UNIVERSITY OF CALIFORNIA, SAN DIEGO

Structural biology of group II intron splicing and retrotransposition

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Daniel B. Haack

Committee in charge:

Professor Navtej Toor, Chair
Professor Timothy Baker, Co-Chair
Professor Andreas Leschziner
Professor Uli Muller
Professor Haim Weizman

2018

The dissertation of Daniel B. Haack is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

Co-Chair

University of California, San Diego

2018

Dedication

To my loving wife Karey, who is my whole world and inspiration.

To my family, who has pushed me to achieve my goals and kept me focused.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to first and foremost thank Professor Navtej Toor for being a great mentor and supporting me through my growth as a scientist. My journey has been wrought by successes and failures. Your even-tempered nature helped guide me through the difficult times and kept me focused on the big picture. Any successes I have earned are due in large part to your patience with me. I would also like to thank my co-advisor and mentor, Professor Timothy Baker. His passion for structural biology is contagious and his willingness to give me a chance changed my graduate career forever.

I want to acknowledge my great Toor labmates. Dr. Aaron Robart, Dr. Russell Chan, Dr. Jessica Peters, Tim Wiryaman, Ana Gomez, and Jason Hingey were all awesome co-workers and I owe them a debt of gratitude. In addition, my Baker labmates were instrumental in developing me as cryo-EM microscopist. The long nights of data collection by Tim Booth and the countless hours a data processing by Dr. Xiaodong Yan were critical to me reaching my goals. It was also great having such a vibrant RNA society at UCSD. I want to thank the Muller lab, the Joseph lab, and the Zid lab for our great conversations over the years and your willingness to share equipment.

Finally, I would like to thank my family, whose love and support made this possible. To my wife, Karey Kowalski, you are my whole world. You were by my side before I started this journey and you are still by my side now. Your belief in me gave me the strength to overcome adversity and your love keeps me smiling. I love you now and always. To my parents, Rick and Becky Haack, thank you for always being there for me. Whether it was a quick phone call or an emergency flight to come visit, you always prioritize me as your son and I love you both so much. I could not have done it without you. Finally, I want to thank my big brother Geoff. Even though we live in different cities, he has made an effort to maintain our relationship and I cherish it every day.

Vita

2008	Bachelor of Science, University of California, Santa Barbara
2013	Master of Science, University of California, San Diego
2018	Doctor of Philosophy, University of California, San Diego

List of Abbreviations

2,6-DAP	2,6-diaminopurine
D	DNA binding
DED	direct electron detector
DSB	double strand break
DTT	dithiothreitol
EBS	exon binding sequence
En	endonuclease
IBS	intron binding sequence
IPTG	isopropyl β-D-1-thiogalactopuranoside
kD	kilodalton
LB	Luria Bertani broth
LTR	long terminal repeat
MBP	maltose binding protein
MPD	(+/-)-2-Methyl-2-4-pentanediol
OD	optical density
ORF	open reading frame
PEG	polyethylene glycol
RNP	ribonuclear/protein complex
RT	reverse transcriptase
snRNP	small nuclear ribonuclear protein complex
TL	tetraloop
TLR	tetraloop receptor
TNC	tandem non-canoncical
TPRT	target primed reverse transcription
XFEL	X-ray free electron laser

Abstract of the Dissertation

Structural biology of group II intron splicing and retrotransposition

by

Daniel B. Haack

Doctor of Philosophy in Chemistry

University of California, San Diego, 2018

Professor Navtej Toor, Chair
Professor Timothy Baker, Co-Chair

Group II introns are DNA sequences that are interspersed throughout the genomes of organisms from all three domains of life. In many cases these intron genes interrupt coding sequences referred to as exons. In order for the exons to be expressed properly, the intervening intron sequence must first be removed. Once transcribed into pre-mRNA, the scattered introns become catalytic ribozymes and are able to efficiently excise themselves and ligate the flanking exons. These RNA molecules contain an active site capable of binding the catalytic metal ions required to perform two sequential transesterification reactions that cut out the intron sequence and paste the exons together. This mechanism is identical to the one used by the spliceosome to process spliceosomal introns from pre-mRNA in eukaryotes. The mechanistic similarities combined with conserved structural elements supports the hypothesis that group II introns and the spliceosome share a common evolutionary ancestor. In addition to their splicing function, group II introns also act as selfish mobile genetic elements known as

retrotransposons. These active retroelements contain an open reading frame for a protein called the maturase. When expressed. this protein acts as a folding chaperone by binding specifically to the intron RNA to promote splicing activity. The maturase is a multifunctional protein containing reverse transcriptase and endonuclease domains allowing the group II intron/maturase complex to invade dsDNA through a target primed reverse transcription mechanism. Sequence homology and mechanistic similarities have evolutionarily linked group II introns to non-LTR retroelements, which make up approximately 34% of the human genome. Some of these retroelements are still active and are the cause of many genetic diseases. The focus of this dissertation is the structural study of group II introns to elucidate the mechanism of splicing and retrotransposition. Using cryo-EM, I have been able to obtain density maps at 5.8 Å resolution for a group II intron while splicing and 4.8 Å resolution for a group II intron actively invading dsDNA. The 4.8 Å map represents the first of its kind for any retroelement.

Chapter 1: Introduction

1.1 Group II intron structure and function

The discovery of catalytic RNA molecules by Tom Cech and Sidney Altman in the 1980s was a ground-breaking step towards understanding the role that RNA has played in the origin and evolution of life [1,2]. Central to Tom Cech's discovery was the identification and characterization of a catalytically active group I intron from the ciliate *Tetrahymena thermophila* [1]. The ability of RNA to perform chemistry as well as to contain inheritable genetic information led to the development of the evolutionary "RNA World Hypothesis" [3]. The basis of this evolutionary model states that life began with an RNA-based organism that relied on its RNA components to perform all the necessary replication and metabolic functions required to live. Molecular artefacts, such as the ribosome, support this hypothesis. The ribosome is a large RNA/protein complex (RNP) found in all kingdoms of life and is responsible for the synthesis of proteins. It is clear from extensive biochemical and structural studies that the chemistry of peptide bond formation within the ribosome is carried out in an RNA active site. RNA is thus absolutely required to synthesize proteins and must have existed prior to the evolution of proteins themselves.

The group I intron discovered by Tom Cech et al. is able to perform self-splicing reactions, allowing it to excise itself from mRNA and ligate the flanking sequences. Shortly after their discovery, an additional unrelated class of intron was identified, later categorized as group II introns[4]. These particular ribozymes are able to catalyze a self-splicing reaction similar to that of group I introns; however, their end product contains a branched lariat structure [5,6] (Figure 1.1). In group II introns, this unique lariat structure is the result of two sequential transesterification reactions. In the first step of splicing, the 2´ hydroxyl of a bulged adenosine near the 3´ end of the intron is activated as the nucleophile and attacks the phosphate backbone at the 5´ splice site [7,8]. The 3´ hydroxyl of the newly freed 5´ exon is then activated as the nucleophile for the second step, attacking the phosphate backbone at the 3´splice site. The end result is the formation of a 2´-5´ phosphodiester lariat bond as well as ligated exons.

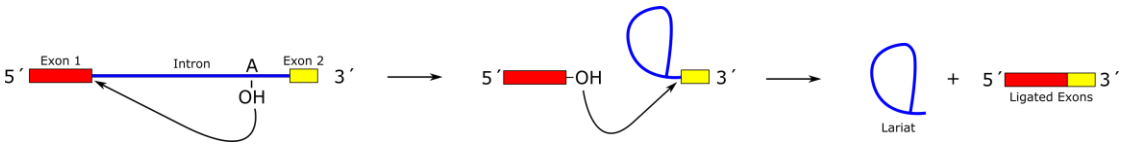


Figure 1.1: **Mechanism of group II intron splicing**. The group II intron undergoes two sequential transesterification reactions to produce ligated exons and excised lariat intron RNA. Exons are denoted by red and yellow while the intron is blue. The bulged adenosine nucleophile is shown.

All group II introns share a characteristic secondary structure and tertiary elements that allow them to fold into a catalytically competent conformation. Group II introns contain six distinct structural domains (Figure 1.2) [9,10]. Domain I (DI) is the largest domain and acts mainly as a folding scaffold. It contains a large number of important tetraloop/tetraloop receptor (TL/TLR) interactions in addition to kissing loops. One primary feature of DI is the exon-binding sequences (EBS). These single-stranded regions within the intron are the reverse complement of the intron binding sequences (IBS) found in the flanking exons. This complementarity allows for base pairing between these two structural elements, which in turn helps to specify the 5´ and 3´ splice sites. Domain II (DII) can vary significantly between introns but contains important tertiary interactions that have been shown to help transition the intron from the 1st step to the 2nd step of splicing. Domain III (DIII), commonly referred to as a catalytic effector, has long been thought to act as an external brace that helps stabilize the active structure of the intron; however, recent structural and biochemical data suggest that this domain is far more dynamic than previously expected and may play a more active role in catalysis (Chapter 3). In many group II introns, Domain IV (DIV) contains an open reading frame (ORF) for a maturase protein. The maturase protein plays an important role in the proper folding and physiological function of group II introns *in vivo*. Domain V (DV) is the catalytic domain, containing the catalytic triad and two-nucleotide bulge (2-nt bulge). These two structural motifs combine to form a triple helix whose phosphate backbone is responsible for forming the active site of the ribozyme [11]. Through precise coordination of catalytic magnesium ions (Mg^{2+}) by DV, the intron is able to perform the chemistry required to undergo splicing. Finally, Domain VI (DVI) contains the ubiquitous bulged adenosine that acts as the nucleophile for the first step of splicing.

In combination, these six domains are able to efficiently fold and precisely determine splice sites for intron excision and exon ligation. Currently available crystal structures of group II introns give insight

into the overall conformation required for splicing [11,12]. They provide a detailed architecture of the splicing active site and reveal that both steps of splicing occur within a single active site, with reactants and products being shuffled as needed.

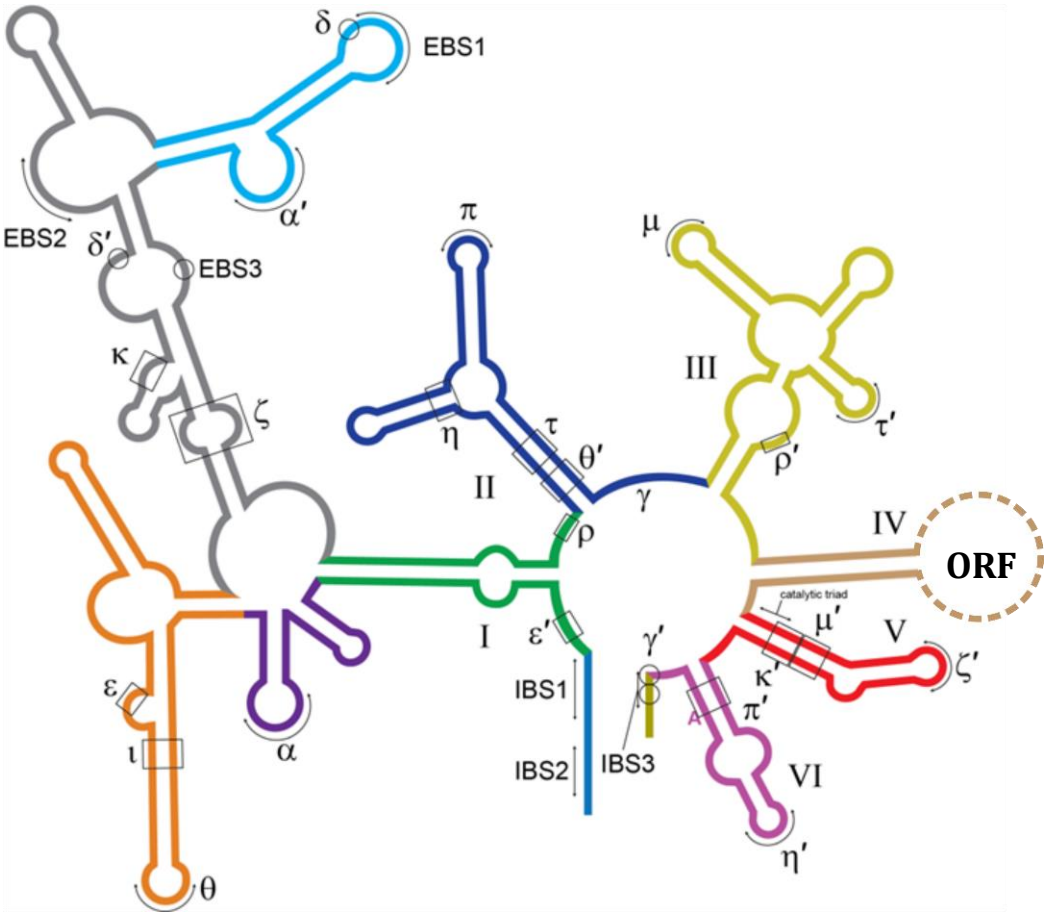


Figure 1.2: **General secondary structure of a group II intron.** Group II introns have a characteristic six domain architecture. Each domain is labeled with a roman numeral (I-VI). Tertiary interactions are labeled with Greek letters. EBS refers to exon binding sequence and IBS refers to intron binding sequence. DIV contains the open reading frame (ORF) for the maturase protein. The position of the catalytic triad is labeled in DV. The bulged adenosine in DVI is also labeled. Courtesy of Jessica K. Peters.

1.2 Group II introns as mobile genetic elements

In certain cases, group II introns act as selfish mobile genetic elements called retrotransposons [5]. This functionality allows group II introns to replicate and insert themselves into double stranded DNA (dsDNA), placing additional copies of themselves in the host genomic DNA.

Essential to this mechanism is the intron-encoded protein called the maturase, which is encoded by an ORF located in DIV [13]. The maturase contains four distinct catalytic domains: a reverse transcriptase domain (RT) [14], a maturase domain (X)/thumb used for RNA binding [15], a DNA binding domain (D), and an endonuclease domain (En) (Figure 1.3) [16-18]. After translation, the protein binds to a unique RNA structure found in DIV, blocking further maturase synthesis. Binding of the intron encoded protein enhances intron self-splicing, producing free, lariat-bound RNPs that function as retrotransposons. Intron mobility begins with the RNP recognizing a suitable target dsDNA sequence via base pairing between the intron EBS sites and the dsDNA target, followed by reverse splicing catalyzed by the RNA. Using the En activity, the bottom strand is then nicked, creating the necessary primer for the reverse transcription of the intron into complementary DNA (cDNA). After ligation and host repair mechanisms, the complete intron sequence is incorporated into a distal portion of the host cell genome. This entire process is broadly referred to as Target Primed Reverse Transcription (TPRT) [19-21]. The structural basis for this process is largely unknown. This is due to the lack of detailed structural information for any of the steps involved in retrotransposition. For example, there is still some debate as to the stoichiometry of the reaction. It is common for RT proteins to act on their substrate as a dimer. In the case of the maturase protein, there is evidence to suggest that dimerization of the maturase is crucial for proper function [22].

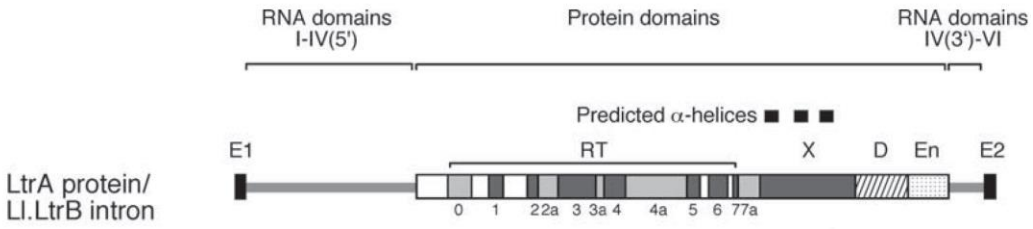


Figure 1.3: **Primary overview of the maturase domains within a group II intron.** The ORF coding the maturase is located within DIV of the intron RNA. The maturase is composed of 4 distinct catalytic domains: reverse transcriptase domain (RT), maturase domain (X)/thumb, DNA binding domain (D), and endonuclease domain (En). The 5´ and 3´ exons are labelled E1 and E2 respectively. Taken from Lambowitz AM and Zimmerly S. 2011 [5].

1.3 Major classes of group II introns

Group II introns are subdivided into three main classes: IIA, IIB, and IIC. These distinct classes are further divided into mitochondrial lineages (IIA1 and IIA2), chloroplast lineages (IIB1 and IIB2) and bacterial lineages (IIA-F) [23,24]. They vary in certain conserved secondary structures and tertiary interaction (Figure 1.4). Class IIA and IIB introns are more structurally complex and some are able to splice efficiently and precisely in the absence of their maturase protein *in vitro*. The current hypothesis for group II intron evolution suggests that all classes originated in bacteria and diverged into the different classes we see today through coevolution with their maturase protein. The appearance of active group II introns in the organelles of lower eukaryotes supports this hypothesis, as these organelles are believed to have originated during the endosymbiont event when an autotrophic prokaryote was endocytosed by a heterotrophic eukaryote [25,26]. Once isolated within the eukaryote, coevolution between the intron RNA and the maturase protein began, leading to the distinct structural classes seen today. The retroelement ancestor hypothesis is supported by the existence of ORF-less group II introns with remnants of RT ORFs, suggesting degeneration from the original bacterial intron[24].

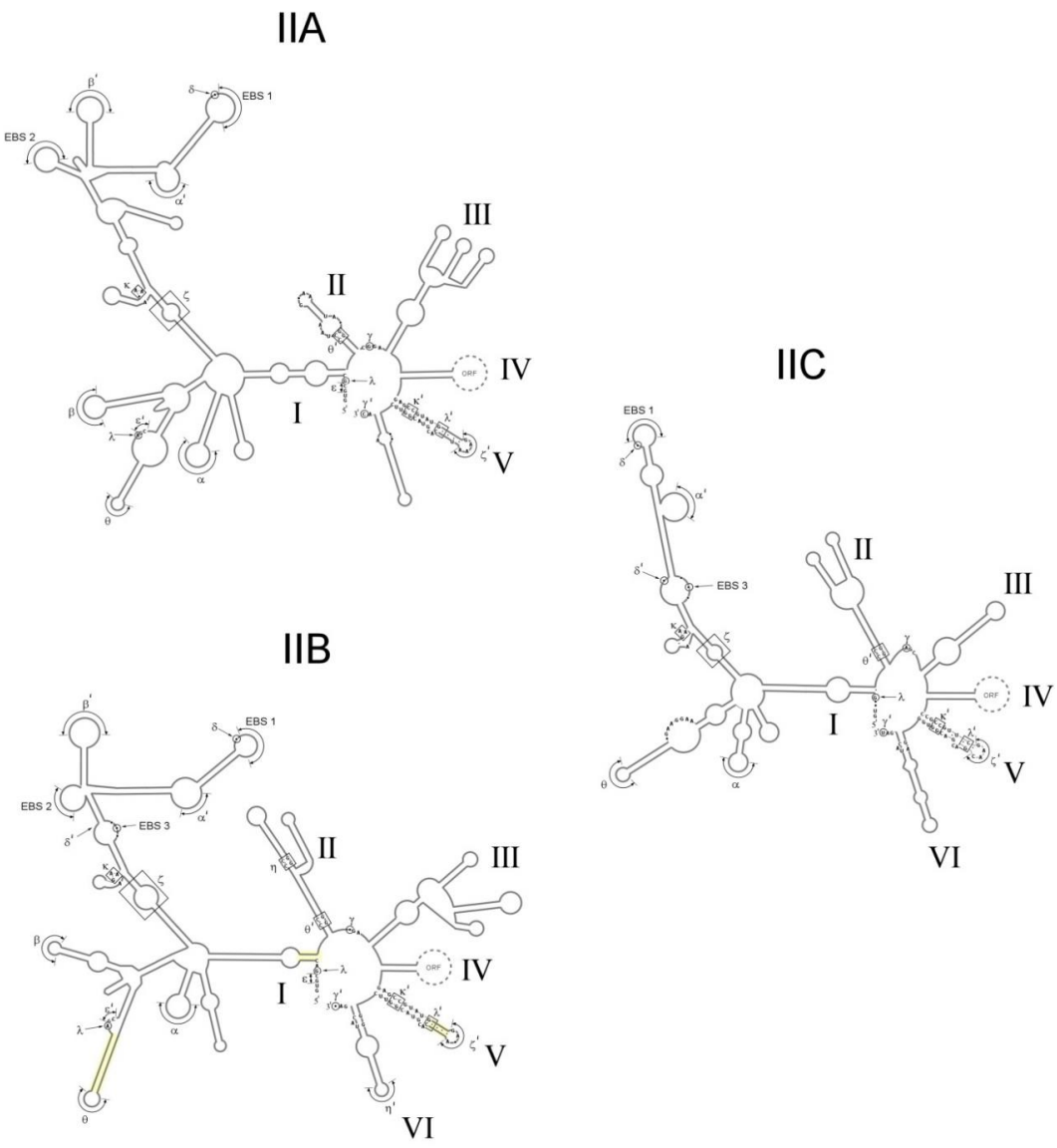


Figure 1.4: **Conserved secondary structure of the three classes of group II introns**. Group II introns are each categorized into one of three main classes. Each class has conserved secondary structure elements and different combinations of EBS-IBS interactions used for splice site selection. In addition, the δ-δ′ interaction, responsible for 3′ splice site selection, varies between classes. Taken from Toor et al. 2009 [24].

1.4 Improving our understanding of the evolution of spliceosomal introns and retrotransposons

It has long been hypothesized that group II introns share a common ancestor with the spliceosome, the multi-megadalton complex required to splice nuclear genes in eukaryotes [4,10,27]. Both share an identical sequential transesterification mechanism along with the unique lariat product (Figure 1.5). There are also conserved secondary structure elements between DV of group II introns and the U2/U6 small nuclear ribonuclear protein (snRNP) found in the spliceosome (Figure 1.6) [28]. Additionally, Mg^{2+} has been shown to be the requisite catalytic metal ion in both active sites [29]. Finally, the bulged adenosine activated as the nucleophile for lariat formation is universally conserved between these systems. Beyond the biochemical data and conserved secondary structures, recent cryo-EM models of the spliceosome have shown there to be structural homology as well [30-33]. When overlaid, the active site architecture of the two splicing machines is identical, providing structural evidence that supports the evolutionary link between group II introns and the spliceosome [34].

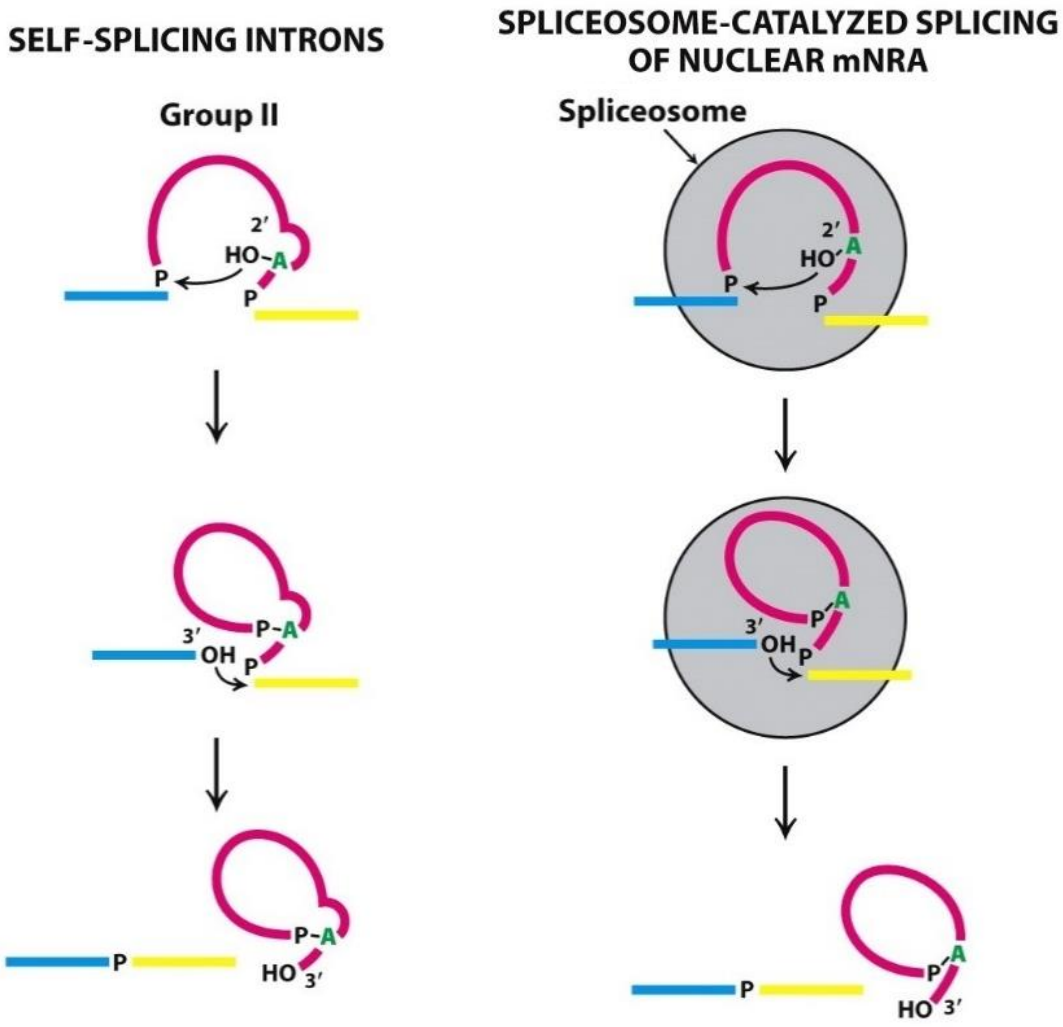


Figure 1.5: **Mechanistic similarities between group II introns and the spliceosome**. Both group II introns and the spliceosome use two sequential transesterification reactions to produce ligated exons and excised lariat RNA. They both use a conserved bulged adenosine as the nucleophile for lariat formation. Exons are depicted in blue and yellow while the intron is shown in pink. The grey circle represents the associated spliceosomal proteins. Taken from Biochemistry, Seventh Edition 2012 W.H. Freeman and Company

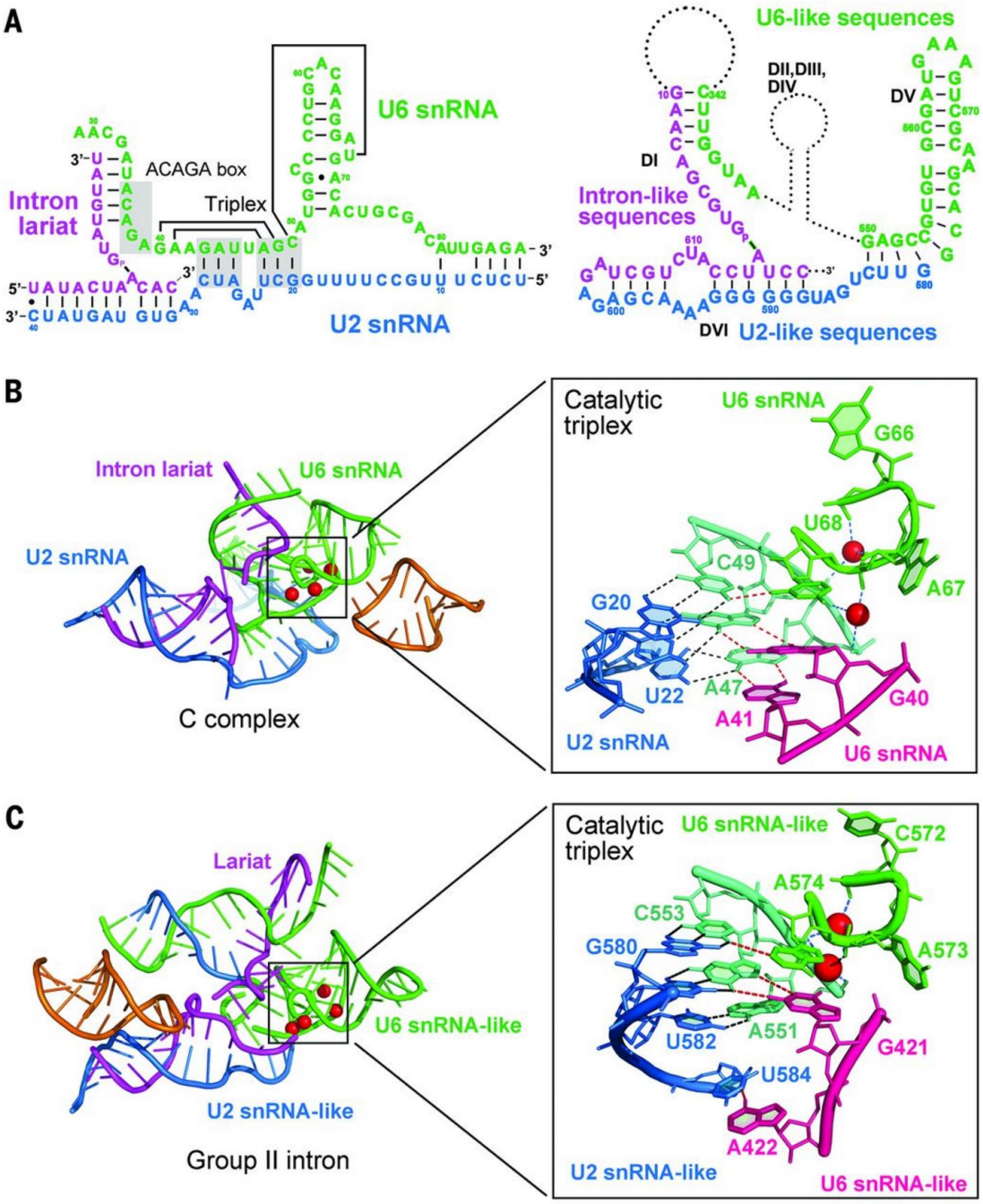


Figure 1.6: **Structural similarities between group II introns and the spliceosome.** A) Secondary structure comparison of the spliceosome (left) and group II intron (right). B) model of the C complex showing the architecture of the splicing active site. C) model of a group IIB intron active site. The catalytic triplex responsible for metal ion binding is nearly identical between the two systems. Taken from Shi et al. 2015 [34].

Group II introns also show surprising similarity to certain retrotransposons, in particular non-long terminal repeat (LTR) retroelements [35,36]. Long interspersed elements (LINEs) exhibit a conserved TPRT mechanism with group II introns and also form an RNP complex between their respective functional proteins and mRNA genes (Figure 1.7). In both TPRT mechanisms, the target DNA is cut by an encoded endonuclease and the encoded RT uses the resulting 3´ hydroxyl to prime cDNA synthesis. One of the most widely studied LINE elements is the R2 element from the silkworm *Bombyx mori* [37]. In this system, the R2 element consists of 5´ and 3´ UTRs interrupted by a single ORF (Figure 1.8). The ORF contains a multifunction protein with nucleic acid binding, reverse transcriptase, and endonuclease activity necessary for TPRT. Sequence homology exists between the maturase protein and the R2 ORF. These mechanism and sequence similarities suggest a common ancestor. Unfortunately, there is currently no structural information for any active retroelement, making detailed structural comparisons impossible.

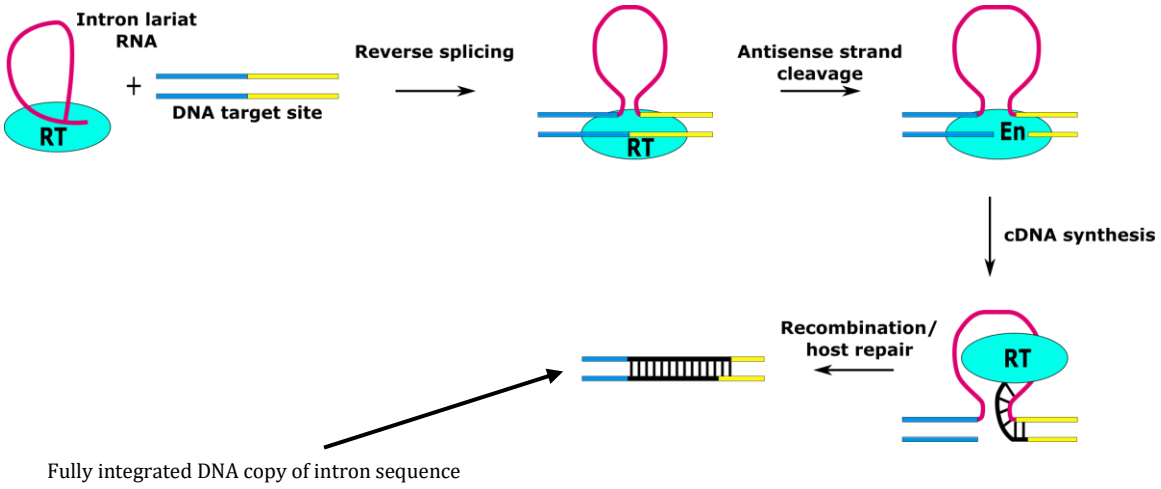


Figure 1.7: **Mechanism of Target Primed Reverse Transcription (TPRT)**. The intron RNA is depicted in pink and the maturase protein is shown in cyan. The target DNA is displayed in blue and yellow representing the dsDNA substrate. RT: reverse transcriptase, EN: endonuclease, cDNA: complementary DNA.

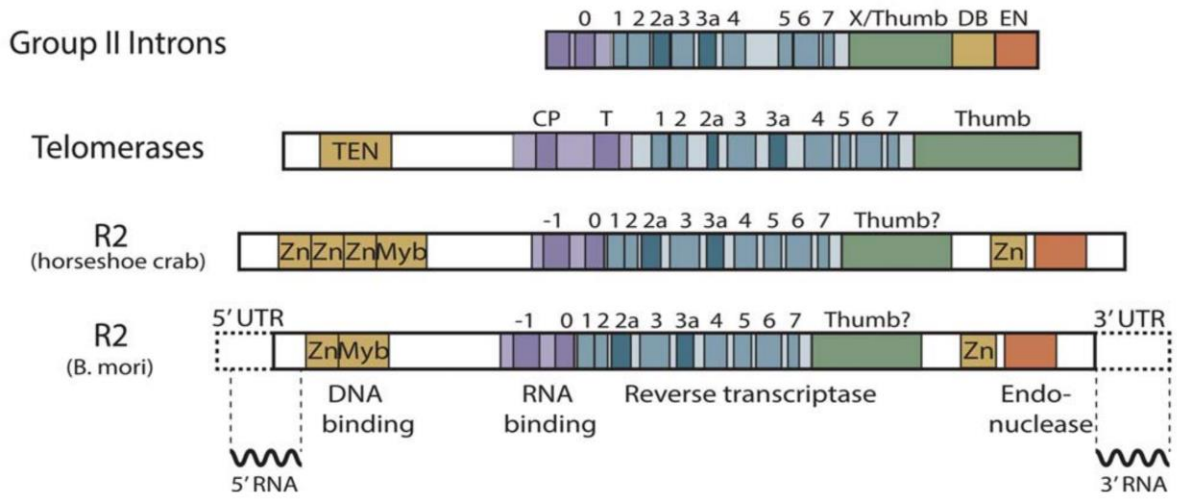


Figure 1.8: **Sequence homology between the group II Intron maturase and non-LTR retroelements.** The domain architecture of the R2 retroelement is identical to the group II intron maturase protein. Both contain an RT domain exhibiting the same motifs (labelled 0-7) followed by a C-terminal EN domain. Taken from Eickbush et al. 2015 [37].

Combined, the retroelement ancestor hypothesis and the endosymbiont event help to clarify the path of evolution both the spliceosome and non-LTR retroelements took in eukaryotes (Figure 1.9). Once endocytosed, the group II introns found in the prokaryotic genome used their mobility mechanism to invade the host eukaryotic genome. These selfish genetic elements then underwent at least one, but possibly many, episodes of proliferation. This likely put a tremendous amount of stress on the genomic stability of the evolving eukaryote. This stress would have created a selective advantage for the degeneration of genetic mobility. The ORFs of most nuclear group II introns therefore degenerated over time, inactivating the TPRT mechanism. They nevertheless maintain their ability to splice efficiently, allowing the transcribed gene to be expressed correctly. These degenerate group II introns evolved into the spliceosomal introns that are spread throughout eukaryotic genomes today; however, a subset of these group II introns located in non-essential regions of the genome did not lose their ability to retrotranspose. They evolved to remain intact retroelements with active RT and En domains, eventually becoming non-LTR retroelements.

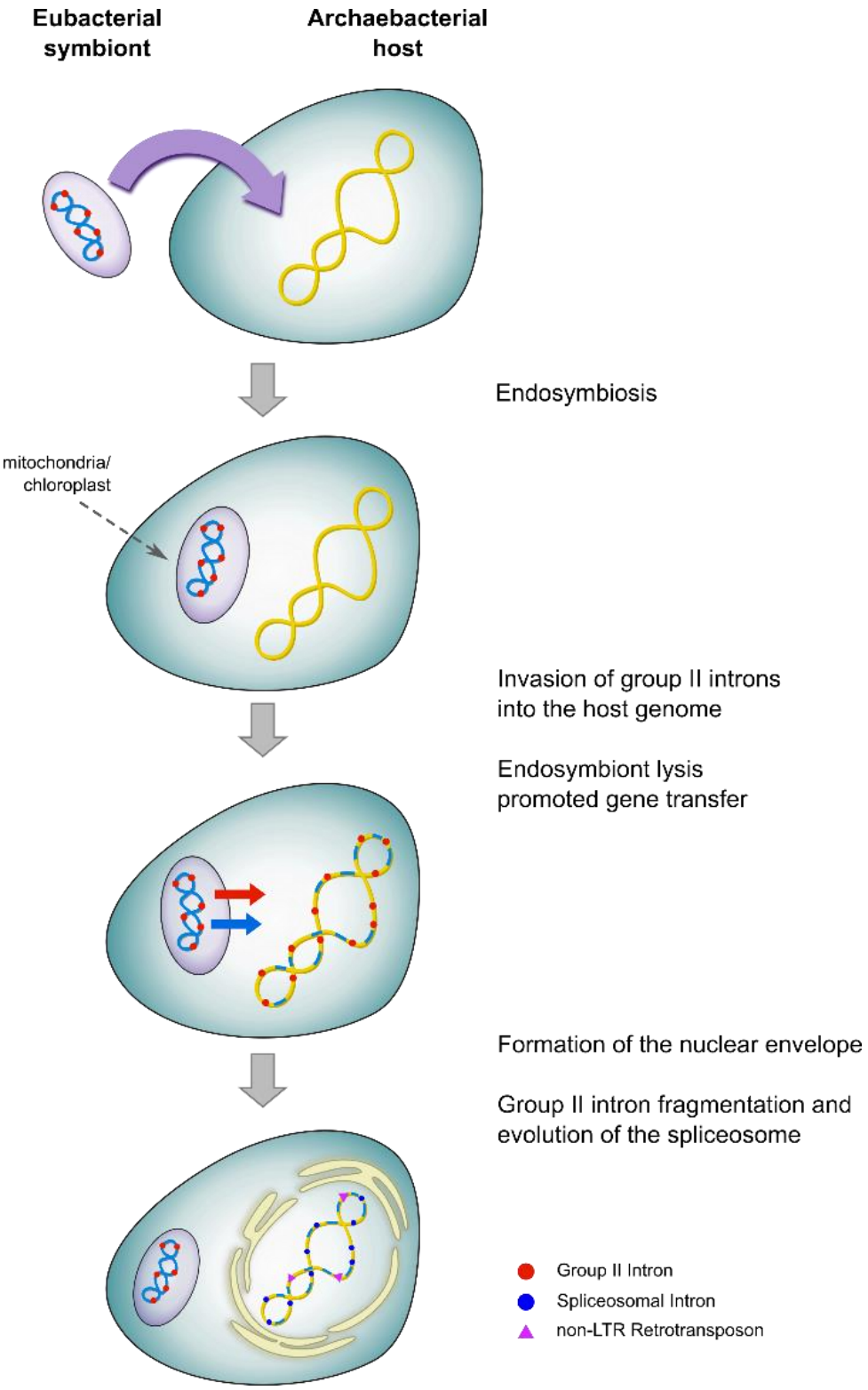


Figure 1.9: **The role of group II introns in the evolution of eukaryotes**. Group II introns are believed to have played a significant role in the evolution of modern-day eukaryotes. After integration into the host genome, they began to proliferate and degenerate into spliceosomal introns and non-LTR retroelements. Courtesy of Jessica K. Peters.

1.5 Significance of group II introns

1.5.1 Splicing

Introns are distributed throughout the genomes of organisms in all three domains of life. Many of these introns disrupt coding regions and, unless removed, cause the translation of aberrant proteins. Even if the intron is removed, if either splice site is chosen incorrectly then the resulting message will have a frame shift leading to errors in gene expression. A vast array of regulatory mechanisms exist in order to ensure that splice sites are chosen correctly; however, the additional complexity conferred by having segmented coding sequences allows eukaryotes to expand their proteome by evolving alternative splicing pathways (Figure 1.10) [38-40]. The ability for transcripts to be processed with the inclusion of different exon sequences has allowed single genes to code for a variety of proteins. Exactly how the spliceosome accomplishes alternative splicing and accurate splice site selection is still largely unknown, but it is understood that incorrect splicing events lead to a variety of diseases.

As previously described, group II introns show strong mechanistic and structural similarities to the spliceosome. Nuclear intron splicing is a very complex and dynamic process. The spliceosome consists of five different small nuclear RNP (snRNPs) particles as well as over one hundred different proteins. In order to splice a nuclear intron, the spliceosome must cycle through an elaborate set of subcomplexes composed of different combinations of the components stated above (Figure 1.11) [41-43]. This coordinated process is difficult to study biochemically and the current cryo-EM structures lack the resolution necessary to reach mechanistic conclusions. On the other hand, group II introns are composed of a single RNA and associated maturase protein. This allows them to be studied more easily and all current detailed structural and biochemical insight into RNA splicing has been derived from group II intron structural biology[11,12,44,45].

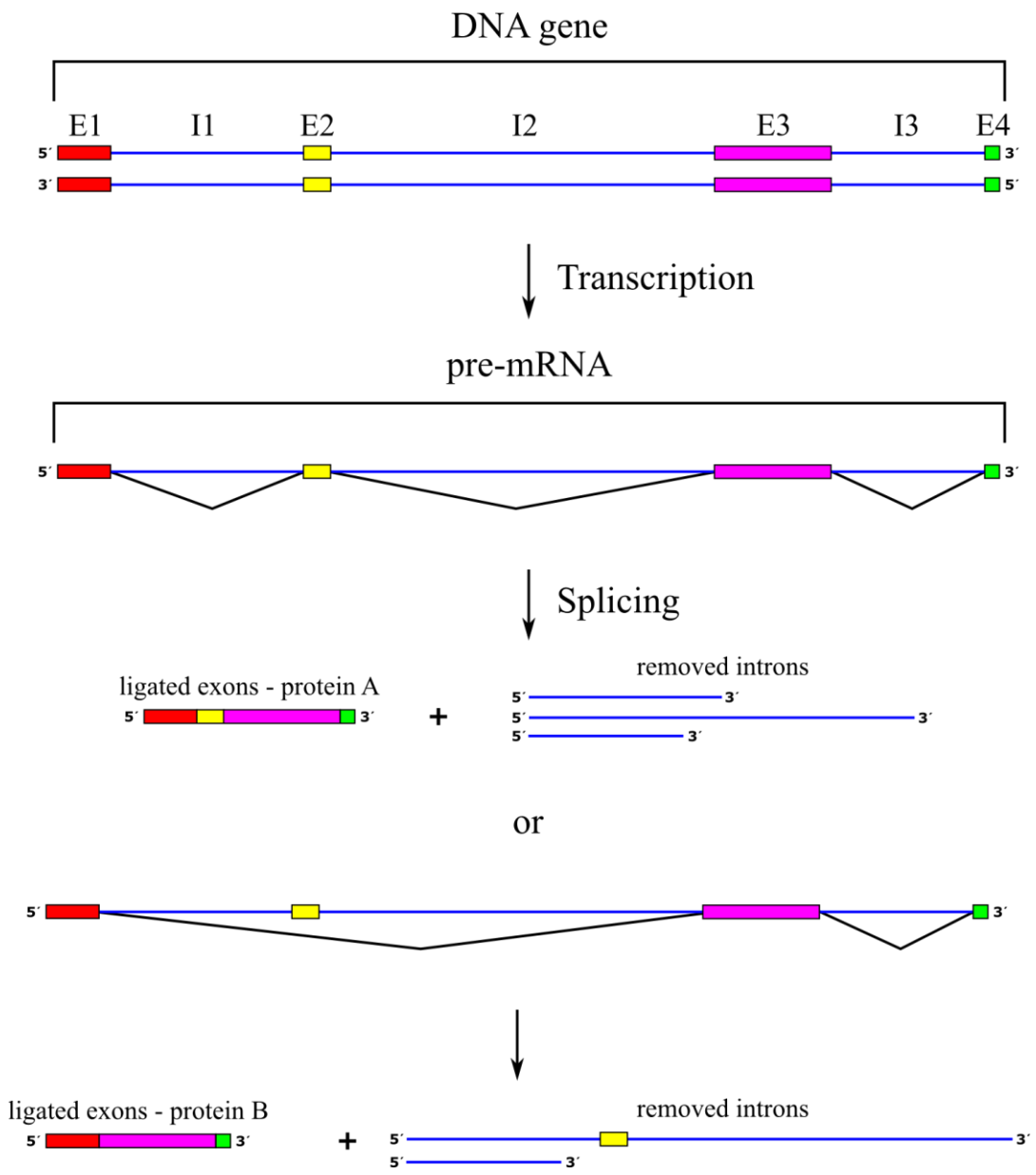


Figure 1.10: **Alternative splicing of nuclear introns**. Spliceosomal introns can be alternatively spliced. In certain cases, some splice sites are silenced or enhanced causing the flanking exons to be excluded or included in the resulting mature-mRNA respectively. This creates a complex proteome in eukaryotes where a single gene can code for multiple proteins.

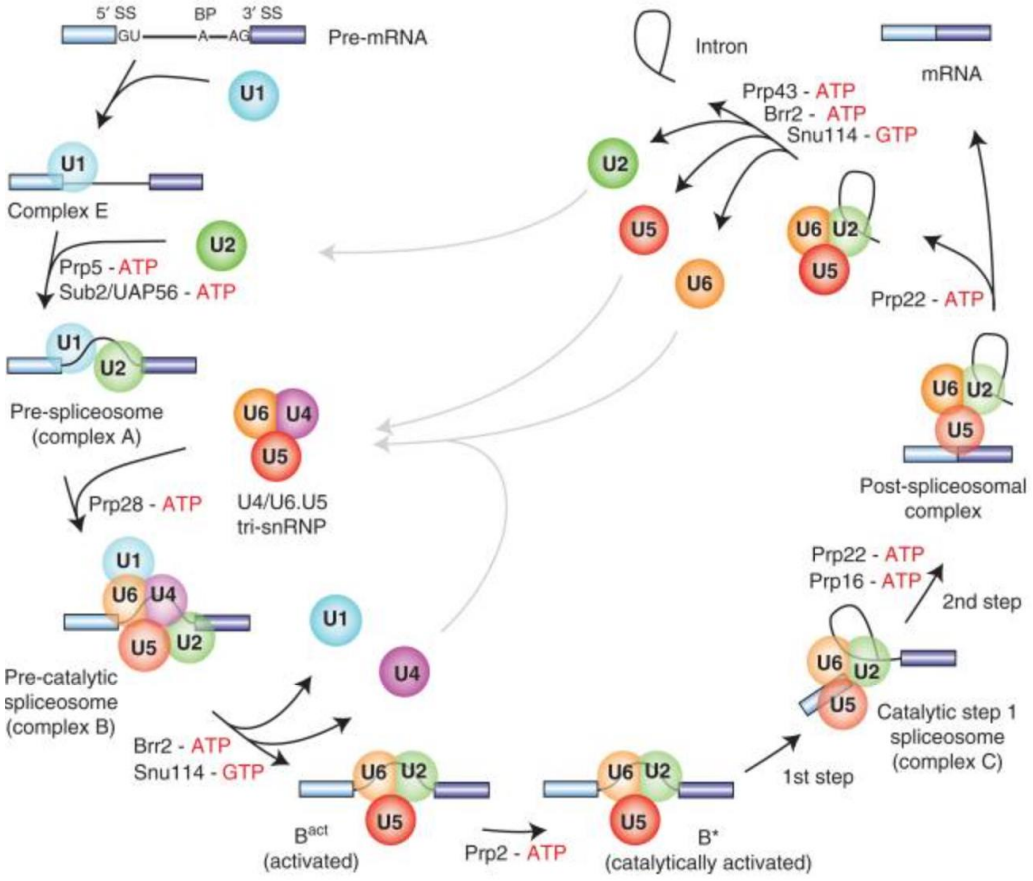


Figure 1.11: **Mechanism of splicing by the spliceosome**. The spliceosome undergoes a complex orchestration of steps to efficiently splice introns. The core of the spliceosome is made up of snRNPs labeled U1, U2, U4, U5, and U6. Throughout the splicing process, different combinations of these subunits assemble along with additional spliceosome associated proteins. Taken from Will et al. 2011 [46].

1.5.2 Mobility

It is estimated that in humans, the non-LTR retroelements L1 and *Alu* make up 28% of the genome by mass [47]. In the case of *Alu*, there are more than 1,000,000 copies distributed throughout the human genome. These prolific genetic elements, which are still actively mobile, undergo novel insertions once every 20 births. They continue to significantly contribute to the genomic diversity seen in humans and play an active role in the creation of novel genes. L1 elements have also been implicated in controlling gene expression in somatic tissue and have been linked with speciation events in primates[48,49]. Retrotransposition events are associated with genetic disease through insertion in coding regions or by causing deleterious double strand breaks (DSBs). It is currently estimated that 0.3% of all mutant phenotypes in humans are caused by non-LTR retrotransposon mobility [47] and lead to severe genetic diseases such as β-thalassemia, cystic fibrosis, and hemophilia (Table 1.1).

Group II introns share the same TPRT mechanism as non-LTR retroelements in addition to strong sequence homology between their RT proteins [37]. This evidence suggests a common evolutionary ancestor. Unfortunately, there is no structural information for any active retroelement and, as a result, the coordination of endonuclease activity, genomic insertion, and cDNA synthesis is not well understood. Due to the significant number of non-LTR retroelements that make up the primate genome, studying the related group II introns could unlock the key to their function and help uncover the role they play in eukaryotic biology.

Table 1.1: **Disease resulting from splicing defects**. This table contains a list of known splicing defects and their associated diseases [50].

Genes	Location	Elements	Diseases	References
FKTN	9q	L1	Fukuyama-type congenital muscular dystophy	(Narita et al., 1993; Kondo-Iida et al., 1999)
DMD	Xp	L1	Duchenne Muscular Dystrophy	(Narita et al., 1993)
APC	5q	L1	Colon cancer	(Mayer et al., 2005)
HBB	11p	L1	Beta-thalassemia	(Kimberland et al., 1999)
RPS6KA3	Xp	L1	Coffin-Lowry syndrome	(Martínez-Garay et al., 2003)
CYBB	Xp	L1	Chronic granulomatous disease	(Meischl et al., 2000)
RP2	Xp	L1	X-linked retinitis pigmentosa	(Schwahn et al., 1998; Ostertag and Kazazian, 2001a)
F9	Xq	L1	Haemophilia B	(Mukherjee et al., 2004)
PDHX	11p	L1	Pyruvate dehydrogenase complex deficiency	(Miné et al., 2007)
EVC, EVC2, C4orf6 and STK32B	4p	L1	Ellis-van Creveld syndrome	(Temtamy et al., 2008)
FAS	10q	Alu	Autoimmune lymphoproliferative syndrom	(Tighe et al., 2002)
F8	Xq	Alu	Haemophilia A	(Ganguly et al., 2003)
F9	Xq	Alu	Haemophilia B	(Vidaud et al., 1993)
CASR	3q	Alu	Hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism	(Janicic et al., 1995)
BRCA2	13q	Alu	Breast cancer	(Miki et al., 1996)
FGFR2	10q	Alu	Apert syndrome	(Oldridge et al., 1999)
GK	Xp	Alu	Glycerol kinase defiency	(Zhang et al., 2000)
OPA1	3q	Alu	Autosomal dominant optic atrophy	(Gallus et al., 2010)
α-galactosidase A	Xq	Alu	Fabry disease	(Kornreich et al., 1990)
HEXB	5q	Alu	Sandhoff disease	(Neote et al., 1990)

1.6 Outline of dissertation

In this dissertation, I discuss my experimental contributions to the field of RNA splicing and retrotransposon mobility. Through the use of X-ray crystallography and cryo-EM, I have investigated group II intron structures in the process of excising themselves from pre-mRNA as well as invading double stranded DNA. In Chapter 2 I describe time-resolved X-ray crystallography experiments, the goal of which was to capture intermediate conformations through the different steps of splicing. This research aims to elucidate important nucleotide rearrangements and potential intermediate catalytic metal ions present in the active site. Chapter 3 serves as an overview of past biochemical experiments and correlates the data to the currently available group II intron crystal structures. By combining the dynamic biochemical data with the static high-resolution structures, new hydrogen bonding networks were identified that assist in the stabilization of the active site. Chapter 4 discusses my work in developing a protocol to purify an active group II intron/maturase complex. This complex was assembled from a recombinant maturase from *E.coli* and *in vitro* transcribed intron RNA. This chapter also discusses the progress of ongoing cryo-EM experiments and current 3D reconstructions of the complex. Chapter 5 examines my work on the purification and structural characterization of an active group II intron retrotransposon in the processes of invading double stranded DNA. This novel structure will give insight into the mechanism of retrotransposition for the first time.

Chapter 2: Time-resolved crystallography of a group IIC intron

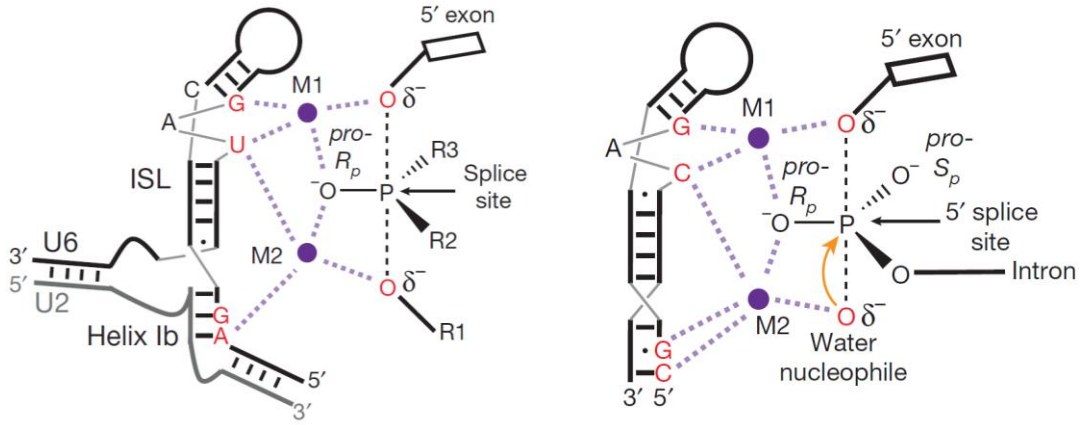
2.1 Determination of a group IIC intron suitable for time-resolved crystallography

Group II introns catalyze splicing via a two-metal-ion mechanism where the divalent metals are spaced approximately 3.9 Å apart [51]. Phosphorothioate recovery experiments and recent crystal structures have confirmed the ideal metal spacing and revealed Mg^{2+} as the ubiquitous catalytic metal (Figure 2.1)[11,12,29]. The general two-metal-ion mechanism is also found at the center of many other biochemical processes from nucleic acid polymerization[52] to the hydrolysis of tRNAs by RNase P [51]. Time-resolved crystallography experiments were performed on DNA polymerase η and were able to capture several structural intermediates[53]. These intermediates were then stitched together to create a molecular movie of phosphodiester bond formation for the first time. One of the major findings of this experiment was the presence of a transient, third Mg^{2+} ion that participates in the chemical reaction by replacing R61 of the polymerase and stabilizes the newly formed phosphodiester bond (Figure 2.2). In fact, additional metal ions have been found and hypothesized in many different two-metal-ion mechanisms [54]. The scope of this research was to use time-resolved crystallography to probe the active site of splicing and determine whether an additional metal is present and playing a similar role to that observed in DNA polymerase η (Figure 2.3).

The initial hurdle was to determine an appropriate intron candidate for the crystallographic experiments. The first requirement is to have an intron crystallized in the pre-catalytic state at sufficient resolution to unambiguously determine metal ion positions within the active site. Secondly, the intron must then be capable of being activated for splicing *in crystallo* and stopped at various time points. To address the first requirement I turned to a pre-determined crystal structure of a pre-catalytic group IIC intron. The intron was discovered in the alkaliphile *Oceanobacillus iheyensis*. This particular bacterial group IIC intron splices through hydrolysis in the absence of its maturase protein [11]. This differs from the canonical splicing mechanism because a water molecule is activated as the nucleophile for the first stepof splicing leading to a linear excised intron as apposed to the lariat product formed when the bulged adenosine is used. Since the bulged adenosine in DVI does not play a role in the hydrolytic splicing

mechanism, the entire domain was removed from the RNA construct used to obtain the pre-catalytic crystals. This intron construct will be referred to as OiΔDVI.

Unfortunately, in order to capture the pre-catalytic state of OiΔDVI, a G to A mutation was made at the central nucleotide of the catalytic triad rendering the RNA catalytically dead [44]. This mutation would preclude my ability to activate splicing *in crystallo* but the intron in general shows a proclivity for crystallization. Therefore, OiΔDVI was chosen as the intron moving forward for crystallization experimetns.



U2/U6 of the spliceosome **Group IIC intron**

Figure 2.1: **Phosphorothioate rescue experiments**. Left) results of phosphorothioate experiments done on the spliceosome. Nucleotides labelled in red showed rescued splicing activity with Mn^{2+}. Purple spheres represent the catalytic Mg^{2+} ions. Right) results of phosphorothioate experiments on a group IIC intron. An identical catalytic metal ion binding architecture is observed between the two systems. Taken from Fica et al. 2013[29].

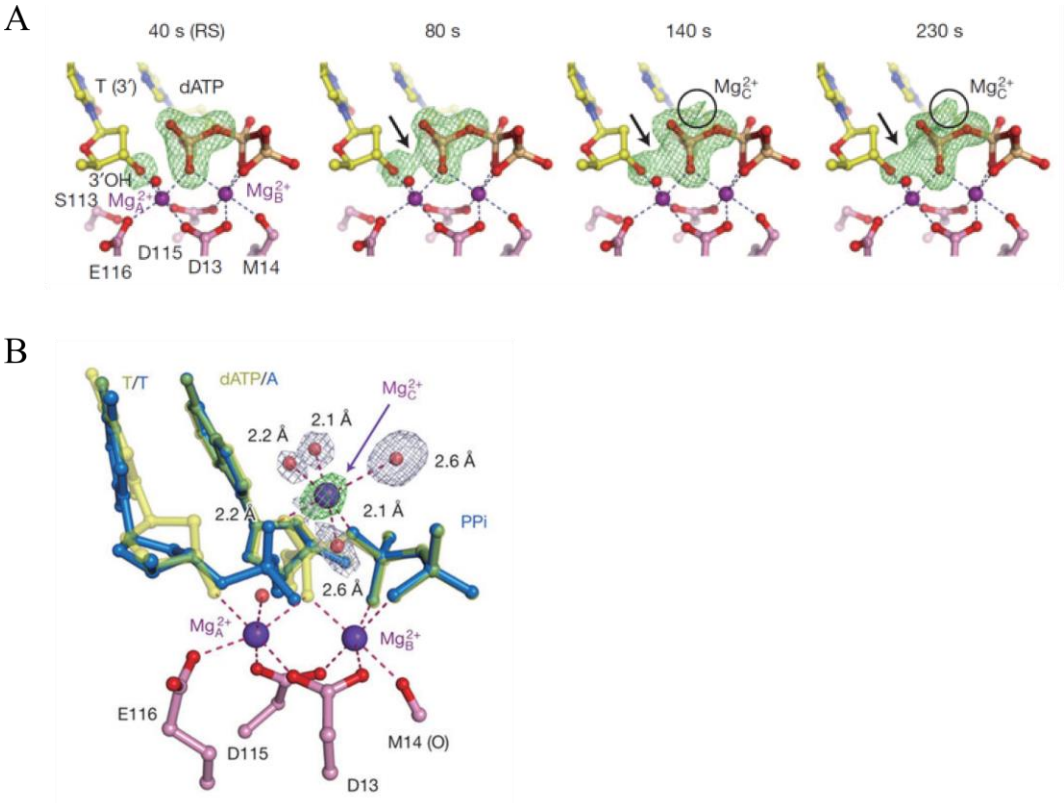


Figure 2.2: **Time-resolved crystallography of DNA polymerase η**. A) Green mesh density showing the appearance of new density at the 230-second time point corresponding to a third catalytic Mg^{2+} ion. B) Zoomed view of Mg_c^{2+} density. At 2.1 Å resolution, the hydration sphere is clearly visible. Taken from Nakamura et al. 2012[53].

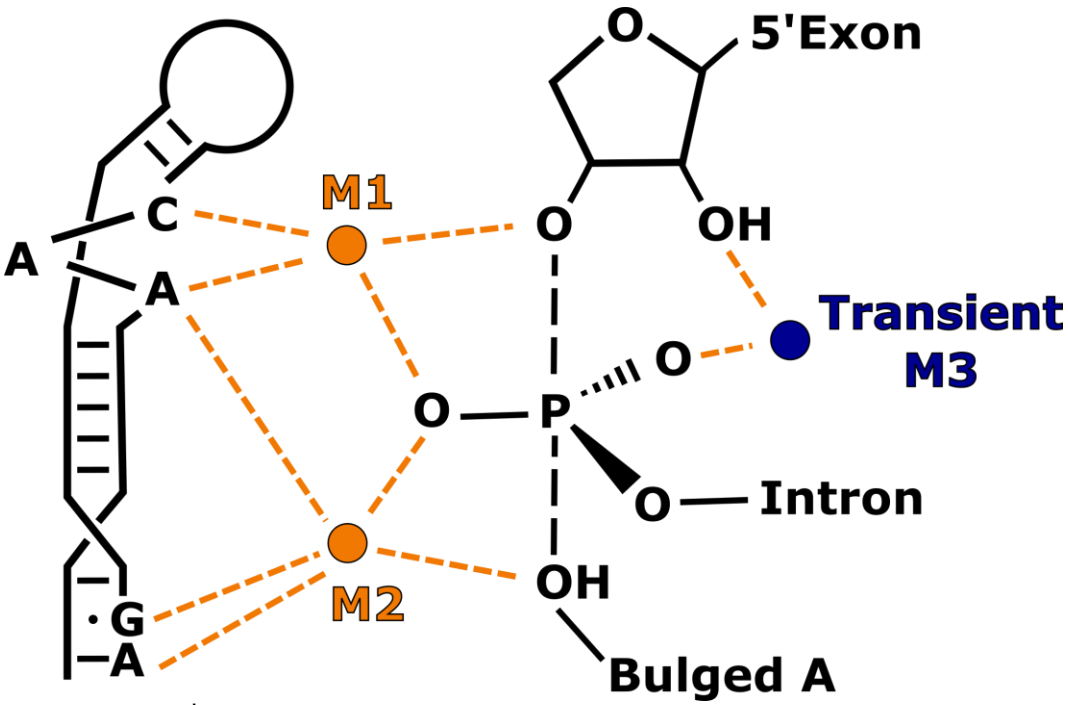


Figure 2.3: **Potential 3rd catalytic metal ion in the group II intron active site**. RNA stem represents DV of group II intron showing the nucleotides that participate in catalytic metal ion binding. Ubiquitous Mg^{2+} ions are shown in orange with the transient Mg^{2+} ion shown in blue.

2.2 Development of a Ca^{2+} crystallization screen

Obtaining a pre-catalytic RNA crystal of OiΔDVI through mutation was not an option for my proposed time-resolved crystallography experiments. Work done by Marcia M. et al showed that RNA crystallized in the absence of Mg^{2+} but in the presence of Ca^{2+} could inhibit splicing while still providing the divalent metal ions required for the RNA to fold properly [45]. Ca^{2+} is a well-documented inhibitor of splicing [55]. In fact, their work had already determined a crystallization condition that provided pre-catalytic crystals of sufficient quality for my needs. Unfortunately, when trying to reproduce their results I obtained only poorly diffracting crystals (approx. 7-8 Å) (Figure 2.4). This resolution was insufficient to evaluate active site density; therefore I began the process of construct optimization and precipitant condition screening.

Knowing that Ca^{2+} is able to adequately replace structural Mg^{2+} but inhibit splicing, I developed a crystallization screen utilizing Ca^{2+} (Table 2.1). Hampton Research develops and manufactures many

different crystallization screens. Their Natrix screens were developed and optimized specifically to promote nucleation of RNA [56]. Therefore, using Natrix I screen as inspiration, I created a Ca^{2+} based screen of 48 unique conditions by simply replacing Mg^{2+} with the corresponding calcium based salt. Using this Ca^{2+} screen, I was able to initiate crystal trials.

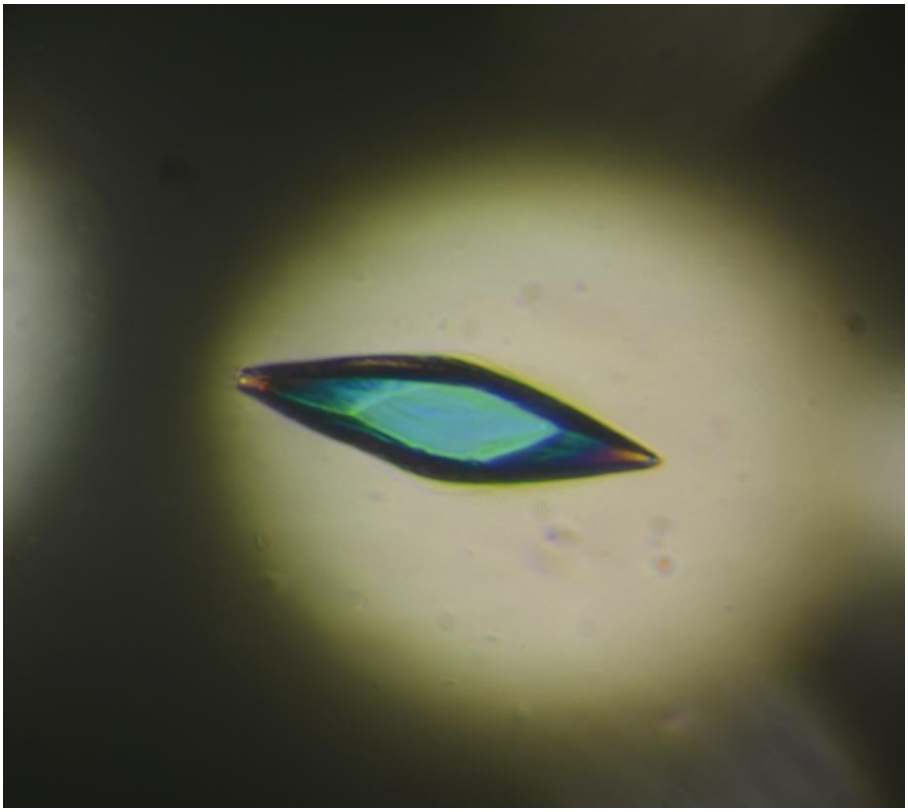


Figure 2.4: **Initial pre-catalytic crystal form**. Crystals were prepared in 0.1 M KOAc, 0.1 M KCl, 0.1 M $CaCl_2$, 0.05 M HEPES-Na pH 7.0, 3% PEG 8,000. Crystals diffracted in a range of 7-8 Å.

Table 2.1: **Ca^{2+} based crystallization screen.** Calcium containing precipitant condition were prepared as shown above. Column 3 indicates what conditions formed RNA crystals. Column 4 indicates the RNA construct and diffraction resolution.

Sol. #	Precipitant solution composition	Crystals (Y/N)	Construct/ Resolution
1	0.01 M $CaCl_2$, 0.05 M MES pH 5.6, 1.8 M Li_2SO_4	Y	WT-3.8 Å /TNC5-3.2 Å
2	0.01 M $Ca(OAc)_2$, 0.05 M MES pH 5.6, 2.5 M $(NH_4)_2SO_4$	Y	TNC4-10 Å /TNC5-3.7 Å
3	0.1 M $Ca(OAc)_2$, 0.05 M MES pH 5.6, 20% v/v (+/-)-2-Methyl-2-4-pentanediol	N	-
4	0.2 M KCl, 0.01 M $CaSO_4$, 0.05 M MES pH 5.6, 10% w/v PEG 400	Y	TNC5-8 Å /TNC6-20 Å
5	0.2 M KCl, 0.01 M $CaCl_2$, 0.05 M MES pH 5.6, 10% w/v PEG 8,000	Y	WT-12 Å
6	0.1 M $(NH_4)_2SO_4$, 0.01 $CaCl_2$, 0.05 M MES pH 5.6, 20% w/v PEG 8,000	Y	TNC5-3.8 Å
7	0.02 M CaC_{l2}, 0.05 M MES pH 6.0, 15% v/v 2-Propanol	N	-
8	0.1 M NH_4OAc, 0.005 M $CaSO_4$, 0.05 M MES p 6.0, 0.6 M NaCl	N	-
9	0.1 M KCl, 0.01 M $CaCl_2$, 0.05 MES pH 6.0, 10% w/v PEG 400	Y	TNC1-15 Å
10	0.005 M $CaSO_4$, 0.05 M MES pH 6.0, 5% w/v PEG 4,000	Y	TNC3-11 Å
11	0.01 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 1.0 M Li_2SO_4	N	-
12	0.01 M $CaSO_4$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 1.8 M $Li2_SO_4$	Y	TNC5-15 Å
13	0.015 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 1.7 M $(NH_4)_2SO_4$	N	-
14	0.1 M KCl, 0.025 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 15% v/v 2-Propanol	N	-
15	0.04 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 5% v/v (+/-)-2-Methyl-2-4-pentanediol	Y	WT-ND /TNC1-8 Å
16	0.04 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 30% v/v (+/-)-2-Methyl-2-4-pentanediol	N	-
17	0.2 KCl, 0.01 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.0, 10% w/v PEG 4,000	Y	WT-10 Å /TNC1-10 Å
18	0.01 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 1.3 M Li_2SO_4	N	-
19	0.01 M $CaSO_4$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 2.0 M $(NH_4)_2SO_4$	Y	WT-3.2 Å
20	0.1 M NH_4OAc, 0.015 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 10% v/v 2-Propanol	Y	TNC2-ND /TNC3-10 Å
21	0.2 M KCl, 0.005 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 0.9 M 1,6-Hexanediol	N	-
22	0.08 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 15% w/v PEG 400	N	-
23	0.02 M KCl, 0.01 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 10% w/v PEG 4,000	Y	WT-8 Å
24	0.02 M NH_4OAc, 0.01 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 10% w/v PEG 4,000	Y	WT-7.5 Å /TNC1-5 Å /TNC5-3.5 Å
25	0.08 M $Ca(Oac)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 30% w/v PEG 4,000	N	-
26	0.2 M KCl, 0.1 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 10% w/v PEG 8,000	N	-
27	0.2 M NH_4OAc, 0.01 M $Ca(OAc)_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 30% w/v PEG 8,000	N	-
28	0.05 M $CaSO_4$, 0.05 M Na-HEPES pH 7.0, 1.6 M Li_2SO_4	Y	TNC5-3.4 Å
29	0.01 M $CaCl_2$, 0.05 M Na-HEPES pH 7.0, 4.0 M Li_2SO_4	Y	WT-3.9 Å /TNC1-3.2 Å
30	0.01 M $CaCl_2$, 0.05 M Na-HEPES pH 7.0, 1.6 M $(NH_4)_2SO_4$	N	-
31	0.005 M CaC_{l2}, 0.05 M Na-HEPES pH 7.0, 25% v/v PEG monethyl ether 550	Y	TNC3-ND
32	0.2 M KCl, 0.01 M $CaCl_2$, 0.05 M Na-HEPES pH 7.0, 1.7 M 1,6-Hexanediol	N	-
33	0.2 M NH_4Cl, 0.01 M $CaCl_2$, 0.05 M Na-HEPES pH 7.0, 2.5 M 1,6-Hexanediol	N	-

Table 2.1 continued.

Sol. #	Precipitant solution composition	Crystals (Y/N)	Construct/ Resolution
34	0.1 M KCl, 0.005 M $CaSO_4$, 0.05 M Na-HEPES pH 7.0, 15% v/v (+/-)-2-Methyl-2-4-pentanediol	Y	TNC2-11 Å
35	0.1 M KCl, 0.01 $CaCl_2$, 0.05 M Na-HEPES pH 7.0, 5% v/v PEG 400	Y	TNC4-11 Å
36	0.1 M KCl, 0.01 CaCl2, 0.05 M Na-HEPES pH 7.0, 10% v/v PEG 400	N	-
37	0.2 M KCl, 0.025 $CaSO_4$, 0.05 M Na-HEPES pH 7.0, 20% v/v PEG 200	Y	WT-ND /TNC2-17 Å
38	0.2 M NH_4OAc, 0.15 M $Ca(OAc)_2$, 0.05 M Na-HEPES pH 7.0, 5% w/v PEG 4,000	Y	WT-ND
39	0.1 M NH_4OAc, 0.02 M $CaCl_2$, 0.05 M Na-HEPES pH 7.0, 5% w/v PEG 8,000	Y	WT-ND
40	0.01 M $CaCl_2$, 0.05 M Tris-HCl pH 7.5, 1.6 M $(NH_4)_2SO_4$	N	-
41	0.1 M KCl, 0.015 M $CaCl_2$, 0.05 M Tris-HCl pH 7.5, 10% v/v PEG monethyl ether 550	Y	TNC3-15 Å
42	0.01 M $CaCl_2$, 0.05 M Tris-HCl pH 7.5, 5% v/v 2-Propanol	N	-
43	0.05 M NH_4OAc, 0.01 $CaCl_2$, 0.05 M Tris-HCl pH 7.5, 10% v/v (+/-)-2-Methyl-2-4-pentanediol	N	-
44	0.2 M KCl, 0.05 M $CaCl_2$, 0.05 M Tris-HCl pH 7.5, 10% w/v PEG 4,000	Y	TNC3-19 Å
45	0.025 M $CaSO_4$, 0.05 M Tris-HCl pH 8.5, 1.8 M $(NH_4)_2SO_4$	N	-
46	0.005 M $CaSO_4$, 0.05 M Tris-HCl pH 8.5, 2.9 M 1,6-Hexanediol	N	-
47	0.1 M KCl, 0.01 M $CaCl_2$, 0.05 M Tris-HCl pH 8.5, 30% v/v PEG 400	N	-
48	0.2 NH_4Cl, 0.01 $CaCl_2$, 0.05 M Tris-HCl pH 8.5, 30% w/v PEG 4,000	N	-

2.3 Identification of a high-resolution pre-catalytic crystal form

Transcribing the intron in the pre-catalytic state was also a concern. If the RNA was already spliced when exposed to the crystallization screen, the Ca^{2+} screen would not reverse the reaction and post-catalytic crystals would form. To ensure the RNA was in the pre-catalytic state, a modified *in vitro* transcription protocol was developed. In this protocol, the Mg^{2+} in the transcription buffer is lowered to the minimum amount required for efficient RNA synthesis. This reduces the amount of free Mg^{2+} in solution that can be used by the intron for catalysis. In addition, the RNA is stored in 10 mM Ca^{2+} once transcribed to further ensure no splicing takes place before setting up crystal trials.

In addition to preparing a calcium-based crystallization screen, RNA construct variation was also attempted. The basis of construct variation involves the incorporation of individual tetraloop/tetraloop receptor (TL/TLR) stems into the RNA of interest. These two loops then specifically interact with one another, potentially creating a high-resolution crystal contact [57]. The position of the TL/TLR in the intron can be varied as well as the length of the stems into which they are inserted. Shortening or lengthening of the stem not only changes its distance from the intron core, but because of the helical twist associated

with an A-form helix, it also changes the face of the TL and TLR that is presented to form the crystal contact. Once a combination of TL length and TLR length produces RNA crystals, an additional round of construct variation can be employed to improve the resolution. The incorporation of tandem non-canonical base pairs in the TL/TLR stems causes micro adjustments to the architecture of the engineered crystal contacts [58]. These small structural changes can lead to stabilization of the contact and improved crystal packing.

In vitro transcribed RNA of the different constructs were then subjected to the calcium-based crystal screen using the sitting drop method. Spermine is added to each drop to help crystal nucleation and each construct was placed at 22°C and 30°C. After 3 days at 30°C and two weeks at 22°C, large crystals formed. Of the six different RNA constructs tested in the 48 different precipitant conditions, 33 wells had RNA crystals. The crystals formed in all precipitant types including: high salt (Li_2SO_4, NH_4Cl, NaCl), polyethylene glycol (400, 4,000, 8,000), MPD, and hexanediol (Table 2.1). This variety in precipitant was a good sign that well ordered crystals were present. All crystal-containing conditions were cryo-protected and frozen in liquid nitrogen. They were then assessed for diffraction quality at the Argonne synchrotron X-ray source (NE-CAT).

Of the 33 different crystal forms, 9 diffracted to better than 4 Å resolution. Molecular replacement of the diffraction data of the crystal forms provided density showing an intact 5´ splice site (Figure 2.6). The presence of an intact splice site confirmed the RNAs were in fact in the pre-catalytic state. The most reproducible crystals formed in Solution 1 (0.01 M $CaCl_2$, 0.05 M MES pH 5.6, and 1.8 M Li_2SO_4) (Figure 2.5). These crystals were chosen as the best candidate moving forward for time-resolved crystallography.

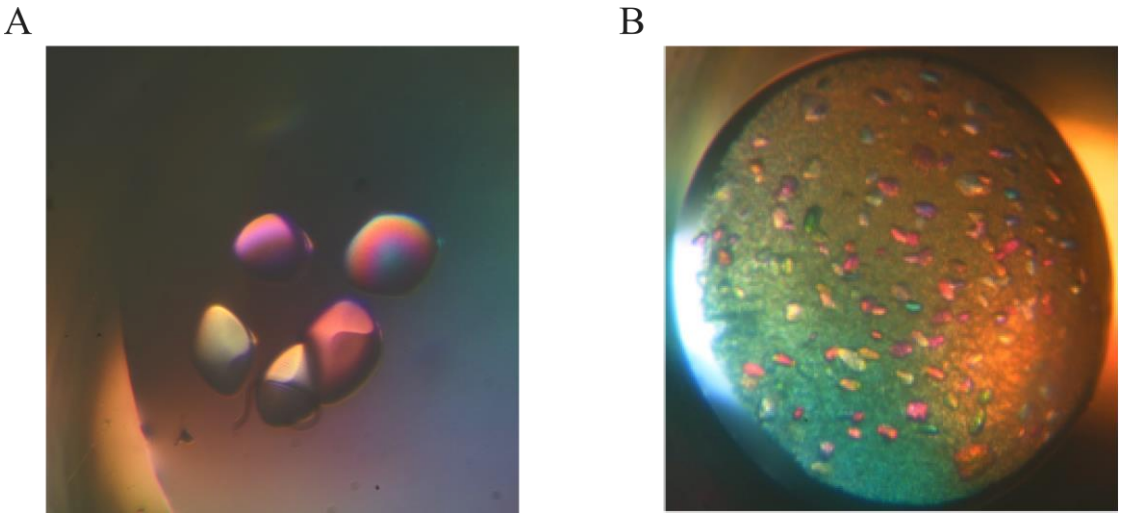


Figure 2.5: **High-resolution pre-catalytic crystal forms**. A) Crystals obtained in 0.01 M $CaCl_2$, 0.05 M MES pH 5.6, and 1.8 M Li_2SO_4. B) Crystals obtained in 0.02 M NH_4OAc, 0.01 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 10% w/v PEG 4,000.

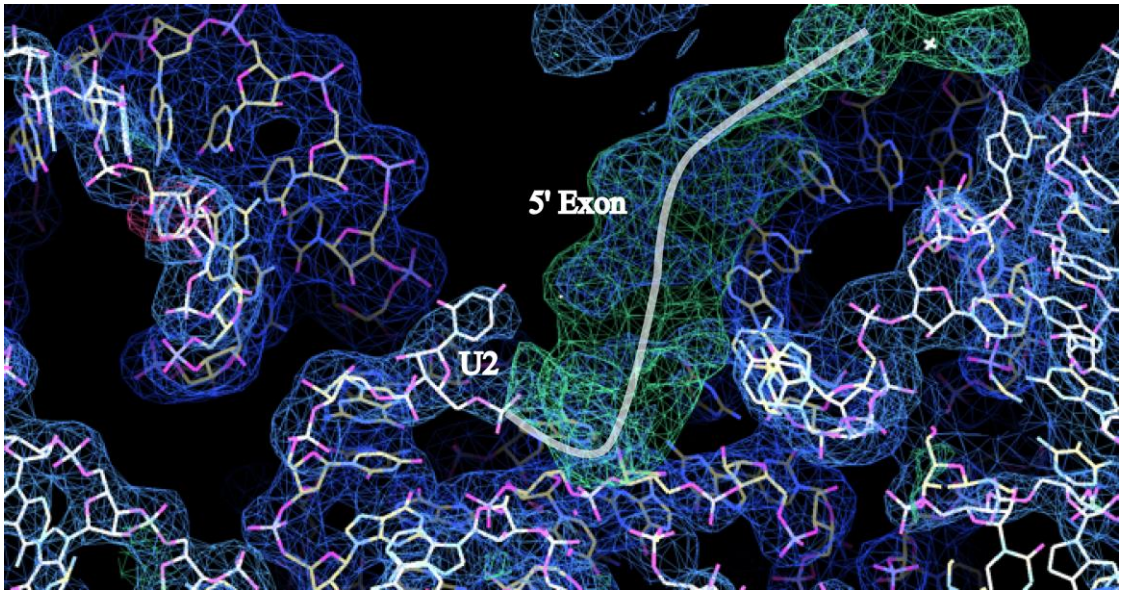


Figure 2.6: **Pre-catalytic density of OiΔDVI**. A molecular replacement omit map showing the 5´ splice site of OiΔDVI. U2 is labeled as well as the density corresponding to the 5´ exon. Green density represents density present in the experimental data that is absent from the model used for molecular replacement. Continuous density is observed in the experimental data for the 5´ scissile phosphate and highlighted by a white line. An intact 5´ splice site indicates the RNA is in the pre-catalytic state.

2.4 Time-resolved crystallography trials and future directions

With pre-catalytic crystals of sufficient quality in hand, I was able to initiate time-resolved crystallography experiments. These experiments involve the activation and controlled termination of splicing at various time points in the attempt to capture the intron at different stages of splicing. To activate splicing, the crystallization solution of interest is replaced by one containing Mg^{2+}, while all other components of the screen remain constant. The initial crystallization condition tried was 0.01 M $CaCl_2$, 0.05 M MES pH 5.6, and 1.8 M Li_2SO_4. This condition yielded 200-600 µm crystals that diffracted in a range of 3.5-4.0 Å. The crystals were transferred to a new and well and splicing buffer was added (0.1 M $MgCl_2$, 0.5 M MES pH 5.6, 1.8 M Li_2SO_4). The crystals were placed at 22°C and 30°C for 1 min, 2 min, 5 min, 10 min, 30 min, 1 hr, 2 hr, 7 hr, etc. and splicing was halted by freezing the crystals in liquid nitrogen. The frozen crystals were then taken to a synchrotron and full diffraction data sets were collected.

After performing molecular replacement of the data sets, the resulting density was assessed to identify any changes to the RNA or metal ions. After close inspection, it was determined that all time points contained intron in the pre-catalytic state. Our lab has previously shown that high salt precipitant solutions inhibit the binding of metal ions in the active site of the intron. This phenomenon makes activation of splicing *in crystallo* impossible given the high salt nature of the crystallization solution chosen. In an attempt to continue, a new PEG containing crystallization condition was selected (Solution 24 – 0.02 M NH_4OAc, 0.01 M $CaCl_2$, 0.05 M $Na(CH_4)_2AsO_2$ pH 6.5, 10% w/v PEG 4,000). This condition was not initially chosen due to crystal quality variability; however, the higher resolution range was adequate for my experimental needs (3.8-8 Å) (Figure 2.5). This new condition was exposed to the same time point scheme as before and evaluated for diffraction quality (splicing buffer: 0.2 M NH_4OAc, 0.15 M $Mg(OAc)_2$, 0.05 M Na-HEPES pH 7.0, 5% w/v PEG 4,000). After molecular replacement, post-catalytic density was observed at the 1 hr time point, with an obviously cleaved 5´ splice site (Figure 2.7). Additionally, metal ion density was observed in the active site, further confirming the efficient diffusion of Mg^{2+} into the intron core. Unfortunately, all intermediate time points had poorly diffracting crystals. It was observed that all intermediate crystals were cracked as the diffusion of Mg^{2+} and the activation of splicing causes a disruption to the crystal packing. This makes determining intermediate structures difficult using time-resolved crystallography.

To overcome the limitations set by Mg^{2+} diffusion and splicing activation, X-ray free electron laser (XFEL) diffraction experiments will be performed [59]. In this methodology, small microcrystals are passed as a stream through an electron laser. Many thousands of diffraction images are taken of single microcrystals that are then computationally combined into a data set for analysis. The benefit of this method is the use of microcrystals. Having small crystals removes the rate of diffusion effects and allows the entire crystal to transition through its mechanism instantaneously, minimizing the effects of lattice cracking. The use of XFEL breathes new life into this work and will be attempted to continue investigating the conformational rearrangements of intron splicing.

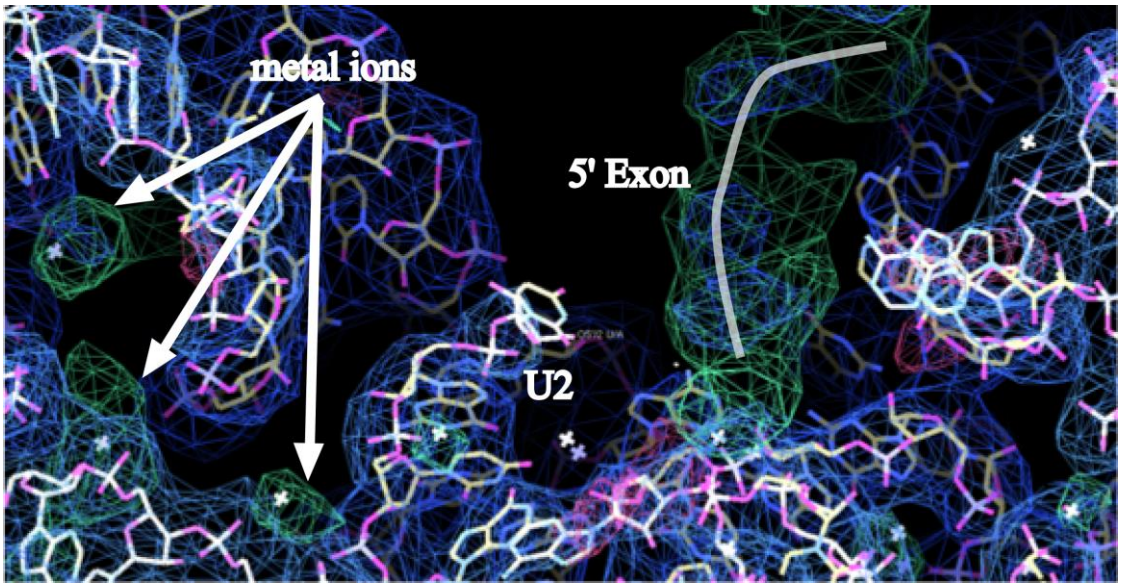


Figure 2.7: **Post-catalytic OiΔDVI density**. An omit map of the 1 hr time point. Green density for metal ions is clearly visible; however, density for the scissile phosphate is absent. Discontinuous density indicates 5´ splice site hydrolysis.

2.5 Materials and methods

Cloning, mutagenesis, and RNA preparation

OiΔDVI DNA was synthesized (Genscript) and cloned into a pUC57 vector using the EcoRV restriction site. All TNC mutant introns were prepared using overlapping PCR and the resulting products were cloned into pUC57 using the EcoRV restriction site. The cloned plasmid was transformed into DH5α cells. Plasmid DNA was isolated using a Qiagen maxi-prep kit. The recovered DNA was linearized using BamHI and RNA synthesis was performed using an *in vitro* transcription protocol with T7 polymerase. Transcriptions were prepared in a 1 mL total volume. 40 µg of linear DNA was added to transcription buffer containing: 50 mM Tris-HCl pH 7.5, 17.5 mM $MgCl_2$, 5 mM DTT, 2 mM spermidine, 0.05% triton-100, and 2 mM of each NTP. In-house purified T7 polymerase is added along with thermophilic inorganic phosphatase to initiate transcription. The reaction is placed at 37°C for 3 hrs. TURBO DNase (20 units) is then added along with 12 uL of 100 mM $CaCl_2$ and allowed to react for 1 hr at 37°C. Proteinase K (200 µg) is then added at 37°C for 1 hr. The mixture is then centrifuged to remove any precipitate. The supernatant is then filtered through a 0.2 µm filter into a 100 Kd cut off filter. The RNA containing solution is buffer exchanged a total of 7 times with 14 mL of filtration buffer (5 mM cacodylate pH 6.5 and 10 mM $CaCl_2$). For the last buffer exchange, the RNA is concentrated to approximately 10 mg/mL.

Crystallization trials

Crystallization trials were initiated using an Art Robbins Gryphon robot for automated condition screening. 0.5 µL of each precipitant solution, 0.5 µL of 10 mg/mL RNA, and 0.5 µL of 0.5 mM spermine were combined in a sitting drop well. The crystal trays were centrifuged at 1000 rpm for 1 minute and then placed at 22°C or 30°C to equilibrate. Trays were evaluated every three days for crystal nucleation.

Structure determination

Crystals of both the pre-catalytic and post-catalytic states of OiΔDVI were solved to 3.8 Å. Phases were determined with molecular replacement using PHENIX[60]. Accession number 4DS6 was used as a reference model. All X-ray data were obtained at NE-CAT's 24-ID-C beamline at the

Advanced Photon Source (Argonne National Laboratory). The diffraction data was processed using HKL2000 [61]. Density analysis and model fitting was performed using COOT [62]. SBgrid provided and complied all the software used [63].

Chapter 3: Conserved nucleotides stabilize the group II intron core though long ranged hydrogen-bonding networks

3.1 Abstract

Group II introns are catalytic RNAs that self-splice to form a lariat product using a two-metal-ion mechanism identical to that used by the eukaryotic spliceosome. The recent crystal structure of a post-catalytic eukaryotic IIB intron provided the first insight into an active site competent for lariat formation. The structure revealed new tertiary contacts and the overall architecture required for the transition between the two steps of splicing. We have identified two conserved structural motifs, the I(i) and DIII internal loops, that modulate catalytic efficiency. These motifs are found to interact with the active site via long-range hydrogen bonding networks that converge on the two-nucleotide bulge of the catalytic DV via a highly conserved adenosine residue from the I(i) loop. Through our analysis, we have also identified the specific role of DIII in functioning as a catalytic effector through stabilization of the ρ-ρ′ interaction. Our findings provide a rationale for the high degree of conservation of residues distant from the group II intron catalytic core. This study reveals the importance of nucleotides that indirectly affect catalysis through long-range hydrogen bonding networks that are not easily identified through an analysis of the crystal structure alone.

3.2 Introduction

Group II introns are large ribozymes that catalyze RNA splicing via two sequential transesterification reactions [5]. They have a characteristic secondary structure comprised of six RNA structural domains, labeled I to VI, radiating from a central hub (Figure 3.1A) [9,10]. The highly conserved domain V (DV) forms the intron active site [11] through coordination of catalytic magnesium ions. A crystal structure of a bacterial group II intron revealed that splicing is catalyzed via a two-metal-ion mechanism (M1 and M2) [11] and a structure of a eukaryotic group II intron lariat identified two auxiliary metals (M3 and M4) [12] important for efficient splicing. In the first step of splicing, the 2′-OH of a bulged adenosine residue from DVI attacks the 5′ splice site to form lariat RNA [7,8]. This is followed by the second step in which the 3′-OH of the newly cleaved 5′ exon attacks the 3′ splice site to produce ligated exons and spliced intron

lariat. Hydrogen bonds between conserved nucleotides fold the RNA into a compact molecule and position the catalytic metal ions in the active site.

The importance of these hydrogen bonds can be probed using nucleotide analogue interference mapping (NAIM) and chemical modification. Two group II introns residing within the mitochondria of the eukaryotes *Pylaieilla litoralis* (*P.li.LSUI2*) [64] and *Saccharomyces cerevisiae* (*aI5*γ) [7] have been extensively studied using these techniques. NAIM allows the global analysis of the individual contribution made by a single functional group of a specific nucleotide on ribozyme activity and/or folding [65,66], and has been used to study the relationship between structure and function for multiple catalytic RNAs. This technique involves the incorporation of a phosphorothioate nucleotide analogue into an RNA molecule through *in vitro* transcription, followed by a selection procedure to purify active or functional molecules of interest. The deficiency of an analogue at specific sites highlights nucleobase functional groups important for catalysis or folding. Only the R_p analogue is incorporated through *in vitro* transcription using T7 RNA polymerase. Due to this incorporation bias, positions where important metal ion coordination occurs with a S_p phosphate oxygen are not visible with NAIM. NAIM analysis of group II introns can be performed by either selecting for catalytic activity or by monitoring the ability of the intron to fold into a compact tertiary structure.

Recently, we reported the crystal structure of the eukaryotic *P.li.LSUI2* intron in the post-catalytic lariat form [12]. This now allows us to correlate NAIM data with a group II intron structure and has resulted in the identification of new interactions and dynamic regions that were not apparent from an inspection of the crystal structure in isolation. As a result, we can now analyze the precise nature of specific tertiary interactions and biochemically probe the role of newly identified hydrogen bonding networks in the splicing reaction.

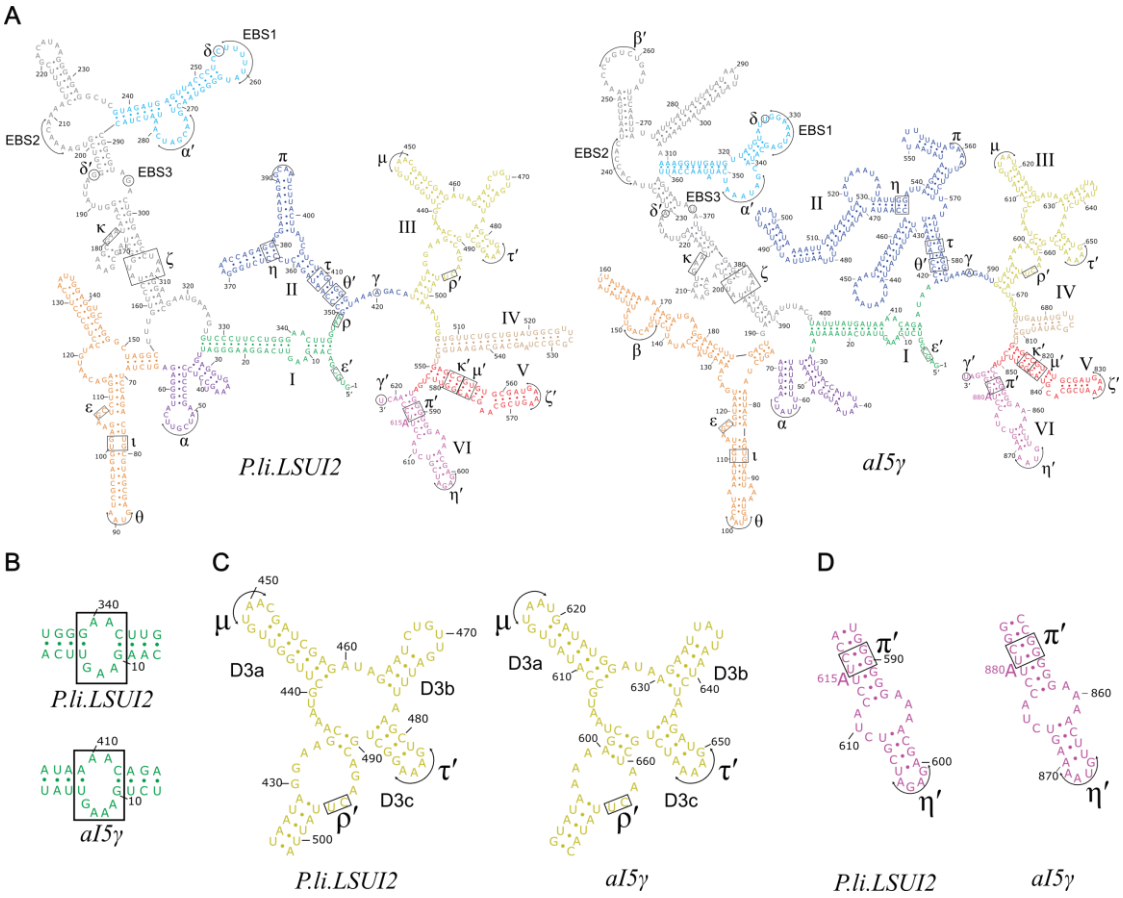


Figure 3.1: **Conservation of sequence and secondary structure between P.li.LSUI2 and aI5γ**. A. Secondary structures of the P.li.LSUI2 construct corresponding to the crystal structure and WT aI5γ outlined with Greek letters corresponding to tertiary interactions. B. The I(i) loops of P.li.LSUI2 and aI5γ are shown with the relevant portion boxed . C. DIII secondary structures of P.li.LSUI2 and aI5γ have maintained a high level of sequence and structural conservation. The lengths of the three helices (D2a, D2b, and D2c) are conserved as well as the structural organization of both the internal loop and four-way junction. D. The secondary structures of DVI in both P.li.LSUI2 and aI5γ are shown. This domain remains highly conserved between the two introns and both contain the η′ and π′ motifs.

3.3 Results and discussion

3.3.1 Patterns of nucleotide conservation

Group II introns are divided into three structural subtypes: IIA, IIB and IIC. IIA and IIB introns are largely found in eukaryotes, while the IIC class is exclusively bacterial [23,24]. The *P.li.LSUI2* and *aI5*γ introns are IIB introns and part of the chloroplast-like class 1 phylogenetic grouping, which has a distinctive pattern of nucleotide conservation within the secondary structure [23,24]. This is especially true of domains III, V, VI and the I(i) loop from domain I, which have essentially identical secondary structures and are the main focus of this study (Figure 3.1B, C, and D). Both introns form lariat during splicing as opposed to the linear hydrolytic product, which dominates in the bacterial IIC splicing reaction [67]. These similarities allow correlation of the cumulative chemical modification data for IIB introns with the *P.li.LSUI2* crystal structure to identify essential hydrogen bonding patterns associated with catalysis and/or folding. The major difference between *aI5*γ and *P.li.LSUI2* is found within domain II (Figure 3.1A); however, it functions in a similar manner and also contains a cryptic π-π′-like interaction with DVI (Supplementary Data and Supplementary Figure S3.1), which promotes the second step of splicing.

With the post-catalytic *P.li.LSUI2* crystal structure, we mapped the location of conserved IIB nucleotides in three-dimensional space (Supplementary Figure S3.2) using the chloroplast-like class 1 consensus secondary structure [23,24]. Apart from peripheral tertiary interactions, the vast majority of conserved residues are found directly within a ~40 angstrom sphere around the active site. These conserved residues also exhibit extensive NAIM interferences that allowed us to generate a short list of candidate nucleotides that we targeted to determine their contribution to the splicing activity of group II introns.

3.3.2 The I(i) loop

The I(i) loop is highly conserved in group IIB introns and is located near the intron 5′ end GUGYG sequence [24]. Essential for the overall fold of this loop is a base triple involving nucleotides G10, A12, and C342. G10 interacts with both A12 and C342 (Figure 3.2A and B) and displays strong inosine and 2′-deoxy interferences. This base triple interaction is crucial in the positioning of A341, which coordinates

the core M4 metal ion via its S_p phosphate oxygen. The M4 binding pocket is completed by the phosphate backbone configuration resulting from A6 and C7 stacking on each other. The net effect of these interactions is to extrude A341 away from the body of the I(i) loop and towards DV.

The extruded A341 forms a base triple with the conserved G575-U558 wobble pair found directly below the two-nt bulge of DV (Figure 3.2C). The importance of the base triple between A341 and DV is highlighted by strong inosine and 2′ deoxy interferences at G575, which disrupt this interaction. The two-nt bulge plays an integral role in coordinating the catalytic metal ions (M1 and M2) [11] and can theoretically exist in two conformations, either as an A573/A574 or A574/G575 bulge. We propose that the base triple involving A341 locks down the position of the DV bulge into the observed AA conformation competent for catalytic metal ion binding.

The significance of the I(i) loop was tested through mutagenesis of nucleotides G10 and A341. G10 is centrally located within the overall fold of the I(i) loop with especially strong interferences. First step branching efficiency was measured for the *in vitro* self-splicing reaction. The G10C mutant has strong negative effects upon the splicing reaction with a 7.3-fold lower formation of branched product compared to wild-type (WT). Mutagenesis of A341 to a guanosine residue (A341G) has an even larger negative effect with a 14.9-fold reduction (Figure 3.2D). This is consistent with the proposed role for A341 in stabilizing the two-nt bulge in the proper conformation to promote the binding of catalytic metal ions.

The cumulative data for A341 is consistent with a model in which it exerts its effects upon the active site via a long-range hydrogen-bonding network spanning from DI to the conserved two-nt bulge of DV. The existence of this network is not readily apparent by looking at the crystal structure in isolation. The hydrogen bonding network (Figure 3.2A) reveals that nucleotides distant from the active site can affect the catalytic metal center through an intricate web of interactions.

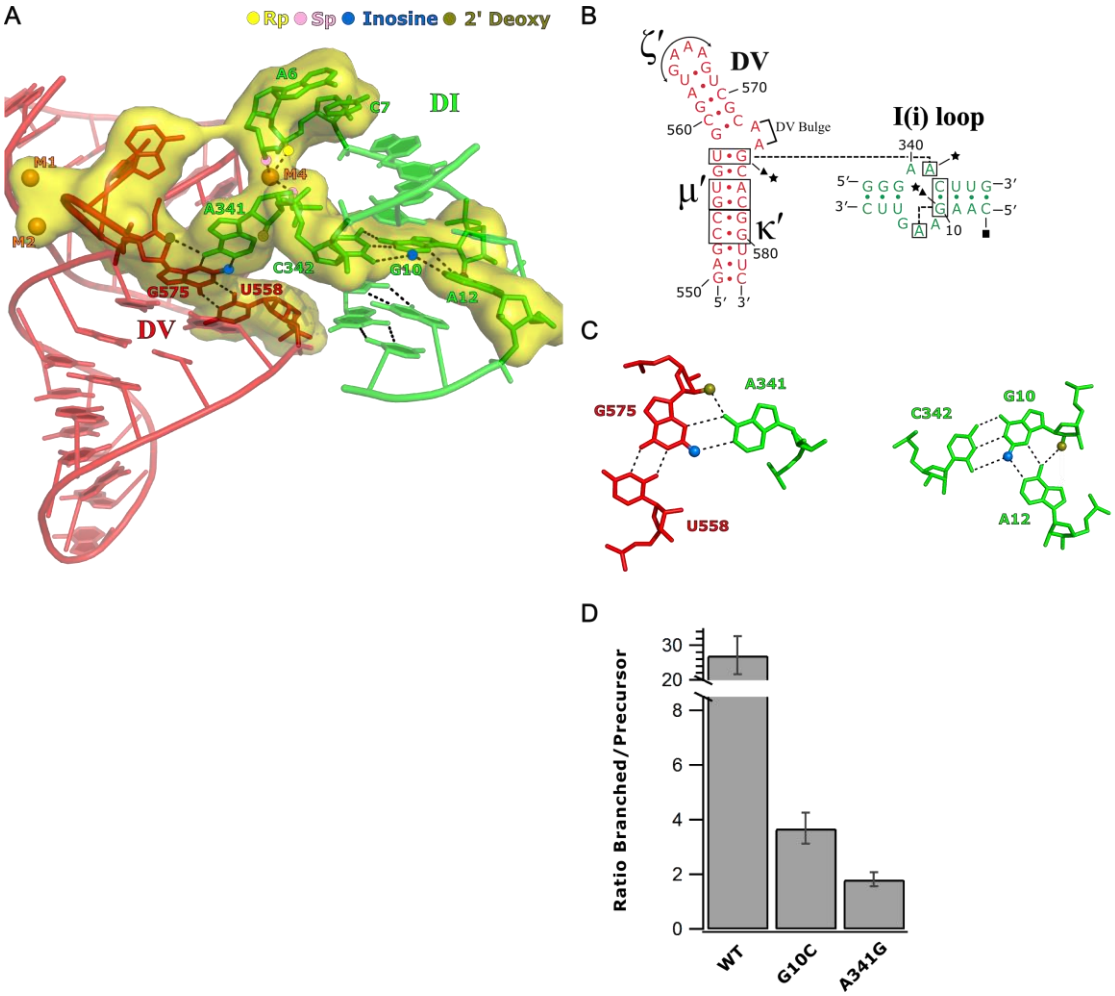


Figure 3.2: **The I(i) loop and A341 help to stabilize the DV bulge**. **A.** The extrusion of A341 from the I(i) loop allows it to participate in a base triple with U558/G575. This in turn reinforces the position of the catalytically essential DV bulge, which is responsible for binding the two catalytic Mg^{2+} ions (M1 and M2). The position of A341 is stabilized by the intricate hydrogen bonding network found within the I(i) loop, as well as M4 and the 5′ end nucleotides A6 and C7. The nucleotides involved in positioning M1 and M2 are shown as a yellow surface representation. NAIM interferences are displayed on the P.li.LSUI2 crystal structure and depicted as spheres, with the colors indicating the type of interference. S_p phosphorothioate interferences for A6 and A341 are predicted based on the crystal structure of P.li.LSUI2. **B.** P.li.LSUI2 secondary structure of DV and the I(i) loop, emphasizing the base triples between A341/G575/U558 and G10/A12/C342 with boxes and dashed lines. NAIM interferences are depicted with symbols: ▲ (inosine), ★ (2′-deoxy), ■ (phosphorothioate). **C.** The base triples between A341/G575/U558 and G10/A12/C342 are shown in an alternate view with the NAIM interferences depicted with the same colored spheres as in Figure 3.2A. **D.** A graph showing the ratio of branched/precursor intron of WT, A341G and G10C.

3.3.3 Domain III: ρ-ρ′ interaction

Domain III has long been considered to be a catalytic effector of group II intron activity [68]. This domain has a highly conserved secondary structure within the IIB1 class and exhibits a large number of NAIM interferences, thus indicating a complex and important role (Supplementary Figure S3.3). Interferences are especially concentrated in the internal loop of DIII [69]. This region contains the ρ-ρ′ interaction that positions the DIII internal loop, J2/3, J1/2, and the 5′ end of the intron in close proximity (Figure 3.3A). J2/3 is thought to engage in a conformational rearrangement between the first and second steps of splicing [70,71] and the 5′ end of the intron is required for accurate 5′ splice site selection [72]. Therefore, ρ-ρ′ is situated in a central point from which it can interact with key regions required for efficient splicing. ρ-ρ′ consists of two base pairs (reverse Watson-Crick A-U and non-canonical A-C) between J1/2 and residues from the internal loop. Disruption of ρ-ρ′ in Δρ and Δρ′ mutants results in a 30.5- and 65.8-fold reduction in branched product, respectively (Figure 3.4). Mutagenesis of both components of this interaction has the largest effect seen in this study with a 326.7-fold reduction for a ΔρΔρ′ mutant (Figure 3.4). The importance of ρ-ρ′ is further strengthened by the presence of strong interferences for the nucleotides involved in this interaction. The internal loop displays 2′-OMe interferences at positions A429 and G430 that cause severe steric clashes with ρ-ρ′ (Figure 3.3B). Residue A494 also displays a strong 2′-deoxy interference that would disrupt a hydrogen bond with C7 leading to destabilization of the M4 metal binding pocket (Figure 3.3C). In addition, 2,6-diaminopurine (2,6-DAP) and 2′-OMe interferences at A494 result in steric clashes with C7. A similar result is observed when 2,6-DAP is substituted at A432 [69]. Taken together, these modifications (Figure 3.3D) affect a crucial metal-binding pocket and the organization of important inter-domain junctions, thus rationalizing the observed negative effect upon splicing as a result of mutation or analogue incorporation.

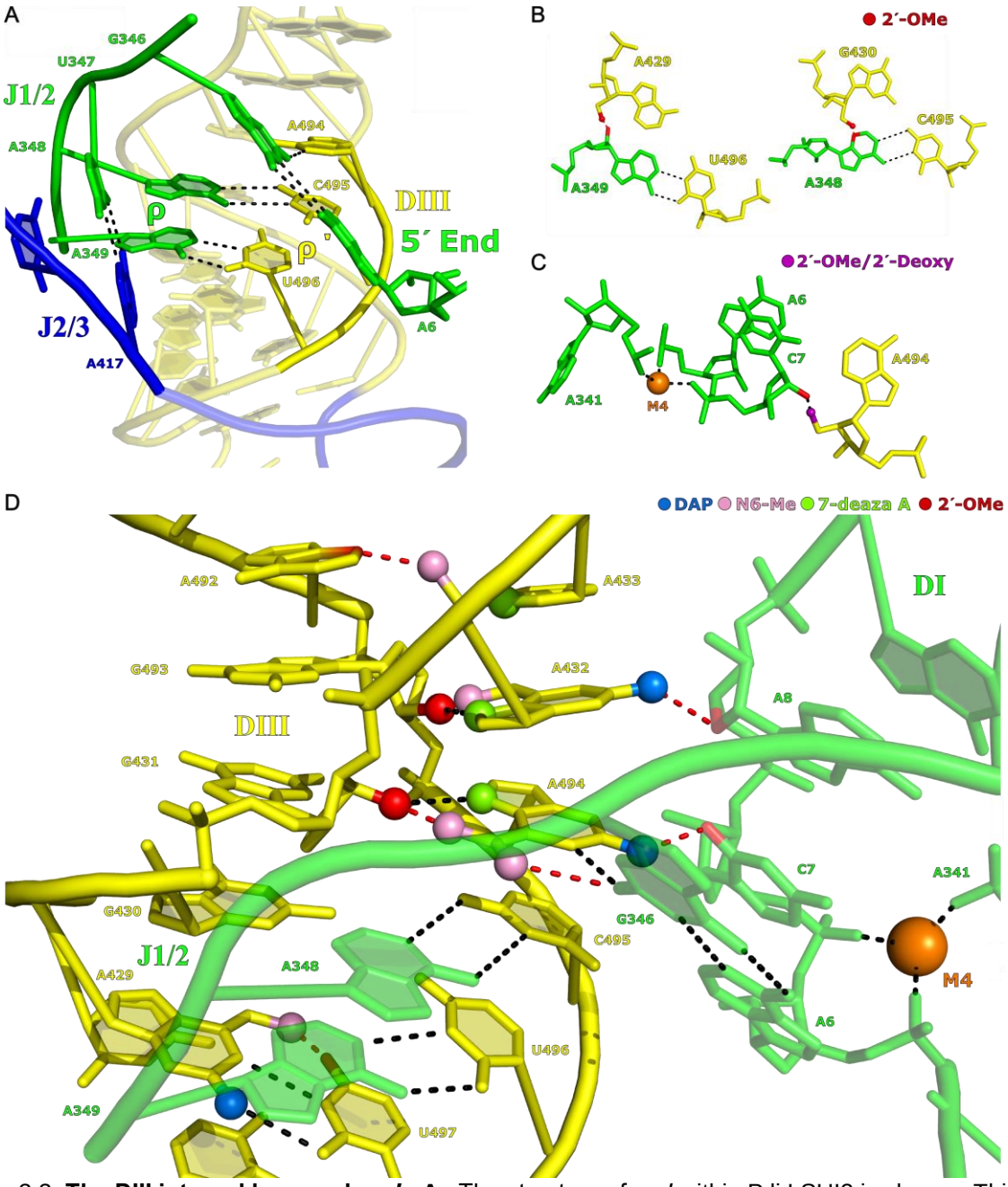


Figure 3.3: **The DIII internal loop and ρ-ρ′**. **A.** The structure of ρ-ρ′ within P.li.LSUI2 is shown. This tertiary interaction organizes four separate regions within the intron: the 5′ end, J1/2, J2/3 and the DIII internal loop. **B.** 2′-OMe interferences that correspond to A429 and G430 in P.li.LSUI2 show a steric clash with ρ-ρ′. **C.** 2′-deoxy and 2′-OMe interferences are shown for A494 in P.li.LSUI2. The 2′-OH of this nucleotide interacts with the 2′-OH of C7 to extend the network from the DIII internal loop into the core of the intron through stabilization of M4 and A341. **D.** The DIII internal loop of P.li.LSUI2I is shown with the corresponding NAIM interferences depicted with colored spheres. An extensive network of hydrogen bonds stabilizes the fold within the loop and allows the proper formation of ρ-ρ′. Red bonds indicate NAIM interferences that are caused by steric clashes. N6-methyl incorporation at A494 results in a clash with either the 2′-OH of G431 or the N2 of G346.

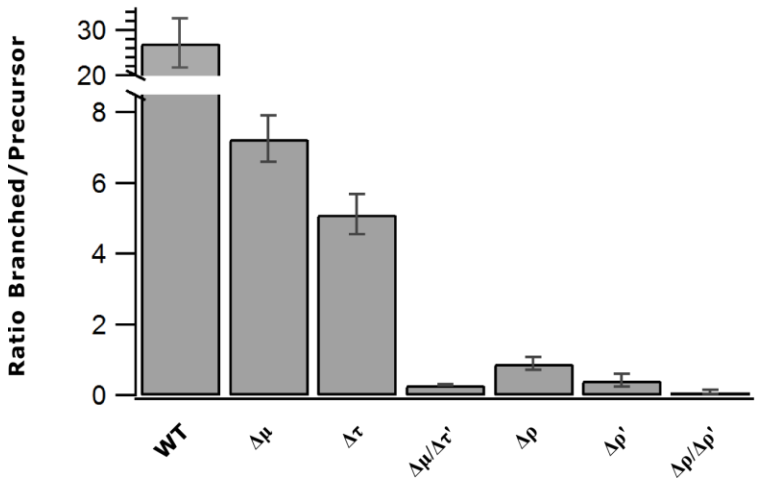


Figure 3.4: **Splicing of DIII mutants**. Mutagenesis was performed on important DIII tertiary interactions. The effect of mutations on the ratio of branched/precursor intron is shown relative to WT.

3.3.4 ρ-ρ′ is dynamic

A surprising pattern emerges when looking at interferences for A429 and G430. In the IIB1 intron subclass, these two conserved purines are predicted to pair with C495 and U496 in the DIII basal stem, consistent with data showing that these residues are protected from DMS modification [73]. However, in the post-catalytic structure, we observe that G430 and U496 are unpaired (Figure 3.3D). The predicted G430- U496 base pair cannot exist concurrently with ρ-ρ′ as U496 is directly involved in forming this tertiary interaction. In addition, both of these nucleotides in *al5*γ (A596 and A597) have 2,6-DAP interferences [69]. Interestingly, 2,6-DAP incorporation strengthens A-U base pairs with the addition of a third hydrogen bond. Weakening of this pair, as shown by N6-methyl interference at A429, has less severe effects on branching than the stabilization of this AU base pair with 2,6-DAP (Figure 3.3D). Based on these data, we hypothesize that the DIII internal loop is a dynamic region, and that adding rigidity through 2,6-DAP incorporation precludes U496 unpairing from G430, thus preventing the formation of ρ-ρ′ and inhibiting splicing. Therefore, the crystallographic and biochemical evidence supports a model in which ρ-ρ′ is dynamic during catalysis and is important for both steps of splicing due its interactions with both J2/3 and the 5′ end.

3.3.5 Domain III: μ-μ′ and τ-τ′

A conserved structural feature of IIB1 introns within DIII is the presence of three stem loops termed D3a, D3b, and D3c. The overall fold of DIII allows the D3a and D3c stems to brace the intron through the formation of the μ-μ′ and τ-τ′ contacts with DV and DII, respectively. The contribution of these two tertiary interactions to the efficiency of splicing was probed through mutagenesis of the D3a and D3c loops to UUCG tetraloops in *P.li.LSUI2.* This resulted in only 3.7- and 5.3-fold reductions in branched product for the Δμ and Δτ′ mutants, respectively (Figure 3.4). The relatively modest effect of these mutations can be rationalized by their proximity to other tertiary interactions (κ-κ′ and θ-θ′), which can compensate for the mutated DIII contacts. In contrast, a ΔμΔτ′ double mutant has a significant effect with a 92.1-fold reduction in splicing, indicating that these two interactions function in a synergistic manner. Due to the importance of ρ-ρ′, we propose that the main role of μ-μ′ and τ-τ′ is to facilitate formation of ρ-ρ′, which in turn positions the 5′ end and J2/3 elements required for efficient splicing.

3.3.6 Domain III: four-way junction

The geometry of the DIII four-way junction positions the τ′ and μ loops that dock into domains II and V, respectively (Figure 3.5A). The base triple between A437, A462 and U470 anchors this junction and stabilizes the D3b stem [12]. Disruption of this base triple either through 7-deaza-A incorporation at A462 or 2,6-DAP at A437 causes interference in the first step [69]. In addition, 2,6-DAP interference at A460 disrupts its non-canonical pairing with G440 and likely prevents the adjacent D3a stem (containing the μ loop) from docking into its receptor in DV (Figure 3.5B). The importance of base stacking architecture within the junction is further supported by N7-deaza interferences seen at A437 and A438 [69]. We have identified, for the first time, a correlation between N7-deaza incorporation and destabilization of nucleobase stacking within an RNA structure (Supplementary Data and Supplementary Figure S3.4); therefore, this analogue reveals additional structural insight beyond the disruption of hydrogen-bonding.

To probe the function of conserved nucleotides within the four-way junction, we used the crystal structure as a guide for the design of point mutants. Mutations in this region would be expected to alter the angles of the stems containing the τ′ and μ loops, thereby preventing docking of these loops into their receptors. Consistent with this hypothesis, an A437G mutation results in a 10.7-fold reduction in splicing,

due to disruption of the central base triple (Figure 3.5C). The junction linker sequences (directly adjacent to the D3a and D3b stems) interact to form a single conserved trans-WC AU pair in the crystal structure, with an A438C mutation causing a 17.2-fold reduction (Figure 3.5C). The non-canonical pair between G440 and A460 is also significant in forming the foundation for the base stack within the junction that includes the base triple. Disruption of the G440-A460 pair with an A460G mutation results in a 3.8-fold reduction in branched product (Figure 3.5C). This is consistent with a model in which μ-μ′ and τ-τ′ stabilize the hydrogen-bonding network that starts in the DIII internal loop.

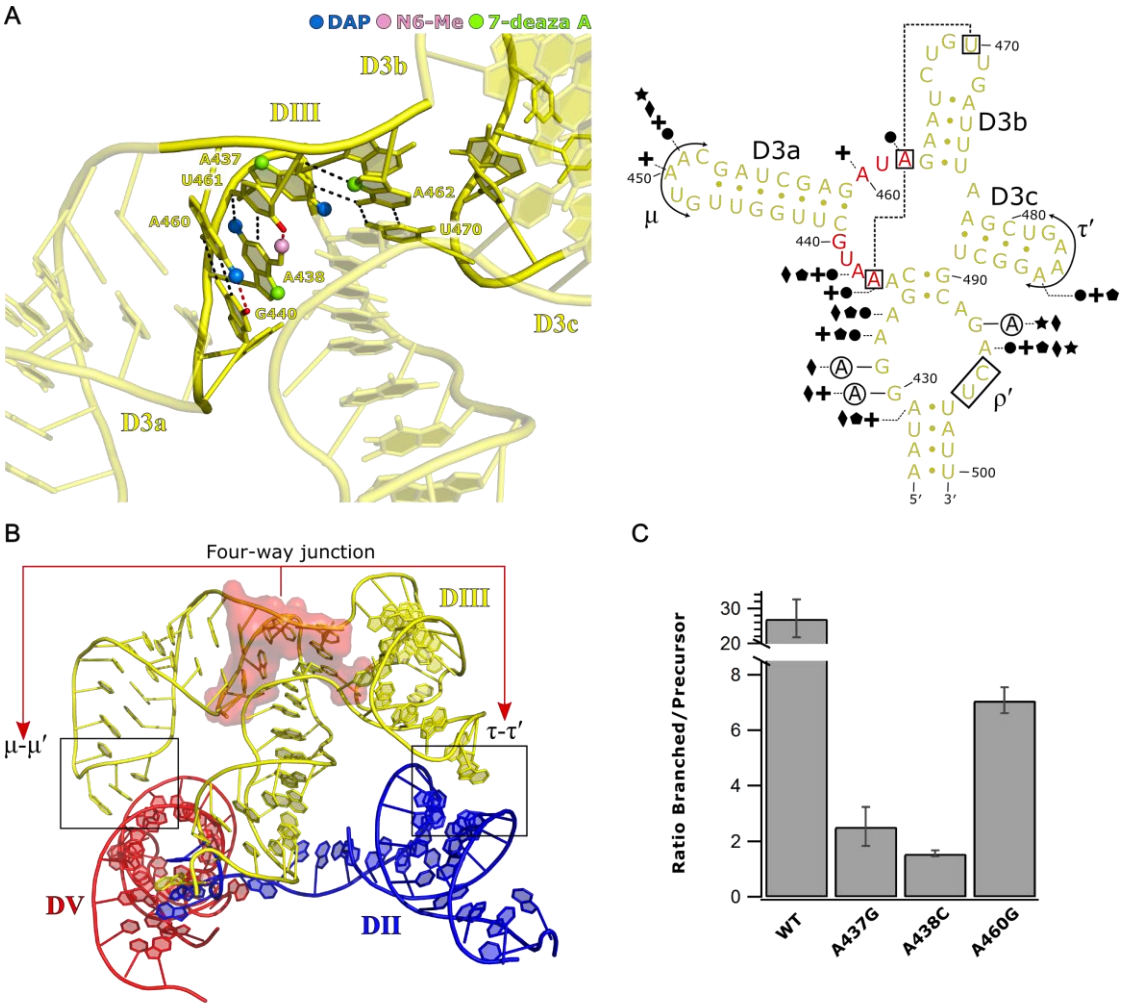


Figure 3.5: **Four-way junction of DIII and its role in splicing**. **A.** The P.li.LSUI2 tertiary structure of the DIII four-way junction is shown with the NAIM interferences depicted as colored spheres. NAIM modifications that cause steric clashes are shown with red bonds. The secondary structure of P.li.LSUI2 is also shown with the conserved junction nucleotides between P.li.LSUI2 and aI5γ in red. The important base triple between A437/A462/U470 is boxed and depicted with dashed lines. G430, G431, and G493 within the I(i) loop of P.li.LSUI2 have diverged from the adenosine nucleotides seen in aI5γ (residues circled). Greek letters correspond to tertiary interactions and NAIM interferences are shown with symbols: ● (7-deaza A), ✚ (2,6-DAP), ⬟ (N6-methyl), ♦ (2′-OMe), ★ (2′-deoxy). **B.** Model showing the effect of destabilizing the four-way junction in DIII. The conserved nucleotides in the junction are shown with a red surface. Destabilization of the hydrogen bonding network within the junction will lead to disorder in the adjacent helices, resulting in misplacement of the μ and τ′ loops. **C.** Graph showing the ratio of branched/precursor intron of P.li.LSUI2 mutants in the four-way junction.

3.3.7 Convergence of two hydrogen-bonding networks

Mapping the location of conserved nucleotides and NAIM interferences in DIII reveals the existence of a second long-range hydrogen-bonding network leading into the catalytic core. This network intersects with the I(i) loop at the highly conserved A341 residue and the M4 metal ion (Figure 3.6A). A341 represents the confluence of the two networks responsible for stabilization and positioning of the active site metal ions and the exon substrates. Kinetic analysis was done on an A341G mutant to test the importance of this residue in propagating the effects of the hydrogen-bonding networks to the catalytic metals. *In vitro* self-splicing data shows that this mutation results in a significant reduction in the initial rate of lariat formation, but is still capable of forming a large quantity of spliced product (Figure 3.6B). We propose that A341 plays a pivotal role in maintaining the correct conformation of the two-nt bulge to prevent toggling between multiple states during splicing (Figure 3.6C). Identification of these networks has greatly expanded the radius of nucleotides known to be responsible for stabilizing the active site architecture beyond the previous, DV-only model.

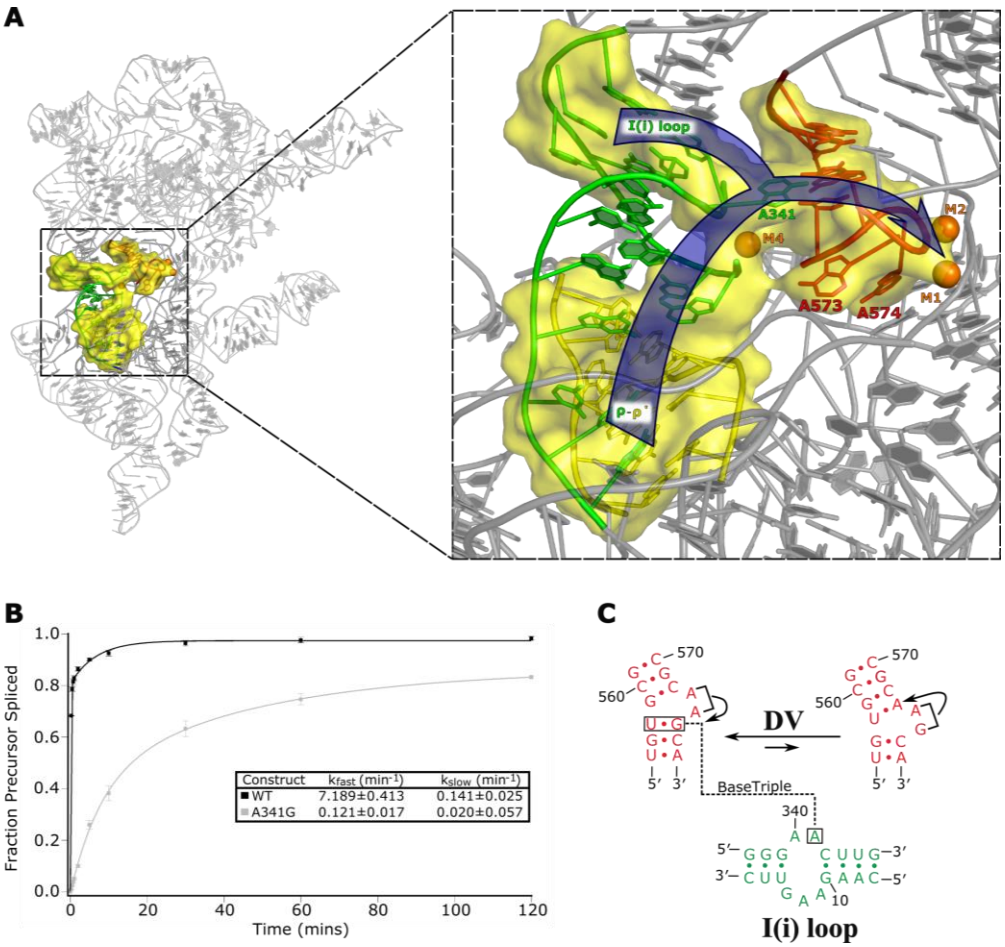


Figure 3.6: **Model for the stabilization of the intron core through long-range hydrogen bonding networks**. **A.** Two separate hydrogen bonding networks (shown as a yellow surface representation) converge on A341 from the I(i) loop. This ultimately leads to the extrusion of A341 from the I(i) loop, which reinforces the position of the DV bulge and allows for the proper binding of the catalytic metal ions (M1 and M2). **B.** Self-splicing kinetic curves for both WT and an A341G mutant. The curves were fit using a double exponential equation. The A341G mutant shows a ~60-fold reduction of k_{fast}. **C.** Depiction of the two possible conformations for the two-nt bulge in DV. Only the A573/A574 conformation binds M1 and M2 in the correct geometry for efficient splicing. A341 from the I(i) loop stabilizes the AA conformation by forming a base triple with U558 and G575.

3.4 Conclusions

This work reveals that nucleotides ~40 Å from the two-metal-ion active site can affect catalysis in a group II intron via networks of hydrogen-bonding interactions between conserved nucleotides. Mutagenesis of nucleotides along these pathways leads to strong negative effects upon splicing, which is consistent with these networks supporting the formation of a competent active site from a distance. This is further corroborated by the comparison between the crystal structure and the chemical modification data for IIB introns. This detailed comparison of NAIM data with the crystal structure of *P.li.LSUI2* also provides insight into the structural basis for the interference effects observed with different nucleotide analogues.

A341 has been identified as an especially important residue for formation of the active site. The base triple involving this residue is conserved between bacterial and eukaryotic group II introns. In addition, the secondary structure of DV is similar to that of the U2/U6 pairing in the spliceosome, and we believe that there is likely an interaction in the spliceosome stabilizing the bulge of U6 in a manner similar to the I(i) loop. This brings up the possibility of similar long-range hydrogen-bonding networks also existing in the spliceosome to modulate splicing.

3.5 Materials and methods

Cloning, mutagenesis, and RNA preparation

Wild-type *P.li.LSUI2* and *aI5γ* intron DNA were synthesized (Genscript) and cloned into pUC57 using the EcoRV restriction site. The open reading frame (ORF) for the maturase in DIV was removed and replaced with a stem loop containing the UUCG tetraloop sequence for both introns. The *P.li.LSUI2* construct also contains a 250-nucleotide 5′ exon and a 75-nucleotide 3′ exon. The *aI5γ* construct contains 250-nucleotide 5′ exon and a 150-nucleotide 3′ exon. All mutant introns were prepared using overlapping PCR and the resulting products were cloned into pUC57 using the EcoRV restriction site. Plasmids were linearized using BamHI (*P.li.LSUI2*) or HindIII (*aI5γ*) and RNA was synthesized through *in vitro* transcription with T7 RNA polymerase. RNA was transcribed and internally labelled using the following reaction conditions: 10 μCi [α-32P]UTP (3,000 Ci mmol-1), 0.5 mM UTP, 1 mM other NTPs, and 5 mM $MgCl_2$. Transcription reactions were then gel purified on a denaturing 4% polyacrylamide

(19:1)/8 M urea gel. Precursor intron RNA was recovered by elution into a buffer containing 300 mM NaCl, 0.01% SDS, and 1 mM EDTA. Eluted RNA was EtOH precipitated and the pellet resuspended in 10 mM Tris-HCl pH 7.5 and 1 mM EDTA and stored at -80ºC.

In vitro branching efficiency assays

Intron RNA was refolded by heating at 90ºC and allowed to cool at room temperature for 15 minutes. *P.li.LSUI2* self-splicing experiments were performed in 10 mM $MgCl_2$, 1 M NH_4Cl, 40 mM Tris-HCl pH 7.5, and 0.02% SDS at 45ºC. Samples were taken at 0 and 30 minutes and the reactions were stopped by EtOH precipitation. Splicing products were analyzed on a denaturing 4% polyacrylamide (19:1)/8 M Urea gel. Each splicing reaction was done in triplicate and the ratio of branched/precursor intron was calculated.

RNA splicing kinetics

Transcribed intron RNA was refolded as previously described. *P.li.LSUI2* self-splicing kinetic experiments were performed in 10 mM $MgCl_2$, 1 M NH_4Cl, 40 mM Tris-HCl pH 7.5, and 0.02% SDS at 45ºC. Samples were taken at 0, 0.25, 0.5, 0.75, 1, 2, 5, 10, 30, 60, and 120 minutes and the reactions were stopped by EtOH precipitation. *aI5γ* self-splicing experiments were performed in 20 mM $MgCl_2$, 1 M NH_4Cl, 40 mM Tris-HCl pH 7.5, and 0.02% SDS at 45ºC. Samples were taken at 0, 1, 5, 10, and 15 minutes and were stopped by ethanol precipitation. The splicing products were analyzed on a denaturing 4% polyacrylamide (19:1)/8 M Urea gel. The *P.li.LSUI2* splicing reaction was done in triplicate and the kinetic curve was fit to a biphasic exponential equation.

P.li.LSUI2 structure refinement

Structure refinement was done with PHENIX [60] using the PDB 4R0D as the starting structure. Coot [62] and the RCrane [74] plugin were used to improve the planarity of base triples and non-canonical pairs. This final structure had an R_{work}=22.9 and R_{free}=27.3. Structure was deposited under accession number 4R0D. Pymol was used to prepare figures.

3.6 Supplementary data

3.6.1 DII in *aI5*γ and *P.li.LSUI2* are equivalent

In the course of our analysis relating chemical modification data to the crystal structure, we gained insight into the positioning of DVI in *aI5*γ. DVI was previously thought to engage in a large-scale conformational change between the two steps of splicing [75]. There are multiple iterations of this model with varying degrees of movement for DVI. All models postulate a 'swinging arm' movement from an 'up' to a 'down' position mediating the transition between the steps of splicing. Initial evidence for this model came from DEPC interference experiments performed on *aI5*γ [75], in which the first and second steps of splicing were individually probed, using constructs in which the EBS-IBS interactions were mutagenized to favor a specific step. As a result, interferences were found for a tetraloop-receptor interaction between DII and DVI called η-η′. In *P.li.LSUI2*, η-η′ consists of a GAGA tetraloop from DVI docking into a receptor in the D2a stem. This interference was found to be specific for the second step, and therefore it was concluded that η-η′ is dynamic and is only engaged during the later stages of splicing. The crystal structure of *P.li.LSUI2* revealed a second tetraloop-receptor interaction between DII and DVI called π-π′ [76]. Like the η-η′ interaction, mutagenesis of π-π′ also has negative effects on the second step. The π-π′ interaction is formed between a GCAA tetraloop in DII of *P.li.LSUI2* and a receptor in DVI directly adjacent to the bulged adenosine nucleophile. Structural and biochemical data indicate that this new tertiary interaction works cooperatively with η-η′ to facilitate the transition from the first to the second step through removal of the lariat bond from the active site. In this revised model, η-η′ is stationary throughout the reaction and π-π′ is dynamic [76].

It was originally assumed that a lack of interference for η-η′ implies that it is not engaged for the first step. However, in light of the crystal structure, we reinterpret this data as indicating that η-η′ is engaged in the first step, but is not required at this stage of splicing thus resulting in no interference. This is consistent with mutagenesis data showing that knocking out the η-η′ interaction has little to no effect on the first step; however, the low-resolution, pre-catalytic crystal structure of *P.li.LSUI2* clearly indicates that η-η′ is engaged at this stage.

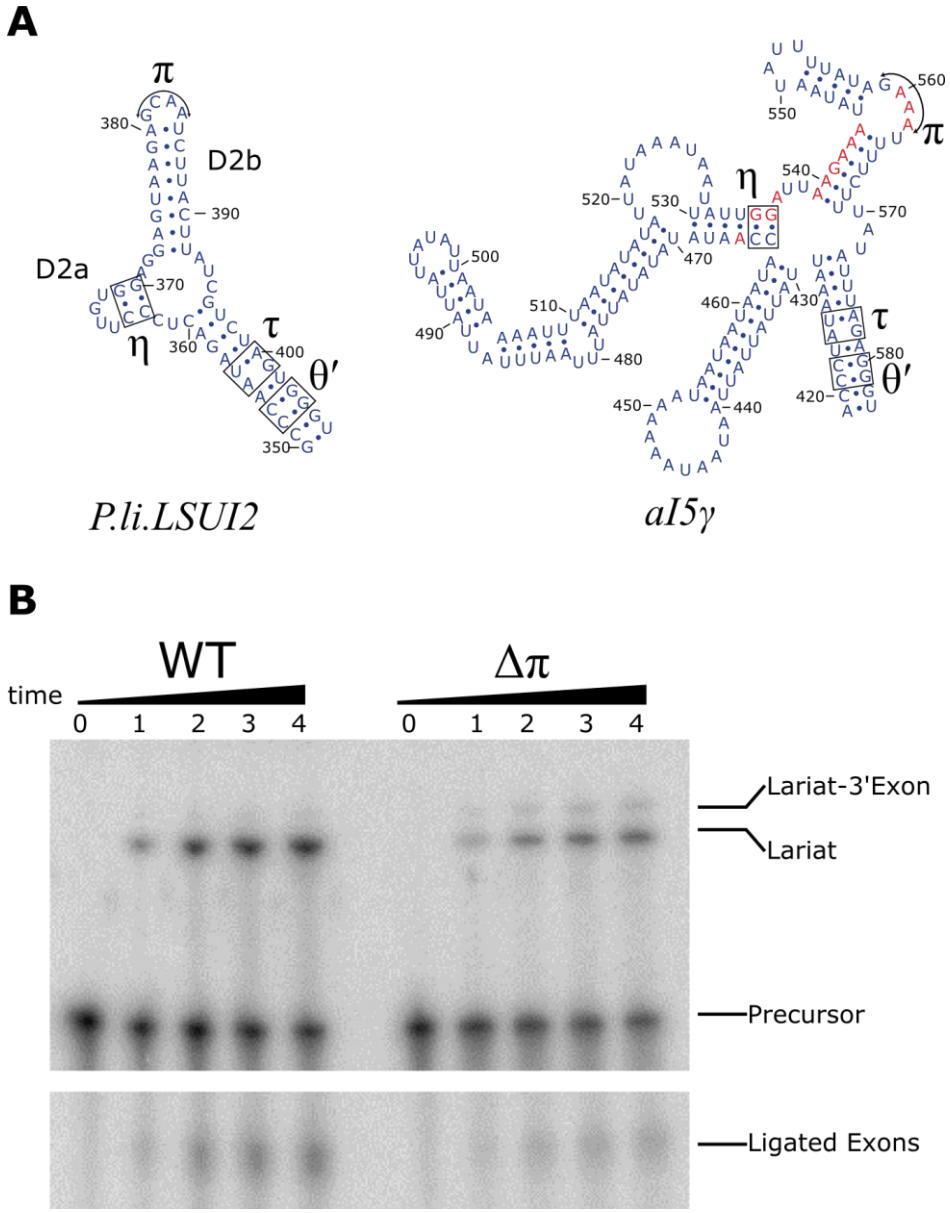
Unknown at the time, Chanfreau et al. (1996) also had biochemical evidence supporting the importance of π-π′ in *aI5*γ. An internal GAAA bulge is present in the third stem of the *aI5*γ DII (analogous

to the D2b stem of *P.li.LSUI2*) (Supplementary Figure S3.1A). This GAAA deviates from a standard stem loop at the secondary structure level, yet it is possible that it adopts a three-dimensional GNRA fold. Therefore, we hypothesize that it functions in the π-π′ interaction to sequester the lariat bond from the active site in the second step. DEPC modification experiments show that this GAAA exhibits strong second-step interferences, consistent with its role as the *aI5*γ equivalent of π. In addition, DEPC modifications disrupting the helix joining η and π also results in a second step splicing defect. This is due to the fact that the spacing between the tetraloops and receptors forming the DII-DVI binding face must maintain a conserved length to allow both η-η′ and π-π′ to form simultaneously during the second step of splicing. Any change in the spacing between η and π by the insertion, deletion, or chemical modification of a single base pair would introduce enough twist in the helix to disrupt both components of this essential domain interface. To test this hypothesis, we mutated the GAAA bulge to an CACA sequence. This mutation was chosen to preserve the secondary structure of DII as determined by MFOLD [77], yet is sufficient to disrupt the GNRA fold necessary for a π-π′ tetraloop-receptor interaction. *In vitro* self-splicing assays of this construct revealed a visible accumulation of lariat-3′ exon and a decrease in ligated exons, which is the same result for the π mutant in *P.li.LSUI2* (Supplementary Figure S3.1B) [76]. This suggests that the π-π′ interaction also exists in *aI5*γ, and that DII has the same functional role in both introns, even though its secondary structure has diverged.

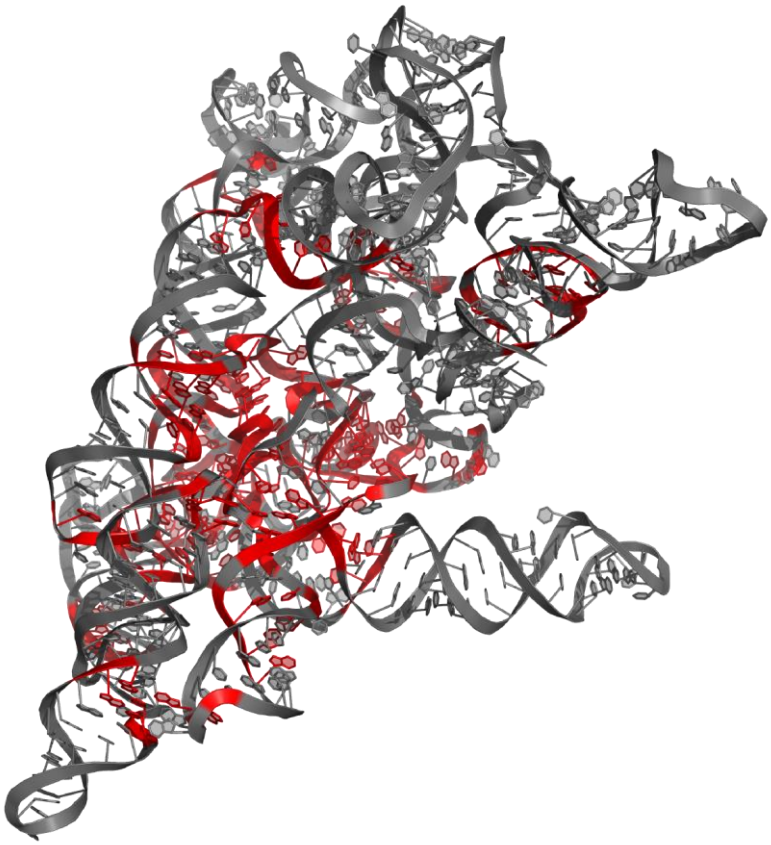
3.6.2 Base stacking interference patterns.

Base stacking adds energetic stability within nucleic acids through overlapping π orbitals, providing rigidity to the helix. High-resolution crystal structures of RNA have shown a common theme of base stacking playing an important role in reinforcement of tertiary interactions. Alterations to the functional groups of nucleobases can affect base stacking energy, leading to the destabilization of important interactions [78]. In particular, there are several positions within the *P.li.LSUI2* crystal structure that contain N7- deaza interferences, with the N7 nitrogen not directly participating in hydrogen bonding (Figure 3.3D and 3.5A). This is unexpected because interference for a base functional group usually indicates important hydrogen bonding interactions with the nucleobase. In these particular cases, we see that the base of the nucleotide is solely engaging in a stacking interaction to stabilize important tertiary

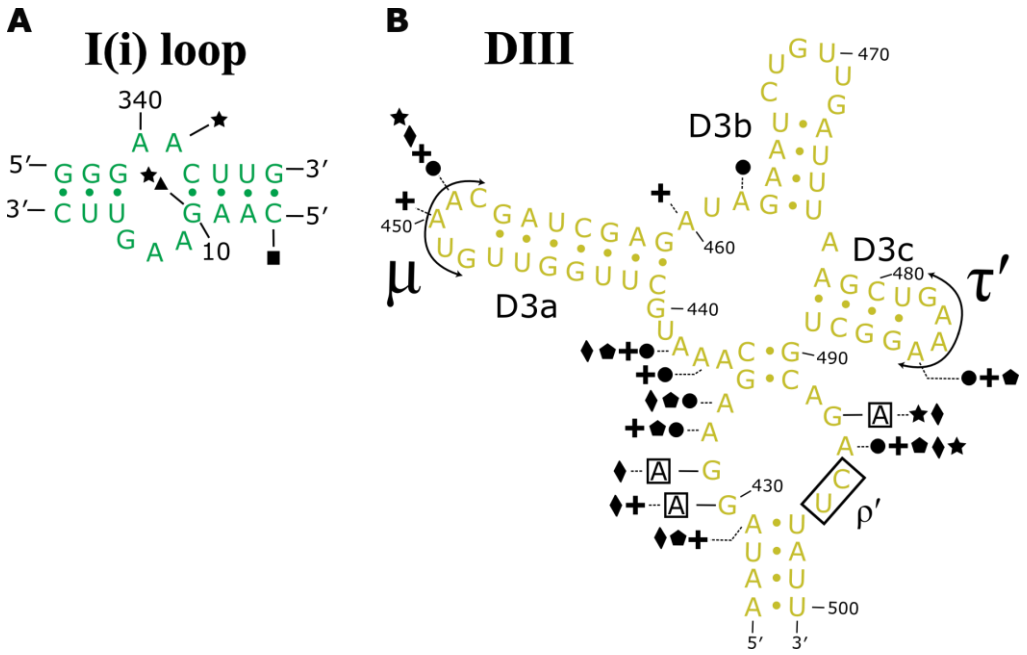
interactions or a vital stem loop. These types of interferences are seen throughout the structure, with the 5′ end sequence of *P.li.LSUI2* being an example of a region containing extensive base stacking interactions. Position A104 stacks above the G3-C107 base pair in ε-ε′ (Supplementary Figure S3.4A), and displays a strong N7-deaza effect. This is also seen in interactions involving DV for both the κ and ζ′ loops, where the second nucleotide of the motif displays N7-deaza interferences (A181 and A565) [76,79]. In these loops, the N7 nitrogen is not hydrogen bonding within the tertiary interaction, but is instead reinforcing the base stacking seen in GNRA and GNRA-like tertaloops. The internal loop in DVI also has several N7-deaza interferences in close proximity to the bulged adenosine required for lariat formation (Supplementary Figure S3.4B). Based on these correlations between the NAIM interferences and the crystal structure, we postulate that the N7-deaza substitution can have large effects upon base stacking of nucleotides. This is the first report of N7-deaza interferences being used to indicate important base stacking interactions in an RNA structure.



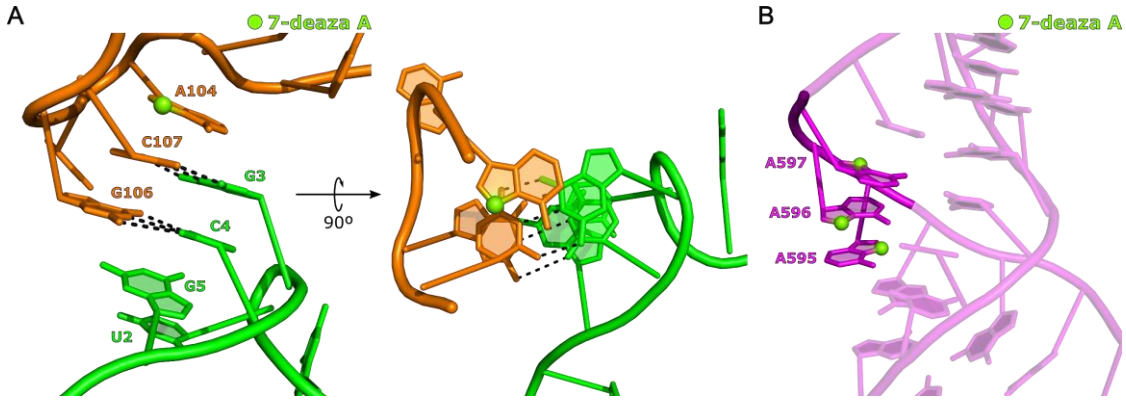
Supplementary Figure S3. 1: **Conservation and role of DII in intron splicing**. **A.** DII secondary structures of P.li.LSUI2 and ai5γ have diverged however the role of DII as a central hub organizing four tertiary interactions (θ- θ′, τ- τ′, η- η′, and π- π′) remains conserved. DII in ai5γ displays several DEPC modifications that are only sensitive in disrupting the second step of splicing. These are depicted in red on the secondary structure and highlight a π- π′ like interaction in DII of ai5γ. **B.** A mutant form of ai5γ was prepared with a G559C/A561C double mutant to determine if a π- π′ like interaction was present in DII. Splicing was done at 20 mM Mg^{2+} and clearly shows an accumulation of lariat-3' exon caused by a stalling of the second step.



Supplementary Figure S3. 2: **A three dimensional view of P.li.LSUI2 showing nucleotide conservation**. In grey are non-conserved nucleotides within the intron. Positions shown in red correspond to nucleotides that share sequence and structure conservation with chloroplast-like class I intron class from the group II intron database (http://www.fp.ucalgary.ca/group2introns/).



Supplementary Figure S3. 3: **P.li.LSUI2 secondary structures and NAIM interferences**. **A.** The I(i)-loop of P.li.LSUI2I is shown annotated with NAIM interferences. **B.** DIII of P.li.LSUI2I is shown annotated with NAIM interferences.



Supplementary Figure S3. 4: **7-deaza analogs to evaluate importance of base stacking**. **A**. A series of base stacking interactions are shown that help to stabilize the architecture of ε-ε′. A 7-deaza interference is shown at A104 which is stacking directly with G3/C107. An additional view of A104 base stacking with ε-ε′ rotated about the x-axis by 90° is also shown. **B.** 7-deaza interferences were modelled onto the internal loop of DVI. 7-deaza interference are observed for A595, A596, and A597.

Supplementary Table S3.1: ***In vitro* self-splicing data**. Ratios of branched/precursor intron were calculated after 30 minutes of splicing. The fold reductions in the ratios relative to WT are shown.

Mutant	Branched/Precursor	Fold Reduction
Wt	27.10	1.0
G10C	3.69	7.3
A341G	1.82	14.9
Δρ′	0.41	65.8
Δρ	0.89	30.5
ΔρΔρ′	0.08	326.7
Δμ	7.27	3.7
Δτ′	5.12	5.3
ΔμΔτ′	0.29	92.1
A437G	2.54	10.7
A438C	1.57	17.2
A460G	7.08	3.8

Chapter 4: Structural biology of a group II intron/maturase complex

4.1 Development of a denaturing purification to isolate a recombinant group II intron maturase protein

Group II intron splicing requires the assistance of a protein called a maturase to function *in vivo*. There are examples of introns splicing *in vitro* in the absence of protein [80]; however, these situations generally require non-biologically relevant ionic conditions (high concentrations of Mg^{2+}). The maturase protein ORF is located in DIV of the intron sequence[5]. The precise role that the maturase protein plays in the mechanism of splicing is unknown but a recent cryo-EM structure suggests that it helps at least in part to stabilize the EBS-IBS interactions for proper 5´ splice site positioning [81]. Nothing more about the structural contributions of the maturase protein is known.

The intron chosen for this study is from the thermophilic cyanobacterium *Thermosynechococcus elongatus* (*T. elongatus*). The genome of *T. elongatus* contains 28 related copies of the same intron gene from the chloroplast-like class 1 intron class [82]. As the intron gene mobilized throughout the organism, the new copies diverged and degenerated to varying degrees. In certain cases, the intron has completely lost its maturase ORF and relies on the maturase of a different intron copy to splice *in vivo*. The specific intron used for this study is *T.el*3c, which belongs to the ORF-less family of *T. elongatus* introns (Figure 4.1). *In vivo*, this intron uses the maturase protein of *T.el*4c to splice. A method had already been developed to purify *T. elongatus* maturase proteins from *E. coli* [83]. The original method involves the synthesis of a maltose binding protein (MBP) *T.el*4c fusion protein gene (104 kD). MBP is N-terminally fused and improves the stability and solubility of the maturase protein. In the absence of the MBP, the maturase protein precipitates in a matter of hours. The synthesized gene is cloned into a pET15b vector using NdeI and BamHI restriction sites. The cloned vector is then transformed into *Rosetta 2* cells and glycerol stocks were made. The cells are grown at 37°C in Luria Bertani (LB) broth supplemented with 2 g/L of glucose until they reach an optical density (OD) of 0.8 at 600 nm. The cells are then induced by the addition of isopropyl β-D-1-thiogalactopuranoside (IPTG) to 1 mM and is incubated at 22°C for 48 hrs. The cells are then harvested by centrifugation and the resulting cell pellets are re-suspended in a lysis buffer containing 0.2 M Tris-HCl pH 7.5, 0.5 mM KCl,, 1 mM EDTA, and 1mM dithiothreitol (DTT). The re-suspended pellets are then lysed by sonication at 60% amplitude. The lysate

is cleared by centrifugation and the supernatant is transferred to a clean tube containing amylose resin. The supernatant is allowed to batch bind for 30 minutes and then it is washed with 5 column volumes of lysis buffer. The bound protein is eluted with lysis buffer supplemented with 10 mM maltose. The eluent is then concentrated on a 50 Kd molecular weight cut off filter. The expected yield of maturase protein from this protocol is approximately 25 mg/L of culture. Unfortunately when I tried to replicate these results, I recovered approximately 10 µg/L of culture. Further investigation revealed that the original protocol was for a truncated version of the maturase protein where the EN domain had been removed. This was not explicitly stated by the authors in the manuscript. Therefore, in order to purify full-length maturase for my structural biology needs, I need to develop a new purification protocol.

After extensive experimentation, a partially denaturing purification protocol yielded the best results. A single 6x-His tag was engineered at the N-terminus of MBP to allow for efficient affinity purification under denaturing conditions. Initial purification attempts used 8 M urea in the lysis buffer to fully denature the protein. After eluting with 250 mM imidazole I recovered approximately 4 mg of maturase per liter of culture. The 8 M urea was then removed via a stepwise dilution; however, this refolding procedure resulted in a white precipitate of aggregated protein (Figure 4.2). In order to avoid this, I performed a urea titration experiment to determine the minimum concentration of urea required to improve my protein recovery. My results indicated that 2 M urea was sufficient for recovery of full length MBP-fusion protein (Figure 4.3). In addition, when the urea was removed by 0.5 M stepwise reductions on-column, the protein remained soluble after elution.

The resulting soluble MBP-maturase protein was evaluated for splicing activity to determine if the refolding procedure yielded functional protein. In this procedure recombinant MBP-maturase protein was combined with *in vitro* transcribed *T.el*3c RNA in splicing buffer solution containing 10 mM $MgCl_2$, 500 mM NH_4Cl, 40 mM Tris-HCl pH 7.5, and 5 mM DTT. The mixture was then heated to 50°C for 10 minutes and the reaction was quenched by performing a phenol/chloroform extraction. The aqueous fraction, containing the spliced RNA, was then isolated and the RNA was recovered through an ethanol precipitation. The RNA pellet was dissolved in formamide and resolved via a 4% 19:1 acrylamide:bis-acrylamide PAGE gel with 8 M urea. Using this splicing assay, the intron/maturase complex splices

efficiently *in vitro* (Figure 4.4). The lariat bond of spliced RNA results in a retarded gel mobility, causing the product band run to slower.

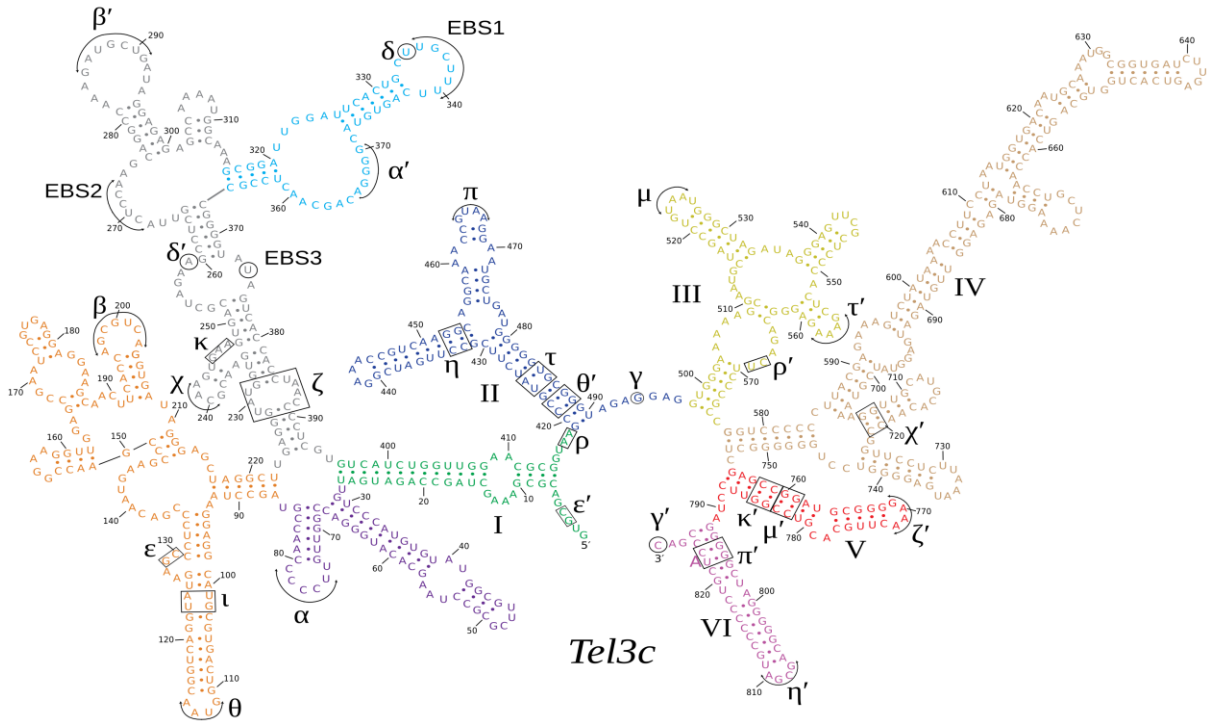


Figure 4.1: **The secondary structure of *T.el*3c**. Secondary structure of the 829-nucleotide RNA component of the Thermosynechococcus elongatus (*T.el*) group II intron retroelement. This intron shows sequence and structure conservation from the chloroplast-like class 1 intron class with six domains radiating outward from a central hub. Domain V is the most conserved and forms the active site (red). Domain VI contains the bulged adenosine required for lariat formation and retrotransposition (magenta)

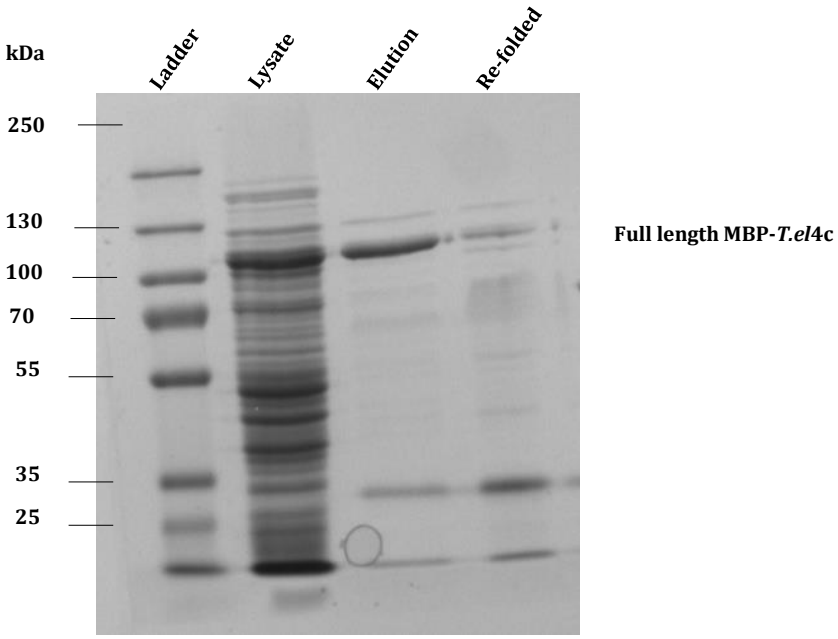


Figure 4.2: **MBP-*T.el*4c denaturing purification and refolding**. After denaturing purification, a large quantity of full length MBP-*T.el*4c protein was recovered in the elution fractions. Upon stepwise re-folding, the protein aggregated and became insoluble.

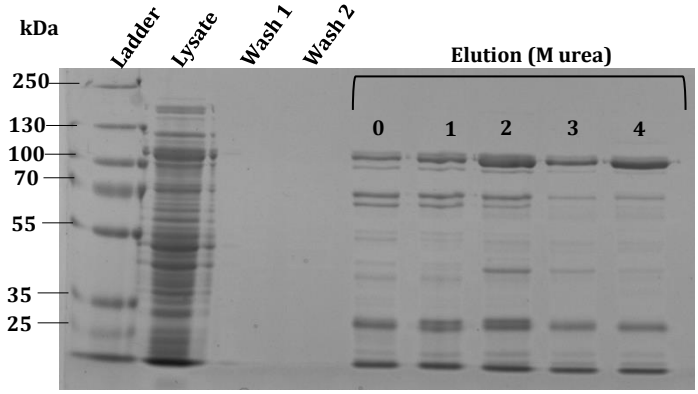


Figure 4.3: **Titration of urea**. Individual 6x-His tag batch purifications were performed. Each lysis buffer contained the concentration of urea noted (0 M, 1 M, 2 M, 3 M, 4 M). 2 M urea recovered the most full length MBP-*T.el*4c.

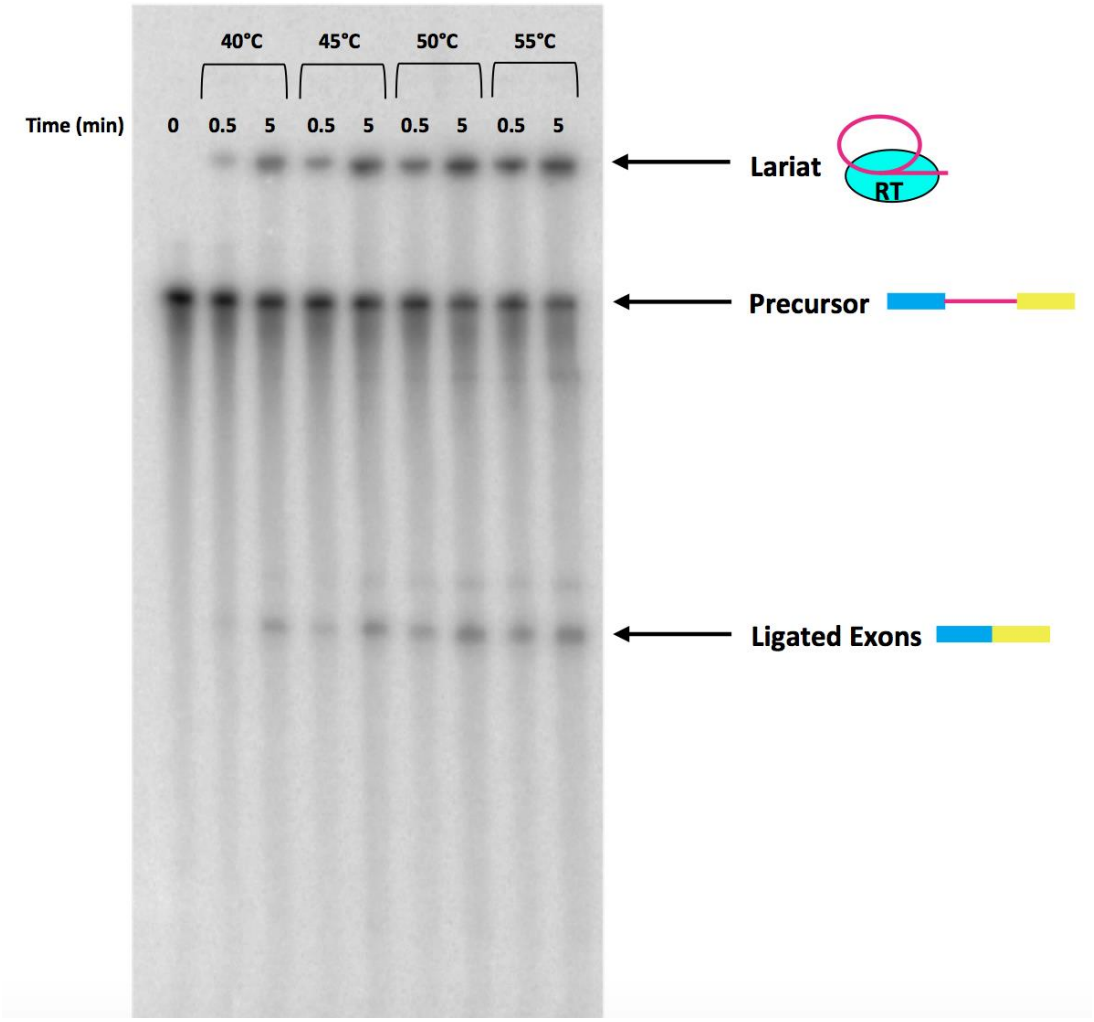


Figure 4.4: **Splicing gel of *T.el*3c/*T.el*4c**. *T.el*3c/*T.el*4c splicing samples were prepared and incubated in splicing buffer at the temperatures indicated. Aliquots were removed at 0.5 and 5 minutes and evaluated on a denaturing 4% PAGE gel with 8 M urea. All relevant bands are labeled.

4.2 Assembly and purification of RNP particles for EM structural studies

In preparation for future cryo-EM experiments, I developed a large-scale native purification to isolate intron/maturase complexes. Using the recombinant MBP-maturase protein and *in vitro* transcribed RNA, I assembled RNP in splicing buffer and allowed the mixture to sit at room temperature for 30 minutes. The solution was then placed at 50°C for 10 minutes followed by a concentration step using a 100 kD cutoff filter. The concentrated solution was then injected onto an AKTA purifier equipped with a Superdex 200 16/60 column. The mobile phase used was composed of 10 mM $MgCl_2$, 40 mM Tris-HCl pH 7.5, 300 mM NH_4Cl, and 1 mM DTT and the flow rate was set to 1 mL/min. The chromatography was monitored by UV absorbance at 280 nm and 1.5 mL elution fractions were collected. Two well resolved peaks were observed (Figure 4.5). Aliquots of each fraction across the two peaks were analyzed for lariat product using the splicing assay previously described in section 4.1 (Figure 4.6). The PAGE gel clearly shows that only peak two contains lariat RNA. *T.el*3c only forms lariat in the presence of *T.el*4c protein; therefore, peak two must contain RNP complex.

The same aliquots analyzed via PAGE gel were also evaluated using electron microscopy. For each sample, a negative stained grid using 2% uranyl acetate was prepared on carbon coated, 200 mesh copper grids. The grids were screened using a 200 kV FEI Tecnai Sphera. After inspection of micrographs from each aliquot, it was readily apparent that peak 1 contained aggregated RNA and peak two contained homogenously dispersed, well-ordered particles (Figure 4.7). In fact, particle quality improved in the later fractions within peak 2. Fraction 31 was chosen moving forward for all future cryo-EM experiments.

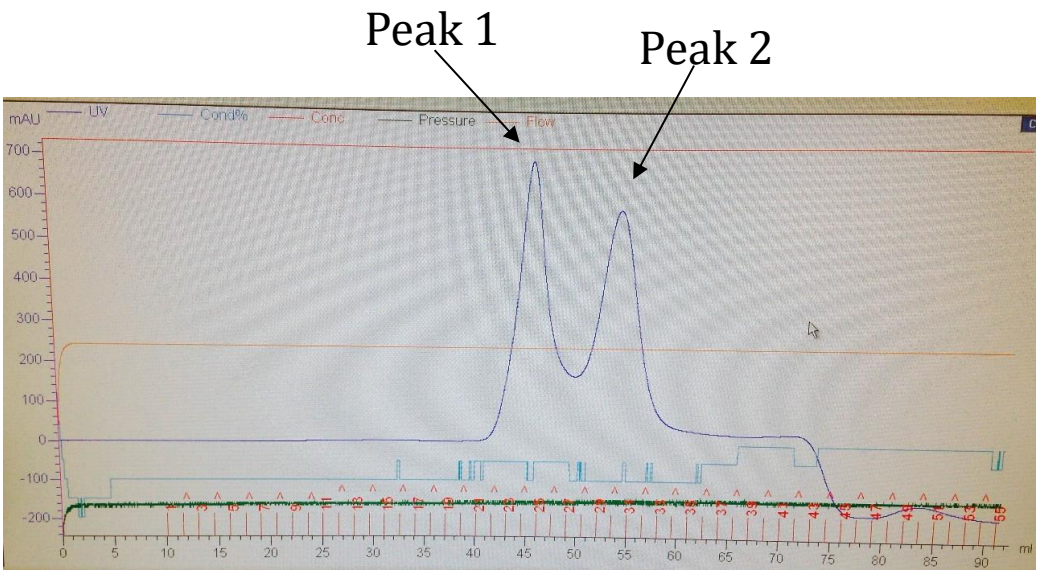


Figure 4.5: **Chromatogram of *T.el*3c/*T.el*4c gel filtration**. Two well resolved peaks are observed. Chromatography was monitored by UV absorbance at A280.

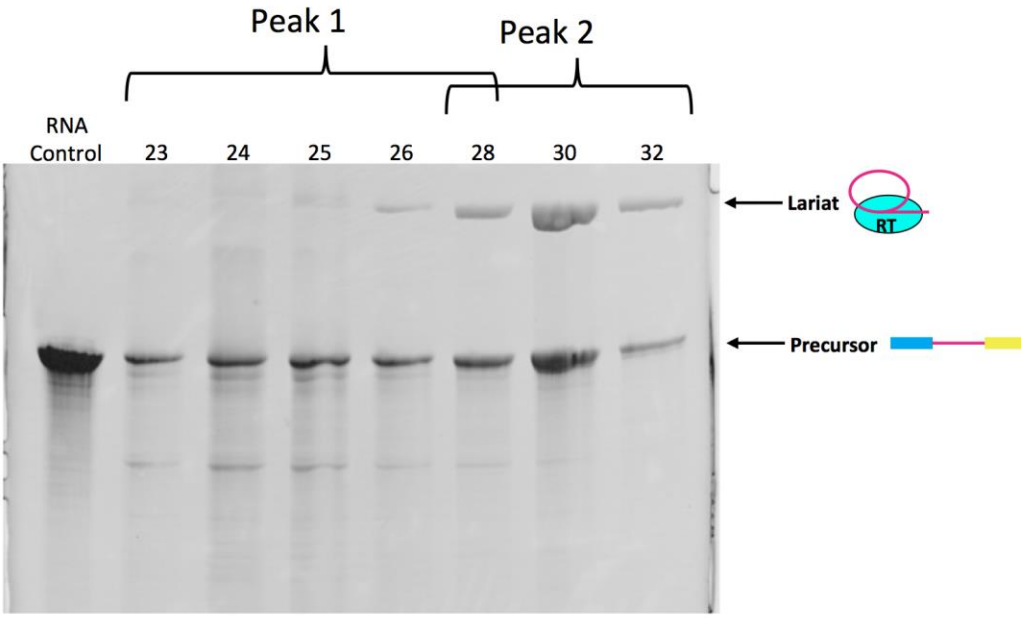


Figure 4.6: **Analysis of gel filtration fractions by gel electrophoresis**. Aliquots of each elution fraction were taken and analyzed on a denaturing 4% PAGE gel with 8 M urea. Only peak 2 contained lariat RNA indicating the presence of RNP particles.

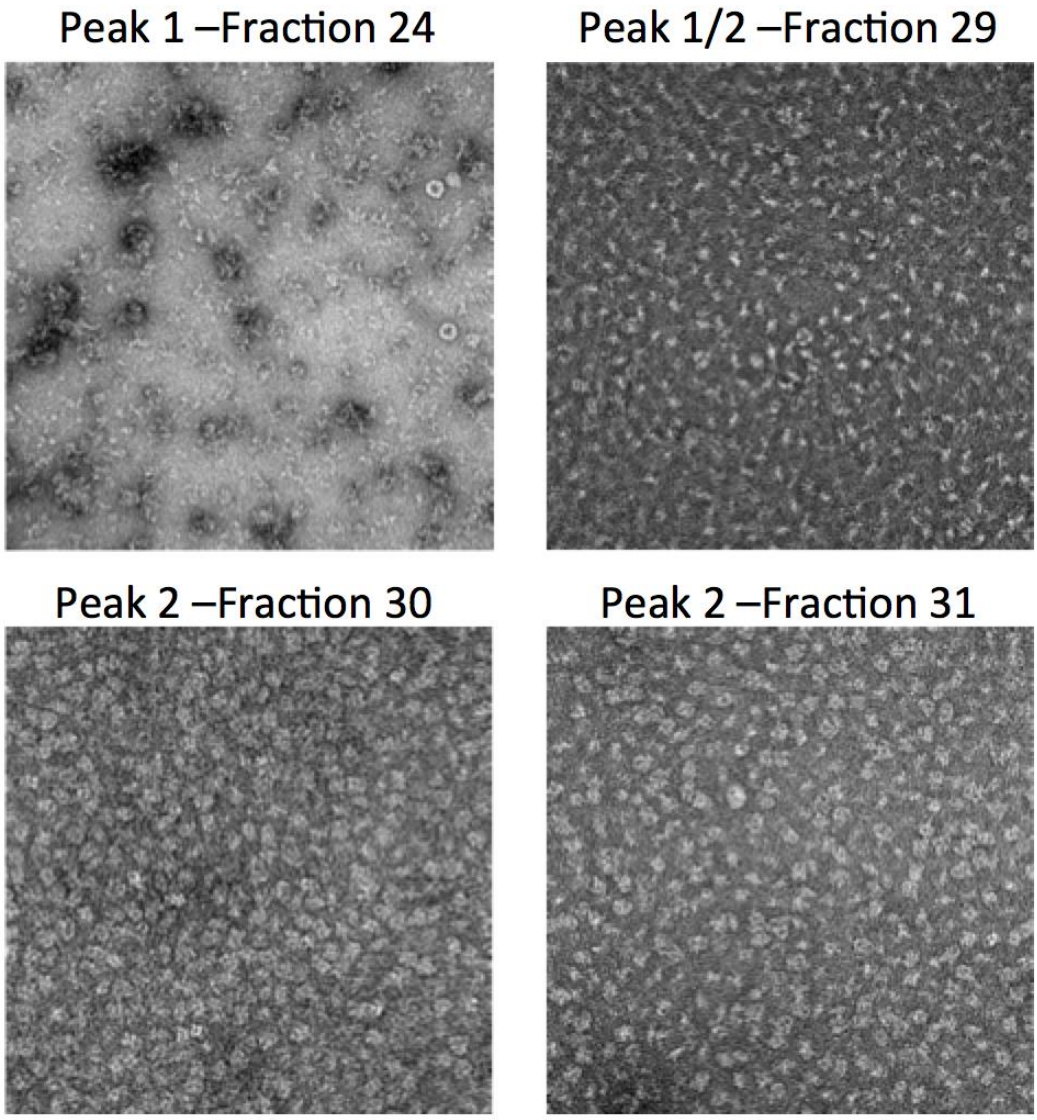


Figure 4.7: **EM grids of negative stained gel filtration fractions**. Elution aliquots from gel filtration were taken to prepare negative stained grids using 2% uranyl acetate. Peak 1 contained aggregated particles and peak 2 contained monodispersed, homogeneous specimen.

4.3 Cryo-EM experiments

Performing electron microscopy at cryogenic temperature with samples frozen in vitreous ice has many benefits [84]. At cryogenic temperatures, the kinetic energy of the atoms making up the molecule of interest is sufficiently low to prevent excessive particle movement during data collection. The low temperature also helps prevent radical diffusion caused by the electron beam that damages the sample. With negative stain applications, the samples are dehydrated and exposed to high concentrations of heavy metals. The grid preparation process therefore leads to molecules with collapsed structures that only vaguely resemble the morphology of the starting particle [85]. For cryo-EM, the samples are suspended in a thin layer of vitreous ice that immobilizes the particles in their native state. This allows researchers to evaluate molecular detail in a biologically relevant context. The specimen is also frozen instantaneously, allowing the resulting grid to be analyzed for intermediate structural states. Using computational methods [86,87], these different molecular states can be separated *in silico* and several different structures can be determined from a single data set. Lastly, the amount of specimen required for cryo-EM is very small compared to other structural biology methods. Specimens with solubility limitations or difficulties with expression and isolation are great candidates for cryo-EM experiments.

The main weakness of cryo-EM for decades was limited resolution. Theoretically, cryo-EM can reach resolutions well beyond 1 Å but in reality lens aberrations and detector quality caused 3D reconstructions to fall well short of its potential. There has been a renaissance in cryo-EM over the last 5 years with the development of direct electron detectors (DED) [88]. These new detectors remove the primary electron signal to light signal conversion that phosphor scintillator detectors use [89]. The ability to directly detect the position and intensity where the electron hits the detector is a tremendous advantage over traditional detectors. This improved detector sensitivity allows DEDs to have a high frame rate when collecting micrographs. With a high frame rate in combination with new motion correction algorithms, 3D reconstruction can overcome the restriction of beam-induced particle motion [90]. By aligning individual frames of a single micrograph, the software can compensate for particle motion. Additionally, higher doses of electrons can be used to improve contrast with this data collection strategy. The frames with higher accumulated does can simply be excluded or down-weighted during data processing after the frames have been aligned. This methodology takes advantage of the higher contrast for initial frame

alignment without having the drawback of including the later frames containing radiation-damaged particles. These technical advancements in cryo-EM have made it possible to achieve near atomic resolution of macromolecules as small at 150 kD [91]. The *T.el*3c/*T.el*4c RNP complex is approximately 360 kD, making it an ideal candidate for cryo-EM experiments.

To initiate cryo-EM experiments, RNP particles of *T.el*3c/*T.el*4c were purified as described in section 4.2. The resulting eluent was diluted to 100 ng/uL and frozen by hand in liquid ethane on R2/2 holey carbon grids (QUANTIFOIL). The grids were evaluated on a 300 KV FEI Tecnai Polara microscope equipped with a Gatan K2 detector for specimen quality and ice thickness. The particles were clearly visible and uniformly distributed within the holes (Figure 4.8). A data set of 1394 micrographs was collected using Leginon automated acquisition software [92]. The individual micrographs were taken with a total exposure of 16 seconds with a frame rate of 300 ms/frame. The total dose was 50 e/A^2. Each averaged micrograph was manually inspected for Thon ring quality and CTF-corrected using GCTF [86,93]. All data processing was done using RELION [86]. Approximately 10,000 particles were hand picked and used as references for automated particle picking. Using all 1,394 micrographs, 914,062 particles were boxed. After several rounds of iterative 2D classification, the best 127,600 particles were used to prepare 3D reconstructions (Figure 4.9). The particles were classified into six structural classes with a distribution as seen in Figure 4.10. Class 5 had the highest resolution (6.9 Å) as well as the best map quality. At this resolution RNA major/minor grooves are visible, which made it possible to begin building a model. Also at this resolution distinct protein features corresponding to the maturase protein were visible. Approximately 75% of the intron RNA was fitted to the model; however, at 6.9 Å *de novo* modeling of the maturase was not possible (Figure 4.11).

To improve the resolution, additional micrographs were collected and added to the initial data set. Another 700 micrographs were taken and when combined with the previous data, a total of 299,100 good particles were used for 3D reconstructions. The merged data sets were processed with cryoSPARC, which uses a newly published *ab initio* algorithm [87]. This new program allows for much faster data processing by significantly decreasing the computational requirements. With the increased data set and new software, the final 3D reconstruction reached 5.8 Å (Figure 4.12).

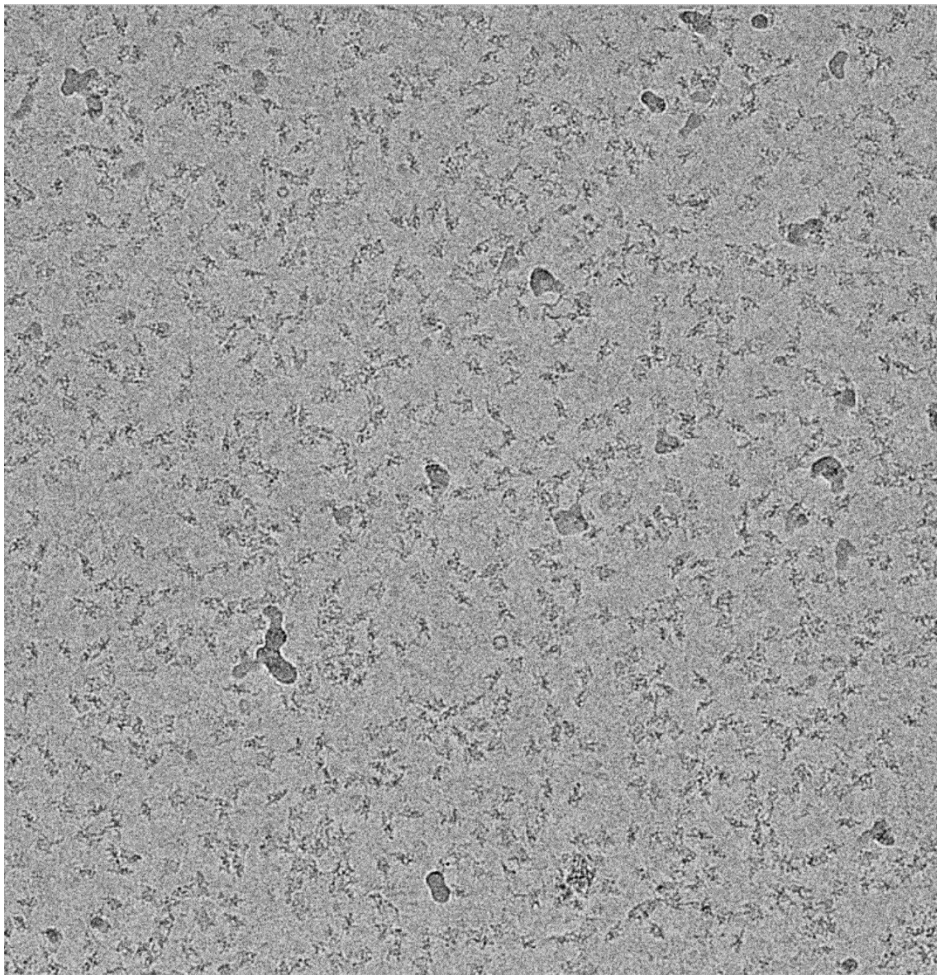
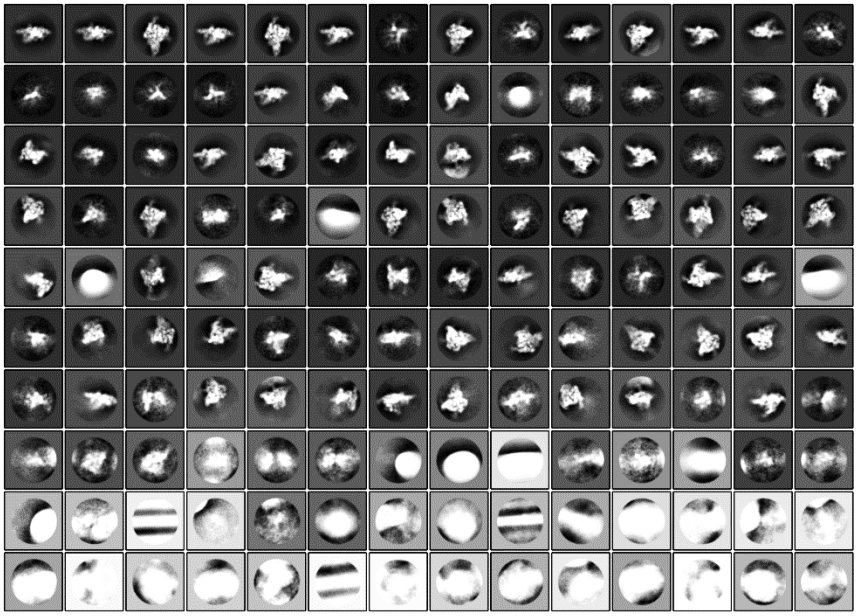


Figure 4.8: **Cryo-EM micrograph of *T.el*3c/*T.el*4c particles over holes**.

Initial 2D classes from
~900,000 particles



After iterative classification
~128,000 particles

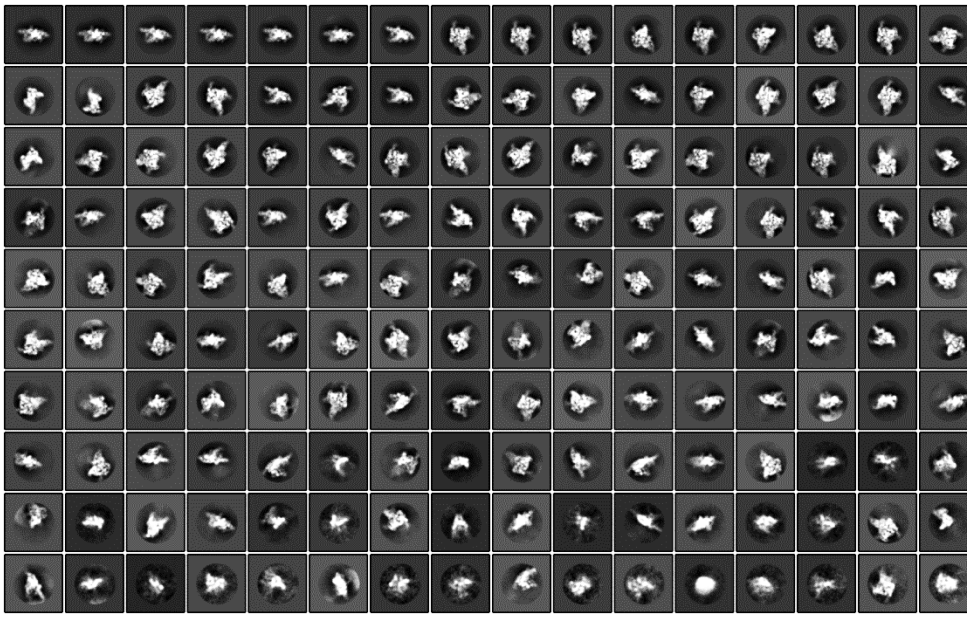


Figure 4. 9: **2D classifications of *T.el*3c/*T.el*4c picked particles**. Top) 2D class averages of entire data set. Bottom) 2D class averages after 3 rounds of iterative 2D classifications.

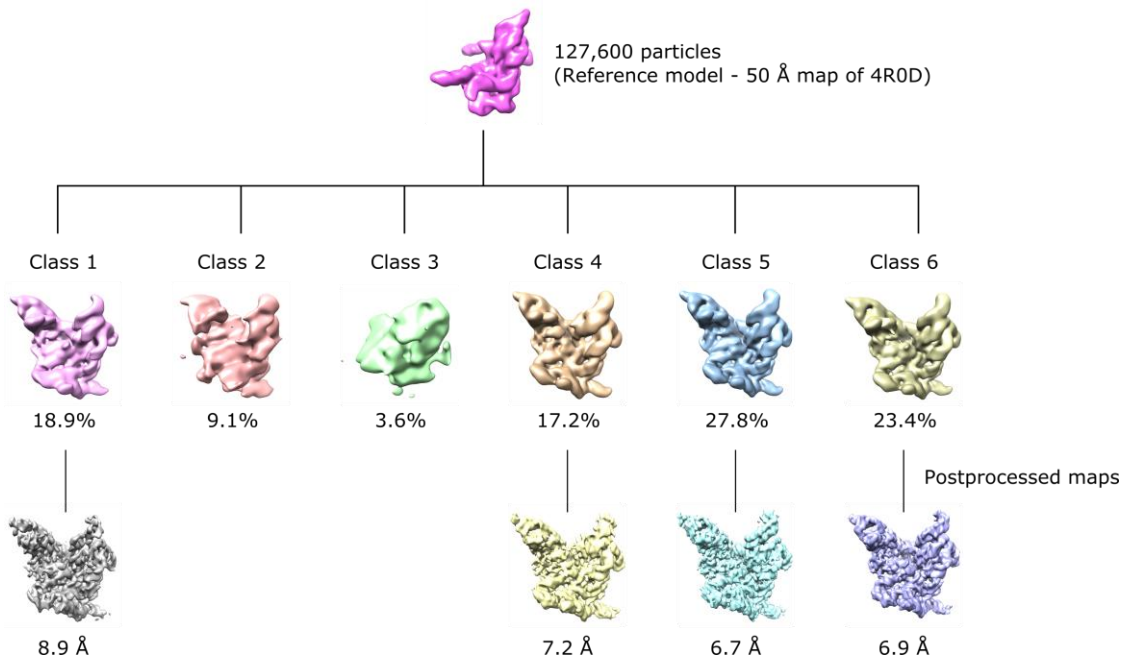


Figure 4.10: **3D reconstruction and classification of *T.el*3c/*T.el*4c**. Particles were separated into 6 classes. After evaluation of the maps, Classes 2 and 3 were poor quality. After comparison of the remaining maps, classes 1, 5, and 6 were in the same conformational state. Class 4 was a unique conformation differing the most in DII and DVI.

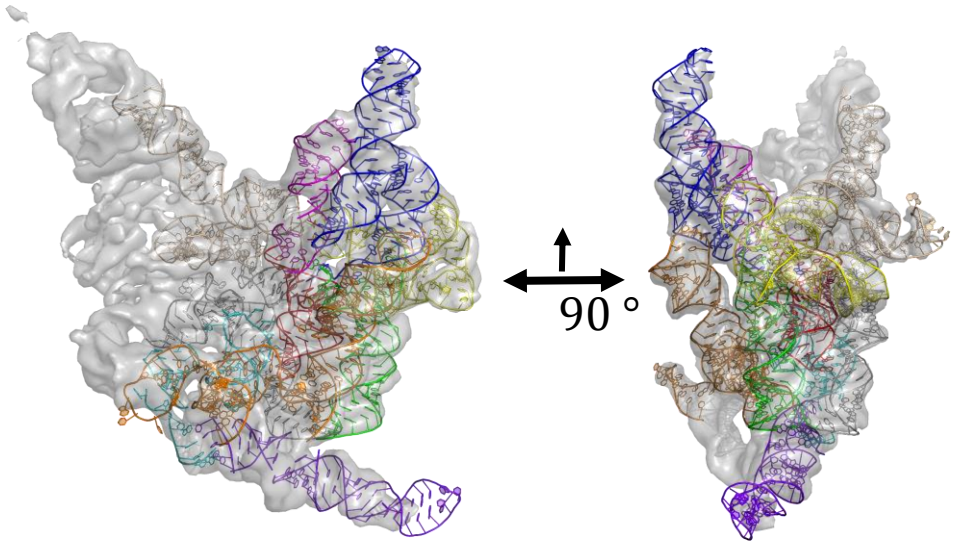


Figure 4.11: **RNA model fitted in Class 5 EM density**. EM density is shown in transparent grey. The RNA is shown in a colored cartoon representation. The RNA helices are colored as seen in Figure 4.1.

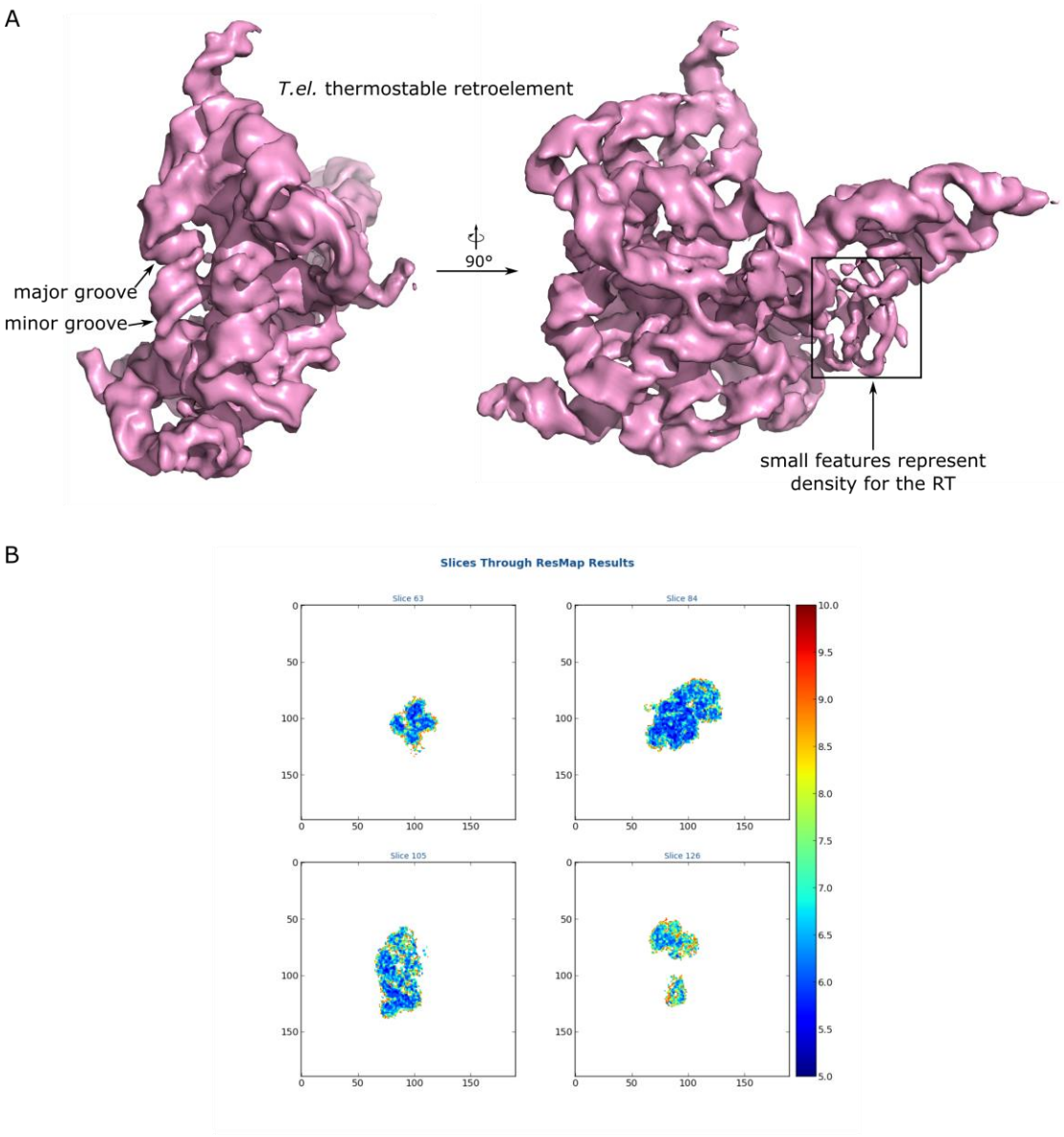


Figure 4.12: **Overall density of *T.el*3c/*T.el*4c EM map at 5.8 Å**. A) At this resolution major and minor grooves are clearly visible. Density corresponding to the maturase protein can be differentiated from RNA features. B) Local resolution map of the *T.el*3c/*T.el*4c density. The resolution is very uniform throughout the map.

4.4 Data analysis and discussion

The bulged adenosine is a universally conserved feature of all intron splicing, from group II introns to the spliceosome [46,94]. It is still unclear why this conservation exists as all ribonucleic acids have the requisite 2´-hydroxyl necessary for lariat formation. It is hypothesized that a conserved adenosine-binding pocket exists in the pre-catalytic state that hydrogen bonds specifically to this nucleotide, correctly positioning it in the active site. Unfortunately, there is no published structural information of a lariat-forming group II in its pre-catalytic state. The only available structure of a group II intron prior to the first step of splicing is a IIC intron from the organism *Oceanobacillus iheyensis* [44]. This structure revealed the kinked phosphate backbone at the scissile phosphate as its presented to the active site for cleavage; however, with the absence of DVI, this intron splices through hydrolysis making it impossible to gain insight into the positioning of the bulged adenosine during the first step of splicing. A lariat-forming group II intron pre-catalytic structure would help to answer this question along with any other structural requirements for lariat formation.

Using cryo-EM, I was able to determine a group II intron/maturase complex to 5.8 Å. Structure classification during the initial 3D reconstruction separated the data into 6 different maps. Classes 2 and 3 were composed of a small portion of the data set (approximately 12.7%) and provided reconstructions of poor quality (Figure 4.10). After comparison of the density, classes 1, 5, and 6 were identified as being in the same conformational state. The high map quality allowed for the accurate positioning of RNA helices to prepare an initial model. Based on this model, it became clear that these three classes represented the intron in the post-catalytic state as strong lariat bond density was observed. Interestingly, when class 4 was overlaid with the post-catalytic classes, the lariat density was absent. *T.el*3c/*T.el*4c does not splice through hydrolysis, meaning that the absence of lariat bond density indicates the intron is in the pre-catalytic state. Therefore, class 4 represents the first structural insight into the pre-catalytic state of group II intron splicing.

Alignment of the two maps also revealed significant difference in the positioning of DII and DVI (Figure 4.13). These two domains contain several important tertiary interactions. Previous structural and biochemical experiments have shown that both π-π´ and η-η´ play an important role in transitioning the intron through the steps of splicing [76]. The π-π´ interaction is located directly adjacent to the bulged

adenosine and is responsible for pulling the lariat bond out of the active site after the first step of splicing. In the post-catalytic state, the π-π´ interaction holds the lariat bond approximately 20 Å away from the intron active site. In order for the lariat formation to take place, this region must be dynamic to allow the bulged adenosine to be positioned correctly to cleave the scissile phosphate. The dynamics of DII and DVI observed in the EM density between the pre and post-catalytic states correlate with this model of splicing. With only a single active site, group II introns must shuffle the 5´ and 3´ splice site substrates efficiently. A dynamic π-π´ and η-η´ support this hypothesis by allowing flexibility in the active site.

At the current resolution, it is clear that π-π´ and η-η´ are playing a role in positioning the scissile phosphate for the first step of splicing. Unfortunately, in order to make any detailed mechanistic insights, the resolution of the map will need to be improved to better than 4 Å. Upon closer inspection of the two data sets, it was recognized that the ice thickness of the prepared grids was limiting our high spatial frequency Thon rings. With the absence of this structural information, it was impossible to improve the resolution beyond 5.8 Å. When attempts at preparing specimen with thinner ice were made, the particles began to aggregate and display severe orientation preference. These two obstacles, when preparing specimen in thin ice, have prevented continued progress and are my current focus of research to push the resolution of *T.el*3c/*T.el*4c beyond 4 Å.

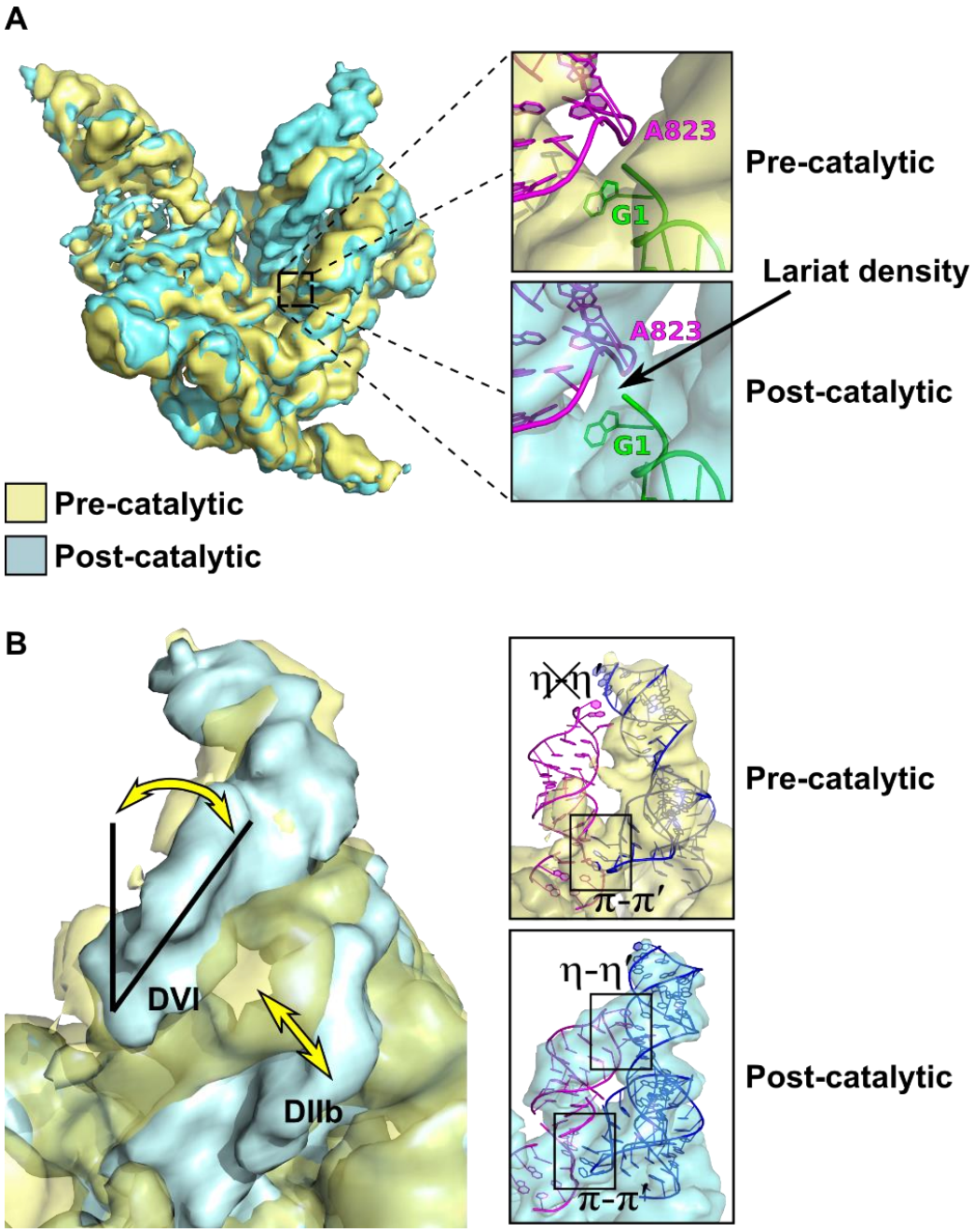


Figure 4.13: **Superposition of pre-catalytic and post-catalytic EM density of *T.el*3c/*T.el*4c**. A) An overview of the pre and post-catalytic EM density maps is shown. The two maps are identical in all domain except for DII and DVI. Additionally, the yellow density shown and obvious absence of lariat density indicating it is in the pre-catalytic state. B) An approximate 25 Å shifts occurs between the two maps. This suggests that DII and DVI are dynamic during the different steps of splicing.

4.5 Materials and methods

Plasmid cloning

A 6x-His-MBP-*T.el*4c maturase DNA gene was synthesized (Genscript) and cloned into a pET15b vector using the NdeI and BamHI restriction sites. The cloned plasmid was then transformed into Rosetta 2 cells (NEB). The *T.el*3c intron gene was synthesized (Genscript) and cloned into a pUC57 plasmid using the EcoRV restriction site. The cloned plasmid was transformed into DH5α cells.

T. elongates denaturing maturase purification

A 100 mL overnight culture of 6x-His-MBP-*T.el*4c is prepared in LB containing carbenicillin. 20 mL of the overnight culture is diluted into 2 L of LB containing carbenicillin. The cells are grown to an optical density of 0.8 and then induced with 1 mM IPTG. The cells are placed at 22°C for 48 hrs. Cells are harvested by centrifugation at 5,000 rpm for 10 minutes at 4°C. The cell pellets are re-suspended in 100 mL of lysis buffer (20 mM Tris-HCl pH 7.5, 500 mM KCl, 2 M urea, 10 mM imidazole, 5 mM 2-mercaptoethanol, and PMSF). The re-suspended cells are lysed using a probe sonicator at 60% amplitude for a total of 80 seconds. The lysate is cleared of cell debris by centrifugation at 12,000 rmp for 45 minutes at 4°C. The resulting supernatant is transferred to a clean tube and 2 mL of Ni-NTA (Qiagen) is added. The mixture is allowed to batch bind for 1 hr at 4°C. The resin is collected in an empty Bio-Rad gravity purification column. The resin is first washed with 5 column volumes of lysis buffer followed by 5 column volumes of a high salt wash (20 mM Tris-HCl pH 7.5, 1.5 M KCl, 2 M urea, 10 mM imidazole, 5 mM 2-mercaptoethanol). The resin is re-equilibrated to 500 mM KCl by rinsing with 5 column volumes of lysis buffer with 10% glycerol. The bound protein then undergoes a re-folding protocol on column by stepwise reduction in urea. The resin is first washed with 5 column volumes urea buffer 1 (20 mM Tris-HCl pH 7.5, 500 mM KCl, 1.5 M urea, 10 mM imidazole, 5 mM 2-mercaptoethanol, 10% glycerol), followed by 5 column volumes of urea buffer 2, 3, and 4 (urea buffer 2: 20 mM Tris-HCl pH 7.5, 500 mM KCl, 1.0 M urea, 10 mM imidazole, 5 mM 2-mercaptoethanol, 10% glycerol), (urea buffer 3: 20 mM Tris-HCl pH 7.5, 500 mM KCl, 0.5 M urea, 10 mM imidazole, 5 mM 2-mercaptoethanol, 10% glycerol), and (urea buffer 4: 20 mM Tris-HCl pH 7.5, 500 mM KCl, 10 mM imidazole, 5 mM 2-mercaptoethanol, 10% glycerol). The bound protein was eluted using urea buffer 4 supplemented with 250 mM imidazole. The

recovered protein is buffer exchanged with filtration buffer (20 mM Tris-HCl pH 7.5, 500 mM KCl, 5 mM 2-mercaptoethanol, 10% glycerol) using a 50 kD cut off filter (EMD-millipore). The protein containing solution is rinsed a total of six times with 14 mL of filtration buffer. For the last buffer exchange, the protein solution is concentrated to 500 uL and brought to 50% glycerol for long term storage at -80°C.

In vito RNA transcription

*T.el*3c plasmid is first linearized using an engineered HindIII restriction site. *In vitro* transcriptions are prepared in a 1 mL total volume. 40 µg of linear DNA template is added to transcription buffer containing: 50 mM Tris-HCl pH 7.5, 17.5 mM $MgCl_2$, 5 mM DTT, 2 mM spermidine, 0.05% triton-100, and 2 mM of each NTP. In-house purified T7 polymerase is added along with thermophilic inorganic phosphatase to initiate transcription. The reaction is placed at 37°C for 3 hrs. TURBO DNase (20 units) is then added along with 12 uL of 100 mM $CaCl_2$ and allowed to react for 1 hr at 37°C. Proteinase K (200 µg) is then added at 37°C for 1 hr. The mixture is centrifuged to remove any precipitate. The supernatant is filtered through a 0.2 µM filter into a 100 Kd cut off filter. The RNA containing solution is buffer exchanged a total of 7 times with 14 mL of filtration buffer (5 mM cacodylate pH 6.5 and 10 mM $MgCl_2$). For the last buffer exchange, the RNA is concentrated to approximately 10 mg/mL.

Chapter 5: Structure determination of a group II intron retrotransposon

5.1 Identification and analysis of a group II intron retrotransposon

The ability of group II introns to act as selfish mobile genetic elements has dramatically shaped the evolution of eukaryotes. At least 46% of the human genome is made up of a class of mobile genetic elements known as non-LTR retrotransposons and their remnants [95,96]. These genetic sequences use a target primed reverse transcription mechanism to copy and paste themselves into new locations within a genome [37]. Group II introns share this mechanism with non-LTR retrotransposons as well as sequence homology between their protein chaperons. These similarities suggest they share an evolutionary ancestry.

In particular, LINE elements in humans have been linked with regulating gene expression through insertions near transcriptional promoters [97]. This phenomenon occurs with regularity in somatic tissue causing neuronal diversity that has affected brain development in primates [48]. LINE elements have proved difficult to study structurally because of their low activity and stability *in vitro* [98]. In order to study the structural biology of retrotransposition, I first developed an activity assay using an intron from the organism *T. elongatus.* As previously discussed, this organism contains 28 related copies of a group IIB intron. These introns have diverged from one another to varying degrees and some have maintained a high level of mobility activity. For the purposes of my work, I selected *T.el*4h, as it is the most active intron for mobility from this organism [82]. Several mutations were made to the wild type (WT) sequence of both the maturase protein and intron RNA to maximize the mobility activity. The WT-maturase protein contains a degenerate RT active site where the required YADD motif is replaced with YAGD. This single D to G mutation significantly affects the RT activity by interrupting the catalytic metal ion binding. For the purposes of this study, the RT active site was restored to the consensus YADD sequence and I will refer to this mutant sequence as the *T.el*4h maturase protein. The *T.el*4h RNA had also undergone mutational degeneration to slow down the retrotransposition activity of the intron. The δ-δ′ interaction has undergone a mutation in the WT-RNA sequence. A simple C236T mutation was made to the RNA gene to restore the AU base pair of δ-δ′.

Using the denaturing purification discussed in section 4.1, I was able to purify milligram quantities of the *T.el*4h maturase protein. When combined with *in vitro* transcribed *T.el*4h intron RNA, the complex

spliced efficiently. A mobility assay was developed to monitor the activity of the various processes during TPRT. Essential to this assay are ^{32}P radiolabeled DNA primers that when annealed mimic the canonical dsDNA target site of insertion. By selectively labeling the top or bottom strand, I can individually monitor RNA target site insertion, En activity, and RT primer extension (Figure 5.1). To perform this assay, *T.el*4h RNA and *T.el*4h maturase protein are first spliced at 50°C in splicing buffer (10 mM $MgCl_2$, 40 mM Tris-HCl pH 7.5, 500 mM NH_4Cl, and 5 mM DTT) for 5 minutes to assemble RNP particles. Radiolabelled dsDNA target is then added along with a 200 µM mixture of dNTPs. After mixing, the reaction is placed back at 50°C for an additional 5 minutes. The reaction is quenched by performing a phenol/chloroform extraction and the RNA is recovered through an ethanol precipitation using linear polyacrylamide as a carrier. The RNA pellet is the re-suspended in formamide and run on a 4% 19:1 acrylamide:bis-acrylamide PAGE gel with 8 M urea. To monitor RNA strand invasion, the top DNA primer is P^{32} labeled on the 5´ end. RNA insertion can be monitored between both steps reverse of splicing as seen in Fig 5.2. After analysis of the radiolabeled bands, approximately 25% of the RNA fully integrated into the target DNA. This is the highest mobility ever reported in an *in vitro* mobility assay. To investigate En activity and cDNA synthesis by the RT, the bottom DNA primer was labeled with ^{32}P on the 5´ end. Unfortunately when performed, no En activity was observed. Without bottom strand cleavage, the RT has no primer to initiate cDNA synthesis. It was hypothesized that perhaps the En domain required a unique metal ion in order to efficiently bind the dsDNA substrate as it contains zinc-finger domains. Several metal ions were chosen and tested at various concentrations (Mn^{2+}, Zn^{2+}, CO^{2+}, $Ni^{2,}$, Fe^{2+}). No conditions tested showed activity for bottom strand cleavage by the En domain. Genetic mobility assays performed in *E.coli* suggest that the En domain is active *in vivo*; however, no direct test of nuclease activity has been shown in the literature for the *T.el*4h maturase protein. It is possible that my denaturing purification causes misfolding at the C-terminus where the En domain exists.

Regardless of the lack of En activity, I still wanted to assess cDNA synthesis by the RT. To overcome the absence of bottom strand cleavage, I used a pre-nicked bottom strand when annealing the target primers. The exact cleavage site is unknown but has been shown to vary between -6 and -11 from the RNA insertion site. Therefore, I prepared several targets with a variety of nicked positions on the bottom strand. Primer extension was clearly visible for all nicked bottom positions (Figure 5.3). The

efficiency of cDNA synthesis differed depending on the position of the nick. Having confirmed the mobility activity using the developed assay, the *T.el*4h/*T.el*4h RNP proved to be a good candidate for future structural studies.

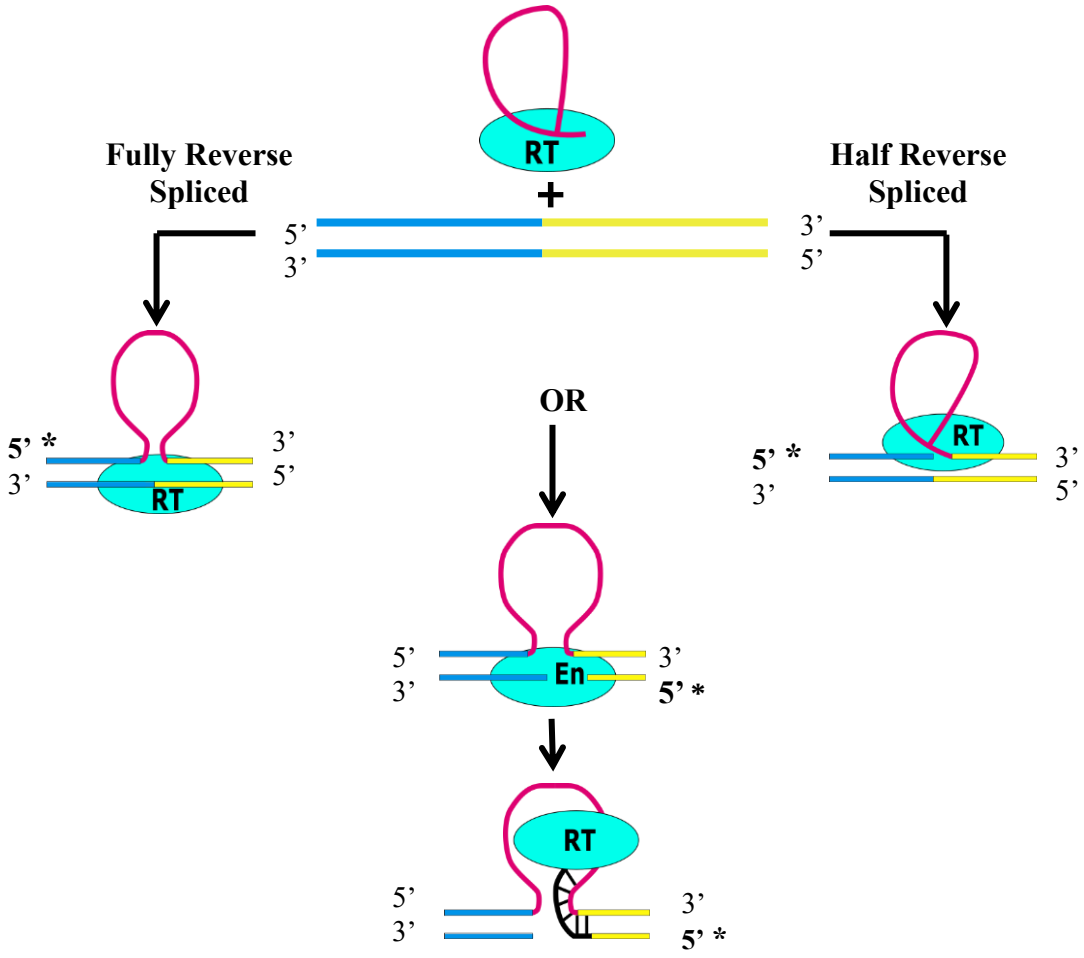


Figure 5.1: **Retrotransposition assay scheme of *T.el*4h/*T.el*4h**. By selectively radiolabeling the 5´ end of either the top or bottom strand, different processes during retrotransposition can be monitored. Top strand labeling monitors RNA insertion into the target DNA. Bottom strand labeling monitor En strand cleavage and cDNA synthesis by the RT.

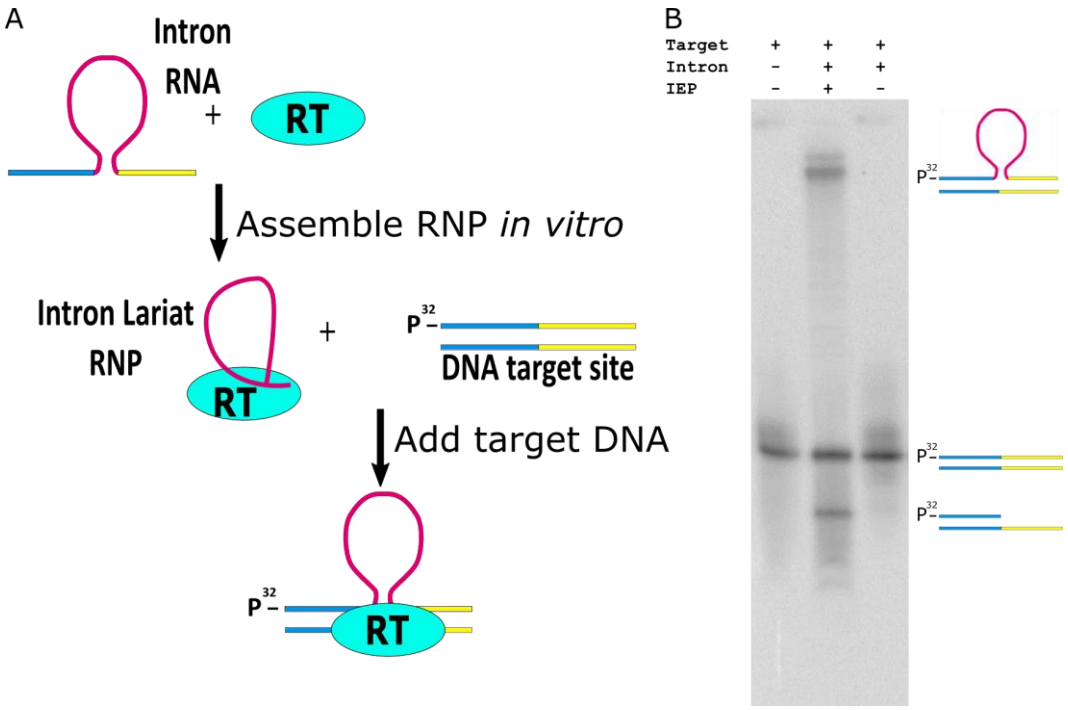


Figure 5.2: ***T.el*4h/*T.el*4h DNA insertion efficiency**. A) a graphical representation of the mobility assay. B) A radioactive gel of the mobility assay. Bands are labeled graphically and mobility activity only occurs in the presence of all three components.

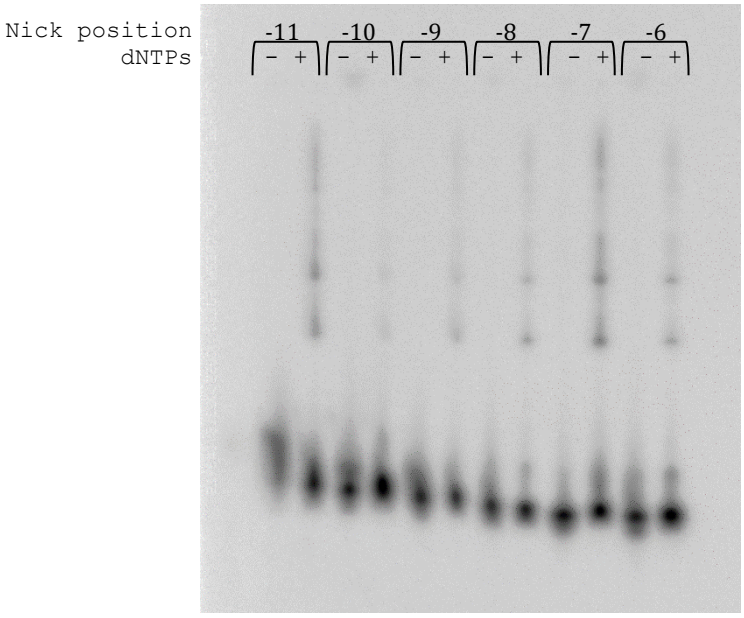


Figure 5.3: **cDNA synthesis of *T.el*4h/*T.el*4h**. Primer extension can be clearly seen in the gel for all nicked targets tested. Nicks at -11 and -7 show the strongest cDNA synthesis.

5.2 Purification of an active group II intron retrotransposon

To initiate cryo-EM experiments, a purification method was developed to isolate active group II intron retrotransposons in the process of invading dsDNA (Figure 5.4). It is important that the assay be able to differentiate RNP particles from RNP particles bound to DNA. Therefore, the target DNA substrate was used as the purification tag. A biotinylated DNA primer was purchased from Integrated DNA Technologies (IDT) and used to prepare the annealed dsDNA target. The same assembly condition as the mobility assay described in section 5.1 was used; however, 6 µg of biotinylated DNA target was used. The resulting reaction mixture was centrifuged to clear precipitant and the supernatant was added to 200 µL of equilibrated streptavidin resin. The solution was allowed to bind for an hour before the solution was rinsed with 6 mL of wash buffer 1 (10 mM $MgCl_2$, 40 mM Tris-HCl pH 7.5, 500 mM NH_4Cl, 1 mM DTT, and 5% glycerol) a total of 4 times. The resin is then rinsed with 6 mL of wash buffer 2 (10 mM $MgCl_2$, 40 mM Tris-HCl pH 7.5, 300 mM NH_4Cl, and 1 mM DTT) a total of 4 times. The bound RNP/DNA particles were finally eluted with elution buffer containing 5 mM biotin. The purification efficiency was monitored by analyzing the different fractions after phenol/chloroform extraction on a 4% 19:1 acrylamide:bis-acrylamide PAGE gel with 8 M urea. After soaking the gel in a solution of ethidium bromide, the elution lane contained no visible RNA (Figure 5.5). Biotin is known to bind very tightly to streptavidin and only elutes under very harsh denaturing conditions [99]. Based on this result, the bound RNP/DNA molecules were attached irreversibly to the resin.

An alternative biotin analog, desthiobiotin, is commercially available and can be purchased conjugated to synthetic DNA primers from IDT. Desthiobiotin has a three orders of magnitude lower affinity for streptavidin, which would improve the elution characteristics of the purification [100]. When this new approach was applied to the purification, the elution lane contained RNA correlating to the *T.el*4h intron (Figure 5.6). Phenol/chloroform extraction of the left over resin showed a large quantity of complex still bound to the streptavidin. In an attempt to shift the elution equilibrium even further towards unbound complex, the purification was repeated with a saturated biotin elution buffer in place of the 5 mM biotin solution used previously (Figure 5.7). This modification improved the elution efficiency further and yielded sufficient RNP/DNA complex for future cryo-EM studies.

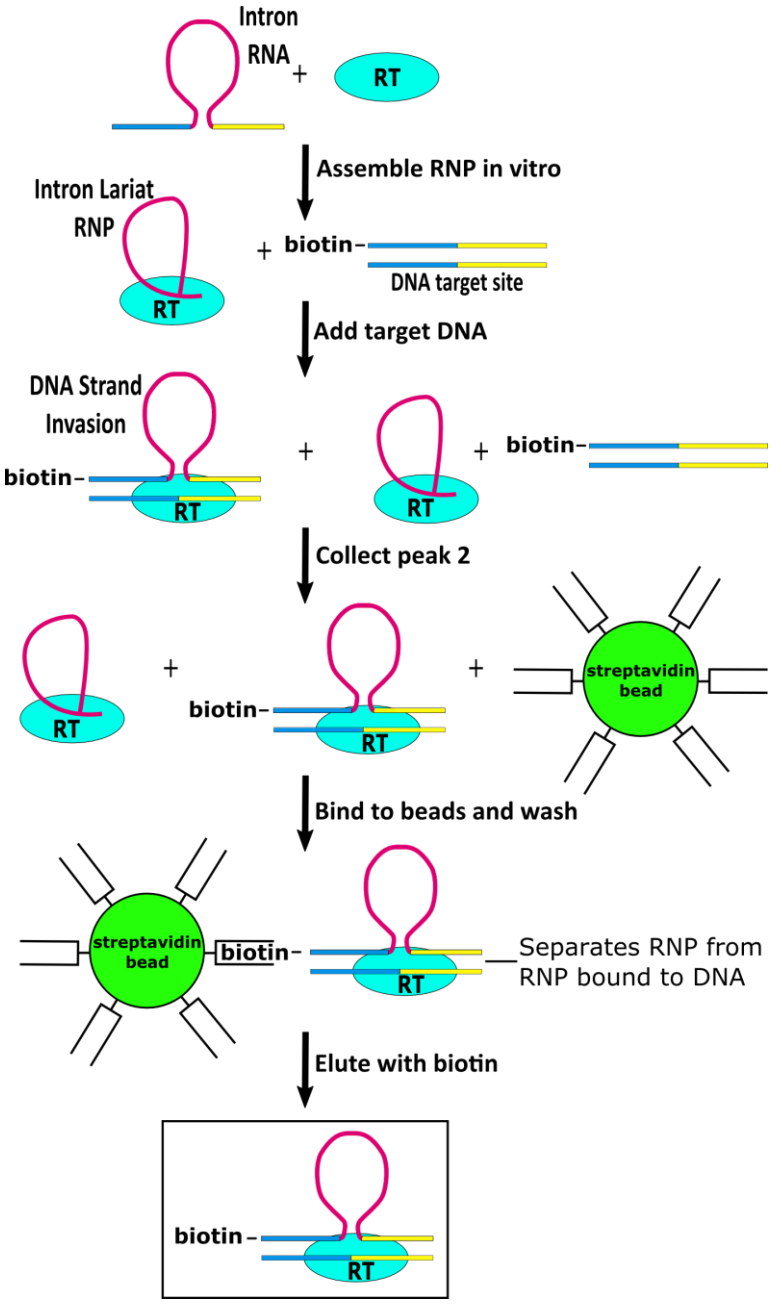


Figure 5.4: **Purification scheme for an active group II intron retrotransposon**. Exons and DNA target are represented in blue and yellow. The intron RNA is represented in pink while the maturase protein is shown in cyan. The target DNA is biotinylated on the 5′ end of the top stand. Streptavidin beads are represented with green spheres.

biotin / 5mM biotin elution

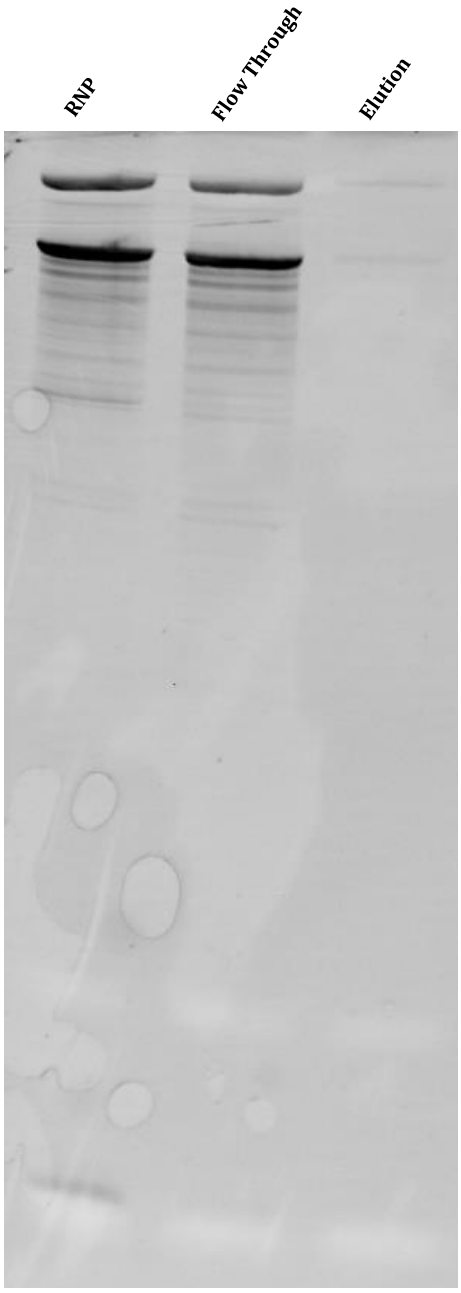


Figure 5.5: **Initial biotin purification results**. A denaturing 4% PAGE gel showing the results of the initial mobility purification. The elution lane is void of RNA due to inefficient release of the bound particles.

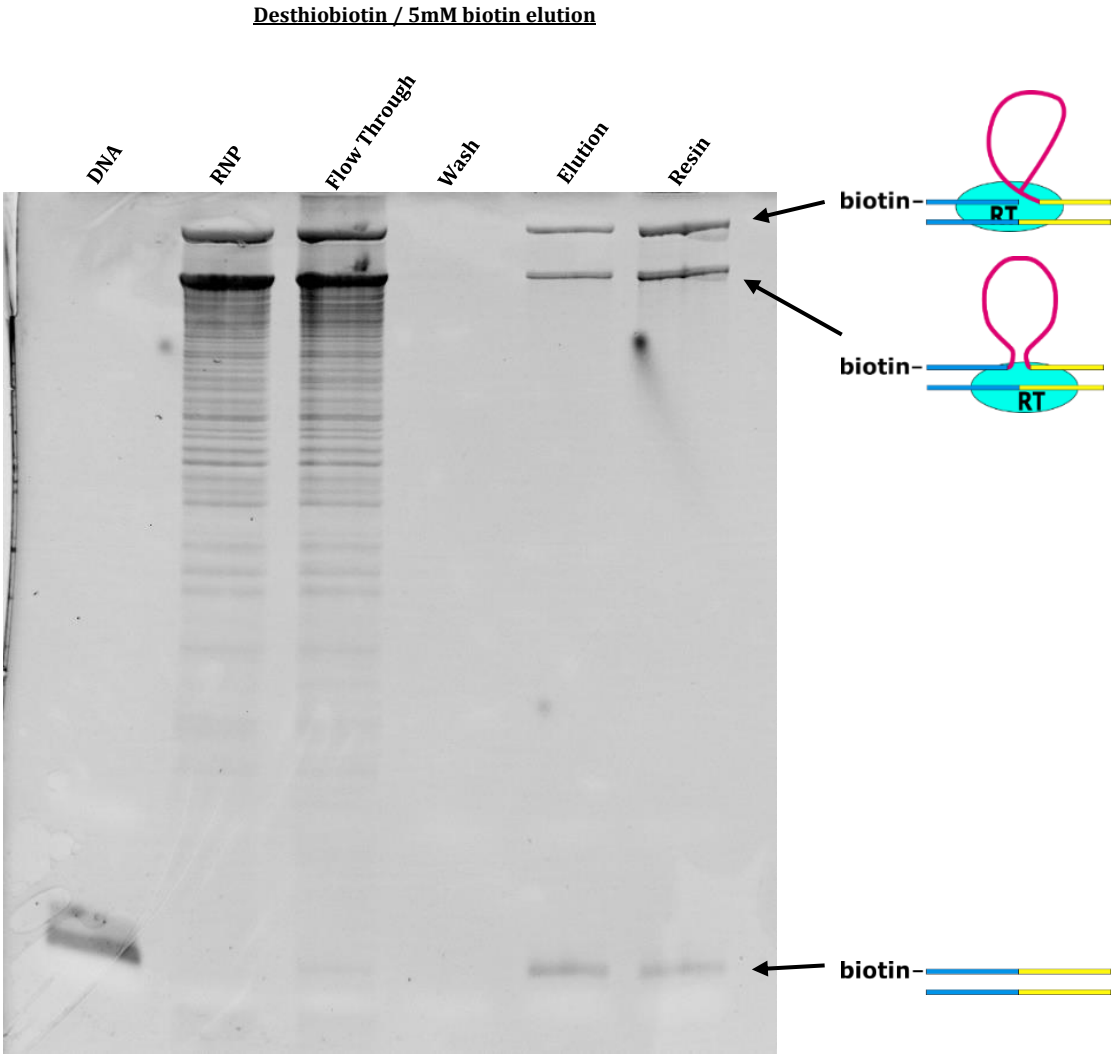


Figure 5.6: **Initial desthiobiotin purification results**. The replacement of biotin with desthiobiotin allows the bound RNP particles to elute. When analyzed after elution, the resin still contained a large fraction of bound specimen.

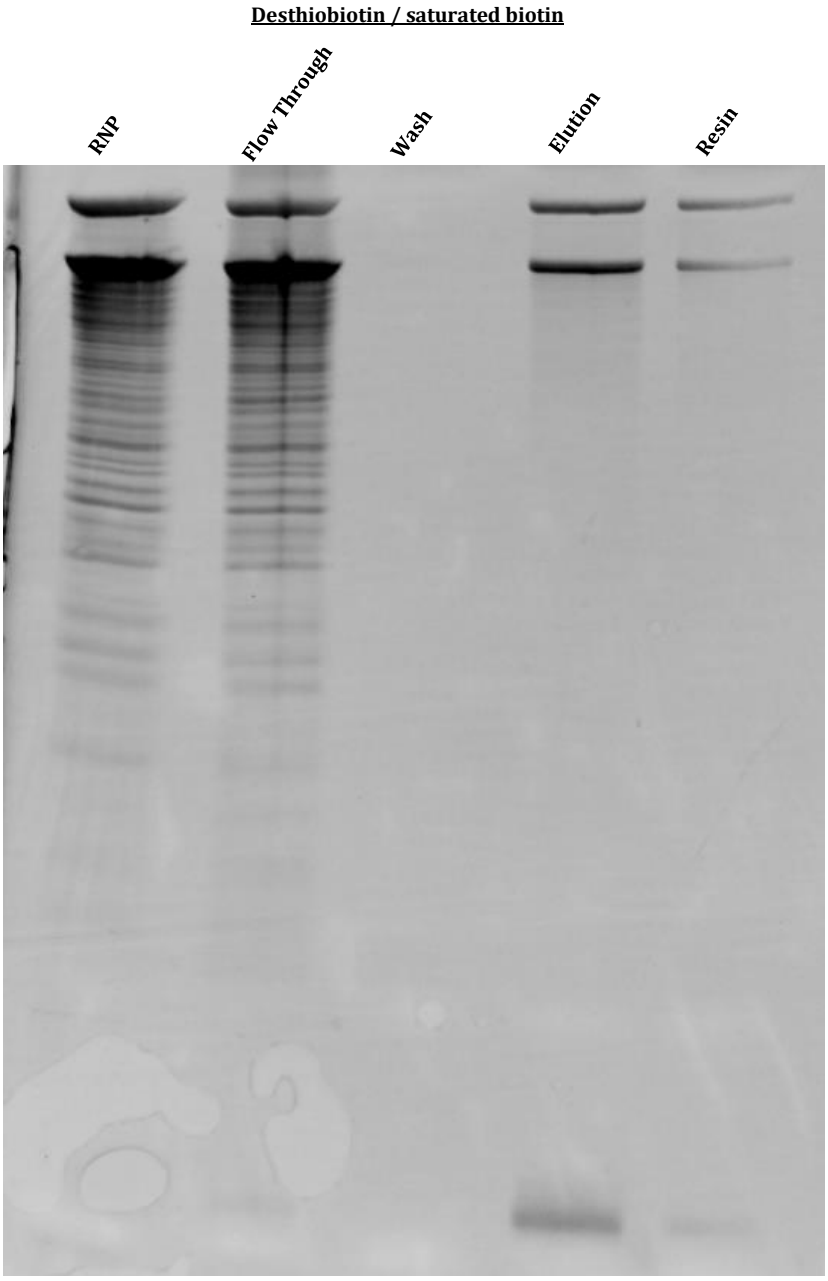


Figure 5.7: **Results of final desthiobiotin purification protocol**. A saturated solution of biotin is used to elute the bound RNP particles from the streptavidin resin. This change causes a drastic shift in the equilibrium between bound and unbound particles. This is evident by comparing the bands in the elution lane to the bands in the resin lane.

5.3 Cryo-EM experiments

Purified *T.el*4h/*T.el*4h RNP bound to a target DNA substrate was frozen in liquid ethane on R1.2/1.2 holey carbon grids (Quantifoil) using a FEI Vitrobot. Grid preparation was switched from hand freezing to the Vitrobot to improve grid reproducibility and to obtain thinner ice. The RNP/DNA solution was concentrated to 100 ng/uL before grid preparation. The frozen grids were screened on a 300 KV FEI Polara microscope equipped with a Gatan K2 detector. Over the course of 2 separate data collection runs, a total of 5,095 micrographs were taken. Micorgraphs were recoreded with a frame rate of 200 ms and a total exposure time of 10 seconds using Leginon automated acquisition software [92]. This led to a total dose of ~ 80 e/Å^2. After manual inspection of micrograph quality, only 3,731 micrographs were selected for automated particle picking. From those micrographs, 201,823 particles were automatically boxed using RELION [86]. After 3 rounds of iterative 2D classification in RELION, 73,770 particles remained in the data set. These particles then underwent *ab initio* 3D reconstruction using cryoSPARC software where six classes were arbitrarily set [87]. Of the six classes, five contained reconstructions of sufficient quality. These five classes contained 92.7% of the particles from the input data set. An additional *ab inito* 3D reconstruction was done using the remaining particles with six classes set arbitrarily. Three classes, containing 54% of the particles were considered good and chosen for a homogenous map reconstruction. The final map resolution is 4.8 Å and includes 36,927 particles (Figure 5.8). The overall map quality is good, but unfortunately the resolution is being limited by an orientation preference of the particles. In thin ice, the particles are associating with the air-water interface causing both the aggregation and orientation bias. Our previous data collection of the *T.el*3c/*T.el*4c specimen was in much thicker ice and did not display this level or particle orientation preference; however, the thicker ice still affected the maximum resolution by limiting my ability to collect the high spatial frequency data in the micrographs.

Attempts at solving this particle orientation issue are underway. A combination of construct variation and support film optimization could lead to an improvement. Additionally, specimen tilt data is currently being collected. In this method, rotating the specimen stage fills in the missing views of the particles [101]. The tilted data is then combined with un-tilted data to form a merged data set with a more complete set of views.

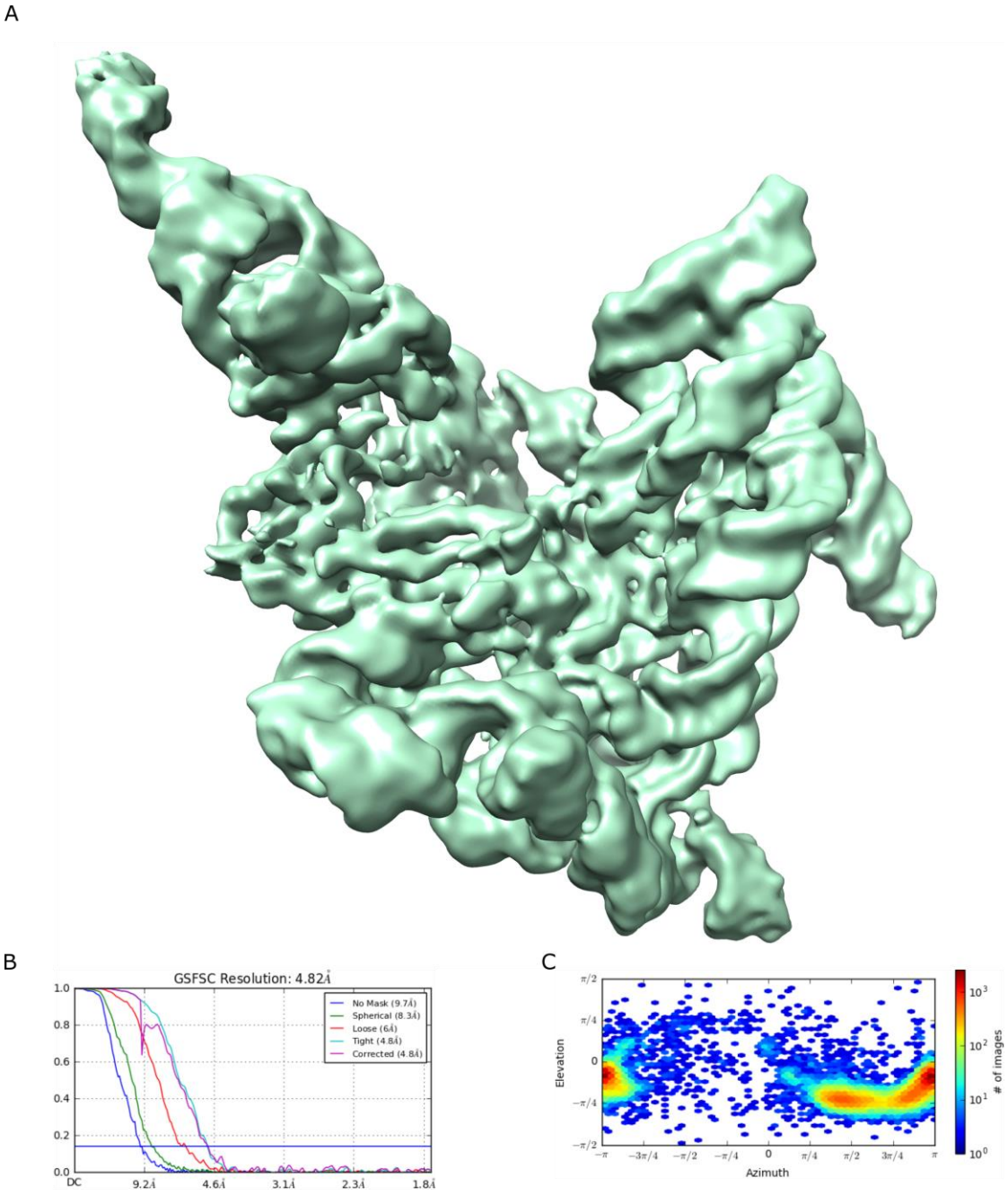


Figure 5.8: **Cryo-EM map for *T.el*4h/*T.el*4h RNP bound to DNA**. A) EM density at 4.8 Å. Clear density for RNA major/minor grooves is visible. Density leading into the active site is visible correlating to the target DNA. B) FSC curve for the 4.8 Å map. C) Heat map displaying the oreientation preference for the *T.el*4h/*T.el*4h specimen.

5.4 Data analysis and discussion

The 4.8 Å map of a group II intron retrotransposon interacting with dsDNA represents the first of its kind. The requirement of thin ice and the resulting orientation bias has created a barrier inhibiting further resolution improvement; however, at the current resolution, the majority of the intron RNA can be built into the model. Solving the challenge of preferred orientation is my current research focus.

Upon closer inspection of the 2D class averages, a dimer class is clearly visible (Figure 5.9). With an initial model built into the EM map, it was determined that the dimerization interface occurs predominantly between the two maturase proteins. Cryo-EM experiments are performed in a native solution of vitreous ice making the presence of a dimer biologically relevant and not experimentally induced, as can be the case with X-ray crystallography. The stoichiometry of the retrotransposition mechanism has been debated in the literature for many years. Biochemical evidence has suggested that the maturase protein associates with the intron RNA as a dimer [22]. The active form of many reverse transcriptases is a dimer, correlating with these biochemical results [102]. Recent crystal structures of two group IIC maturase RT domains show the proteins packing in the crystal as a dimer [103]. This observation could be due to forced crystal contacts during nucleation; however, the researchers followed up these finds by performing sedimentation velocity analysis by analytical ultracentrifugation (SV-AUC) and multiangle light scattering coupled to size-exclusion chromatography (SEC-MALs) experiments of the RT domain bound to DIV of the intron RNA. The results suggest that the dimer complex actually forms between two RNP particles instead of a single RNA bound to a dimer of maturase proteins. The SAXS data only contains partial RNP particles of the RT domain of the maturase with DIV of the intron RNA. The dimer observed in my EM data set is of a complete group II intron retroelement. The combination of these two results supports the biological relevance of the dimer in the mechanism of TPRT. The maturase binds to the intron with picomolar affinity and it has long been a question of mine as to how the RT separates itself from its high affinity-binding site in DIV [104]. By having retrotransposition occur through a dimer of RNPs, the maturase protein no longer has to disengage from the RNA to synthesize cDNA. With this finding, I propose a new mechanism for retrotransposition with dimerization of retroelements as a key feature (Figure 5.10).

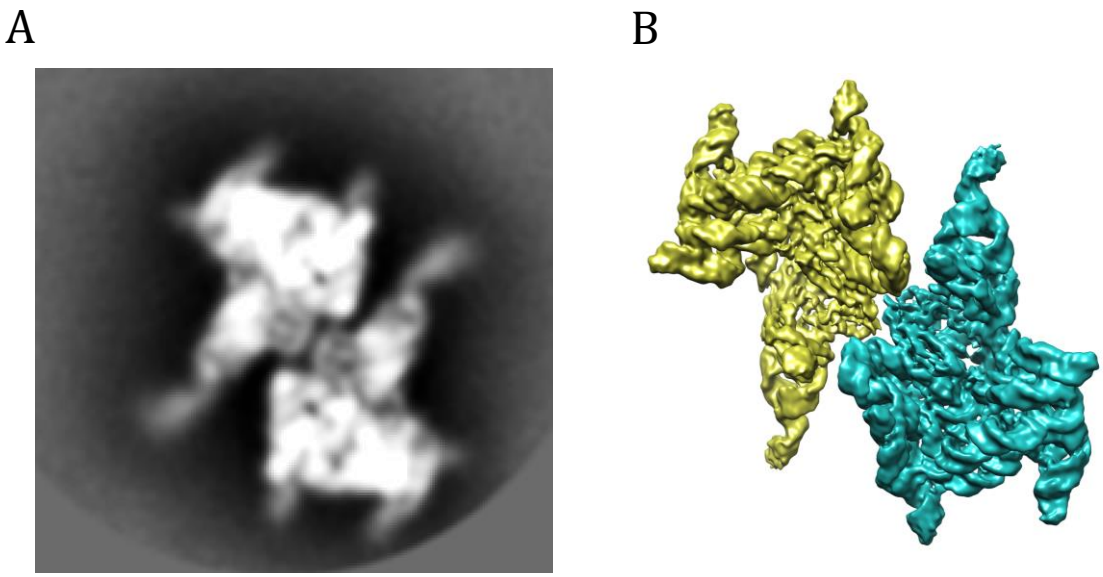


Figure 5.9: **Presence of a dimer 2D class average**. A) A dimer is clearly visible in the 2D class average. B) 3D representation of the 2D class average. This model suggests the dimer interface is comprised mostly between the two maturase proteins.

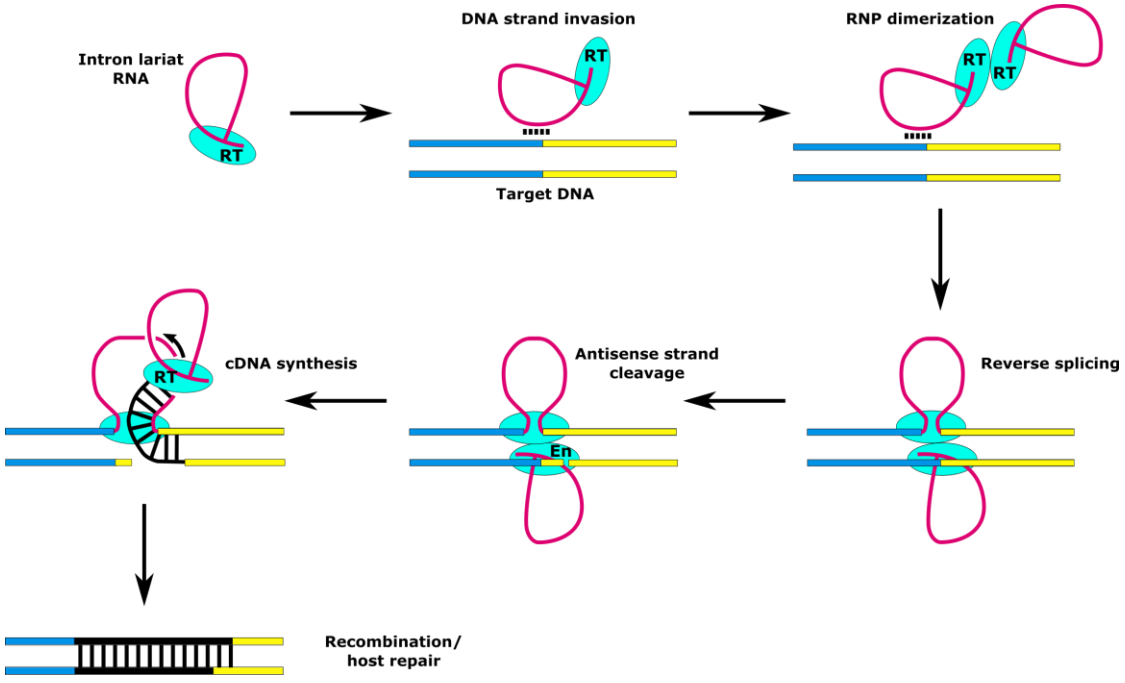


Figure 5.10: **Proposed mechanism for retrotransposition by TPRT**. Dimerization of the RNP particles during retrotransposition allows the maturase protein to stay bound to the intron during the process. This give flexibility to the mechanism without requiring the disruption of the high affinity binding site between the maturase and DIV of the intron.

5.5 Materials and methods

Plasmid cloning and DNA primers

A 6x-His-MBP-*T.el*4h maturase DNA gene was synthesized (Genscript) and cloned into a pET15b vector using the NdeI and BamHI restriction sites. The cloned plasmid was then transformed into Rosetta 2 cells (NEB). The *T.el*4h intron gene with the ORF removed from DIVb stem was synthesized (Genscript) and cloned into a pUC57 plasmid using the EcoRV restriction site. The cloned plasmid was transformed into DH5α cells. The target DNA primers were synthesized by IDT, with desthiobiotin (retrotransposon purification) or without (mobility and RT assay).

Top DNA target seq.: GATAGAGATTTTCCCAGGGTTGGCCGAGCGGATGAGGCAGCGAAC

Bottom DNA target seq.: GTTCGCTGCCTCATCCGCTCGGCCAACCCTGGGAAAATCTCTATC

T. elongates denaturing maturase purification

The *T.el*4h maturase protein is purified using the same protocol as shown in section 4.5.

In vito RNA transcription

*T.el*4h plasmid is linearized using an engineered HindIII restriction site. *In vitro* transcriptions are prepared in a 1 mL total volume. 40 µg of linear DNA template is added to transcription buffer containing: 50 mM Tris-HCl pH 7.5, 17.5 mM $MgCl_2$, 5 mM DTT, 2 mM spermidine, 0.05% Triton-X-100, and 2 mM of each NTP. T7 polymerase is added along with thermophilic inorganic phosphatase to initiate *in vitro* transcription. The reaction is placed at 37°C for 3 hrs. TURBO DNase (20 units) is then added along with 12 uL of 100 mM $CaCl_2$ and allowed to react for 1 hr at 37°C. Proteinase K (200 µg) is then added at 37°C for 1 hr. The mixture is centrifuged to remove any precipitate. The supernatant is filtered through a 0.2 µm filter into a 100 Kd cut off filter. The RNA containing solution is buffer exchanged a total of 7 times with 14 mL of filtration buffer (5 mM cacodylate pH 6.5 and 10 mM $MgCl_2$). For the last buffer exchange, the RNA is concentrated to approximately 10 mg/mL.

***In vitro* mobility and RT assay**

To perform mobility experiments, the target DNA top strand is 5´ radiolabeled with ^{32}P. To perform reverse transcription experiments, the target DNA bottom strand is 5´ radiolabeled with P^{32}. In either case, 50 pmol of DNA primer is mixed with 50 pmol of [γ-^{32}P] ATP, 5 uL of reaction buffer and 20 units of T4 polynucleotide kinase (New England BioLabs) in a total reaction volume of 50 µL. The solution is incubated for 30 minutes at 37°C. The mixture is then passed through a desalting gel filtration column to capture the unreacted ATP. The labeled primers are ethanol precipitated using NH_4OAc and linear polyacrylamide as a carrier. The resulting pellet is re-suspended in water. In the case of analyzing RNA insertion, radiolabeled top DNA is annealed to cold bottom DNA by heating the mixture to 90°C for 2 minutes and allowing the solution to cool at RT for 15 minutes. In the case of analyzing cDNA synthesis, cold top DNA is annealed to radiolabeled bottom DNA in the same way. *T.el*4h intron RNA is then spliced with *T.el*4h protein in 10 mM $MgCl_2$, 40 mM Tris-HCl pH 7.5, 500 mM NH_4Cl, and 5 mM DTT for 5 minutes at 50°C. Annealed target DNA is added along with 200 µM dNTPs and allowed to react at 50°C for an additional 5 minutes. The samples are then quenches by phenol/chloroform extraction and the nucleic acids are recovered by perming an ethanol precipitation using linear polyacrylamide as a carrier. The pellets are re-suspended in formamide and analyzed on a denaturing 4% 19:1 acrylamide:bis-acrylamide PAGE gel with 8 M urea.

References

1 Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147-157 (1982).

2 Stark, B. C., Kole, R., Bowman, E. J. & Altman, S. Ribonuclease P: an enzyme with an essential RNA component. *Proc Natl Acad Sci U S A* **75**, 3717-3721 (1978).

3 Gilbert, W. Origin of life: The RNA world. *Nature* **319**, 618, doi:doi:10.1038/319618a0 (1986).

4 Ferat, J. L. & Michel, F. Group II self-splicing introns in bacteria. *Nature* **364**, 358-361, doi:10.1038/364358a0 (1993).

5 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).

6 Robart, A. R. & Zimmerly, S. Group II intron retroelements: function and diversity. *Cytogenet Genome Res* **110**, 589-597, doi:10.1159/000084992 (2005).

7 Peebles, C. L. *et al.* A self-splicing RNA excises an intron lariat. *Cell* **44**, 213-223 (1986).

8 van der Veen, R. *et al.* Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. *Cell* **44**, 225-234 (1986).

9 Michel, F., Umesono, K. & Ozeki, H. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**, 5-30 (1989).

10 Michel, F. & Ferat, J. L. Structure and activities of group II introns. *Annu Rev Biochem* **64**, 435-461, doi:10.1146/annurev.bi.64.070195.002251 (1995).

11 Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-splicing group II intron. *Science* Vol. 320 77-82 (2008).

12 Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature*, doi:10.1038/nature13790 (2014).

13 Zimmerly, S. *et al.* A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* **83**, 529-538 (1995).

14 Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *Embo j* **9**, 3353-3362 (1990).

15 Mohr, G., Perlman, P. S., Departmentof Biochemistry University of Texas Southwestern Medical CenterCenter, D., TX 75235-9038, USA & Lambowitz, A. M. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Research* **21**, 4991-4997, doi:10.1093/nar/21.22.4991 (2017).

16 San Filippo, J. & Lambowitz, A. M. Characterization of the C-Terminal DNA-binding/DNA Endonuclease Region of a Group II Intron-encoded Protein. *Journal of Molecular Biology* **324**, 933-951, doi:https://doi.org/10.1016/S0022-2836(02)01147-6 (2002).

17 Blocker, F. J. *et al.* Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* Vol. 11 14-28 (2005).

18 Dai, L. *et al.* A Three-Dimensional Model of a Group II Intron RNA and Its Interaction with the Intron-Encoded Reverse Transcriptase. *Mol Cell* **30**, 472-485, doi:10.1016/j.molcel.2008.04.001 (2008).

19 Zimmerly, S., Guo, H., Perlman, P. S. & Lambowltz, A. M. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**, 545-554, doi:https://doi.org/10.1016/0092-8674(95)90027-6 (1995).

20 Cousineau, B., Lawrence, S., Smith, D. & Belfort, M. Retrotransposition of a bacterial group II intron. *Nature* **404**, 1018, doi:doi:10.1038/35010029 (2000).

21 Martínez-Abarca, F., Toro, N. RecA-independent ectopic transposition in vivo of a bacterial group II intron. *Nucleic Acids Research* **28**, 4397-4402, doi:10.1093/nar/28.21.4397 (2017).

22 Rambo, R.P. & Doudna, J.A. Assembly of an Active Group II Intron−Maturase Complex by Protein Dimerization. *Biochemistry* doi:S0006-2960(04)09912-X (2004).

23 Dai, L., Toor, N., Olson, R., Keeping, A. & Zimmerly, S. Database for mobile group II introns. *Nucleic Acids Res* **31**, 424-426 (2003).

24 Toor, N., Hausner, G. & Zimmerly, S. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**, 1142-1152 (2001).

25 Cavalier-Smith, T. The origin of nuclei and of eukaryotic cells. *Nature* **256**, 463, doi:doi:10.1038/256463a0 (1975).

26 Margulis, L. & Bermudes, D. Symbiosis as a mechanism of evolution: status of cell symbiosis theory. *Symbiosis* **1**, 101-124 (1985).

27 Nilsen, T. W. RNA-RNA Interactions in Nuclear Pre-mRNA Splicing. *35*, doi:https://cshmonographs.org/index.php/monographs/article/view/3876 (2009).

28 Seetharaman, M., Eldho, N. V., Padgett, R. A. & Dayie, K. T. Structure of a self-splicing group II intron catalytic effector domain 5: Parallels with the spliceosome. *RNA* Vol. 12 235-247 (2006).

29 Fica, S. M. *et al.* RNA catalyzes nuclear pre-mRNA splicing. *Nature* **503**, 229-234, doi:10.1038/nature12734 (2013).

30 Fica, S. M. *et al.* Structure of a spliceosome remodelled for exon ligation. *Nature* **542**, 377-380, doi:10.1038/nature21078 (2017).

31 Galej, W. P. *et al.* CryoEM structure of the spliceosome immediately after branching. *Nature* **537**, 197-201, doi:10.1038/nature19316 (2016).

32 Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617-621, doi:10.1038/nature22799 (2017).

33 Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349**, 1182-1191, doi:10.1126/science.aac7629 (2015).

34 Hang, J., Wan, R., Yan, C. & Shi, Y. Structural basis of pre-mRNA splicing. *Science* doi:10.1126/science.aac8159 (2015).

35 Zimmerly, S. & Semper, C. Evolution of group II introns. *Mobile DNA* **6**, 7, doi:doi:10.1186/s13100-015-0037-5 (2015).

36 Malik, H. S. *et al.* The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution* **16**, 793-805, doi:10.1093/oxfordjournals.molbev.a026164 (2017).

37 Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, Mdna3-0011-2014, doi:10.1128/microbiolspec.MDNA3-0011-2014 (2015).

38 Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802-813, doi:10.1261/rna.876308 (2008).

39 Matlin, A. J., Clark, F. & Smith, C. W. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**, 386-398, doi:10.1038/nrm1645 (2005).

40 Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457-463, doi:10.1038/nature08909 (2010).

41 Brow, D. A. Allosteric cascade of spliceosome activation. *Annu Rev Genet* **36**, 333-360, doi:10.1146/annurev.genet.36.043002.091635 (2002).

42 Matlin, A. J. & Moore, M. J. Spliceosome assembly and composition. *Adv Exp Med Biol* **623**, 14-35 (2007).

43 Staley, J. P. & Woolford, J. L., Jr. Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines. *Curr Opin Cell Biol* **21**, 109-118, doi:10.1016/j.ceb.2009.01.003 (2009).

44 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).

45 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).

46 Will, C. L. & Luhrmann, R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3**, doi:10.1101/cshperspect.a003707 (2011).

47 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703, doi:10.1038/nrg2640 (2009).

48 Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903, doi:doi:10.1038/nature03663 (2005).

49 Kano, H. *et al.* L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. doi:10.1101/gad.1803909 (2009).

50 Jung, Y. D. *et al.* Retroelements: molecular features and implications for disease. *Genes Genet Syst* **88**, 31-43 (2013).

51 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 (1993).

52 Steitz, T. A. A mechanism for all polymerases. *Nature* **391**, 231-232, doi:10.1038/34542 (1998).

53 Nakamura, T., Zhao, Y., Yamagata, Y., Hua, Y. J. & Yang, W. Watching DNA polymerase eta make a phosphodiester bond. *Nature* **487**, 196-201, doi:10.1038/nature11181 (2012).

54 Hougland, J. L., Kravchuk, A. V., Herschlag, D. & Piccirilli, J. A. Functional identification of catalytic metal ion binding sites within RNA. *PLoS Biol* **3**, e277, doi:10.1371/journal.pbio.0030277 (2005).

55 Erat, M. C. & Sigel, R. K. Divalent metal ions tune the self-splicing reaction of the yeast mitochondrial group II intron Sc.ai5gamma. *J Biol Inorg Chem* **13**, 1025-1036, doi:10.1007/s00775-008-0390-7 (2008).

56 Scott, W. G. *et al.* Rapid crystallization of chemically synthesized hammerhead RNAs using a double screening procedure. *J Mol Biol* **250**, 327-332, doi:10.1006/jmbi.1995.0380 (1995).

57 Wiryaman, T. & Toor, N. Structure determination of group II introns. *Methods* **125**, 10-15, doi:10.1016/j.ymeth.2017.06.020 (2017).

58 Heus, H. A. & Hilbers, C. W. Structures of non-canonical tandem base pairs in RNA helices: review. *Nucleosides Nucleotides Nucleic Acids* **22**, 559-571, doi:10.1081/ncn-120021955 (2003).

59 Stagno, J. R., Bhandari, Y. R., Conrad, C. E., Liu, Y. & Wang, Y. X. Real-time crystallographic studies of the adenine riboswitch using an X-ray free-electron laser. *Febs j* **284**, 3374-3380, doi:10.1111/febs.14110 (2017).

60 Zwart, P. H. *et al.* Automated structure solution with the PHENIX suite. *Methods Mol Biol* **426**, 419-435, doi:10.1007/978-1-60327-058-8_28 (2008).

61 Otwinowski, Z. & Minor, W. [20] Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* **276**, 307-326, doi:10.1016/s0076-6879(97)76066-x (1997).

62 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* Vol. 66 486-501 (2010).

63 Morin, A. *et al.* Collaboration gets the most out of software. *Elife* **2**, e01456, doi:10.7554/eLife.01456 (2013).

64 Costa, M., Fontaine, J. M., Loiseaux-de Goër, S. & Michel, F. A group II self-splicing intron from the brown alga Pylaiella littoralis is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. *J Mol Biol* **274**, 353-364, doi:S0022283697914169 [pii] (1997).

65 Strobel, S. A. & Shetty, K. Defining the chemical groups essential for Tetrahymena group I intron function by nucleotide analog interference mapping. *Proc Natl Acad Sci U S A* **94**, 2903-2908 (1997).

66 Waldsich, C. & Pyle, A. M. A folding control element for tertiary collapse of a group II intron ribozyme. *Nat Struct Mol Biol* Vol. 14 37-44 (2007).

67 Toor, N., Robart, A. R., Christianson, J. & Zimmerly, S. Self-spliicng of a group IIC intron: 5' exon recogniation and alternative 5' splicign events implicate the stem-lopp motif of a transcriptional terminator. *Nucleic Acids Res* Vol. 34 6461-6471 (2006).

68 Podar, M., Dib-Hajj, S. & Perlman, P. S. A UV-induced, Mg(2+)-dependent crosslink traps an active form of domain 3 of a self-splicing group II intron. *Rna* **1**, 828-840 (1995).

69 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).

70 Mikheeva, S., Murray, H. L., Zhou, H., Turczyk, B. M. & Jarrell, K. A. Deletion of a conserved dinucleotide inhibits the second step of group II intron splicing. *RNA* **6**, 1509-1515 (2000).

71 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* Vol. 16 1-9 (2010).

72 Jacquier, A. & Michel, F. Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *J Mol Biol* Vol. 213 437-447 (1990).

73 Fedorova, O., Waldsich, C. & Pyle, A. M. Group II intron folding under near-physiological conditions: collapsing to the near-native state. *J Mol Biol* **366**, 1099-1114, doi:10.1016/j.jmb.2006.12.003 (2007).

74 Keating, K. S. & Pyle, A. M. Semiautomated model building for RNA crystallography using a directed rotameric approach. *Proc Natl Acad Sci U S A* Vol. 107 8177-8182 (2010).

75 Chanfreau, G. & Jacquier, A. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J* **15**, 3466-3476 (1996).

76 Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* Vol. 514 193-197 (2014).

77 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).

78 Kowal, E. A. *et al.* Altering the Electrostatic Potential in the Major Groove: Thermodynamic and Structural Characterization of 7-Deaza-2′-deoxyadenosine:dT Base Pairing in DNA. *J Phys Chem B* **115**, 13925-13934, doi:10.1021/jp207104w (2011).

79 Boudvillain, M. & Pyle, A. M. Defining functional groups, core structural features and inter-domain tertiary contacts essential for group II intron self-splicing: a NAIM analysis. *EMBO J* **17**, 7091-7104, doi:10.1093/emboj/17.23.7091 (1998).

80 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 (1996).

81 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).

82 Mohr, G., Ghanem, E. & Lambowitz, A. M. Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol* **8**, e1000391, doi:10.1371/journal.pbio.1000391 (2010).

83 Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *Rna* **19**, 958-970, doi:10.1261/rna.039743.113 (2013).

84 Hurbain, I. & Sachse, M. The future is cold: cryo-preparation methods for transmission electron microscopy of cells. *Biol Cell* **103**, 405-420, doi:10.1042/bc20110015 (2011).

85 Thompson, R. F., Walker, M., Siebert, C. A., Muench, S. P. & Ranson, N. A. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods* **100**, 3-15, doi:10.1016/j.ymeth.2016.02.017 (2016).

86 Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530, doi:10.1016/j.jsb.2012.09.006 (2012).

87 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).

88 Bammes, B. E., Rochat, R. H., Jakana, J., Chen, D. H. & Chiu, W. Direct electron detection yields cryo-EM reconstructions at resolutions beyond ¾ Nyquist frequency. *J Struct Biol* **177**, 589-601, doi:10.1016/j.jsb.2012.01.008 (2012).

89 Wu, S., Armache, J. P. & Cheng, Y. Single-particle cryo-EM data acquisition by using direct electron camera. *Microscopy (Oxf)* Vol. 65 35-41 (2016).

90 Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332, doi:10.1038/nmeth.4193 (2017).

91 Yang, G., Zhou, R. & Shi, Y. Cryo-EM structures of human gamma-secretase. *Curr Opin Struct Biol* **46**, 55-64, doi:10.1016/j.sbi.2017.05.013 (2017).

92 Carragher, B. *et al.* Leginon: an automated system for acquisition of images from vitreous ice specimens. *J Struct Biol* **132**, 33-45, doi:10.1006/jsbi.2000.4314 (2000).

93 Zhang, K. Gctf: Real-time CTF determination and correction. *J Struct Biol* **193**, 1-12, doi:10.1016/j.jsb.2015.11.003 (2016).

94 Lehmann, K. & Schmidt, U. Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol* **38**, 249-303, doi:10.1080/713609236 (2003).

95 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).

96 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

97 Richardson, S. R. *et al.* The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, Mdna3-0061-2014, doi:10.1128/microbiolspec.MDNA3-0061-2014 (2015).

98 Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *Embo j* **21**, 5899-5910 (2002).

99 Rybak, J. N., Scheurer, S. B., Neri, D. & Elia, G. Purification of biotinylated proteins on streptavidin resin: a protocol for quantitative elution. *Proteomics* **4**, 2296-2299, doi:10.1002/pmic.200300780 (2004).

100 Hirsch, J. D. *et al.* Easily reversible desthiobiotin binding to streptavidin, avidin, and other biotin-binding proteins: uses for protein labeling, detection, and isolation. *Anal Biochem* **308**, 343-357 (2002).

101 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).

102 Divita, G., Restle, T. & Goody, R. S. Characterization of the dimerization process of HIV-1 reverse transcriptase heterodimer using intrinsic protein fluorescence. *FEBS Lett* **324**, 153-158 (1993).

103 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).

104 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).