig. 1a–d). Here, we used single-particle cryogenic electron microscopy (cryoEM) to obtain the structures of the RNP at each sequential stage during splicing. By capturing the RNP in the state immediately before branching (3.0 Å overall), we visualized how the branchpoint adenosine (bpA) and the splice site (SS) are held in place through molecular interactions between the branch helix and conserved regions of the RNP. Our structural observations reveal a close resemblance between group II intron RNPs and the spliceosome in terms of branchpoint recognition and branch helix positioning. We also gained unique insights into the strategy by which the attacking 2'-OH is precisely positioned in activation distance to the catalytic metal M1, ready for nucleophilic attack. This high-resolution view enables us to construct a complete and catalytically relevant molecular picture of the splicing active site before branching, which has largely eluded structural characterization despite earlier hints⁹.

In addition to the prebranching RNP structure, we present the RNP structures preceding and following the exon ligation step. These structures allow us to visualize large movements of the branch helix, local movements of the branchpoint and the SS exchange that occurs between the two steps of splicing. The conformational dynamics of the spliceosome branch helix recapitulates that of the group II intron, thereby demonstrating that branch helix dynamics are a conserved physical attribute that is codified in splicing machines.

Capturing group II intron RNP in action

To elucidate the mechanism of forward splicing through branching, we sought to visualize structures of the group II RNP complex at each stage along the branching pathway. The two chemical steps of group

¹Howard Hughes Medical Institute, Chevy Chase, MD, USA. ²Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA. ³Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. ⁴Department of Chemistry, Yale University, New Haven, CT, USA. ⁵These authors contributed equally: Ling Xu, Tianshuo Liu, Kevin Chung. [✉]e-mail: ling.xu@yale.edu; anna.pyle@yale.edu

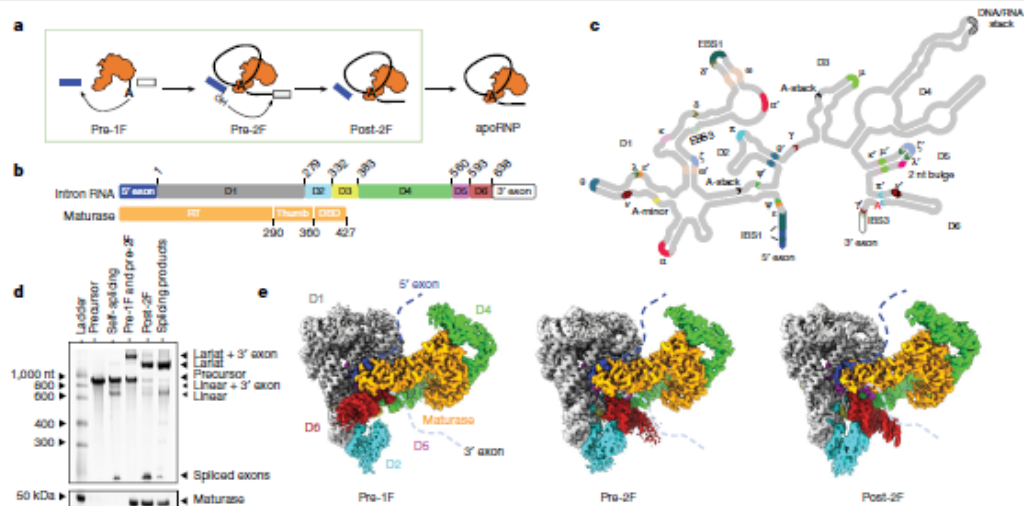


Fig. 1 CryoEM reconstructions of a group II RNP undergoing the branching reaction. **a**, Cartoon of group II RNP splicing. **b**, Domain organization of the intron RNA and its maturase. **c**, Secondary structure of the intron, with annotated tertiary interactions. **d**, A GelRed-stained 5% urea–polyacrylamide gel electrophoresis (PAGE) gel (top) and a SYPRO Ruby-stained SDS–PAGE gel (bottom) showing various conditions used to obtain samples for cryoEM. Lane

1 is the size marker for RNA (top) and protein (bottom). Lane 2 is the marker showing migration of the precursor RNA. Lanes 3 and 6 are the reaction ladders showing migration of the linear and lariat products, respectively. Lanes 4 and 5 are independent cryoEM samples captured at various reaction stages. **e**, Composite cryoEM maps of the prebranching (pre-1F), preligation (pre-2F) and postligation (post-2F) RNP complexes.

II intron splicing are spontaneous and do not require energy input or step-specific factors. It is therefore challenging to stall the reaction without disrupting the active site and earlier attempts caused conformational distortions that made it difficult to discern the exact molecular mechanism of branching²⁰. To resolve this, we incubated splicing precursor constructs in the presence of maturase protein, replacing Mg^{2+} with Ca^{2+} to yield complexes stalled in the precursor and branching intermediate states (Fig. 1d and Extended Data Fig. 1e,f). To obtain the postligation RNP, we assembled the lariat apoRNP²⁷ with an oligonucleotide equivalent to the ligated exon, thereby enabling us to investigate the structural changes upon completion of intron splicing (Fig. 1d).

The corresponding RNP samples were vitrified on grids and appeared as monodispersed particles on cryoEM micrographs, suitable for structure determination. As these samples show preferred orientation, we used the Chameleon system (Spotiton)^{21,22} and combined this with tilted datasets²³ to obtain a uniform angular distribution (Extended Data Figs. 2 and 3). The increased diversity of particle orientations allowed us to generate two distinct, isotropic maps. Upon inspection of the two reconstructions, we assigned the corresponding maps to the prebranching (pre-1F) and preligation (pre-2F) states, respectively (Fig. 1e). The resolution of these maps approaches 2.8 and 2.9 Å, respectively, for the catalytic core (Extended Data Fig. 3). We obtained a three-dimensional (3D) reconstruction for the postligation RNP (post-2F) and the resolution of the catalytic core was determined to 2.9 Å (Fig. 1e and Extended Data Fig. 4). This collection of structures allows us to present a full molecular picture of the group II intron RNP as it proceeds along the branching pathway.

Positioning of the branch helix

To splice through the branching pathway, the group II RNP forms intramolecular RNA and intermolecular RNA–protein interactions that

precisely arrange the branch helix (D6) in the branching-competent conformation (Fig. 2a and Supplementary Video 1). The intron scaffold domain, D1, contributes to this by forming an extended, interlocked interaction network between D1c and D6 (denoted ν – ν') (Fig. 2b). This network features a long-range base pair between G86 and C601, both of which are bulged nucleotides with strong conservation signatures (Fig. 2c and Extended Data Fig. 5a,b). Consistent with their significance in D6 positioning, deletion of G86, C601 or both, leads to branching defects, whereas substituting this GC pair for an AU pair partially rescues branching (Fig. 2d). A wobble pair between G84 and U104 in D1c anchors another intricate molecular network around C601 to further restrain the conformational sampling of D6. In agreement with our structural observations, a G84A/G86A dual mutation, which does not alter D1c secondary structure, has the most pronounced deleterious effect (Fig. 2d). Hence, our results highlight the active role of interdomain RNA interactions in proper positioning of D6.

On the opposite side of the D6 helix, the thumb and DBD domains of the maturase protein compose another extensive RNA–protein intermolecular interface, where we visualized three clusters of interactions. The first cluster (Trp310, Ser313 and Gln359) is located at the basal stem of D6 (Fig. 2e) and its disruption, through mutation to alanine, abolishes intron branching (Fig. 2f). At the junction loop between the thumb and DBD domain a second cluster of residues (Thr362 and Asn365) grasps the central section of D6 adjacent to the branch site and the ribozyme active site. In addition to phosphate backbone interactions, a highly conserved lysine residue (Lys361) inserts into the main groove of D6 and makes direct contact with the 5'SS (G1N7), juxtaposing the first-step nucleophile with the scissile phosphate (Fig. 2e and Extended Data Fig. 5c). Consistent with this structural observation, a Lys361 single alanine mutant is sufficient to eliminate branching activity, as does the Lys361/Thr362/Asn365 triple mutant (Fig. 2f). Finally, the side chains of Lys372 and Arg377 in the DBD domain grip the distal, upper stem of D6 and mutations introduced at these sites compromise branching,

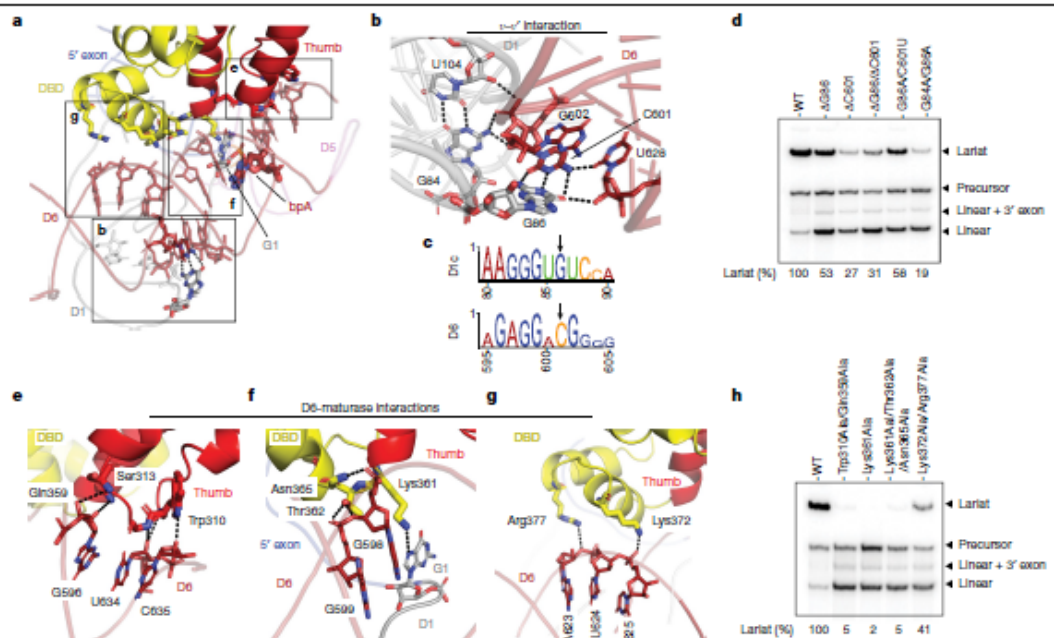


Fig. 2 | Interactions of D6 with the Intron RNP before lariat formation. **a**, Interaction network surrounding the D6 helix before branching (boxed elements are labelled with the figure panel designations described next). **b**, Newly discovered long-range RNA interaction ($v-y'$) that positions D6. **c**, Conservation of nucleotides involved in **b**. **d**, A denaturing radioanalytical splicing gel showing effects of Intron mutants in the presence of WT maturase protein. Individual data points representative of $n = 4$ in vitro splicing assays are

shown. **e**, Intron-maturase interaction cluster at the basal region of the D6 helix. **f**, Intron-maturase interaction cluster at the central region of the D6 helix. **g**, Intron-maturase interaction cluster at the distal region of the D6 helix. **h**, A denaturing radioanalytical splicing gel demonstrating effects of maturase mutants on promoting branching of WT intron construct. Individual data points representative of $n = 4$ in vitro splicing assays are shown.

Using this vast molecular network, the maturase protein stabilizes the branching-competent conformation of D6 and brings the 5'SS adjacent to the branchpoint, thereby explaining why it is indispensable for promoting branching (Extended Data Fig. 1).

In the D6-docked state, one observes an interplay between the intron 5'SS and the branch helix, where the first two nucleotides expose their Watson-Crick edges to engage in tertiary interactions with D6. Specifically, G1:O6 engages the 2'-OH of C633, the nucleotide next to the branchpoint, which secures D6 and brings the branchpoint close to the 5'SS. U2 further strengthens contacts between the 5'SS and D6 through a base triple interaction with the G599-C629 base pair (Extended Data Fig. 6a). The pre-IF structure therefore provides a glimpse into the group II intron 5'SS and establishes its role in branch helix positioning, explaining the conservation signature of the group II intron 5'SS (ref. 24).

Branchpoint recognition and dynamics

Having elucidated the mechanism by which the branch helix is docked in the prebranching conformation, we next sought to unveil the branchpoint recognition strategy. This question is of vital importance, as stringent branchpoint use is a hallmark of group II intron splicing²⁵ and yet the interaction network that recognizes and activates the branchpoint nucleotide for catalysis has eluded structural characterization.

Our prebranching map, with a local resolution of 2.8 Å around the branchpoint, allows confident model building which reveals the

structural basis of branch site recognition. We show that the bpA (A632) is recognized by means of a base triple interaction with the G598-C630 base pair. The exocyclic amine of the bpA (bpA:N6) forms a crucial hydrogen bond with O2 of C630 (Fig. 3a), consistent with previous chemical genetics studies²⁵. The interaction partner (C630) is located two nucleotides upstream and, based on covariation analysis of group II introns²⁶, it is almost exclusively a pyrimidine. The 2'-OH of C630 forms an extra hydrogen bond with bpA:N1. This interaction serves as an extra molecular lock to hold the branch site in place. Intriguingly, the molecular recognition pattern we observe in group II introns is identical to that reported in structures of the spliceosome^{6,27}. Recent models for the yeast C complex⁶ (postbranching) revealed the same molecular interaction between the bpA and a highly conserved uridine located two nucleotides upstream (Fig. 3b), which can reasonably fit in the density of the yeast B* complex (prebranching). A similar bpA recognition strategy has also been proposed in the C complex of the human spliceosome²⁷. Our structure therefore unveils a mechanistic parallel in splicing machines for defining the branchpoint, which seems to be hard-coded by molecular evolution.

Next, we sought to visualize the local conformational dynamics of the bpA along the branching pathway. In the pre-IF state, the bpA adopts an unusual conformation that causes it to point toward the main groove of the branch helix, through a base triple interaction. This conformation leads to significant distortion of the bpA sugar-phosphate backbone (Fig. 3a,d), which places the 2'-OH next to the scissile phosphate. After branching, as the complex enters the preligation state, the bpA flips to

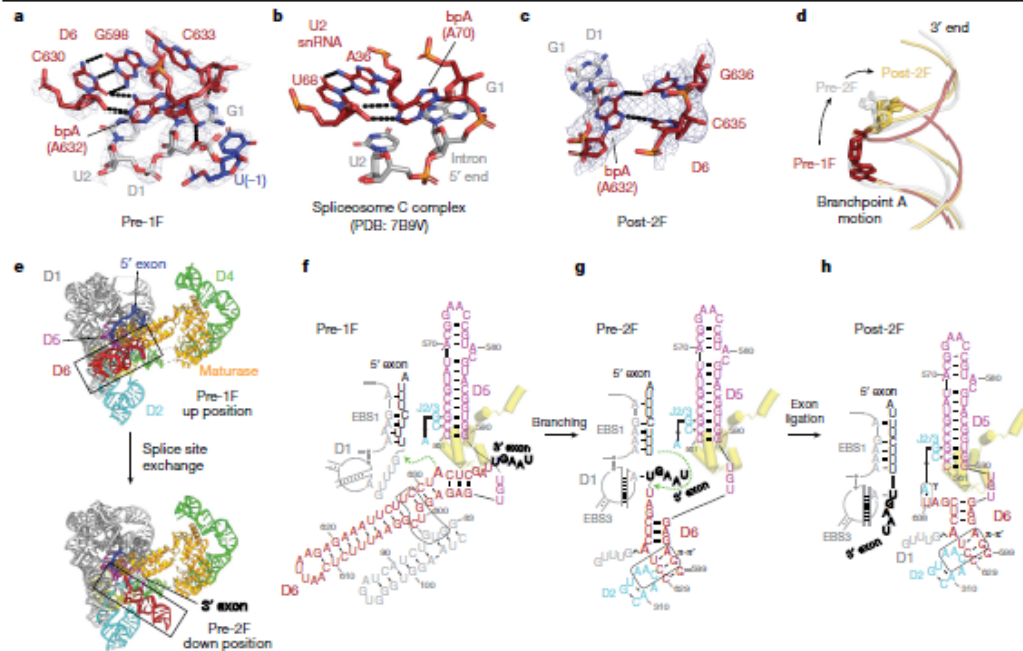


Fig. 3 | Molecular recognition of the branchpoint A and D6 dynamics. **a**, Interactions that specify the bpA before branching. **b**, Positioning of the bpA in the yeast spliceosome C complex. **c**, Interaction network surrounding the bpA postligation. **d**, Aligned D6 helices showing conformational movement of

the bpA during stages of splicing. **e**, Conformational rearrangement of D6 from branching to exon ligation. **f–h**, Secondary structure schematic with annotated tertiary contacts of the RNP in the pre-1F (**f**), the pre-2F (**g**) and the post-2F state (**h**). Yellow cartoon batons indicate the position of the maturase DBD helices.

the opposite side of the branch helix and points toward the 3'-end of D6 (Fig. 3d and Supplementary Video 2). This dramatic conformational rearrangement of the bpA relaxes the backbone distortion, potentially releasing free energy that compensates for the energetic cost of disengaging interactions that originally anchored the bpA⁵. Upon exon ligation, the 3'-end of the intron (C635 and G636) moves further towards the branch site and makes direct contact with the Hoogsteen face of the bpA (Fig. 3c). These extra-molecular interactions, formed after exon ligation, limit the conformational flexibility of the bpA and mark the termination of the splicing pathway.

Branch helix conformational dynamics

In addition to the local dynamics of the bpA, comparison of the intron RNP structures reveals a set of tertiary contact rearrangements needed to coordinate the two sequential steps of splicing (Fig. 3e–h and Supplementary Video 3). In the pre-1F state, the intron recognizes the 5'-exon through the EBS1–IBS1 interaction and the branch helix adopts the D1c- and maturase-docked conformation (Fig. 2a), which we refer to as the 'up' position hereafter (Fig. 3e). In this arrangement, an array of long-range interactions form between J4/5 (A559 and A560) and J5/6 (U591, G592, U593) which participate in a coordinated series of interactions (Extended Data Fig. 6c). This network begins with a canonical base pair (A560–U591) and continues with a non-canonical pairing, in which the Hoogsteen edge of A559 interacts with the sugar edge of G592 and is capped by the final nucleotide of J5/6, U593, which stacks beneath A559. Owing to the interactions that pull D6 into the up position and the constraints imposed by the J5/6 interaction network, the phosphate

backbone connecting D5 and D6 adopts a bent conformation, flipping adjacent nucleotides to opposing sides (Extended Data Fig. 6d).

After branching, D6 undergoes a substantial structural rearrangement that involves an approximately 90° swing, to the 'down' position (Fig. 3e). Through this process, the intron pulls the 5'SS and the newly formed lariat bond about 21 Å out of the active site and exchanges it for the 3'SS, thereby preparing the active site for exon ligation. During this transition, the $v-v'$ tertiary interaction and D6–maturase contacts are disrupted. In the resulting pre-2F structure, D6 docks onto D2, engaging $\pi-\pi'$, which latches onto the branch helix, thereby pulling D6 and the covalently linked 3'-exon into position (Fig. 3g). This allows formation of the EBS3–IBS3 base pairing that defines the 3'SS (U(+)-A231) (Fig. 3f,g). Comparison of the catalytic D5 helix in the pre-1F and pre-2F structures reveals that it remains stationary within the D1 scaffold (root mean square deviation of 0.5 Å). Instead, movement of the D6 helix hinges on the J5/6 linker and appears as motion of the branch helix relative to a fixed RNP body. Swinging of D6 into the plane of the RNP relaxes the bent conformation of J5/6 (Extended Data Fig. 6e), enabling an exchange of substrates in the active site and driving the branching reaction forward⁴. The structural importance of J5/6 in branching is consistent with mutational studies that investigated its biochemical function in positioning of the branch helix^{28,29}. Further movement of D6 is observed on completion of splicing, where there is minor motion of the D6 3'-end, which tucks inwards, allowing engagement of $\gamma-\gamma'$ (A327–U638) (Fig. 3h). Hence, our structures provide detailed molecular insights into the conformational rearrangements and sequential transitions that are required for branching and SS specification during group II intron splicing.

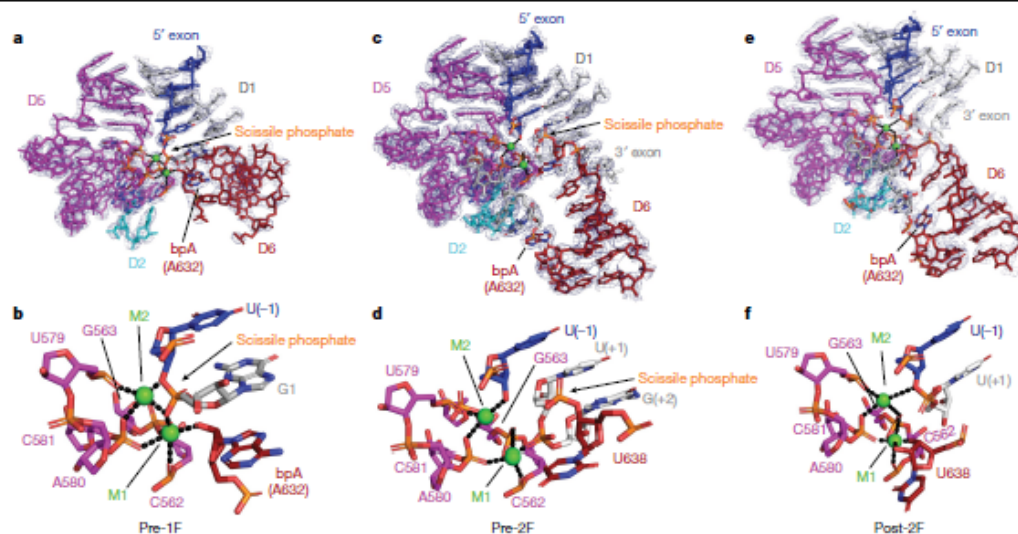


Fig. 4 | Molecular mechanism of group II RNP branching and exon ligation. **a**, Organization of catalytic elements before branching. The bpA is juxtaposed to the 5'SS and poised for lariat formation. **b**, Zoomed in view of (a). **c**, Active site configuration before exon ligation. The 5'SS is primed for attack to ligate

the exons. **d**, Zoomed in view of (c). **e**, Positioning of active site elements immediately after exon ligation. **f**, Zoomed in view of (e). Divalent metal ions are shown as green spheres.

Catalytic mechanism of intron splicing

Having revealed the dynamical strategies used by the group II intron RNP throughout the branching pathway, we next sought to visualize the chemical catalytic mechanisms for each step. Catalysis of the branching reaction is potentiated by a heteronuclear metal ion core organized around the catalytic triplex and the two-nucleotide bulge of D5 (Fig. 4a,b). Through precise positioning of D6 and formation of intra-D6 interactions (Fig. 2a and Fig. 3f), the first-step nucleophile (2'-OH of the bpA) is precisely positioned by catalytic metal M1 and placed in the activation distance (2.3 Å), where it is poised for the nucleophilic attack (Fig. 4b). Remarkably, the attacking 2'-OH nucleophile in the pre-1F structure occupies an identical position to that of the water nucleophile in an earlier pre-hydrolytic structure (Fig. 5a), which provides an unambiguous explanation for the competitive nature of the two splicing pathways³⁰. The scissile phosphate of the 5'SS, between U(-1) and G1, adopts the same sharply kinked conformation previously observed for the hydrolytic, precatalytic state⁷. The pro-Rp oxygen of the scissile phosphate coordinates both M1 and M2 whereas the 3'-bridging oxygen is in direct contact with M2, facilitating departure of the 3'-oxyanion leaving group. This high-resolution view of the active site in the prebranching state hence provides direct visualization of the two-metal ion mechanism for group II intron branching proposed three decades ago²⁸. In addition, we identified two strong, globular densities around the divalent metal core, whose positions correspond to the previously identified monovalent ions, K1 and K2 (ref. 7) (Extended Data Fig. 7a,b). Our findings therefore highlight the formation of a heteronuclear metal ion core as a general catalytic strategy fundamental to RNA splicing^{6,7,32}.

Upon cleavage of the 5'SS, D6 movement brings the first-step nucleophile and the now covalently linked G1 out of the active site (Fig. 3e and Fig. 4c,d). The first-step leaving group, the U(-1):3'-OH, remains tightly coordinated with catalytic metal M2 and becomes the activated second-step nucleophile. The second-step scissile phosphate between

U638 and U(+1) then becomes visible in the pre-2F state, adopting the same precleavage kinked configuration (Fig. 4c,d). These data establish that the same active site is used for both splicing steps without modifying the catalytic ion configuration nor the metal-binding platform (Fig. 4b,d and Extended Data Fig. 7). Moreover, our structure of the post-2F state with the ligated exon bound (Fig. 4e,f) shows that the metal catalytic core remains well organized, whereby the 3'-bridging oxygen of U(-1) remains associated with M2 and the 3'-OH of U638 is coordinated with M1 (Fig. 4f).

Discussion Splicing at the RNP interface

As revealed by our study, group II intron and spliceosome not only share structural and chemical components (Extended Data Fig. 7c) but also a conserved dynamical strategy for sequential rearrangement between the steps of splicing. Direct parallels can be drawn between the motions of the D6 helix in the group II intron and the branch helix in the spliceosome. Comparison of their identical branching states reveals the same 90° swing of the U2-intron branch helix during the transition from the branching B^a complex to the exon ligation C^a complex (Fig. 5b,c). Analogous conformational dynamics are observed in the group II intron holoenzyme, as it swaps SSs without disrupting the catalytic core, when transitioning between the steps of splicing. Remarkably, the branch helix swinging motion has equivalent centres of rotation to that of the spliceosome, whereby the J5/6 linker in the group II intron acts as a hinge, much like the corresponding U2/U6 linker in the spliceosome²⁹. Intriguingly, as in the spliceosome (U2/U6), there are no conformational rearrangements of the catalytic triplex (Fig. 5b,c), which remains static through the stages of branching²⁹. We now have direct evidence of a conserved dynamical mechanism of SS exchange by group II introns that has direct parallels with the spliceosome, strengthening the argument that group II introns and spliceosome share the same ancestry.

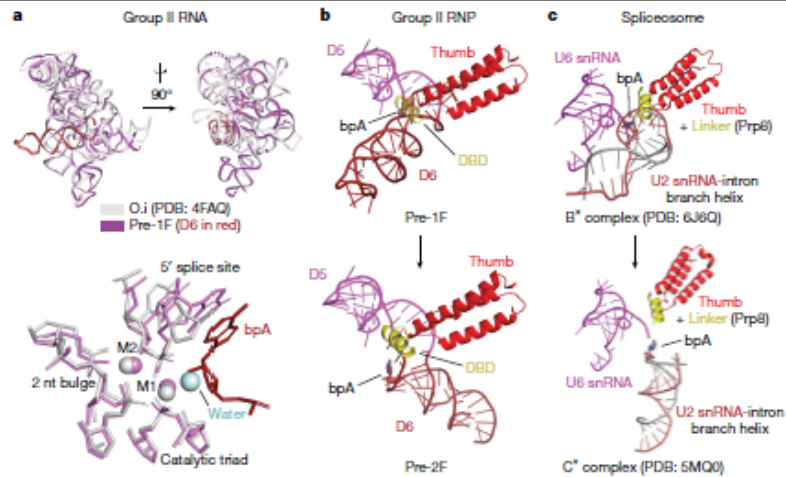


Fig. 5 | Mechanistic comparison of group II introns and the spliceosome. **a**, Comparison of the overall fold of group II introns (top) and the aligned active sites of the *Oceanobacillus ihayensis* intron before hydrolysis (nucleophilic water

in light blue) and the *E. recifei* intron before branching (bpA in dark red) (bottom). **b, c**, Conserved RNP interface and branch helix dynamics in group II RNPs (**b**) and spliceosomes (**c**) during the first- to second-step transition.

Despite the many features in common with group II introns, we identified a functional difference that provides an extra layer of regulation for the spliceosome. The first short α -helix located in the maturase DBD domain has a positively charged surface (Extended Data Fig. 8a) that is indispensable for spontaneous group II intron RNP branching (Fig. 2e, f). In contrast, whereas the equivalent helix in the linker domain of Prp8 adopts a highly similar pose (Fig. 5b, c), the contact surface is negative to neutral in charge (Extended Data Fig. 8b). This marked difference between the maturase and Prp8 has evolutionary implications. On one hand, the maturase is the lone protein cofactor necessary for proper positioning of the D6 branch helix. However, the spliceosome requires recruitment of step 1 specific factors, such as Yju2, to activate branching¹⁹. We can now structurally rationalize the need for Yju2 by comparing the maturase and Prp8 surfaces. The N terminus of Yju2 may compensate for the positive charges that are lost during molecular evolution from the maturase to Prp8 by forming a highly positively charged contact surface that interacts with the branch helix (Extended Data Fig. 8b). Intriguingly, the maturase side chain, Lys361, shown to be essential for intron branching in our study (Fig. 2f) has no equivalent in Prp8; whereas a highly conserved residue (Arg3 in *Saccharomyces cerevisiae* and *Homo sapiens*)²⁰ at the N terminus of Yju2 plays a similar role in contacting G1 of the 5'SS. We therefore observe hints of a molecular evolutionary strategy that fragmented the single-protein RNP into a multiprotein splicing machine, which allows for fine tuning of RNA splicing as a regulated biological process.

Conservation of molecular recognition

The prebranching RNP structure presented in this study reveals the 5'SS and branchpoint recognition strategy used by group II introns, thereby providing critical insights into how splicing machinery maintains precise SS and branchpoint definition during molecular evolution.

We present the pre-attack conformation of the bpA in group II introns (Fig. 4a). This high-resolution view unambiguously explains the branch site recognition strategy used by group II introns. Instead of canonical base pairing, the intron resorts to a base triple (*cis* Watson–Crick/sugar edge interaction) formed between the bpA and a CG base pair located two nucleotides upstream (Fig. 3a). The interaction also serves

to hold the bpA inwards, toward the major groove of the D6 branch helix, thereby limiting its conformational flexibility and correctly positioning its 2'-OH relative to the catalytic metal for activation. The same molecular recognition strategy is used by the spliceosome to anchor its bpA (Fig. 3b). This striking similarity provides molecular evidence that there is minimal change to the strategy for branchpoint definition during evolution from group II introns to the spliceosome.

Also, we revealed the molecular basis of group II intron 5'SS recognition. Through base–sugar interactions originating from G1 and a base triple interaction from U2 (Extended Data Fig. 6a), the intron 5'SS interlocks with the branch helix and closely contacts the branchpoint, preparing the system for branching. The abundance of molecular interactions surrounding the 5'SS also enforces stringent nucleotide identity requirements. Given the mechanistic parallel with the spliceosome (Extended Data Fig. 6b), we can now justify why the same 5'-GU motif⁴ has persisted through time, highlighting that the strategy to define the 5'SS is so robust that it has withstood the forces of molecular evolution.

Group II RNP life cycle

By combining the cryoEM structures obtained in this study with previous mechanistic and structural work done on group II introns^{7,12,17,24}, we can now propose a mechanism for the group II intron splicing life cycle (Supplementary Video 4), including excision from the flanking exons and retrohoming into DNA sites (Extended Data Fig. 9 and Supplementary Video 5). After translation of the maturase from the ORF, the protein facilitates RNA folding by binding to the D4a arm and interacting with D1, which folds first and acts as a scaffold^{25,34} for assembly of downstream domains. The D6 branch helix docks onto D1 through the intramolecular $v-v'$ interaction and engages the thumb/DBD domains of the maturase to stabilize the helix in the up position. Specific molecular interactions distinguish and lock the bpA and 5'SS into place, juxtaposing the 2'-OH against the scissile phosphate for nucleophilic attack through the heteronuclear metal ion active site. During the first stage of branching, the 2',5'-phosphodiester bond is formed, exposing the 3'-OH of the 5'-exon for ligation. To exchange substrates, the branch helix then disengages $v-v'$ and the maturase–D6 interactions, permitting the bpA to pivot around its phosphate and the

D6 helix to swing downwards (Supplementary Videos 2 and 3), where it forms the $\pi-\pi'$ tertiary interaction with D2. This pulls theariat out of the active site and replaces it with the 3'SS, demarcated by the EBS3–IBS3 base pairing, thereby positioning the 3'-exon for ligation. Using the same active site, the 3'-OH is activated for nucleophilic attack, resulting in splicing of the exons. Following exon ligation, the D6 3'-tail tucks inward and the terminal nucleotide is secured by the $\gamma-\gamma'$ interaction. The ligated exons are then released and the liberated apoRNP retains its overall architecture, enabling it to remain primed for binding DNA substrates on the basis of shape and sequence complementarity for engagement in reverse splicing¹⁷.

To undergo retrotransposition, we postulate that equivalent, conserved D6 and branch site motions are used¹⁶ to achieve intron integration and substrate exchange using a persistent heteronuclear metal ion core. Given the proximity of the maturase reverse transcriptase active site to the 3'-end of the integrated intron, a logical hypothesis is that the 3'-end of the fully reverse spliced product is threaded into the maturase reverse transcriptase domain. Here, the protein, using an exogenous primer, begins target primed reverse transcription, unravelling the base pairing, disassembling the elaborate intron tertiary structure²⁷ and generating a complementary DNA strand to effectively copy and paste the RNA sequence into a new genomic site, thereby completing the intron life cycle. Further biochemical and structural work will be needed to evaluate this hypothesis and address the remaining mechanistic aspects of the group II RNP life cycle after branching.

Online content

Any methods, additional references, Nature Portfolio 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/s41586-023-06746-6>.

- Shi, Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat. Rev. Mol. Cell Biol.* **18**, 655–670 (2017).
- Galej, W. P., Toot, N., Newman, A. J. & Nagai, K. Molecular mechanism and evolution of nuclear pre-mRNA and group II intron splicing: insights from cryo-electron microscopy structures. *Chem. Rev.* **118**, 4156–4176 (2018).
- Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nat. Rev. Genet.* **23**, 697–710 (2022).
- Rogalska, M. E., Vioroli, C. & Valcarlos, J. Regulation of pre-mRNA splicing: roles in physiology and disease and therapeutic prospects. *Nat. Rev. Genet.* **24**, 251–269 (2023).
- Zhao, C. & Pyle, A. M. Structural insights into the mechanism of group II intron splicing. *Trends Biochem. Sci.* **42**, 470–482 (2017).
- Wilkinson, M. E., Fica, S. M., Galej, W. P. & Nagai, K. Structural basis for conformational equilibrium of the catalytic spliceosome. *Mol. Cell* **81**, 1439–1452 (2021).
- Marcia, M. & Pyle, A. M. Visualizing group II intron catalysis through the stages of splicing. *Cell* **151**, 497–507 (2012).
- Zhao, C. & Pyle, A. M. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat. Struct. Mol. Biol.* **23**, 558–565 (2016).
- Galej, W. P., Oubridge, C., Newman, A. J. & Nagai, K. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* **493**, 638–643 (2013).
- Belfort, M. & Lambowitz, A. M. Group II intron RNPs and reverse transcriptases: from retroelements to research tools. *Cold Spring Harb. Perspect. Biol.* **11**, a022375 (2019).
- Eickbush, T. H. Mobile introns: retrohoming by complete reverse splicing. *Curr. Biol.* **9**, R11–R14 (1999).
- Toot, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II intron. *Science* **320**, 77–82 (2008).

- Robert, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toot, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* **514**, 193–197 (2014).
- Chan, R. T. et al. Structural basis for the second step of group II intron splicing. *Nat. Commun.* **9**, 4676 (2018).
- Qu, G. et al. Structure of a group II intron in complex with its reverse transcriptase. *Nat. Struct. Mol. Biol.* **23**, 540–557 (2016).
- Hsack, D. B. et al. Cryo-EM structures of a group II intron reverse splicing into DNA. *Cell* **178**, 612–623 (2019).
- Chung, K. et al. Structures of a mobile intron retroelement poised to attack its structured DNA target. *Science* **378**, 627–634 (2022).
- Zhao, C., Liu, F. & Pyle, A. M. An ultrasensitive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183–195 (2018).
- 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).
- Liu, N. et al. Exon and protein positioning in a pre-catalytic group II intron RNP primed for splicing. *Nucleic Acids Res.* **48**, 11185–11198 (2020).
- Dandey, V. P. et al. Spotiton: new features and applications. *J. Struct. Biol.* **202**, 161–169 (2018).
- Nobis, A. J. et al. Reducing effects of particle adsorption to the air–water interface in cryo-EM. *Nat. Methods* **15**, 793–795 (2018).
- Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–795 (2017).
- Zimmerly, S. & Semper, C. Evolution of group II introns. *Mob. DNA* **6**, 7 (2015).
- Liu, Q. et al. Branch-site selection in a group II intron mediated by active recognition of the adenine amino group and steric exclusion of non-adenine functionalities. *J. Mol. Biol.* **267**, 163–171 (1997).
- Chu, Y. T., Adami, C., Liu, Q., Perlman, P. S. & Pyle, A. M. Control of branch-site choice by a group II intron. *EMBO J.* **20**, 6866–6876 (2001).
- Bertram, K. et al. Structural insights into the roles of metazoan-specific splicing factors in the human step 1 spliceosome. *Mol. Cell* **60**, 127–139 (2020).
- Boulanger, S. C. et al. Length changes in the joining segment between domains 5 and 6 of a group II intron inhibit self-splicing and alter 3' splice site selection. *Mol. Cell Biol.* **16**, 6906–6904 (1996).
- Fica, S. M. et al. Structure of a spliceosome remodelled for exon ligation. *Nature* **542**, 377–380 (2017).
- Dankels, D. L., Michels, W. L. Jr & Pyle, A. M. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. *J. Mol. Biol.* **256**, 31–49 (1996).
- Steltz, T. A. & Steltz, J. A. A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl. Acad. Sci. USA* **90**, 6408–6402 (1993).
- Genna, V., Colombo, M., De Vivo, M. & Marcia, M. Second-shell basic residues expand the two-metal-ion architecture of DNA and RNA processing enzymes. *Structure* **26**, 40–50 (2018).
- Liu, Y. C., Chen, H. C., Wu, N. Y. & Chang, S. C. A novel splicing factor, Yju2, is associated with NTC and acts after Prp2 in promoting the first catalytic reaction of pre-mRNA splicing. *Mol. Cell Biol.* **27**, 5403–5413 (2007).
- Pyle, A. M. Group II intron self-splicing. *Annu. Rev. Biophys.* **45**, 183–205 (2016).
- Zhao, C., Rajashankar, K. R., Marcia, M. & Pyle, A. M. Crystal structure of group II intron domain 1 reveals a template for RNA assembly. *Nat. Chem. Biol.* **11**, 967–972 (2015).
- Qin, P. Z. & Pyle, A. M. Stopped-flow fluorescence spectroscopy of a group II intron ribozyme reveals that domain 1 is an independent folding unit with a requirement for specific Mg²⁺ ions in the tertiary structure. *Biochemistry* **36**, 4718–4730 (1997).
- Guo, L. T., Olson, S., Patel, S., Gravelley, B. R. & Pyle, A. M. Direct tracking of reverse-transcriptase speed and template sensitivity: implications for sequencing and analysis of long RNA molecules. *Nucleic Acids Res.* **50**, 6980–6989 (2022).

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023